UNIVERSITY OF CALIFORNIA,
IRVINE

Towards Probative Foundations for Bayesian Statistics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

David Mghanga Mwakima

Dissertation Committee:
Chancellor's Professor James O. Weatherall, Chair
Chancellor's Professor Simon M. Huttegger
Professor P. Kyle Stanford
Jean-Claude Falmagne Chair of Mathematical Psychology and Professor Jeffrey N. Rouder
Professor of Philosophy of Physics Samuel C. Fletcher

2024

DEDICATION

To my fellow under-laborers who share de Finetti's vision
of a Bayesian 21st Century

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank Penelope ("Pen") Maddy, Jeremy Heis and JB Manchak. They believed that the Department of Logic and Philosophy of Science (LPS) would be an excellent program for me and that I would flourish here, although I had not taken a single class in the philosophy of science! Pen's fascinating and original work that engaged with the indispensability arguments in the philosophy of mathematics and that articulated for the first time a "Second Philosophy of Metaphysics" provided the initial impetus for some of the work that I ended up doing in my dissertation. For the training and illustrations of how to do meticulous scholarship in the history of philosophy of science, I am grateful to both Pen and Jeremy. Thank you JB for helping me find my place at LPS after wandering around, significantly I hope, during my first and second year!

Secondly, I would like to thank my colleagues in LPS for creating a vibrant collegial community where we could all do our best work. Special thanks to members of my cohort (which I think was the largest ever cohort of admitted students by LPS as of 2018): Jingyi Wu, David Freeborn, Nathan Gabriel, Margaret Farrell, Saira Khan, Adam Chin and Jason Chen. The fact that we *all* stayed and completed our degrees is quite remarkable. But I believe that it is a testament to the collective and deliberate effort we all put in to build community, to show support for each other and to foster a healthy academic community. They are truly the "first-goats"! I would also like to thank Kino Zhao, Alysha Kassam, Will Stafford, Greg Lauro, Chris Mitsch, Elliott Chen, Mike Schneider (my host during the 2018 LPS Prospective Student Weekend), Sarita Rosenstock, Kaetlin Taylor, Aydin Mohseni, Daniel Herrmann, Bruce Rushing, Marian Gilton, Jeffrey Schatz, Tom Colclough, Stella Moon, Charles Leitz, Lee Killam, Travis LaCroix (for this UCI thesis LaTeX template), Helen Meskhidze, Josef Kay, Kevin Kadowaki and Jessica Gonzalez, who were between one to five years ahead of me in their respective Ph.D. programs, for the peer-to-peer mentorship. They were like my "older siblings" in the program and I am so thankful for the genuine interest in my progress, the support and the encouragement (especially during those times when things seemed to be falling apart). Among my "younger siblings", I would like to thank Ellen Shi and Sophio Machavariani (who were also my Teaching Assistants for Intro to Inductive Logic in Spring 2023), Rebecca Korf, True Gibson, Josiah Lopez-Wild, Tori Cotton, Christian Torsell, Neil Crawford, Matthew Coates, Benjamin Genta, Clara Bradley, Jessica Lauman-Lairson, Aidan Carter, Curtis Mason, Jaehyun Lee, Ainsley May, Jack vanDrunen, Antoine Mercier, Frank Hu, Ryan Chen and especially Elijah Spiegel for the stimulating conversations around our recently discovered shared interest in the philosophy of statistics!

Outside the Department, I would like to thank Andre Lebrun, Louis Doulas, Nicholas Smith, Natalia Nealon, Annabelle Tada, Anna Pederneschi, Itzel Garcia, Oscar Piedrahita, Evan Sommers, Dylan Popowicz, Taylor Dunn, Cherrish Hardy Jones, Darby Vickers, Rachel Cooper, Rena Goldstein, Tanuj Raut, Eyob Zewdie and Sayid Bnefsi from the Department of Philosophy for expanding my intellectual community, for the close friendships and for the good will and cheer they spread all around. From the Department of Statistics, I want to thank my colleagues Alexandria Lee Richardson, Jizhi Zhang, Christina Magana-Ramirez, Kevin Li, Borna Bateni (now at UCLA), Adriana Chávez (in UCI Cognitive Science), Jie

Wan (in UCI Cognitive Science), Thanasi Bakis, Bianca Brusco (as well as the participants in my talk for the Department of Statistics Brown Bag Series) and Mikaela Nishida for the friendship and collaborative environment around the study of statistics that they created. Deserving special gratitude are Faculty members within the Statistics Department who warmly welcomed me into their program and saw me through until I earned my degree: Michele Guindani (now at UCLA), Annie Qu, Veronica Berrocal, Tianchen Qian, Weining Shen, Yaming Yu, Volodymyr Minin and Zhaoxia Yu. I would also like to thank Knut Solna — Faculty member in the Mathematics Department at UCI — for directing and supervising my study of Measure Theory and Stochastic Processes. It was a pleasure working with him.

But my progress through the Ph.D. program has not just been due to the intellectual community I surrounded myself with, there are also those people who enriched my life outside academia. This list is incredibly large, so I fear that I may inadvertently leave some people out. Still, let me shout out and say thank you to Ronald Rogo and to Emmy Rono, two people with whom we go way back, and who have truly touched my life. Thanks to my friends, teachers and mentors from Starehe: Paul, Elvis, Edwin, Alfonce, Stephen, Hiram, Pascal, Moses, Felix, Mrs. Oduor, Sr. Gracy, Sr. Clara, Mrs. Omulama, Mr. Kasili, Mr. Mukangai, Mr. Mbugua, Mr. Ndung'u, Mr. Mwangi, Mr. Ndegwa and Mr. Kithyaka; my friends and teachers from Deerfield: Brandon, David Kongo and family, Sue Carlson and family, Mr. Bakker and family, Dr. Ross, Mrs. Heise and especially Ms. Martha Lyman. Thank you to my mentors and friends from my time at Harvard: Alison Simmons, Bernhard Nickel, Peter Koellner, Mark Richard, Jeremy Mullen (and members of the Harvard Undergraduate Fellowship, especially Ellen and Curtis), Zeynep Soysal (and the members of her Analyticity class from 2015), Sara Sussman, Pierre Julien, Gergana Hunt (and the I.T. team at the DCE Computer Lab at Harvard). I am grateful to Dr. Mary A. Ochieng, Dr. Elvice Ongong'a and Dean Godfrey Madigu for generously hosting me as a Visiting Fellow at the Strathmore Institute of Mathematical Sciences in Nairobi, Kenya in Summer 2023 and between March – May 2024 and to my friends Kui, Zena, Essy, and Rachel, whom I met during this time. Thank you to my African community here at Irvine: Gaelle, Safia, Sibs, Chimee, Munyao and Modeste (who also welcomed me into the loving Cameroonian community around Irvine) for the strong show of support and friendship — I still wish I had met all of you wonderful people sooner than I did! For the contentment I have found following our serendipitous connection, I am grateful to Martha.

I would like to thank my family. Daddy and Mummy, thank you for the safe and nurturing environment you raised me in. The older I get, the more I realize how truly fortunate and extremely privileged I have been to have been brought up by you. For the sacrifices (both personal and financial) and commitment that made it possible for me to get the best life and education possible, thank you. For some of the values I learned by being raised in a loving Christian household, thank you. Eddah, my sister, I am also grateful to have grown up with you. I still remember that time in 9th grade, when I almost quit being a librarian at Starehe because of the "significant" time commitment it required. Your advice to learn how to compartmentalize different aspects of my life, to realize that life doesn't get any less complicated as we grow (we just have to find balance and to cultivate resilience) convinced

me not to quit; and that advice has stuck with me to this day when things get overwhelming. I believe that's one of the many reasons I am here today. Thank you.

I would also like to thank some Faculty at LPS and some scholars around the world who have shaped my interests and work. Among the Faculty members at LPS, I would like to thank Cailin O'Connor and Jeffrey Barrett who have followed my work with interest and have guided me in making it both clear, accessible and correct. Thanks to Toby Meadows for his advice during my first year to broaden my work and interests by taking classes on traditional topics in philosophy (history, Kant and Kai Wehmeier's C-ALPHA Logic Seminar) in order to be a well-rounded philosopher. For the community of scholars around the world who have played a significant part in how this dissertation has turned out, special thanks to Jan-Willem Romeijn. Jan-Willem's warmth, honesty and collegiality has been lovely. Thanks to Stathis Psillos for some initial suggestions from Summer 2020 about where I could go with my interests around evidence and its quality. I would also like to thank Christopher Smeenk and the participants of the Philosophy of Cosmology Seminar from Fall 2020 (co-taught with Jim) for the lively discussion around Perrin's work on Brownian Motion, which I was just beginning to seriously dig into. My gratitude also goes out to Branden Fitelson, whom I met at PIKSI Logic 2024. He has also entered the conversation towards the end, but I believe ours is just the beginning. My gratitude also goes out to Maureen Eckert, the organizers and participants of the 2016 Summer Diversity Program in Logic at UMass Dartmouth. I am grateful to David Boonin, the organizers and participants of the 2017 Summer Seminar in Philosophy at the University of Colorado, Boulder. Some of my friends from these two programs: Mahmoud Jalloh, Katie Deaven, Nichi Smith, Bailey Thomas, Emma Hardy, Sara Purinton, Yasha Sapir, Anthony Nguyen and Alex Pho are also some of the best philosophers I have met thanks to these programs. Thank you for the diverse community and resourceful network you've given me access to. Thank you to the organizers and participants of the Summer 2024 Philosophy of Statistics Early Career Summer Workshop at the University of Minnesota, Twin Cities for sharing your insights and feedback on my work. It was delightful to meet such promising and talented scholars, and I look forward to future collaborative work with many of them.

Finally, I would like to thank my committee. The "best thing" about having in your committee such preeminent experts of their respective fields, is the hope that their grandeur somehow rubs off on you. But the "worst thing" is realizing that they are going to hold your work to a very high standard! Thank you Jim, Simon, Kyle, Jeff and Sam for being in my committee. Working under your supervision, I have strived for excellence by emulating the quality of your own work. Whether I have attained these lofty goals, I leave it up to my readers. What I can say is that any remaining errors are on me, not them. They have done their part to keep my writing and my thinking clear — but more importantly, honest. Special thanks to Jim for believing in the fruitfulness of the project, for believing in my ability to see it through, and for trusting my judgment to drive the project forward in the directions that I chose. Thank you for your perceptive engagement with my work, for getting me to ask, and to answer, the difficult but right questions (that a critical reader would have), which had not even occurred to me as I wrote. Thank you, Kyle for your sound counsel. I owe a lot to Kyle for the direction I have taken this work since my first class in philosophy of science,

where I took “my vitamins” under his tutelage! I fondly remember that day in 2019 when I met Kyle in his office after I had just come off a phone call with Pen. At the time, I was struggling to find my place in LPS. Among the many things he said, Kyle told me to pick a topic that wasn't “so easy” that it's boring; but also not “so hard” that it frustrates me. What an impossible task Kyle set me! Hopefully, this topic has been just right. Jeff, thank you for introducing me to Bayesian statistics. In the summer of 2020 I was grappling with the question: what, if anything, does statistics have to say about “the quality of evidence”? Thank you for orienting me towards an answer. Simon, and I should add Brian Skyrms, thank you for welcoming me into the community of under-labourers in the tradition of de Finetti. Given the depth and the technical aspect of your work, it took me a while (a little over 5 years!) to get to the point where I could have a meaningful intellectual conversation with you. My hope is that the last two months of engagement are just the beginning of our fruitful and on-going collaboration towards a Bayesian 21st century. Finally, thank you Sam for being quite the trail-blazer both at LPS and in the philosophy of statistics community! Thank you for the keen interest you've taken in my work, for your painstaking attention to detail, for the gift of sincere and constructive criticism, as well as mentorship. My work is all the better thanks to you.

VITA

David Mghanga Mwakima

EDUCATION

Ph.D. in Philosophy — **June 2024**
University of California, Irvine — *Irvine, California*

M.Sc. in Statistics — **March 2024**
University of California, Irvine — *Irvine, California*

M.A. in Philosophy — **December 2021**
University of California, Irvine — *Irvine, California*

A.B. in Philosophy (Hons.), Mathematical Sciences — **May 2017**
Harvard College — *Cambridge, Massachusetts*

SELECTED HONORS AND AWARDS

School of Social Sciences Instructional Fellow — **2020**
School of Social Sciences, University of California, Irvine

UNIVERSITY SERVICE

Inaugural UROP Research Discovery Program Mentor — **2022**
University of California, Irvine

DECADE Representative for LPS — **2020 – 2022**
University of California, Irvine

DECADE PLUS Leadership Coach — **2019 – 2021**
University of California, Irvine

Associated Graduate Students (AGS) Council Member — **2019 – 2020**
University of California, Irvine

ABSTRACT OF THE DISSERTATION

Towards Probative Foundations for Bayesian Statistics

By

David Mghanga Mwakima

Doctor of Philosophy in Philosophy

University of California, Irvine, 2024

Chancellor's Professor James O. Weatherall, Chair

My dissertation addresses the question of how scientists evaluate the evidence they have for their claims. In the first chapter, I demonstrate the viability of using Bayesian methods in statistics to evaluate statistical evidence using an episode from the history of science involving Jean Baptiste Perrin, who was a French chemical physicist working in the early 20th century. This episode has fascinated philosophers of science because Perrin's experimental work that confirmed the atomic hypothesis (the view that matter is composed of atoms) has been cited as an illustration of the impact that strong evidence can have. For this reason, numerous accounts have been offered for why Perrin's evidence was so distinctive. Bayesian accounts of this episode have been quite influential. However, they have been criticized by philosophers because they face: (1) the "Catch-all hypothesis" problem (which is the problem of exhaustively specifying, in the space of hypotheses, the logical complement of a given hypothesis in order to compute the marginal likelihood function of the data — this is the denominator in Bayes' theorem); and (2) the problem that any specification of priors in the Perrin case is ad hoc. In view of these difficulties, I provide a novel and more precise Bayesian account of this episode than those that have been offered and I argue that my account avoids these problems. In doing so, I contribute to the philosophy of statistics by showing the viability of using Bayes Factor (which is a Bayesian measure of the relative strength of evidence for two competing models or hypotheses) to quantify statistical evidence;

and to the philosophy of science, where prominent realists and anti-realists today are charting a middle path forward in the realism-antirealism debate.

The other chapters of my dissertation address different aspects of the following question: "How reliable are coherent Bayesian methods for evaluating statistical evidence in science?" Coherent Bayesian methods are those Bayesian methods that satisfy the Likelihood Principle. This principle states that parametric statistical inference should be based on the equivalence class of functions of the parameter within a given statistical model in which the data are fixed (this equivalence class is also known as the likelihood function). For example, using the Bayes Factor to quantify statistical evidence is a coherent Bayesian method. Some statisticians and philosophers of statistics who argue against using coherent Bayesian methods argue that these methods conflict with other important desiderata that scientists have. These desiderata include: (1) calibrating inferences and predictions (where this involves providing an objective measure, or guarantee, of how often the inferences and predictions are verifiably correct), and (2) model assessment (where this involves probing or testing statistical models to determine their compatibility with the observed data). These desiderata are important because, taken together, they reflect the healthy skepticism scientists typically have towards their claims. This attitude involves probing those claims and quantifying the reliability of the inferences that they make supporting or disproving those claims. The lack of tools for satisfying these needs using coherent Bayesian methods is a serious indictment of those methods and is the primary reason given for why they are not yet widely adopted in practice — this is the probativist criticism (from the word 'probative' which means to test or to try).

Most philosophers who are sympathetic to the Bayesian approach either misconstrue the force of the probativist criticism or dismiss that criticism by rejecting some underlying assumption made by those who advance it. For example, one underlying assumption that is often rejected is that statistical modeling should be aiming at the truth. So, in chapter two, I sharpen the probativist criticism and I argue that it cannot be dismissed by rejecting one or more

of its underlying assumptions. In chapter three, I turn to the Likelihood Principle and the normative constraints it places on coherent inference. Here I argue that the scope of the Likelihood Principle should be restricted to parametric inferences that involve point estimation. If the scope of Likelihood Principle can be so restricted, then my work here will contribute to laying the groundwork for introducing tools within the Bayesian framework for model assessment and will advance the debate regarding the possibility of probative foundations for Bayesian statistics.

Chapter 1

On the Quality of Perrin's Evidence

1.1 Introduction

Evidence occupies a central role in philosophy. Within philosophy of science specifically, evidence, and its quality, constitutes the basis on which the debate between realists and anti-realists as understood *today* is, or should be, adjudicated.[1] Within this debate, an informal notion of the "quality of evidence" from Jean Perrin's experimental work on Brownian motion has for decades been offered as an explanation of the shift in epistemic attitudes toward the atomic hypothesis among prominent scientists at the turn of the 20th century.[2] The shift in attitude was from viewing the atomic hypothesis as a merely instrumentally useful hypothesis (anti-realism) to viewing it as a well-established theory (realism). Assuming that the shift in epistemic attitudes was caused by the distinctive quality of Perrin's evidence, one goal for philosophers of science has been to characterize what makes evidence, such as Perrin's, good.

[1] See especially Psillos (2018), Psillos (2021) and Stanford (2021).

[2] See Glymour (1980), Salmon (1984), Mayo (1996), Maddy (1997), Achinstein (2002), Maddy (2007), Stanford (2009), Psillos (2011), Psillos (2014) and Smith and Seth (2020) for a sample of some of the recent views that have been offered on this topic.

It is helpful to think of the various accounts that have been offered for evaluating evidence as falling along a spectrum. On one end of this spectrum, there are informal accounts provided by Quine (1976) and Maddy (2007, 398f, 406). On the other end, there are highly formal presentations in inductive logic, confirmation theory and formal epistemology.[3] In between these two extremes, there are semi-formal proposals due to Achinstein (2001) and Roush (2005), which approach the evaluation of evidence probabilistically, but independently of statistics. Finally, there are statistical approaches to the evaluation of evidence, which have slowly began penetrating philosophical discussions with some authors calling on philosophers to pay more attention to these statistical approaches to evaluating evidence.[4]

Here my goal is to characterize, from a *statistical perspective*, what made the evidence from Perrin's experiments on Brownian motion good. In order to accomplish this goal, I will focus specifically on Perrin's granule-displacement experiments that confirmed Einstein's diffusion model of the motion of suspended particles in a dilute solution. One reason for just focusing on the granule-displacement experiments is that the confirmation of Einstein's model by these experiments has been lauded by many as some of the most convincing evidence that came out of Perrin's laboratory.[5] Another reason for focusing on these experiments is that the observations can be modeled using well-known *statistical models* (see the next section). I will argue that the quality of Perrin's *statistical* evidence that confirmed Einstein's model was good because it was highly specific and discriminating. The specificity and discriminating

[3]See Carnap (1962), Fitelson (2007) and Pettigrew (2016) for examples of some of the work that is done here. See Earman (1992) and Sprenger and Hartmann (2019) for a comprehensive overview and detailed bibliographies.

[4]See Edwards, Lindman, and Savage (1963), Edwards (1992), Royall (1997), Mayo (2000), Royall (2004), Forster and Sober (2004), Fitelson (2007), Sober (2008), Mayo and Spanos (2011), Gelman and Shalizi (2013), Mayo (2013), Morey, Romeijn, and Rouder (2013), Reid and Cox (2015), Morey, Romeijn, and Rouder (2016), Gelman and Hennig (2017), Mayo (2018), Rouder and Morey (2019), and Fletcher and Mayo-Wilson (forthcoming) for a representative sample.

[5]See Smith and Seth (2020, 154). It is worth mentioning that Perrin performed at least three types of experiments on Brownian motion: vertical-gradient experiments, granule-displacement experiments and granule-rotation experiments. See Nye (1972) and Smith and Seth (2020, Ch. 4) for a detailed discussion of these experiments. In Psillos (2011) and Psillos (2014), Psillos focuses on the vertical-gradient experiments. I do not wish to claim that my analysis of the experiment I focus on extends to these experiments. I believe that these other experiments would require a different account to make sense of how or whether they provided strong evidence and what this evidence was for.

character of his evidence can be understood using *Bayes Factors* (see section 1.4 below for a discussion of Bayes Factors).

Nevertheless, one may have the following worry about my strategy. The worry is that Perrin himself makes no statistical arguments using Bayes Factors or the other common measures of statistical evidence, which I discuss in the next section. In fact, Perrin did not even calculate probable errors. So isn't my account ahistorical in the sense that it leaves out the actual context of Perrin's experimental work? This context involved "eye-balling" the data for conclusions as was done in almost all 19th century chemical and epidemiological research, despite the widespread knowledge of Laplace's work on probability and of least squares and its connection with Gaussian distributions.

In response to this worry, let me say that my goal is to show what a Bayesian statistical account, using Bayes Factors, of Perrin's experimental work *would look like* and not to argue that, in fact, this is what Perrin *did*. The advantage of the account which I intend to provide is that it can illuminate some of the existing accounts of the quality of Perrin's evidence that have been offered in the literature while avoiding some of their shortcomings. Consider, for example, Salmon's argument for scientific realism about atoms and molecules. According to Salmon (1984, 213 – 227), Perrin provided strong evidence for the atomic hypothesis by compiling, in *Les Atomes* (1913), converging values for Avogadro's Number from a variety of independent experiments. He argues that it would be "an utterly astonishing coincidence" to have values for Avogadro's Number from different independent experiments all converging to approximately the same value unless there was a common causal explanation — atoms and molecules. As I will show in what follows (see section 1.6 below), one can use Bayesian statistical analysis to illuminate the force of, and intuition underlying, Salmon's argument. The specific improvement to what Salmon did is that not only do I single out all the experiments that Salmon cites, but I also provide explicit statistical models for them and show how the experiments are linked using a Bayesian *meta-analysis*.

Now consider more recent authors. On the one hand, Mayo (1996, Ch. 7) has a compelling discussion, from a Frequentist perspective, of the severe testing and statistical reasoning involved in Perrin's confirmation of Einstein's model. She eschews any Bayesian characterization of this episode because Perrin's reasoning did not involve any explicit specification of prior credences.[6] On the other hand, Achinstein (2002), Psillos (2011) and Psillos (2014)'s arguments for scientific realism about atoms and molecules involve assumptions about what the value of the prior credences must be in order for their arguments to work. Smith and Seth (2020, 81, n. 13) find these assumptions ad hoc.[7] In what follows, unlike Mayo, I provide a Bayesian statistical perspective of Perrin's confirmation of Einstein's model. It is not my intention to criticize Mayo's illuminating error-statistical/severe testing perspective of Perrin's evidence; nor am I interested in the question of realism about atoms. My intention is to provide a correct retrospective Bayesian *statistical* perspective of Perrin's evidence for the first time (to the best of my knowledge) as an *alternative* perspective to Mayo's influential account of the same episode. At the same time, I will show how my approach avoids the objection of ad hoc specification of priors that has been raised against Achinstein and Psillos.

Here's how I have organized the rest of my paper. In the following section, I set the stage for what I mean by statistical evidence and why this matters for my account. This section is followed by another section with a detailed scientific and philosophical analysis of the reasoning or arguments involved in *specifying* the relevant theoretical models and statistical models that are involved in evaluating Perrin's evidence. In section 1.4, I give a quick overview of the Bayesian approach to statistical inference and Bayes Factors, which I use to characterize the quality of Perrin's evidence. In section 1.5 and 1.6 I briefly discuss the pros and cons of using Bayes Factors to quantify statistical evidence and draw some lessons

[6]See Mayo (1996, 232, 242).

[7]The problem of specifying prior credences is related to the problem of unconceived alternatives. It has been discussed by Roush (2005) and Stanford (2009) in connection to the problem of the "Catch-all Hypothesis". See section 1.6 below for how my account can avoid this problem.

for philosophy of science. Remaining questions for my account are addressed in section 1.7 before I conclude.

1.2 What is Statistical Evidence?

Before proceeding, let me say more about what I mean by "statistical evidence" and why I want to characterize Perrin's evidence from a Bayesian statistical perspective. Statistical evidence is a form of evidence, somewhere between the semi-formal and formal accounts on the spectrum just mentioned. First, it is *evidence* because it shares what is common to the genus of evidence, namely, a capacity to impact our epistemic attitudes towards a claim or our dispositions to act, which are influenced by claims that we accept. The disjunction in the preceding sentence is important because on a widely-held, standard and non-technical understanding of the meaning of "evidence", evidence has to do with belief and must always impact our epistemic attitudes or beliefs about a claim. I am not disputing this understanding. Rather, I am claiming that it is too narrow because: (i) it leads us to preclude certain items in the world as evidence; and (ii) it prevents us from making certain comparisons we would like to make, for example, between statistical evidence in classical statistics and statistical evidence in Bayesian statistics. To broaden our discussion of evidence, I propose to refer to the widely-held, standard account of evidence as the *epistemic* role of evidence and distinguish this from the *indicative* role of evidence such as making decisions using statistical evidence. Second, it is a *form* or *kind* because it can be distinguished as a species under the broader genus of evidence. The species of evidence have differentiating features, which are suggested by the modifier or adjective. Some of the species (non-exhaustive) of this genus are: direct evidence, historical evidence, indirect evidence, legal evidence, observational evidence, and propositional evidence. One view in epistemology, in fact, is that evidence is

propositional.[8] Propositional evidence is, roughly, a fact p — where p is a proposition — that confirms or justifies a given belief token.[9]

Of course there is some overlap between these species. Legal evidence is often propositional evidence since legal evidence is a fact that is admissible in a legal context such as a court of law. At the same time, legal evidence can include some pieces of historical evidence such as eye-witness reports or testimonies. Direct evidence often includes observational evidence such as seeing a smoking gun. Statistical evidence itself can offer indirect evidence for facts, i.e., propositional evidence.[10] One can try to make distinctions between all these kinds of evidence precise. But making these distinctions lies beyond the scope of this paper. I mention these various ways of talking about evidence to justify speaking of evidence more broadly and also insofar as it allows me to focus entirely on statistical evidence within *statistics*, as opposed to: (i) informal ordinary-life discussions of these other kinds of evidence in epistemology, and (ii) the formal discussions of evidence E, hypothesis H and theory T (where E, H and T are unqualified) that is the bread and butter of inductive logic or formal epistemology.

What distinguishes statistical evidence from evidence E in formal epistemology are two things: (i) the random character of statistical evidence, and (ii) the requirement of a statistical model of the observed data. In fact, these two distinguishing features of statistical evidence are linked. The random character of statistical evidence depends on the statistical model for data. Here's what I mean. A measure of statistical evidence is a real-valued function whose inputs are statistics. A statistic, by definition, is any function $t(\boldsymbol{x})$ of actual data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. In the context of *parametric* statistical inference, the data is modeled as a realization of a finite vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ of observations, measurements etc. For i in $1, 2, \ldots, n$, each X_i is a random variable that has a probability distribution function

[8] See Williamson (2000, Ch. 9, 10) and Brown (2015) for the notion of propositional evidence. See Joyce (2004) for a helpful summary and appraisal of Williamson's account of evidence.

[9] Williamson (2000, 194) writes, "Propositions are the objects of propositional attitudes, such as knowledge and belief; they can be true or false; they can be expressed relative to contexts by 'that' clauses."

[10] Compare with Mayo (2018, 435)'s discussion of direct and indirect uses of probabilities.

$f(X_i;\boldsymbol{\theta})$ that belongs to a *parametric family*, members of which are *identified* by the specific value of the parameter vector $\boldsymbol{\theta}$ they take. The model of the data is known as the *sampling model*. In the Frequentist approach to statistics, a statistical model just consists of the sampling model for the data. As we shall see in section 1.5 below, a Bayesian statistical model requires not just a sampling model for the data, but also a *prior model* on anything the data conditionally depend on that we would like to incorporate into our analysis. For example, suppose the experiment is to determine the coefficient of thermal expansion of a steel rod. Here the results of independent repeated measurements of the length of the rod (at a given temperature) can be modeled as the components of a vector $\boldsymbol{X}$ from the Normal or Gaussian parametric family of distributions. In the case of the normal or Gaussian parametric family, $\boldsymbol{\theta} = (\mu, \sigma^2)$ is a vector consisting of the mean (μ) and variance (σ^2) of the distribution. If $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ are the actual outcomes of those measurements, a statistic such as the mean $t(\boldsymbol{x}) = \bar{x} = \sum_i^n x_i/n$ is clearly a function of the data, which can be used in statistical inference: (i) to estimate μ; or (ii) to test hypotheses regarding μ.

Since a function of a random variable is a random variable, a statistic is a random variable. This means that a measure of statistical evidence is a random variable insofar as it takes random variables as inputs. This also means that regardless of what one's attitude towards the *evidential value* of p-values, confidence intervals, likelihood ratios, odds-ratios, Bayes Factors, and Mayo's Severity Function is; it remains the case that these are all measures of *statistical* evidence in the various schools of statistics on this picture.[11] To be sure, some of these measures do not have an *epistemic* evidential role.[12] The p-value, for example, is not a conditional probability assuming the null hypothesis is true. This is because on the Frequentist interpretation of probability, which is the interpretation that p-values are based on, statistical hypotheses are not repeatable events (so, statistical hypotheses cannot be modeled as random variables). Moreover, confidence intervals do not indicate our degree of belief or

[11] Mayo's Severity Function is discussed in Mayo (2018, 143ff.).
[12] Compare with Fletcher and Mayo-Wilson (forthcoming).

epistemic confidence about the boundaries within which a parameter might lie. The reason is that confidence intervals (which are based on the Frequentist interpretation of probability) are random variables while parameters (on the Frequentist approach to statistics) are unknown but not random.

The foregoing discussion is intended to restrict the scope of my paper and to provide some background to what I mean by "statistical evidence". By focusing on the evaluation of the quality of Perrin's evidence from a Bayesian statistical perspective, I am signaling two things. Firstly, I am signaling that I will be interested in Bayesian statistics not Bayesian formal epistemology. On the one hand, Bayesian statistics is one approach to statistical inference. Among the things that distinguishes Bayesian statistics, from say, classical Frequentist statistics, is that on the Bayesian approach one can consider prior distribution functions on parameters.[13] On the other hand, Bayesian formal epistemology or Bayesianism is a philosophy or school of thought that addresses questions in the theory of knowledge and confirmation theory. This philosophy has two distinguishing features: (i) an epistemic interpretation of probability as coherent graded beliefs or credences and (ii) the use of Bayes' theorem as an inductive rule through one form or another of conditionalization.[14] One reason for restricting my interest here is pragmatic, i.e., I don't have much to say about Bayesian formal epistemology. Another reason for focusing on Bayesian statistics is that Bayesian methods in epistemology sometimes mask the subtleties that underly actual Bayesian and non-Bayesian statistical modeling and inference. Given these reasons, I am signaling, secondly, my agreement with Mayo (2018, 73) who writes:

> [T]he Bayesian epistemologist invites trouble by not clearly spelling out corresponding statistical models. They seek a formal logic, holding for statements about radiation, deflection, fish, or whatnot. I think this is a mistake. That

[13]Barnett (1999) gives a good overview of the various paradigms of statistical inference. Compare with Bernardo and Smith (2000) and Part III of Bandyopadhya and Forster (2011). For the information-theoretic approach to model selection and statistical inference see Burnham and Anderson (2002).

[14]See Sprenger and Hartmann (2019).

doesn't preclude a general account for statistical inference; it just won't be purely formal.[15]

In actual Bayesian statistical modeling, judicious choices of suitable prior probability distributions (to represent prior ignorance, for example) and Markov chain Monte Carlo methods to compute posterior distributions, make Bayesian methods in statistics very subtle business.[16] Elaborate tools for model checking and diagnostics are also being advocated for.[17] This means that care must be taken in going from the famous example, which serves to motivate Bayes' theorem by exposing base-rate fallacies in diagnostic medical testing, and the use of "Bayesian methods" in philosophy of science, where the distinction between Bayesian statistics and Bayesian epistemology isn't always made.[18]

Finally, by focusing on evaluating evidence from a statistical perspective I can distinguish between the following levels:

(1) Substantive or fundamental theories

(2) Theoretical models of these substantive theories

(3) Statistical models of the theoretical models

I will return to this three-level distinction in section 1.4 and 1.7 below. For now, I summarize the description and relationships between these levels. Sentences in a theory (substantive or not) specify constraints. For example, in the theoretical model of a substantive theory such

[15] A point related to the one Mayo makes here can be made using Woodward's distinction between *data* and *phenomena*. Woodward argues that one needs to use statistical methods, at least in science, to analyze data before one can infer that they have evidence for the existence or nonexistence of a phenomenon. See Woodward (2011), Bogen and Woodward (1988), and especially Woodward (1989, 409). More recently, Norton in Norton (2003) and Norton (2021) has also expressed his opposition to an entirely formal account of inductive inference.

[16] See Efron and Hastie (2016, Ch. 13) for a good discussion.

[17] See Morey, Romeijn, and Rouder (2013).

[18] See Magnus and Callender (2004) for discussion of the base-rate fallacy in the context of the realism-antirealism debate in philosophy of science.

as kinematics, a constraint can be that there is linear relationship between velocity and time for a body in uniform motion. A statistical modeler will use data to evaluate the validity of the constraints using a statistical model that captures those theoretical constraints. Because data collected from measurement processes typically involve errors or uncertainties, we need to use statistical methods to handle these uncertainties using the statistical models before drawing inferences about the theoretical models or substantive theories. Typically, at least in science, it is statistical evidence at the level of statistical models that (indirectly) impacts our beliefs about theoretical claims or hypotheses at level (1) and level (2).

In what follows, I make use of this three-level distinction in characterizing what made the quality of Perrin's evidence for the discontinuity of matter good (see section 1.4 and section 1.7 below). There is a hydrodynamical theoretical model given by the Langevin equation at the level of theoretical models, and a statistical model for the granule displacements based on Einstein's diffusion model for Brownian motion at the level of statistical models. It is the statistical evidence supporting the Gaussian distribution of granule displacements that I use to argue that Perrin had obtained strong statistical evidence for the discontinuous structure of matter, which is assumed in the derivation, and solution, of the hydrodynamical model of Brownian motion. The discontinuous structure of matter in this hierarchy will be at the level of substantive or fundamental theories. I discuss all these interrelated parts of my account more fully in what follows, especially in section 1.7.

1.3 Perrin's Evidence

1.3.1 Einstein's Diffusion Model

In the previous section, I have said what I mean by statistical evidence and why that matters for my account. Here I want to say what Perrin's statistical evidence was. For this I need to

say what the statistical model is and how it was supported by the data or observations that Perrin made. In this subsection I discuss what the statistical model is, in the next subsection I discuss what theoretical model it is a model of. In the next section I say why the evidence for this statistical model provided strong evidence for the theoretical model it is a model of.

The statistical model in Perrin's granule experiments on Brownian motion was suggested by Einstein's 1905 diffusion model for Brownian motion.[19] Assuming the equipartition of energy among the three degrees of freedom and Van't Hoff's Law that extends the ideal gas law to dilute solutions; Einstein's diffusion model made a prediction for how many granules would be displaced from mean position (the origin) after a given time due to osmotic pressure. The prediction was that the number n_i of suspended granules between two fixed points a and b on the the x-axis that would be displaced from the mean (the origin) after a given time t would be given by the following formula.[20]

$$n_i = n \times \int_a^b \frac{1}{\sqrt{2Dt}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x^2}{2Dt}\right)} dx$$

The integrand is the familiar Gaussian or normal probability distribution function. This formula says that the predicted n_i is given by multiplying the area under this function (the definite integral between a and b), which is a probability, with the total number of suspended granules n. In other words, Einstein's diffusion model for Brownian motion predicted that the statistical model of the displacements of the n suspended granules is Gaussian or normal with a mean of 0 and a variance or mean squared displacement $\xi^2 = 2Dt$.[21] D is the diffusion coefficient and t is the time interval. Using Einstein's model, Perrin and his graduate student M. Chaudesaigues made the calculation for $t = 30$ seconds of the predicted number n_i of

[19] See Einstein (1905)'s "On the Movement of Small Particles Suspended in a Stationary Liquid Demanded by the Molecular-Kinetic Theory of Heat" translated by A. D. Cowper and reprinted in Einstein (1956, 1 – 17).

[20] Although the displacements happen on the surface of the liquid, the equipartition of energy allows us to take projections of this displacement on the x-axis and analyze the displacement there.

[21] The notation of ξ^2 instead of σ^2 for mean squared displacement follows Smith and Seth (2020).

Projections (in μm) comprised between	First Series		Second Series	
	n_i (found)	n_i(calculated)	n_i(found)	n_i(calculated)
0 and 1.7	38	48	48	44
1.7 and 3.4	44	43	38	40
3.4 and 5.1	33	40	36	35
5.1 and 6.8	33	30	29	28
6.8 and 8.5	35	23	16	21
8.5 and 10.2	11	16	15	15
10.2 and 11.9	14	11	8	10
11.9 and 13.6	6	6	7	5
13.6 and 15.3	5	4	4	4
15.3 and 17.0	2	2	4	2

Table 1.1: Calculated and observed n_i in two series of experiments

gamboge particles that would be displaced within intervals that were multiples of $1.7\mu m$. They then recorded the number n_i of particles observed within these intervals alongside their predicted values in Table 1.1.[22]

The data above show a close match between predicted values by Einstein's model and observed values from Perrin's experiments. They confirm a statistical model. Specifically, the data confirmed that the statistical model for the displacements of n suspended granules on the x-axis after time t is Gaussian with a mean of 0 and variance (or *mean squared displacement*) of $\xi^2 = 2Dt$. It is standard practice among statisticians (both Bayesian and Frequentist) to write this statistical model using the following shorthand:

$$x(t) \sim N(0, 2Dt)$$

[22]This table is taken from Perrin (1910, 64 – 65).

Here $x(t)$ denotes the displacement on the x-axis after time t. This shorthand is read as "The displacements on the x-axis are normally distributed with a mean of 0 and variance of $2Dt$." I shall make use of this shorthand freely in what follows.

But what is the significance of this confirmation? Since Perrin's experimental work was geared towards confirming the molecular-kinetic theory, the significance of this confirmation lies in what can be inferred to exist on the basis of this evidence. The questions now are: (1) what can be inferred to exist? And (2) how can we understand the quality of the evidence that justifies us in making those inferences? The answer to the first question involves showing that there must be some force sustaining the motion of the particles in Brownian motion. The answer to the second question involves making assumptions about the nature of this force and checking whether the nature of this force, as assumed, is supported by Perrin's evidence for Einstein's diffusion model. It will emerge in what follows that assumptions regarding the nature of this force can be used to specify *two* competing hydrodynamical models (models based on forces due to the surrounding liquid on which the Brownian particles are suspended). One of these hydrodynamical models is compatible with the liquid behaving like a continuum fluid at a scale immediately below that of Brownian particles, while the other is incompatible with the liquid behaving like a continuum fluid at that scale. It is this specification of competing models that I will use to support my argument that Perrin's evidence was good because it was highly specific and discriminating.

For these reasons, confirmation of the predicted statistical model for granule displacements from Einstein's diffusion model is important for my argument because the statistical evidence Perrin obtained supporting the statistical model, can be construed as specific and discriminating evidence for a *hydrodynamical* model, which is formulated assuming the existence of the force sustaining Brownian motion. To this end, in subsections 1.3.2 – 1.3.3, I will be concerned with answering the first question by analyzing the reasoning involved in making

the inference about what exists. In section 1.4 I will give my answer to the second question, which is: "why was the evidence that warranted this inference good?"

1.3.2 There are highly localized and irregular pressure fluctuations

In order to show that there must be some force sustaining the motion of the particles in Brownian motion, I will refer to Einstein's 1907 paper written in response to Svedberg's 1906 publication of the results he had obtained concerning Brownian motion.[23] This paper is important for my argument because one of the arguments in it is that there must exist random impulsive forces acting on suspended granules if hydrodynamics is to be reconciled with the kinetic theory of heat.[24] From a hydrodynamical point of view — the appropriate level of description of Brownian motion — the best candidate for these impulsive forces are *highly localized and irregular pressure fluctuations*.[25] Therefore, this paper contains the answer to our first question, which is: what can be inferred to exist? In this subsection I summarize the reasoning involved in Einstein's argument.

[23]The Einstein paper is reprinted in Einstein (1956, 63 – 67) as "Theoretical Observations on the Brownian Motion."

[24]Compare with Munson, Young, Okiishi, and Huebsch (2009, 97) for the modern hydrodynamical modeling of pressure.

[25]Einstein was indeed using background information in order to arrive at an informative prior specification of the model space of Brownian motion. Some of this background information can be traced back to the work of the French physicist Léon Gouy. In 1889 and 1895, Gouy had performed careful experiments and published his results of these experiments thereby considerably narrowing down the space of theoretical explanations for Brownian motion. Nye writes, "Gouy excluded all exterior causes except the internal agitation of a liquid, and stated that [Brownian movement] is a direct and visible proof of the modern hypothesis of the nature of heat." See Nye (1972, 27 – 29). As will emerge below, the model space for Brownian motion can be narrowed down even further into a mutually exclusive and exhaustive set of two models: a continuity of matter model vs. a discontinuity of matter model.

From the kinetic theory of heat, the mean velocity $\bar{v}$ of a suspended particle of mass m can be determined using

$$m\frac{\overline{v}^2}{2} = \frac{3}{2}k_B T \tag{1.1}$$

$k_B = \frac{R}{N_A}$ is Boltzmann's constant, R is the ideal gas constant, N_A is Avogadro's Number, and T is the absolute thermodynamic temperature. Equation (1.1) expresses the familiar idea that temperature is proportional to average kinetic energy. For particles in colloidal platinum solutions such as the ones Svedberg had prepared $\bar{v} = 8.6\text{cms}^{-1}$.[26]

Now suppose that the *only* force acting on the suspended particle undergoing Brownian motion is a viscous drag force, i.e., a force due to liquid friction that decelerates the particle. Newton's second law of motion for this particle of mass m is:

$$m\frac{dv}{dt} = -\zeta v \tag{1.2}$$

$\zeta = 6\pi r\eta$ is a damping term (i.e., a term that determines the rate of deceleration) that depends on η the viscosity of the liquid and r the radius of the spherical particle. This equation ignores the inertia of the particle and says that the dynamics of the suspended particle is only governed by Stokes' Law.[27]

If (1.2) is the hydrodynamical law governing Brownian motion, one can show that for a colloidal platinum particle suspended in water, it would take 3.3×10^{-7} seconds for it to lose one-tenth of its velocity. This means that at the macroscopic time-scale of a laboratory

[26]The assumptions in Einstein's calculations here are: $m = 2.5 \times 10^{-15}g$, $k_B = 1.38 \times 10^{-23}m^2kgs^{-2}K^{-1}$ and $T = 292K$.

[27]See Munson, Young, Okiishi, and Huebsch (2009, 493 – 500) for a derivation of Stokes' Law and the validity of the assumption that inertia can be ignored.

measurement (about 30 seconds), the particle would have lost almost all of its velocity through hydrodynamic friction or viscous drag.

But since Brownian motion was known to be incessant, equation (1.2) is not the true dynamical law governing the motion of a particle undergoing Brownian motion. This means that in order to maintain the average $\overline{v}$ demanded by the kinetic theory of heat (8.6cms^{-1}), the suspended particle must experience rapid impulses from *somewhere*. Here's how Einstein (1956, 66) puts it:

> We have to modify this conception [equation (1.2) above], we must assume that the particle gets new impulses to movement during this time by some process that is the inverse of viscosity, so that it retains a velocity which on an average is equal to $\sqrt{\overline{v^2}}$.

This modification was implemented by Langevin in 1908 leading to the celebrated *Langevin equation*.[28]

$$m\frac{dv}{dt} = -\zeta v + F(t) \tag{1.3}$$

In arguing that $F(t)$ exists, one cannot beg the question by assuming a priori that it is due to molecular impacts.[29] Notice as well that Einstein only says that we must assume "new impulses by some process." Compare this with how Smith and Seth (2020, 236) describe the situation:

> Pressure-gradients must be present in the liquid even though it is in thermodynamic equilibrium. The local pressure-gradients must be associated with and

[28] See the translation of the Langevin paper in Langevin (1997).

[29] See Smith and Seth (2020, 237ff) for some of the arguments why one must not assume this. Compare with Stanford (2009, 257) who discusses Roush (2005)'s modest atomic hypothesis.

hence arising in conjunction with highly localized, extraordinarily rapid pressure fluctuations occurring continually throughout the liquid.

Therefore, the most one can say, from a hydrodynamical point of view, is that the source of $F(t)$ are *pressure impulses* or fluctuations on the suspended granules from the ambient fluid. These fluctuations happen locally, on extremely short time-scales, and haphazardly.

1.3.3 Evidence for discontinuity

In the previous subsection, I have answered the first question, namely: what can be inferred to exist? I showed that $F(t)$ must be included in an accurate hydrodynamical model for Brownian motion. I now want to give an answer to the second question, namely: why was the quality of Perrin's evidence good? This involves showing that the confirmation of the Gaussian distribution variance $\xi^2 = 2Dt$ of the statistical model provides strong statistical evidence for a hydrodynamical model of Brownian motion that includes $F(t)$. But such a hydrodynamical model implies the discontinuity of matter at a level *immediately* below that of the suspended granules. Therefore, the main goal of this subsection is to show how Perrin's statistical evidence confirming the statistical model is also evidence for the discontinuity of matter.[30] I will use the implied discontinuity of matter by a hydrodynamical model that includes $F(t)$ in my argument that Perrin's evidence was specific and discriminating in section 1.4.

[30]See Stein (2021) where my framing of the Perrin case is very close to Poincaré's understanding of the atomic debates. The issue, as Poincaré saw it, was to decide between continuous and discontinuous approaches to chemical physics.

First, let me rewrite equation (1.3) above explicitly in terms of $x(t)$, i.e., displacement in the x-direction.

$$m\frac{d^2x}{dt^2} = -\zeta\frac{dx}{dt} + F(t) \tag{1.4}$$

Langevin solved this equation for the mean squared displacement $\overline{x(t)^2} = \xi^2 = 2Dt$ in an "infinitely more simple" way than Einstein. The solution of this equation obtained by Langevin himself (and more rigorously by others after him) was obtained by making two kinds of assumptions. One kind was based on the kinetic theory of heat, namely, that at thermal equilibrium the distribution of the velocities of suspended granules will be the Maxwell-Boltzmann distribution. This was a simplifying assumption since it meant that one could use $m\frac{\overline{v}^2}{2} = \frac{1}{2}k_BT$ instead of individual velocities for each particle. The other assumptions were statistical assumptions about $F(t)$. In fact, in solving the eponymous equation by making the following statistical assumptions about $F(t)$, Langevin was the first to employ methods which were later used for solving what are now called *stochastic differential equations*.[31]

(a) $\overline{F(t)} = 0$

Since $F(t)$ is a fluctuating force on the surface of a fluid at rest in thermal equilibrium, assumption (a) says that $F(t)$ must have zero mean even though it can vary widely and wildly across the surface of the fluid on very short timescales.

(b) $\overline{\langle F(t), F(t')\rangle} = 2\zeta k_BT\delta(t'-t)$

[31]For a rigorous discussion of the mathematics and methods involved in solving the Langevin equation, see Mazo (2002) and Coffey, Kalmykov, and Waldron (2004).

$\overline{\langle \cdot, \cdot \rangle}$ is the autocorrelation function in Mazo (2002)'s notation. By making it proportional to the Dirac delta function ($\delta(t' - t)$), assumption (b) says that the fluctuating forces are sharp, rapid and uncorrelated at different but very short timescales.[32] Some authors take condition (a) and (b) together as implying that $F(t)$ is a Gaussian white noise process.[33]

(c) $\overline{\langle F(t), x(t) \rangle} = 0$

Finally, assumption (c) says that the fluctuating impulsive pressure forces are independent of position.

The key point here is that by assuming the kinetic theory of heat and using these three statistical assumptions about $F(t)$, Langevin solved equation (1.4) for the mean squared displacement $\overline{x(t)^2}$ obtaining $\overline{x(t)^2} = \xi^2 = 2Dt$. Now recall that Perrin and M. Chaudesaigues used Einstein's diffusion model to predict n_i. The very close match between observed n_i and predicted n_i that Perrin tabulated was evidence that the distribution of n_i on the x-axis is Gaussian with mean 0 and variance $\xi^2 = 2Dt$. Since the variance of this distribution can be obtained from the Langevin dynamical equation, evidence for Einstein's model is evidence for a hydrodynamical model of Brownian motion in which $F(t)$ has the properties (a) – (c).[34] But such a hydrodynamical model implies that the substructure of the ambient fluid at a scale immediately below that of the granules is discontinuous.

In order to see this discontinuity, recall that according to continuum fluid mechanics, the pressure $F(t)$ depends on, or is related to, position $x(t)$. This follows because the Navier-Stokes Equations together with the Continuity Equation are a set of four equations in four unknowns: three equations involving flow velocity components in the three spatial directions

[32] See Arfken and Weber (2001, 84 – 88) for discussion of the Dirac delta function.

[33] See Gardiner (1983, 69) and compare with Mazo (2002, 59 – 63) and Coffey, Kalmykov, and Waldron (2004, 12).

[34] Actually only properties (b) and (c) are required to infer the discontinuous structure of matter at a scale immediately below those of the granules.

(Navier-Stokes) and one equation involving pressure gradients (the Continuity Equation).[35] This means the system of equations is determined. In particular, given $x(t)$ in a fluid and the flow velocities at $x(t)$ from Navier-Stokes, one can solve for $F(t)$ although a solution cannot always be obtained by analytical means. The existence of this solution assumes that $F(t)$ and $x(t)$ are *related* in the way specified by the system of equations. Since this system of equations characterizes the behavior of *continuous* fluids, the assumption that $F(t)$ is uncorrelated with $x(t)$ (assumption (c) above) is *incompatible* with the substructure of the ambient fluid surrounding a particle undergoing Brownian motion *having* a continuous description. Therefore, this substructure, at a scale immediately below the granules, must be discontinuous.[36]

Further, from the Continuity Equation, we know that pressure at one point in a fluid is transmitted to adjacent points continuously according to the pressure gradients within that fluid. This means that given some pressure gradient, the pressure $F(t)$ at a point $x(t)$ and time t *determines* the pressure $F(t')$ at a different but adjacent point $x'(t')$ at time t'. But the assumption that $F(t)$ and $F(t')$ are uncorrelated (assumption (b) above) at short-time scales is incompatible with the substructure of the ambient fluid being continuous at a scale below that of the granules.[37] Thus, obtaining data verifying $\xi^2 = 2Dt$ would be very unlikely assuming continuity of the ambient fluid at that scale. The fluid must be discontinuous at that scale.

Before moving on to the next section, let me be clear in order to avoid potential misunderstanding. It is known that hydrodynamics assuming continuum fluid mechanics depends on the type of fluid and the scale of the physical processes being modeled. Hydrodynamics may be formulated assuming a continuous fluid description (where Navier-Stokes and the Continuity Equation apply), if the associated *molecular mean free path* λ is small compared

[35] See Munson, Young, Okiishi, and Huebsch (2009, 42, 271).
[36] Compare with Smith and Seth (2020, 237)
[37] Compare with Smith and Seth (2020, 238).

to a *typical length scale of the problem L*. The mean free path is the mean distance traveled by a molecule of the fluid between collisions with other molecules. A quantitative measure for identifying the scales at which continuity applies is provided by the dimensionless ratio $Kn = \frac{\lambda}{L}$ known as *Knudsen number*. So in saying that the assumption of continuity of the ambient fluid at that scale is incompatible with the data verifying $\xi^2 = 2Dt$, what I mean is that scales or levels of description matter. At the scale at which Brownian Motion is happening the assumption of continuity is no longer valid because Kn approaches 1, which is considered the upper limit of the continuum hypothesis.[38] The non-zero $F(t)$ is introduced at the microscale where Brownian motion is happening by the Langevin equation and the statistical assumptions for $F(t)$ only apply at this scale. But at this scale or level of description, because the pressure fluctuations $F(t)$ are uncorrelated with position $x(t)$ and at different times, a shift of perspective to incorporate discontinuity is required to explain the data from the experiments.

A concern one could still have here is that continuous models could allow for forms of autocorrelation and spatial distribution of $F(t)$ (assumption (b) and (c) above) that are very close to those assumed in solving the Langevin equation. If so, these models would provide a similar fit for the data as Einstein's diffusion model for Brownian motion did. My response to this concern is that it will depend on what the continuous models actually propose. Coming up with this proposal is not an easy task. For at the scale where Kn approaches 1, the continuity assumption is no longer valid. At best, what we can do here is speculate about the possibilities. For historical reasons, we have no account of an explicit alternative statistical model that is formulated on the assumption of continuity at the scale of Brownian motion. If we had one, then it is plausible that the evidence would bear them out equally well — this was one reason for Poincaré's conventionalism about the atomic hypothesis, after all.[39] But we do not know what the model of Brownian motion assuming

[38] See section 1.2.2 in Katopodes (2018) and compare with section 9.9.2 in Rapp (2022).
[39] See Stein (2021).

continuity of the ambient fluid at that scale would *actually* look like. What we know for sure is that it cannot have D, the diffusion coefficient, in it because D is a function of Avogadro's number.

1.4 Why was Perrin's evidence good?

The previous section was helpful in specifying the hydrodynamical model space, which I will use to characterize the quality of Perrin's evidence using Bayes Factors. Before offering this characterization, let me first give an overview of what Bayes Factors are.

On a Bayesian statistical approach, the first thing to note is that we are justified (by the theorems of de Finnetti and later Diaconis and Freedman (1980) on *exchangeability*) to speak of probability distributions on *parameters* — effectively treating parameters as random variables. This turns out to have a huge payoff because it naturally leads to *hierarchical* models, where we use higher level probability models of the parameters, which appear on lower level sampling models of the data, to capture dependencies in our data, especially where such dependencies seem reasonable enough to capture. Using a hierarchical model, a Bayesian statistician will not only use the sampling model for the data $f(X_i|\boldsymbol{\theta})$ for $i = 1, 2, \ldots, n$; but also a prior model $\pi(\boldsymbol{\theta})$ for all the parameters $\boldsymbol{\theta}$ in the sampling model. Such a statistician may even have a prior on the models $p(M_j)$ for $j = 1, 2, \ldots, k$ themselves, possibly continuing this hierarchy as high up as they please using *hyperparameters*. We shall see in section 1.5 below how a Bayesian hierarchical model can be used to analyse Salmon (1984)'s argument.

For now, in order to illustrate the fundamental ideas of a Bayesian statistical model and for the sake of my subsequent argument, suppose that we have observed data $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Suppose also that we are just comparing two theoretical models M_1 and M_2. For each model

M_1 and M_2 there is a corresponding Bayesian statistical model, which has two parts. The first part, which is common to both Frequentist and Bayesian approachers, is a sampling model for the data. We write the sampling models as:

$$X_i \,|\, \boldsymbol{\theta}_1, M_1 \sim f(X_i \,|\, \boldsymbol{\theta}_1, M_1) \quad \text{for } i = 1, 2, \ldots, n$$
$$X_i \,|\, \boldsymbol{\theta}_2, M_2 \sim f(X_i \,|\, \boldsymbol{\theta}_2, M_2) \quad \text{for } i = 1, 2, \ldots, n$$

Here $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are a set of parameters in the statistical models associated with model M_1 and model M_2. The second part of a Bayesian statistical model, which is its distinguishing feature from a Frequentist statistical model, is a prior model on the parameters in the respective sampling models. We write the prior models as:

$$\boldsymbol{\theta}_1 \,|\, M_1 \sim \pi(\boldsymbol{\theta}_1)$$
$$\boldsymbol{\theta}_2 \,|\, M_2 \sim \pi(\boldsymbol{\theta}_2)$$

Now suppose that $p(M_1)$ and $p(M_2)$ are the prior probabilities on the theoretical models, then Bayesian statistical inference uses Bayes' theorem to find the posterior probabilities on the theoretical models as:

$$p(M_1 \,|\, \boldsymbol{\theta}_1, \boldsymbol{X}) = \frac{L(\boldsymbol{X} \,|\, \boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1 \,|\, M_1)p(M_1)}{\sum_{i=1}^{2} L(\boldsymbol{X} \,|\, \boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i \,|\, M_i)p(M_i)}$$
$$p(M_2 \,|\, \boldsymbol{\theta}_2, \boldsymbol{X}) = \frac{L(\boldsymbol{X} \,|\, \boldsymbol{\theta}_2, M_2)\pi(\boldsymbol{\theta}_2 \,|\, M_2)p(M_2)}{\sum_{i=1}^{2} L(\boldsymbol{X} \,|\, \boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i \,|\, M_i)p(M_i)}$$

The terms $L(\boldsymbol{X} \mid \boldsymbol{\theta}_1, M_1)$ and $L(\boldsymbol{X} \mid \boldsymbol{\theta}_2, M_2)$ that appear in the expression for finding the posterior probabilities are the joint distributions of the data $\boldsymbol{X}$ under the respective models. These terms are called *likelihood functions* by statisticians.[40] Marginal likelihood functions involve integrating out the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in their respective parameter spaces Θ_1 and Θ_2. That is:

$$L(\boldsymbol{X} \mid M_1) = \int_{\Theta_1} L(\boldsymbol{X} \mid \boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1 \mid M_1)d\boldsymbol{\theta}_1$$
$$L(\boldsymbol{X} \mid M_2) = \int_{\Theta_2} L(\boldsymbol{X} \mid \boldsymbol{\theta}_2, M_2)\pi(\boldsymbol{\theta}_2 \mid M_2)d\boldsymbol{\theta}_2$$

So we may write the posterior probabilities of M_1 and M_2 as:

$$p(M_1 \mid \boldsymbol{X}) = \frac{L(\boldsymbol{X} \mid M_1)p(M_1)}{\sum_{i=1}^{2} L(\boldsymbol{X} \mid M_i)p(M_i)} \tag{1.5}$$
$$p(M_2 \mid \boldsymbol{X}) = \frac{L(\boldsymbol{X} \mid M_2)p(M_2)}{\sum_{i=1}^{2} L(\boldsymbol{X} \mid M_i)p(M_i)} \tag{1.6}$$

when we are interested in the posterior probabilities just given the data $\boldsymbol{X}$.

Dividing equation (1.5) by equation (1.6) we get:

$$\frac{p(M_1 \mid \boldsymbol{X})}{p(M_2 \mid \boldsymbol{X})} = \frac{L(\boldsymbol{X} \mid M_1)p(M_1)}{L(\boldsymbol{X} \mid M_2)p(M_2)} \tag{1.7}$$

[40]Note that for a Frequentist, the likelihood functions are not conditional distributions; rather the likelihood functions are joint distributions of the data for a fixed value of the parameters $\boldsymbol{\theta}$. For this reason, a Frequentist will write the likelihood function as $L(\boldsymbol{X}; \boldsymbol{\theta}, M)$ to distinguish her approach from the Bayesian approach.

Rearranging the terms in equation (1.7) suitably we get:

$$\frac{\frac{p(M_1 \mid \boldsymbol{X})}{p(M_1)}}{\frac{p(M_2 \mid \boldsymbol{X})}{p(M_2)}} = \frac{L(\boldsymbol{X}|M_1)}{L(\boldsymbol{X}|M_2)} \tag{1.8}$$

The right hand side of equation (1.8) can be used to quantify the *relative predictive accuracy* of our models.[41] This quotient is the *Bayes Factor*. The left hand side of equation (1.8) quantifies the ratio with which our credences for each model *updated* given some data $\boldsymbol{X}$. The equality between the left and right hand side of equation (1.8) connects the *relative strength of evidence* — how much we are led to update our credences for the competing models given some data — with Bayes Factor, which is the relative predictive accuracy of our models.

A more intuitive way of thinking about the Bayes Factor is this: Bayes Factor quantifies the relative strength of our evidence for a given model in terms of the *specificity* and *discriminating character* of that evidence. This way of thinking is valid because there are two ways that the Bayes Factor can be high: either the numerator is high relative to the denominator or the denominator is low relative to the numerator. On the one hand, we can say that a relatively high value for the numerator quantifies the specificity of our evidence given M_1. A relatively high value is in effect saying that this is the sort of evidence that is *very likely* given M_1 when comparing it to M_2. The evidence is more specific to M_1 than M_2. On the other hand, we can say that a relatively low value for the denominator quantifies how discriminating our evidence is. A relatively low value indicates that this sort of evidence or data is *very unlikely* given M_2 when comparing it to M_1. The evidence discriminates against, or rules out, M_2. In sum, we can say that since the strength of our evidence is reflected by how much we are led to update our relative credences in light of it (the left hand side), the

[41] See Rouder and Morey (2019).

evidence in favor of some model is strong or good if it is specific and discriminating (the right hand side).

Here's how I apply Bayes Factor to evaluate the quality of Perrin's statistical evidence for the Gaussian distribution of displacements. Let M_1 be a hydrodynamical model for Brownian motion which includes $F(t)$ and in which the ambient fluid is discontinuous. The alternative M_2 is a hydrodynamical model for Brownian motion which does not include $F(t)$ with its assumed properties, i.e., the substructure of the ambient fluid even at a level immediately below that of the granules is a continuum. M_1 leads to the statistical model discussed in section 1.3.1. With this model specification, let the marginal likelihoods under M_1 and M_2 be $L(\boldsymbol{X} \mid M_1)$ and $L(\boldsymbol{X} \mid M_2)$, respectively.

On the one hand, $L(\boldsymbol{X} \mid M_1)$ is very high relative to $L(\boldsymbol{X} \mid M_2)$ because Perrin's evidence confirming that the statistical model of the Gaussian distribution of displacements has variance $\xi^2 = 2Dt$ was very specific under the assumption that matter is discontinuous at the scale of the particles undergoing Brownian motion, i.e., under M_1. By "specific" here I am referring to the close match between predicted values by the model and actually observed values in the two series of experiments (see Table 1.1). The specificity also arises because the Langevin dynamical equation leads to the variance $2Dt$ of the statistical model of displacements. Since D is related to Avogadro's number N_A by

$$D = \frac{RT}{N_A 6\pi r\eta}$$

the specificity of Perrin's evidence for the discontinuity of matter comes from here as well since we can use the estimated D from the distribution of displacements to estimate N_A and compare it to its theoretical value (see section 1.6 below).

On the other hand, because the Langevin dynamical equation in which $F(t)$ has the properties (a) – (c) is *incompatible* with the ambient fluid being a continuum, $L(\boldsymbol{X} \mid M_2)$ is very low

relative to $L(\boldsymbol{X} \mid M_1)$. It is also relatively low because one of the parameters in the associated statistical model, D, is a function of N_A — a key quantity in the molecular-kinetic theory of gases and heat. But since we are assuming that M_2 is formulated within an alternative theoretical framework to the molecular-kinetic theory of gases and heat, M_2 cannot include a function of N_A as a parameter. And any theoretical model that did not include a function of N_A as a parameter would be incompatible with the statistical data on displacements, meaning the likelihood of that data given that model would be very low relative to an alternative that includes N_A.

Let me say a bit more here. My claim is that there is no other way to get the predicted mean square displacements unless one uses a hydrodynamical model with an $F(t)$ term satisfying conditions (a) – (c). Whether this term is needed to make sense of the data is the entire question at issue, and it is by showing that this is needed that Perrin establishes the relevant facts, summarized by "atoms exist" or the determination of N_A. In other words, I am claiming that the close match between theoretical predictions of N_A and actual estimates of N_A from statistical experiments would be very unlikely on theoretical models for Brownian motion that either don't include an $F(t)$ term or explicitly make use of a function of N_A in their derivations. An example of a theoretical model in this category would be Ostwald's *energetics*.[42]

At the same time, I believe that a continuum theorist (working in, say, continuum fluid dynamics) can still accept the statistical assumptions on $F(t)$ because she recognizes that Knudsen number is close to one at the scale of Brownian motion. So nothing I have said impugns continuum fluid dynamics. The pressure gradients at appropriate levels of descriptions required by continuum fluid dynamics are in fact smooth (or at least continuous) and the hydrodynamical phenomena that can be modeled faithfully by the Navier-Stokes and the Continuity Equation at macroscales can still be modeled as such. So it is true that one both

[42]Smith and Seth (2020, 333 – 341)

has a continuum description of fluid dynamics at the macroscale, but also hydrodynamical models at the scale of Perrin's experiments on Brownian motion which, in fact, show that matter is discontinuous.

Taking together these two points about the relatively high and relatively low values of $L(\boldsymbol{X} \mid M_1)$ and $L(\boldsymbol{X} \mid M_2)$ respectively, it follows that $\frac{L(\boldsymbol{X} \mid M_1)}{L(\boldsymbol{X} \mid M_2)} \gg 1$, i.e., the resulting Bayes Factor comparing M_1 to M_2 is much greater than 1. But this large Bayes Factor means that the quality of Perrin's evidence for the discontinuity of matter at a scale immediately below that of the suspended granules was good because it was highly specific and discriminating.

Before moving on to the next section, let me address an apparent tension that may occur to a reader who has followed my account so far. In the beginning I claimed that what is novel in my account is that I approach the evaluation of Perrin's evidence from a statistical perspective using Bayes Factors. But the details of the Bayes factor "calculation" I have given here don't involve any actual *computation* of statistics. It seems that all that matters to my account is the qualitative, even logical, compatibility of the data with different hypotheses. While I acknowledge that there are priors over the diffusion coefficient that get integrated out in getting the marginal likelihood functions $L(\boldsymbol{X} \mid M_1)$ and $L(\boldsymbol{X} \mid M_2)$, I have not said what these priors would be *for a continuous model*. Could the value of the Bayes Factor change so dramatically, depending on what the priors are, that the continuous model is equally supported by the data?

My response to this apparent tension is to return to the three-level distinction I introduced in section 1.2:

(1) Substantive or fundamental theories

(2) Theoretical models of these substantive theories

(3) Statistical models of the theoretical models

The continuous structure of matter vs. discontinuous structure of matter is at the level of substantive or fundamental theories (Level (1)). The hydrodynamical model given by the Langevin equation is a theoretical model at Level (2) of the substantive theory that matter is discontinuous. By checking the predictions made by the theoretical model using a statistical model at Level (3), we can tell whether the substantive theory is supported by the data.

Now the reasoning using Bayes Factors is for M_1 and M_2 at Level (2). Without an explicit theoretical model (at Level (2)) of the substantive theory that matter is continuous; no statistical model, which would include prior specifications, can be formulated to check this statistical model's prediction using data. It is for this reason that there are no actual statistics that can be computed using Bayes Factors from this historical case. Although no computation can be done, I have reconstructed what the *reasoning* during this episode could have been like from a Bayesian perspective. In my discussion above I have said what the Bayes Factor *would be* for dichotomous alternatives M_1 and M_2 given the actual data $\boldsymbol{X}$ in Table 1.1. What is important for my paper is that data in this *table* is statistical. It shows a close match between predicted values and observed values for the *statistical model* of M_1. I have argued that the data from this table is primarily statistical evidence for M_1 against an alternative M_2. So the tension is apparent and it is not quite true that I have abandoned the initial framing of my paper. Whether this statistical evidence translates into evidence *simpliciter* for the discontinuity of matter at the scale of Brownian motion depends on whether one believes that M_1 is an adequate theoretical model of that substantive theory.

1.5 Pros and Cons of using Bayes Factors

Although the Bayes Factor looks like a *likelihood ratio statistic* commonly found in Frequentist hypothesis testing, it is important to emphasize that Bayes Factor is *not* simply a likelihood ratio statistic. I mention this partly to justify my choice of analyzing Perrin's evidence

from a Bayesian perspective as opposed to a non-Bayesian or Frequentist perspective. The first distinguishing feature is that we got Bayes Factor from the ratio of *marginal* likelihoods — not likelihoods — by integrating out the parameters. This technique of marginalization is not possible within a Frequentist framework where prior models on parameters, which are required in order for this to work, are not considered. This technique turns out to offer the Bayes Factor more flexibility to compare all sorts of models with each other than is possible within the Frequentist framework. There are thus several advantages to using Bayesian methods to quantify the relative evidence we have for *any* theoretical models. The first advantage is that the probability distributions in the sampling models given the theoretical models do not necessarily have to belong to the same parametric family as is typically the case in the case of likelihood ratio based statistics in the Frequentist framework. Secondly, the vector of parameters are not necessarily nested, again, as is typically the case in other likelihood ratio based test statistics in the Frequentist framework.

These clear advantages have to be tempered with some of the well known disadvantages of using Bayes Factors for model comparison or quantifying the strength of evidence. First, Bayes Factors clearly depend on the prior model on parameters. We can see this by looking at how we calculated the marginal likelihoods for each model M_i for $i = 1, 2$

$$L(\boldsymbol{X} \mid M_i) = \int_{\Theta_i} L(\boldsymbol{X} \mid \boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

Now the problem for Bayes Factors has to do with using *uninformative* priors on the parameters. An uninformative prior is a prior that is chosen in such a way that its influence on the posterior distribution is as small as possible. With an uninformative prior, statisticians want to eliminate as much bias as they can from their analysis. A typical uninformative prior on a parameter such as the mean μ of a continuous random variable X is a uniform distribution $\pi(\mu)$ over the entire real line $\mathbb{R}$, also known as a flat prior. The problem is that

$B_{12} = \frac{L(\boldsymbol{X}\|M_1)}{L(\boldsymbol{X}\|M_2)}$	$2\log_e(B_{12})$	Evidence against M_2
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

Table 1.2: Guidelines for interpreting the magnitude of Bayes Factor

the uninformative prior in this case turns out to be an *improper prior*. The sense in which it is improper is that it is not a probability distribution function since it is not normalized, i.e., $\int_{\mathbb{R}} \pi(\mu) d\mu = \infty$. So the marginal likelihood functions are undefined in this case and so is the Bayes Factor. For this reason, Bayes Factors are highly sensitive to the prior model on *parameters*.

Another problem with Bayes Factors is that Bayes Factors are not *calibrated*. What this means, among other things, is that there is no way to tell what it means to say that a Bayes Factor is "large" in the same principled way that one can say a given likelihood ratio based statistic in Frequentist settings is large, where one considers the sampling distribution of the statistic. To be sure, there are *guidelines* based a scale given by Jeffreys (1961) (see Table 1.2) for how to *intepret* the magnitude of a Bayes Factor but this not the same thing as calibrating Bayes Factor.[43] In order to calibrate Bayes Factors one would need, not only to address the question of "largeness" but also to specify *how often* a given value of Bayes Factor is expected to occur with a certain choice of a statistical model.[44]

[43]See Kass and Raftery (1995).

[44]The calibration of Bayes Factors is related to the problem raised by Mayo (2018) and others (see the discussion following O'Hagan (1995)'s paper) that Bayes Factor can be used to find evidence for a "wrong" model. See Mwakima (2024c).

1.6 Some Lessons

In discussing the pros and cons of using Bayes Factors, I believe now is a good time to return to something I mentioned in the introduction, namely, that Bayes Factors can be used to draw out the force of Salmon (1984, 224)'s "coincidence argument" to a common cause — atoms and molecules. The best way (I can think of) to understand Salmon's argument from a Bayesian statistical perspective is to use a hierarchical Bayesian model for meta-analysis.

Here is why think a Bayesian hierarchical model for meta-analysis is appropriate here. The first reason involves what the goal of a meta-analysis is. As a method for summarizing and integrating the findings of research studies in a particular area, meta-analysis aims to provide a combined analysis of related studies in order to indicate the overall strength of the evidence for, say, a beneficial effect of a treatment under study, or the value of important parameters found in theoretical models. Essentially, meta-analysis involves pooling information across multiple studies each designed to address the same scientific question with the goal often being to estimate a single effect measure common to all studies.

This brings me to the second reason why the Bayesian hierarchical model is appropriate here. What is crucial for Salmon's argument is that the different experiments (5 in total) all lead to converging values of Avogadro's Number N_A. Referring to the four papers that Perrin published in 1908 in the *Comptes rendus* of the Académie des Sciences, Salmon (1984, 216) writes:

> It is of the greatest importance to *our* story to note that these papers included not only the precise value of Avogadro's number ascertained on the basis of his study of Brownian movement, but also a comparison of that value with the results of several other determinations based upon entirely different methods, including Rutherford's study of radioactivity and Planck's work on blackbody radiation.

He writes, later, that:

> If there were no such micro-entities as atoms, molecules, and ions, then these different experiments designed to ascertain Avogadro's number would be genuinely independent experiments, and the striking numerical agreement in their results would constitute an utterly astonishing coincidence...We can say, very schematically, that the coincidence to be explained is the "remarkable agreement" among the values of N_A that result from independent determinations.

Given Salmon's own emphasis on the relatedness (they all lead to approximately the same value of N_A) and independence of the experiments, it makes sense to use a Bayesian *meta-analysis* using hierarchical models. The point I wish to make for understanding Salmon's argument (bracketing the issue of a common cause) is that a Bayesian hierarchical model for meta-analysis has a distinct advantage when it comes to explaining this agreement between the different experiments whose goal is to estimate N_A; and when it comes to saying why the body of evidence constitutes strong evidence for theoretical models which include specific predictions for N_A.

Let us consider the five theoretical models that Salmon considers (See Table 1.3).[45] Let X_i for $i = 1, \ldots, 5$ denote the observed values of Avogadro's number from each study. We can then write the sampling model as:

$$X_i|\boldsymbol{\theta}_i \overset{ind}{\sim} f(X_i \,|\, \boldsymbol{\theta}_i) \qquad \text{for } \; i = 1, 2, \ldots, 5$$

[45]The figures in the table are taken from Perrin (1910, 90). Compare with Smith and Seth (2020, 260), and Smith and Seth (2020, 369, n.8) for the values of N_A from the experiments on X-ray diffraction by the Braggs.

Source of Theoretical Model	$N_A = g(\boldsymbol{\theta})$	Estimate for $N_A(\times 10^{22})$
Brownian Motion	$N_A = \frac{RT}{D6\pi r\eta}$	65 - 70.5
Alpha Decay	$N_A = \frac{F}{2e_0}$	62 - 71
X-ray diffraction	$N_A = \frac{8M}{\rho a^3}$	61.5
Blackbody Radiation	$N_A = \frac{R}{k_B}$	61 - 62
Electrochemistry	$N_A = \frac{QM}{me_0\nu}$	60 - 90

Table 1.3: The Five Ways of Determining N_A

This sampling model says that within each study $i = 1, 2, \ldots, 5$ the observed values of Avogadro's number X_i are jointly but independently distributed according to a distribution $f(X_i \,|\, \boldsymbol{\theta}_i)$ that depends on $\boldsymbol{\theta}_i$. For example, in the Brownian motion displacement experiments, $\boldsymbol{\theta} = D$, the diffusion coefficient, which for given t can be estimated from the variance $\xi^2 = 2Dt$ of the normal distribution. In the blackbody radiation experiments the key parameter to be estimated is $\boldsymbol{\theta} = k_B$.

Another way of thinking about what the sampling model is saying is this. For each theoretical model, N_{A_i} is the key *theoretical* parameter whose value is to be determined using the X_i. N_{A_i} in turn is some function $g(\boldsymbol{\theta}_i)$ of the parameters $\boldsymbol{\theta}_i$ for $i = 1, \ldots, 5$, which appear in the statistical models associated with each theoretical model. Therefore, for each experiment the joint distribution of the observed Avogadro number X_i is $f(X_i|N_{A_i})$ where N_{A_i} is the "true value" for Avogadro's number in experiment i. This way of thinking is valid since $g^{-1}(N_{A_i})$ gives $\boldsymbol{\theta}_i$. See Table 1.3.

I do a meta-analysis using a Bayesian hierarchical model by placing a prior on N_{A_i} conditional on a theory that unifies all the studies. One candidate theory is the theory that matter is discontinuous. Call this theory M_1. According to this theory, the N_{A_i} are from a common distribution with unknown parameter N_A, which is the true value for Avogadro's number that

we want to estimate by pooling from all the five experiments.[46] To complete the hierarchical Bayesian model I write:

$$N_{A_i} \overset{iid}{\sim} h(N_A) \qquad \text{for } i = 1, 2, \ldots, 5$$
$$N_A \,|\, M_1 \sim \eta(N_A \,|\, M_1)$$

Now suppose that $p(M_1)$ represents our epistemic probability that matter is discontinuous. In meta-analysis of the experimental estimates for N_A, we are interested in the posterior distribution:

$$p(N_A \,|\, \boldsymbol{\theta}_i, X_i, M_1) \qquad \text{for } i = 1, 2, \ldots, 5$$

By Bayes' theorem the posterior distribution is proportional to the joint distribution of N_A, X_i, $\boldsymbol{\theta}_i$ and M_1 for $i = 1, \ldots, 5$. By marginalizing out $\boldsymbol{\theta}_i$ from this joint distrubution we can write

$$p(N_A \,|\, \boldsymbol{X}, M_1) \propto L(\boldsymbol{X} \,|\, N_A, M_1)\eta(N_A \,|\, M_1)p(M_1)$$

where $L(\boldsymbol{X} \,|\, N_A, M_1)$ is now the marginal likelihood of the data.

[46] As of 2018 the Committee on Data for Science and Technology (CODATA) places the value of Avogadro's Number at $60.2214076 \times 10^{22}$ mol^{-1}.

I am now in a position to explain how my Bayesian meta-analysis explains why Salmon would find this convergence of values of N_A to be a very compelling argument for a common cause and how that impacts our epistemic beliefs. For fixed $\eta(N_A \mid M_1)$ and $p(M_1)$, the posterior distribution $p(N_A \mid \boldsymbol{X}, M_1)$ depends on the data through $L(\boldsymbol{X} \mid N_A, M_1)$. If the true N_A is 60×10^{22}, then the posterior distribution will have more mass around values of N_A that are around that number since all the X_i's are in fact close to 60×10^{22}.

Now suppose that we are interested in comparing the estimate of N_A conditional on M_1 to the estimate of N_A conditional on an alternative theory M_2. M_2 here can be a different unifying theory, which we want to use to carry out the same meta-analysis we have done using M_1. To achieve a jointly exhaustive set of theories, let M_2 be the theory that says matter is continuous at the microscale. Suppose we also assign the same prior probabilities $\eta(N_A \mid M_2)$ and $p(M_2)$ to M_2 as we did for M_1. In this case, it is precisely the Bayes Factor $\frac{L(\boldsymbol{X} \mid N_A, M_1)}{L(\boldsymbol{X} \mid N_A, M_2)}$ that will be used to compare the two models. On the one hand, the theoretical models that M_1 unifies in the meta-analysis all have very close and specific values for X_i conditional on N_A being around 60×10^{22}. In my account, I have shown that this specificity comes in because the parameters in the statistical models $\boldsymbol{\theta}_i$ are different functions of N_A. So using these parameters in the statistical models unified by M_1, we are *more likely* to observe values of X_i in Table 1.3 than we would if we used a different theory like M_2. A quantitative and precise way of saying this is that Bayes Factor $\frac{L(\boldsymbol{X} \mid N_A, M_1)}{L(\boldsymbol{X} \mid N_A, M_2)}$ in favor of M_1 would high. This is the explanation for why Salmon would find this convergence of values of N_A to be a very compelling argument for a common cause. It is important to emphasize, however, that I have offered this explanation without appealing to *any causal explanation*. Because my explanation has focused entirely on considering the relevant *statistical models*, I believe that I can avoid the problem of unconceived alternatives which face accounts of causal explanations such as Salmon's. I believe that this constitutes a significant virtue of my account. I return to this point below.

I can now also say how I avoid the objection of ad hoc specification of priors. I do this by distinguishing the question I am asking from that which Achinstein and Psillos are asking. My question is: what made the quality of Perrin's statistical evidence good? Achinstein and Psillos's question is: how were prominent scientists convinced of the truth of the atomic hypothesis in light of Perrin's evidence? Both Achinstein and Psillos answer their question in terms of how prior credences were updated in light of the data. Let M_1 be the atomic hypothesis and M_2 be an alternative to it. They need $p(M_1)$ to be 0.5 or "not too low" in order that $p(M_1|\boldsymbol{X}) > p(M_2|\boldsymbol{X})$ after updating. These are the prior specifications that Smith and Seth find ad hoc. Because I am asking a different question, the answer to my question depends entirely on Bayes Factor $\frac{L(\boldsymbol{X} \mid M_1)}{L(\boldsymbol{X} \mid M_2)}$. Thus, by asking a different question and characterizing the strength of evidence in terms of relative predictive accuracy using Bayes Factor, I can avoid making ad hoc specification of priors.

1.7 Some remaining questions

There are several questions one may ask at this point. First, one may ask: how faithful is my specification of the model space to this historical episode? In performing his experiments, Perrin was certainly focused on confirming the molecular-kinetic theory or the existence of atoms and molecules. In fact, philosophers who discuss Perrin are also often concerned with the implications of Perrin's work for the existence question. So why am I not presenting the models involved in this historical episode in terms of those that assert the existence of atoms and molecules and those that don't?

Here are my reasons. Perrin cannot possibly have provided strong evidence for atoms and molecules because the actual details of atomic structure require quantum mechanics, which was unavailable to Perrin in 1908 – 1913. This means that if Perrin was providing definitive evidence for anything, it is for something that is compatible with both classical physics

(including the kinetic theory) and quantum physics. My specification of the model space has this compatibility.

Essentially, by emphasizing the need to focus on statistical evidence, I am capable of making fine distinctions between the following levels, which I also described briefly in section 1.2:

(1) Substantive or fundamental theories

(2) Theoretical models of these substantive theories

(3) Statistical models of the theoretical models

The action in the atomic debates I have zoomed in on happens at level (3) — statistical models of the theoretical models. My argument is that Perrin provided strong statistical evidence for the statistical model of the Gaussian distribution of displacements. Using a meta-analysis of the five ways for determining N_A, I have also argued that Perrin could rightly claim to have provided strong evidence that N_A is around 60×10^{22} mol^{-1}. The strength of evidence for the theoretical models is inherited *upwards* if one believes that the statistical models are adequate for capturing the theoretical predictions and constraints demanded by the theoretical models at level (2). For example, if one accepts that the variance of the statistical model for the displacements is $2Dt$, then one accepts the statistical model as an adequate model for checking the theoretical constraints imposed by Einstein's diffusion model and the hydrodynamical model for Brownian motion that leads to the Langevin equation. In this case, the strong statistical evidence accrues to the theoretical model higher up in level (2). Moving up the hierarchy to substantive theories is more complicated for reasons that Stanford (2009) has discussed having to to do with the problem of unconceived alternatives and the "Catch-all Hypothesis".

In considering how to move the strength of Perrin's statistical evidence up to the level (1), I chose to specify the model space at level (1) as M_1 — discontinuity of the ambient fluid im-

mediately below the granules, and M_2 — continuity of the ambient fluid immediately below the granules; because this is the only retrospective, and consistent, way of saying that Perrin provided strong evidence for *something* without begging the question or failing to consider all serious alternatives. This model specification at level (1) has the advantage of including dichotomous and exhaustive models, which means it avoids the "Catch-all Hypothesis" problem (i.e., the problem of exhaustively specifying, in model space, the *logical complement* of a given hypothesis in order to compute $L(\boldsymbol{X})$) discussed in Stanford (2009) in relation to the problem of unconceived alternatives.[47]

Second, and finally, one may ask: if in order to compute Bayes Factors one needs to specify at least *two* competing models, what is (or are) the alternative statistical model(s) that I am comparing *pairwise* with Einstein's statistical model? Einstein's statistical model says $\xi^2 = 2Dt$ and I say that the evidence for this model gives a high Bayes Factor in favor of M_1. But what would the Bayes Factor be if we considered explicit alternatives to Einstein's statistical model? For example, why not have $\xi^2 = 2Dt^2$ or $\xi^2 = 2D\sqrt{t}$ or $\xi^2 =$ constant for all t or $\xi^2 = 2Dt^{-1}$ and so on? The challenge raised by this last question is pressing when one realizes that there are infinitely many *explicit* alternative statistical models that one can specify and that this list of explicit alternative models cannot be exhaustively specified or enumerated. If so, then I have still not shown that the "Catch-all Hypothesis" problem can be avoided.

Here's why I believe the problems raised by this question for my account are only apparent. Recall that at level (2) both Einstein and Langevin were led to their derivations by assumptions about the nature of $F(t)$ that sustains Brownian motion. Both found that assuming discontinuity, which is exhibited by the random pressure fluctuations $F(t)$ at the micro-scale; the mean squared displacement of colloidal particles ξ^2 would have to *increase* with time. This means that on alternative models, which did not include $F(t)$, the mean

[47]Compare Smith and Seth (2020, 238 – 239).

squared displacement of the colloidal particles would have to either (i) *decrease* over time; or (ii) *remain constant*. Thus, the hydrodynamical model space can be used to partition the statistical model space as follows:

1. Models with increasing ξ^2 over time t
2. Models with decreasing ξ^2 over time t
3. Models with $\xi^2 = 0$ or some constant for all t

As a partition, it is exhaustive of the space and we can thus avoid the "Catch-all Hypothesis" problem. Further, statistical models with decreasing ξ^2 over time t would require ξ^2 to be a monotonically decreasing function of time. This would mean that rather than spread out over time according to Einstein's diffusion model (from regions of high concentration to regions of low concentration), the particles in Brownian motion would be clustering or moving closer to the mean position (origin). This is very improbable given the facts we know about diffusion. So models with decreasing ξ^2 over time t can be ruled out. This leaves models with increasing ξ^2 over time t and models with $\xi^2 = 0$ or some constant for all t. Models with $\xi^2 = 0$ or some constant are impossible because it would mean no concentration gradient or osmotic pressure for diffusion to take place. This means that the only possibilities for *any* model in the statistical model space for Brownian motion are ones in which ξ^2 increases over time t. One does not need to specify *explicitly* what the form of ξ^2 in these models has to be as the alternative statistical model to the actual one in which $\xi^2 = 2Dt$. The reason is that the only *free parameter* to be estimated in these statistical models is D, the diffusion coefficient. Let M_1^* be another statistical model for the displacements in which ξ^2 is a *different* increasing function of time and D^* be the estimated diffusion coefficient for this model that is within ε distance of the estimated diffusion coefficient D in the statistical model M_1 with $\xi^2 = 2Dt$. Then Perrin's statistical evidence shows that M_1^* will be practically indistinguishable from M_1. This means that they will both have high a Bayes Factor in their favor compared to

any model in which ξ^2 is constant or a decreasing function of time and we may choose either M_1 or M_1^*.

1.8 Conclusion

In conclusion, let me recapitulate the main points of my paper. I have argued that the quality of Perrin's statistical evidence can be characterized as specific and discriminating using Bayes Factors. My argument involved focusing on the data involved in Perrin's confirmation of the statistical model of displacements of particles in Brownian motion according to Einstein's diffusion model. While focusing on this statistical model, I also analyzed the space of hydrodynamical models for Brownian motion carved out by the Langevin dynamical equation, which leads to the variance, $\xi^2 = 2Dt$, of the statistical model. This specification of the theoretical model space was the crucial step in arguing that Perrin provided strong evidence for the discontinuity of matter, while also avoiding the "Catch-all Hypothesis" problem and the ad hoc specification of priors objection that has been directed at Bayesian perspectives of this historical episode.

Chapter 2

Coherence, Calibration and Severity

2.1 Introduction

Some statisticians and philosophers of statistics argue that coherent Bayesian methods conflict with other important desiderata that scientists have.[1] These desiderata include: (1) calibrating inferences and predictions (where this involves providing an objective measure, or guarantee, of how often the inferences and predictions are verifiably correct), and (2) model assessment (where this involves probing or testing statistical models to determine

[1]The various Bayesian methods currently in use within statistics can be distinguished on the basis of coherency or admissibility arguments from decision theory (Wald, 1949), (Wald, 1950), (Ferguson, 1967), (DeGroot, 1970, Ch. 7), (Berger, 1985) and (Robert, 2007). In decision theory, a decision rule is inadmissible, or incoherent, if there is a rule with a better outcome (in some sense) than it. The relevant sense of incoherence and inadmissibility for statistical inference is nicely discussed for a philosophical audience in Sudderth (1995). Skyrms (1990, 125) summarizes the extensive literature on coherence from the 20th century as "coherence is embeddability in a classical Bayesian model." Skyrms's work (and the references cited therein) is important because it emphasizes a philosophical distinction that it is often made between *static coherence* and *dynamic coherence*. See also Huttegger (2017, Ch. 5 and Ch. 6). Using some of the results discussed in Sudderth (1995), Berger (1983) has shown that coherent Bayesian methods are those Bayesian methods that satisfy the Likelihood Principle. Although the literature discussed in Sudderth (1995) characterizes coherent inferences and predictions as conditional distributions based on finitely additive priors, recent work by Kelley (2023) has some discussion on the possibilities of extending accuracy dominance criteria and coherence to countable and uncountable settings.

their compatibility with the observed data).[2] These desiderata are important because, taken together, they reflect the healthy skepticism scientists typically have towards their claims. This scientific attitude involves probing their claims and quantifying the *reliability* of the inferences that they make supporting or disproving those claims.[3] This criticism of coherent Bayesian methods is known as the *probativist criticism*.[4]

In advancing the probativist criticism, Mayo has claimed that coherent Bayesian methods fail to meet the *minimum requirement for severity*:

> One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have [B]ad [E]vidence, [N]o [T]est. (BENT) (Mayo, 2018, 5)

Without getting into the weeds of what the minimal requirement for severity amounts to, let me make a few remarks.[5] First, the minimal requirement for severity is a general *meta-statistical/meta-methodological* principle that Mayo thinks all approaches to statistics should aim to satisfy. The qualifier 'meta', is meant to indicate that this principle involves the actual and counterfactual assessment of statistical methods and procedures, "one level removed" (as Mayo likes to say), from the methods themselves.[6] As a general meta-statistical principle, the error probabilities it asserts about methods are *quasi-formal*. Mayo thinks we can say of a method or procedure, in a less than fully formal way, that it is probably wrong because of

[2]See Box (1980), Box (1982), Diaconis and Freedman (1983), Diaconis and Freedman (1986), Rubin (1984), Gelman, Meng, and Stern (1996), Cox (2006) and Reid and Cox (2015).

[3]Compare with Fletcher (2024) who has recently developed measures of (dis)confirmational (un)reliability that acknowledge these scientific goals.

[4]See Mayo (2018, Ch. 6).

[5]See Mwakima (2024d) where I elaborate on the error and statistical testing rationale for the minimal requirement for severity and defend it against recent criticisms by Maranda (2023) and van Dongen, Sprenger, and Wagenmakers (2023).

[6]The helpful reading of Mayo's account as a counterfactual account is given by Fletcher (2020).

how it was actually carried out *or* how counterfactually it could have led us to error. At the same time, the minimal requirement for severity can also be understood at the ground-level (as opposed to meta-level) of statistical hypothesis testing and interval estimation *formally* using the sampling distribution of a test statistic and classical confidence interval estimators.[7]

The reaction by some coherent Bayesians to Mayo's minimal requirement for severity is to ask, "Why should we take the minimal requirement for severity seriously? And why does it matter whether or not Bayes Factors satisfy this requirement?"[8] Those who push back at this requirement also raise at least the following philosophical questions. First, what does "false" or "flaws" mean? Second, what does "practically guaranteed" mean? Third, what is a "capability of finding flaws"? By raising the first question ("What does "false" or "flaws" mean?"), some people deny that the claims made by statistical models are the sorts of things that can be true or false.[9] Insofar as Mayo ties the requirement of severity to assessments of truth made by claims in statistical models, the minimal requirement for severity is not a requirement that should be taken seriously. The second and third questions are related. According to Mayo, who follows the tradition of classical hypothesis testing going back to Neyman and Pearson (1928) and Neyman and Pearson (1933), there is always the possibility of erroneous inference. For this reason the principles of statistical inference in this school are formulated to control the rate of incidents of these errors. So, Mayo's minimal requirement for severity is one way of expressing the view that methods of statistical inference cannot ignore the possibility of error. It says that a method of statistical inference must have tools: (1) to engage in model assessment ("capability of finding flaws"); and (2) to calibrate its inferences and predictions ("no practical guarantee of agreement"). The phrase "no practical guarantee of agreement" is tied to calibration because when one calibrates

[7] A helpful and excellent summary is Fletcher (2020).

[8] Compare Richard Morey's slides *Bayes Factors, p-values, and The Replication Crisis* from Session 1 of The Statistics Wars and Their Casualties Workshop held on 22nd September 2022. `https://cardiffunipsychstats.co.uk/statswars2022/`

[9] See Morey, Romeijn, and Rouder (2013, 71), Rouder, Morey, and Wagenmakers (2016) and compare with Box and Jenkins (1976, 285) and Box (1979).

an instrument or procedure, one is acknowledging that the instrument or procedure is not perfectly accurate when compared to a standard and when used in actual measurements. Consequently, tolerances are always quoted with most manufactured measuring instruments and a measurement using these instruments typically includes some margin of error.

Two ways have been proposed in the applied Bayesian statistics literature to address the probativist criticism. The first way adopts an eclectic approach to statistical practice by fusing what is good about Bayesian inference (it is fundamentally sound) with what is good about non-Bayesian inference (its tools for model assessment and calibration) even at the cost of coherency. The defenders of this proposal call themselves *Calibrated Bayes*, *Frequentist Bayes* or *Pragmatic Bayesians*.[10] What is puzzling, *philosophically*, is characterizing precisely what background or foundational principles they are committed to, in what sense their position is Bayesian or if these unifications are coherent.[11] According to Mayo (2018, 395 – 436) the foundations of Bayesian statistics are "in flux" when it comes to pinning down precisely what the foundations of contemporary Bayesian statistics are. In fact, in a recent influential paper representing the Pragmatic Bayesians, Gelman and Shalizi (2013, 10) remark, "most of [the] received view of Bayesian inference is wrong." In making this remark, they seek to align the modern practice of Bayesian statistics with the error-statistical approach of Mayo. To this alignment or unification, Mayo (2018, 28 – 29) says, "[T]he idea of error statistical foundations for Bayesian tools is not as preposterous as it may seem." But adds that whether this can be done is open.[12]

The second way is to show (somehow) that coherent Bayesian methods can address the probativist criticism on their own. Morey, Romeijn, and Rouder (2013), who call themselves "Humble Bayesians", have argued against calibrating inferences and predictions by emphasizing, among other things, that statistical models are neither true nor false. One can avoid

[10]Representatives or advocates of these positions are Little (2006), Wasserman (2006), Kass (2006), Little (2011), Gelman (2011) and Kass (2011).

[11]Compare Mayo (2011), Mayo (2013) and for a recent discussion see Gelman and Yao (2020).

[12]Compare Fletcher (2020).

the demand for calibration by denying that our models and the evidence we have for them should be tracking some objective truth about the real world in order to be reliable in practice. There is nothing to be "wrong" about when assessing the relative strength of evidence using Bayes Factors among *specified* models so long as the model specifications are theoretically motivated, judiciously chosen and sufficiently justified.[13] According to some of these authors, what is crucial is whether Bayes Factors are *interpretable* in comparing just the specified models. It is a mistake to demand they be calibrated in the space of *all* statistical models (which somehow includes the "true model"— which we never know).[14] They further emphasize the interplay between coherent Bayesian statistics and (psychological) science, by thinking about "objectivity" in a different way in terms of transparency and the researcher's role in "adding value" to statistical analysis by bringing in fact-based domain knowledge and expertise to specify and to fit *informative* priors that are predictively accurate.[15] "Adding value" means contributing positively to a given scientific research agenda by *justifying* one's modeling assumptions. The process of justification invites dialogue among researchers that then advances fruitful discussion of issues within the field. It is this *public* and *intersubjective* conversation that constitutes objectivity on their view.

While my sympathies are certainly with the Humble Bayesians, I don't think they have *fully* addressed Mayo's probativist criticism.[16] Here's how Mayo puts it:

> How do I criticize your prior degrees of belief? As Savage said, "[T]he Bayesian outlook reinstates opinion in statistics — in the guise of the personal probabilities of events..." (Savage 1961, p. 577). Or again, "The concept of personal

[13]Compare with the quotation from Lindley in Mayo (2018, 228).

[14]This is the position defended by Morey, Wagenmakers, and Rouder (2016) and Rouder and Haaf (2019). See especially Haaf, Klaasen, and Rouder (2021, 13) and compare with section 5.1 and 5.2 of Gelman and Hennig (2017).

[15]See Rouder, Morey, and Wagenmakers (2016) and Haaf, Klaasen, and Rouder (2021). The emphasis on 'informative' means they are distancing themselves from Bayesians who seek default, reference or uninformative priors. These priors are appealing to the Frequentist Bayes and Pragmatic Bayesians because they have "good frequentist properties". Kass and Wasserman (1996) is an excellent discussion of *objective priors*. Compare this paper with the discussion in Gelman and Hennig (2017).

[16]See Mayo (2018, §§4.1 and 4.2).

> probability...seems to those of us who have worked with it an excellent model for the concept of opinion" (ibid., pp. 581 – 2). That might be so, but what if we are not trying to model opinions, but instead insist on meeting requirements for objective scrutiny? For these goals, inner coherence or consistency among your beliefs is not enough. One can be consistently wrong, as everyone knows (or should know). (Mayo, 2018, 228)

> Either your methodology picks up on influences on error probing capacities of methods or it does not. If it does, then you are in sync with the minimal severity requirement. We may compare our different ways of satisfying it. If it does not, then we've hit a crucial nerve. If you care, but your method fails to reflect that concern, then a supplement is in order. Opposition in methodology of statistics is fighting over trifles if it papers over this crucial point. If there is to be a meaningful "reconciliation," it will have to be here. (Mayo, 2018, 270)

From these two passages, I take it that the real question raised by the probativist criticism is this: *how do coherent Bayesian methods account for the possibility of error?* The set of models singled out for comparison using Bayes Factors can be misspecified and the model comparison process using interpretable Bayes Factors (even with accumulating evidence) can still turn out to be miscalibrated (in the sense of being "consistently wrong").[17] Perhaps this is our unavoidable epistemic situation and it is in this sense that we should be humble or open-minded in light of our own fallibility.[18] Still, I claim that coherent Bayesians would want to find ways to say objectively how, and to what extent, they are miscalibrated (i.e., "meeting requirements for objective scrutiny"). This is what scientists and some Federal

[17] I take this to be the point that Hoijtink, van Kooten, and Hulsker (2016b) and Hoijtink, van Kooten, and Hulsker (2016a) are making in insisting that Bayes Factors have "frequentist properties". But I do not think that granting this point commits one to a frequentist interpretation of probability. Compare Stern (2016).

[18] Compare the discussion in Morey, Romeijn, and Rouder (2016, 9 – 10).

Regulatory agencies want.[19] But this will involve appreciating the real question raised by the minimal requirement for severity, then using or developing tools within the coherent Bayesian framework to meet that challenge.

There are two related ways that I wish to make this question salient in this paper. The first way is to single out the relevant sense of "calibration" that a coherent Bayesian ought to take seriously. Making these distinctions matters because when one looks at the literature on calibration, there are on the one hand some authors who advocate for calibration; while on the other hand there are authors who are reject it. Often it is not clear whether those who disagree within this debate have the same sense of calibration. It is quite possible that the authors would, in fact, agree with each other regarding the relevance of calibration in one sense of calibration; but are simply talking past each other if they associate calibration with different senses.

The second way is to bring de Finetti into the conversation regarding the possibility of probative foundations for Bayesian statistics. This is the main contribution that I intend to make here. Gelman and Shalizi, as I have mentioned, reject "the received view" of Bayesian statistics while Mayo (2018, 227) gives de Finetti's apparent "logical positivist" view short shrift. Little, Kass and other contemporary applied Bayesian statisticians seek pragmatic unifications between Bayesians and non-Bayesians, who hold on to a limit of relative frequencies interpretation of probability. These proposed "unifications" by Calibrated Bayes, Frequentist Bayes and Pragmatic Bayesians are not genuine unifications because they still maintain a dualism. The dualism is between the subjective interpretation of probability and objective chances interpreted as limits of relative frequencies. Therefore, the advantage of bringing de Finetti into this conversation, I will argue, is that from de Finetti we have:

[19]Compare the discussion in van Dantzig (1957, 11), Cox (2006) and US Food and Drug Administration (2010).

1. A subjective Bayesian account of how to genuinely unify without a dualism between the different interpretations of probability.[20]

2. Principles that a coherent subjective Bayesian *will* accept that together imply the salience of the probativist criticism. These principles are:

   (a) Probabilities are special cases of previsions.

   (b) The use of proper scoring rules to provide an alternative operational definition of probabilities.[21]

   (c) The posterior distribution is an estimator.

   (d) de Finetti's General Representation Theorem asserts the existence of a prior distribution on *function spaces*.

One upshot of the view that I defend in this paper will be to draw out the connection between the minimal requirement for severity and the theorems that establish the statistical consistency of Bayes estimators as one way to emphasize the importance, and the difficulty, of addressing the probativist criticism. Not only do I think that this connection is a new and plausible way to fruitfully engage with Mayo's work by seeking common ground; but there is also recent philosophical interest for drawing out this connection. This interest has arisen from work by philosophers of science who have shown that the demand for calibration in the sense of statistical consistency cannot be set aside so easily as some coherent Bayesians think.[22] The work by these philosophers takes aim at both the martingale convergence

[20]The unification here has sometimes been called *pragmatic* (see especially de Finetti (1974a)). The pragmatic reading of de Finetti is also developed by Skyrms (1984) and Galavotti (2019). Compare with Jeffrey (1984). de Finetti himself also acknowledged the influence of pragmatism on his own views (see de Finetti (1974b, vii)).

[21]See the footnote in the translation of de Finetti's earlier work (de Finetti, 1937) in Kyburg and Smokler (1964, 62). In the footnote, de Finetti refers to an earlier paper (de Finetti, 1962) and joint work-in-progress, de Finetti and Savage (1962), on how to elicit initial probabilities, the English summary of which is in de Finetti (1972, Ch. 8). This later (post 1964) de Finetti would revisit the same idea in de Finetti (1965), de Finetti (1969), de Finetti (1970) and de Finetti (1974b, Ch. 5) as well as de Finetti (1981).

[22]See Belot (2013a) and Belot (2013b). One can find a critique along similar lines to those of Belot in Earman (1992, 41 and Ch. 6). van Fraassen also has a series of blog posts on this issue with a nod to Mayo. See `https://basvanfraassensblog.home.blog/2019/08/11/what-is-bayesian-orgulity-1/`

theorems, which are taken by some coherent Bayesians as providing the Bayesian agent with "a guarantee" of statistical consistency or perfect calibration; and the objectivity of Bayesianism in the sense of merging of opinions by dint of (dynamic) coherence.[23]

Here's how I have organized the rest of the chapter. In the next section I distinguish different senses of calibration and I identify statistical consistency as the relevant sense of calibration for the rest of my paper. In section 2.3 I begin developing the first principle of de Finetti's view — 2(a) above — that probabilities are special cases of previsions. The discussion here will be the first step in understanding de Finetti's unification program. In section 2.4 I show how by endorsing the use of Brier's Score (a proper scoring rule) for the operational definition of subjective probabilities, de Finetti is, in fact, committing himself to calibration in the sense of statistical consistency not forecast calibration. This section will be important for by-passing objections by Kadane and Lichtenstein (1982), Seidenfeld (1985) and Lad (1996) against forecast calibration by showing that while their objections are successful in downplaying the relevance of forecast calibration for coherent Bayesians; it does not follow that the same kinds of objections apply to statistical consistency. Section 2.5 then explains how the operational definition of probability from Brier's Score and the assumption of exchangeability implies the existence of priors — this is de Finetti's representation theorem. I show that it is really de Finetti's general representation theorem that informs Belot's critique and I explain how that critique casts the probativist criticism in a new light. Before I conclude in section 2.7, I use section 2.6 to reflect on how optimistic applied Bayesian statisticians can expect to be in addressing that criticism.

[23]For the perfect calibration results often cited by Bayesians see Dawid (1982). For merging of opinions and objectivity see Huttegger (2015a), Huttegger (2015b) and Huttegger (2017, Ch. 8).

2.2 The varieties of calibration

To capture the relevant sense of calibration that applies to the considerations of severity that I am interested in, let me distinguish the following kinds of calibration:

- forecast calibration
- instrumental calibration
- statistical coverage
- statistical consistency

Forecast calibration refers to the reliability of a subjective probabilistic forecast. In fact, the term 'calibration', as it is discussed in some statistical and philosophical circles, traces its origin to meteorological contexts (although the term is broad enough to capture stock market forecasts, sporting events forecasts, medical prognoses and so on).[24] To illustrate the general idea, say a weather forecaster announces, "There is a 70% chance of rain tomorrow". If it turns out not to rain, she isn't necessarily a "bad" forecaster. But if she repeatedly announces a 70% chance of rain for each day (in a row) in a sequence of 10 days but it only rains on 2 of these days, then we'd be led to question the reliability of her forecasting abilities for 70% chances of rain. On what basis would we assess or appraise a forecaster such as her? Brier (1950)'s idea was to measure, what from the contemporary point of view we would call, the mean squared error of the forecast, leading to what has come to be known as *Brier's Score*.[25] Forecast calibration found its way into the statistical literature with the

[24]The literature on this sense of calibration is vast and extremely rich. Good places to start are Lichtenstein, Fischhoff, and Phillips (1977), Kadane and Lichtenstein (1982), Seidenfeld (1985) and Lad (1996).

[25]See Pettigrew (2016) for an up to date discussion of (proper) scoring rules. Pettigrew characterizes them and shows that the Brier Score is a member of a class of (in)accuracy measures known in the statistical literature as *Bregman divergences*. See especially Gneiting, Balabdaoui, and Raftery (2007), and Gneiting and Raftery (2007).

publication of Dawid (1982).[26] Within philosophy of science, van Fraassen (1983) has largely influenced how the subsequent issues have been framed within the philosophy of probability and formal epistemology.[27]

Instrumental calibration is perhaps the most familiar kind of calibration with physicists and engineers. Here one compares the measurements of a measuring instrument, say a grocer's weight of 1kg, to measurements taken by a *standard* platinum-iridium alloy cylinder securely held in Paris. Discrepancies between the measurements taken by this weight and this standard can be used to recalibrate the weight to attain the required accuracy. When Cox (2006, 197) writes:

> Frequentist analyses are based on a simple and powerful unifying principle. The implications of data are examined using measuring techniques such as confidence limits and significance tests calibrated, as are other measuring instruments, indirectly by the hypothetical consequences of their repeated use [...] The objective is to recognize explicitly the possibility of error and to use that recognition to calibrate significance tests and confidence intervals as an aid to interpretation.

the talk of measuring instruments is metaphorical. But it suggests applying the qualifier 'instrumental' to this kind of calibration in order to distinguish it from other senses of calibration.

There is another way to think about instrumental calibration. Here an illustration might help. It is known that temperature can be measured by either a mercury thermometer or a thermocouple. A thermocouple is a device that detects voltage whenever there is a temperature gradient. Since most people don't understand what a voltage change means for

[26]But, as we shall see, Dawid's result is really a result regarding the *statistical consistency* of a Bayes estimator under some assumptions. It has affinities with the martingale convergence results.

[27]See van Fraassen (1984), Joyce (1998), Lange (1999), Joyce (2009), Weirich (2011), Hájek (2012), Hoefer (2012) and for a recent and comprehensive discussion of the state of the art see Pettigrew (2016).

them with regard to temperature, a thermocouple has to be calibrated to a known scale (the Celsius or Fahrenheit scale) on a thermometer for us to *interpret* the thermocouple readings. There are thermocouple voltage-to-temperature equations for doing this conversion.

It is not obvious how this second way of thinking about instrumental calibration concerns the same concept as that involving accuracy of weights. In the former case there is no issue of setting the relevant correspondence, while in the latter case there is no issue of discrepancy. I have chosen to call both of them instrumental calibration because in both cases one is dealing with *scales of instruments*. In the first case, it is about the *accuracy* of the scale, in the second case it is about *the interpretation* of the scale. Consider how this applies to a recent discussion in the psychological science literature of whether or not Bayes Factors should be calibrated in order to be useful in psychology.[28] Bayes Factors are instruments or tools within the Bayesian framework for quantifying evidence. So, when Hoijtink, van Kooten, and Hulsker (2016b) are talking about calibrating Bayes Factors, they are thinking of instrumental calibration, say as one would think of converting from a thermocouple scale to a Celsius scale. But Morey, Wagenmakers, and Rouder (2016) disagree because they believe Bayes Factors are interpretable as is. Bayes Factors are like the temperature scale in Celsius or Fahrenheit in this analogy.

Within the statistical literature one often hears calibration talked about in terms of finding estimators with "good frequentist properties" (Cox, 2006, 198). One of these properties is the *statistical coverage* of confidence interval/set estimators. If in the infinite limit of hypothetical repetitions of the same random experiment a *nominal* 95% confidence interval/set *actually* covers the true parameter(s) 95% of the time, then we say the confidence interval/set estimator is calibrated in the sense that nominal coverage coincides with actual coverage, otherwise it is miscalibrated.[29]

[28]See the papers by Hoijtink, van Kooten, and Hulsker (2016b), Hoijtink, van Kooten, and Hulsker (2016a) and Morey, Wagenmakers, and Rouder (2016).

[29]Wasserman (2006) has an amusing story about why this is a desirable property for estimators to have. By insisting on "frequentist properties" of Bayes Factors, Hoijtink, van Kooten, and Hulsker (2016a) can

Finally, this brings me to the sense of calibration in terms of *statistical consistency*. Jeffrey (1984, note 36) and Gneiting and Raftery (2007, §9) think about calibration in terms of statistical consistency of estimators.[30] An estimator $T(\boldsymbol{X})$ is a function of the data $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ that is used to estimate a parameter θ of interest. An estimator $T(\boldsymbol{X})$ is *unbiased* if $\mathbb{E}_\theta[T(\boldsymbol{X})] = \theta$.[31] The *mean squared error* ($\mathrm{MSE}_\theta(T(\boldsymbol{X}))$ of an estimator $T(\boldsymbol{X})$ for θ is $\mathbb{E}_\theta[(T(\boldsymbol{X}) - \theta)^2]$. Using the known identity regarding the variance[32]

$$\mathbb{E}_\theta[(T(\boldsymbol{X}) - \theta)^2] = \mathrm{Var}_\theta(T(\boldsymbol{X})) + (\mathbb{E}_\theta[T(\boldsymbol{X})] - \theta)^2$$

$$\mathrm{MSE}_\theta(T(\boldsymbol{X})) = \mathrm{Var}_\theta(T(\boldsymbol{X})) + \mathrm{Bias}^2_\theta(T(\boldsymbol{X}))$$

In other words, the mean squared error is the sum of the variance and square of the bias of an estimator.

Now, a sequence of estimators $T_n(\boldsymbol{X})$ is said to be *consistent* for θ if it converges in probability to θ. And a sequence of estimators $T_n(\boldsymbol{X})$ is said to be *consistent in quadratic mean* if $\mathrm{MSE}_\theta(T_n(\boldsymbol{X})) \to 0$ as $n \to \infty$. Using Chebyshev's Inequality, it can be shown that a sequence of estimators is consistent if it is consistent in quadratic mean, i.e., if and only if the variance and bias are both asymptotically zero.

Two important and well-known results within a non-Bayesian setting that concern the consistency of estimators are the Weak Law of Large Numbers and the Strong Law of Large Numbers.

---

sometimes be read as indicating that they are after statistical coverage. Because having good frequentist properties might come at the cost of coherence, they rightly invite disagreement from Morey, Wagenmakers, and Rouder (2016).

[30] This is the relevant sense in which I wish to discuss the issue of calibration for the rest of my paper. I discussed the other senses of calibration above in some detail because it will allow me to by-pass some objections from some prominent philosophers. So when we get to that part of my paper, it will be helpful to have the distinctions I am making in mind.

[31] The subscript under the expectation operator is to emphasize that the expectation is a function of θ.

[32] $\mathrm{Var}(X) = \mathbb{E}[X^2] - [\mathbb{E}[X]]^2$

Weak Law of Large Numbers

Let $\boldsymbol{X} = (X_1, X_2, X_3, \dots)$ be independent, identically distributed random variables each with mean μ and variance σ^2. Let $T_n(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} X_i}{n}$, then $T_n(\boldsymbol{X})$ is consistent for μ.

The Weak Law of Large Numbers was proved by Jacob Bernoulli in 1713 for 0-1 random variables (nowadays called Bernoulli random variables). Poisson proved a generalization of this result and first called it the Law of Large Numbers.[33] It is Aleksandr Khintchin who proved the general form of the Weak Law of Large Numbers using Chebyshev's Inequality in 1929.

Strong Law of Large Numbers

Let $\boldsymbol{X} = (X_1, X_2, X_3, \dots)$ be independent, identically distributed random variables each with mean μ and $\mathbb{E}[X_i^2] < \infty$. Let $T_n(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} X_i}{n}$. Then $P\left(\left\{\lim_{n\to\infty} T_n(\boldsymbol{X}) = \mu\right\}\right) = 1$.

Borel formulated and proved the first variant of the strong law of large numbers in 1909 for 0-1 variables. Cantelli in the early 1910s, Khintchin in the 1920s and Kolmogorov in 1930 stated sufficient conditions extending the applicability of the Strong Law of Large Numbers.[34]

2.3 Previsions and Frequencies

The starting point for de Finetti's operational definition of probability is the notion of a *prevision*. This is a term that is unique to de Finetti and is meant to contrast with

[33] See Lecture 4 in von Mises (1957).
[34] See von Mises (1964, 236).

prediction.[35] It is helpful to think of a prevision in contemporary terms as the expected value or weighted average of any quantity.[36] In his work de Finetti often illustrates a prevision by drawing an analogy with the center of mass and the minimum of the moment of inertia.[37] It is no coincidence. It is known that the center of mass of a system of masses is the weighted average of the masses with weights the distance from the origin in some coordinate system. It is also known that by the Parallel Axis Theorem, the minimum of the moment of inertia of a body is along an axis through the center of its mass.

Importantly, the relevant prevision in the case of probabilities is the expected payoff of a gamble/lottery ticket. It is by assuming that the fair price of a gamble/ticket is its expected payoff that de Finetti is able to define subjective probabilities from previsions. For example, if \$1 is the fair price for you for a lottery ticket that pays \$2 if an event E occurs and \$0 if the event does not occur, then operationally it is *as if* $P(E) = \frac{1}{2}$ for you. Further, since a proposition or event X can be identified with its indicator function $\mathbb{I}(X)$, probabilities are previsions of indicator functions. This means that probabilities are special cases of previsions — principle 2(a) above.

Suppose that an agent is in a state of uncertain knowledge and wants to *estimate* the relative frequencies of a sequence of *observables* $X_1, X_2, \ldots, X_n$. I use the term observables as a generic term covering mundane events or outcomes of measurements in science. These observables do not have to be related in any way. de Finetti thinks that this estimation in a state of uncertain knowledge can be accomplished with a prevision of the frequencies. According to de Finetti this is "the first and most elementary link in the long chain of conclusions which, as we proceed, will clarify and enrich our insight into the relationship that

[35]de Finetti's position is that previsions are operational *estimates* of subjective probabilities relative to an individual. As estimates, they don't have truth values but are constrained by coherence and calibration (i.e., how accurate they are). A prediction, on the other hand, can be falsified by facts. Hence de Finetti's insistence on distinguishing between previsions and predictions.

[36]See de Finetti (1974b, §§3.1.4 – 3.1.5)

[37]See de Finetti (1974b, §3.3.4) for example.

holds between *probability* and *frequency*" (de Finetti, 1974b, §3.10.3). What de Finetti says here is not an exaggeration; for he adds in de Finetti (1974b, §5.8.2):

> It is in this very thing [the identity below] — and in nothing else — that the value of any theorem in the calculus of probability lies, and it cannot be otherwise. It is to tell us whether, in making the same evaluation in two different ways, we arrive at different conclusions, and, in this case, to invite us to think again and to rectify the situation by modifying one or the other.

Let $\mathbb{I}(X_i)$ be the indicator for X_i, then $Y = \sum_i^n \mathbb{I}(X_i)$ is the number of X_i that occur and $\frac{Y}{n}$ is the frequency.

$$\mathbb{E}\left[\frac{Y}{n}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{I}(X_i)}{n}\right] \tag{2.1}$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \mathbb{I}(X_i)\right] \tag{2.2}$$

$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{I}(X_i)] \tag{2.3}$$

$$= \frac{1}{n}\sum_{i=1}^n P(X_i) \tag{2.4}$$

Despite its generality (in the sense that nothing has been assumed about the distribution of the X_i or any relationship between them), this sequence of identities has been referred to as "de Finetti's Law of Small Numbers" by Jeffrey (1984, 81). It says that *the prevision of actual frequency is the average of the probabilities.* According to de Finetti, this is the most general thing one can say about the relation between subjective probabilities (understood as previsions) and actual frequencies in the state of uncertain knowledge. In the special case that the X_i are identically distributed with probability p we get:

$$\frac{1}{n}\sum_{i=1}^{n} P(X_i) = \frac{np}{n} \tag{2.5}$$

$$= p \tag{2.6}$$

For example, suppose I want to estimate the bias of a coin θ, i.e., the frequency of heads and tails that I should expect but I am in a state of uncertainty regarding whether the coin is fair. In fact, let us suppose that I am in the epistemic state such that the observable $X_1 =$ heads is more probable than the observable $X_2 =$ tails to *me*. So I assign probability 0.6 to $X_1 =$ heads and 0.4 to X_2 by coherency. According to de Finetti, my prevision of the bias θ is the average of these probabilities $\frac{0.6+0.4}{2} = 0.5$, which is a good estimate if the coin is, in fact, fair.

Why does going through all of these seemingly trivial and well known definitions and facts matter? Well, for two reasons. First, it is by starting with previsions that de Finetti first shows how to deal with actual relative frequencies within a subjective framework as illustrated in the previous example. The Law of Small Numbers is de Finetti's definitive statement that he did not deny that there are actual frequencies; what he denied or rejected is the *definition* of probabilities as limits of relative frequencies. Second (and this is something that I go into more detail in the next section), while discussing the foundations of Bayesian statistics, Mayo (2018, 406 – 407) writes:

> Cashing out Bayesian uncertainty with betting seems the most promising way to "operationalize it." Other types of scoring functions may be used, but still, there's a nagging feeling they leave us in the dark about what's really meant.

I sympathize with Mayo on the last sentence although de Finetti and I would disagree with the first sentence. I sympathize with Mayo because, like several other philosophers have also

pointed out, de Finetti is notoriously difficult to read![38] Here's how Lindley (1986) puts it in his comments on Efron (1986)'s "Why isn't everyone a Bayesian?":

> It may be that we Bayesians are poor writers, and certainly the seminal books by Jeffreys (1961) and de Finetti (1974, 1975) are difficult reading, but it took Savage (see the preface to Savage 1972) several years to understand what he had done; naturally, it took me longer.

Add to this the fact that a lot of his original work was published in Italian, his tone is often curt, impatient and only *a bit* provocative, then you can appreciate what one has to deal with here to understand what de Finetti meant! So, my goal in this paper is to provide some illumination in the best way I can. Fortunately for me, a lot of de Finetti's work is now available in translation including de Finetti (2008). This book is a collection of transcribed audio recordings of philosophical lectures on probability by de Finetti from 1979 at the National Institute for Advanced Mathematics in Rome. The lectures were recorded with the aim of having them published. They were published in Italian ten years after the death of de Finetti. The lectures are a treasure trove of insight into understanding the mature thought of de Finetti and can be fruitfully read alongside de Finetti (1972) and de Finetti (1974b). In particular, these lectures provide strong evidence to reject what Mayo claims about "the most promising way to operationalize subjective probabilities". According to de Finetti (2008, 4 – 5):

> There is a way, and I believe it is the only one, which allows us to say exactly what we mean by "probability." Such is the method of *proper scoring rules*, which consists in asking a person (let us call her "A") what is the probability she assigns to an event E, A being warned that she will receive a score depending

[38] See Diaconis and Skyrms (2018, 125) for example.

on the answer she will provide and on the value of the "objective probability" of E (in the sense specified above: 0% if E is false, 100% if E is true).

The idea behind proper scoring rules, why and how they work for operationalizing subjective probabilities only makes sense when used together with previsions of actual relative frequencies. It is for these reasons that I have spent this section explaining these concepts. In the next section I show how this works and draw the connection to statistical consistency. The bearing of this connection on the interpretation of de Finetti's representation theorem is then discussed in section 2.5.

2.4 Proper Scoring Rules and Statistical Consistency

Following Lad (1996, 335ff) but adapting his characterization to my notation and approach in this paper, let θ be any quantity whose value we wish to estimate in the state of uncertain knowledge. Let $\delta(\boldsymbol{X})$ be your estimate of it. δ is a function of observables $\boldsymbol{X}$, which could include actual data, your background knowledge, information or expertise that inform your opinion about θ. Any real valued function $S(\theta, \delta)$ is a *scoring rule* for δ if it is convex and attains its minimum at $\delta = \theta$. The squared error loss or the quadratic score function (of which Brier's score is an instance) is a common example of a scoring rule, although many functions satisfy the criteria for a scoring rule.[39] Scoring rules were initially proposed for assessing weather forecasters (but they do not have to be used solely for forecast calibration).[40] In the case of meteorology, θ could be the chance of precipitation and δ the quoted probability of it. If δ is a prevision, then a *proper scoring rule* is a scoring rule such that the expected loss (i.e., risk) of quoting an estimate $\delta^* \neq \delta$ is greater than or equal to the risk of δ. The characterization of proper scoring rules is nicely illustrated if we think of the expected loss

[39] See Savage (1971).
[40] See Gneiting and Raftery (2007, §9).

under a quadratic score function. In this case, it is known that the expected value minimizes the expected loss. So if δ is a prevision of θ, i.e., $\delta = \mathbb{E}[\theta]$, then it minimizes the expected loss.

It is this previous sentence that de Finetti uses to connect previsions to proper scoring rules in providing an operational definition of probability based on Brier's Score. Surprisingly, de Finetti claims that the criterion of proper scoring rules does not have anything to do with honesty, which is contrary to what many in the literature have claimed.[41] In Lecture II of the Philosophical Lectures we read:

> BETA: I have to admit, I have grasped proper scoring rules very vaguely. I have understood that there is a certain penalization, but I have not been able to understand why is it appropriate to indicate the probability which one deems more reasonable. Why is it inappropriate to "cheat"?[42]
>
> DE FINETTI : It is not "inappropriate"; this has nothing to do with cheating. One is allowed to indicate whatever number she likes.

de Finetti then adds, "Since she is subject to penalization, X [i.e., the agent] will suffer a loss in any case; her best interest is to minimize the prevision of it." The prevision of a penalization is the expected loss. So de Finetti is telling us here that the choice of Brier's score as a scoring rule is that it leads back to a prevision as the best estimate of any quantity in the state of uncertain knowledge because it minimizes the expected loss (or risk).[43] Now

[41] See Lad (1996, 336) and many of the references cited earlier under forecast calibration.

[42] The five participants in de Finetti's course, with their comments and questions have been named 'Alpha', 'Beta', 'Gamma', 'Delta', and 'Epsilon' according to the order in which they participated for the first time in the series of lectures. Beta is Anna Martellotti, who is now a professor of Mathematical Analysis at the University of Perugia. See de Finetti (2008, xiii).

[43] See also Lecture IV of de Finetti (2008, 28).

it can be shown that under a quadratic score function, the *Bayes Rule*

$$\delta(\boldsymbol{X}) = \mathbb{E}[\theta \mid \boldsymbol{X}] = \int \theta \pi(\theta \mid \boldsymbol{X}) d\theta$$

as an estimator for θ, minimizes the posterior expected loss. But δ is the posterior *mean* as an estimator for θ, i.e, a prevision. This is exactly what principles 2(b) "The use of proper scoring rules to provide an alternative operational definition of probabilities" and 2(c) "The posterior distribution is an estimator" say.

What does all of these have to do with statistical consistency and why does it matter? There are two reasons. First, the reading of de Finetti suggests that the appeal to proper scoring rules is not intended to assess the reliability of forecasters. Rather proper scoring rules are meant to operationally define probabilities as previsions. This means that many of the objections against forecast calibration in the literature do not apply to the project that I am undertaking in this paper.[44] For what is common to all of these objections is that they are addressed at forecast calibration especially if it is intended to compare the reliability of different forecasters. Here I am not making any comparisons between probability forecasters, since the focus has been on previsions as estimators. With respect to estimators, statistical consistency is an important, and desirable, property.

The second reason is that the conclusion about the posterior mean as an estimator for θ raises the question about its statistical consistency because the Bayes Estimator is, in general, not unbiased.[45] If the bias-squared of an estimator has to asymptotically go to zero for consistency in quadratic mean, does this mean that sequences of Bayes Estimators can never be consistent? No, consistency in quadratic mean is only a sufficient condition, not a necessary one. Therefore, it is a remarkable fact that despite not being unbiased the statistical consistency of a sequence of Bayes Estimators was demonstrated.

[44] See Kadane and Lichtenstein (1982), Seidenfeld (1985) and Lad (1996) for the sorts of objections I have in mind.

[45] See Ross (1996, 34 – 35)

The proofs here make use of what is known as a *martingale* (which is an abstract generalization of a fair gambling process). The proof depends on showing that a sequence of Bayes Estimators can be thought of as an instance of *Doob's Martingale* (which is a martingale with special properties).[46] Using these properties, Doob (1949) proved that Bayes Estimators are almost surely consistent. Since almost sure consistency implies consistency, this has a consequence that Bayes Estimators are calibrated in the sense of statistically consistent.

Now, a crucial assumption made by Doob (1949, 25) in the theorems is that the "true parameter" is in the support of the prior $\pi(\theta)$. He writes:

> Note that these are "probability one theorems". The estimate of the value of θ drawn may not be good for a θ set of probability 0.

Doob's assumption raises the following question: can a Bayesian always *know* that the true θ is within the support of her prior to guarantee calibration in the sense of statistical consistency of Bayes Estimators? The answer, it turns out, is complicated. In order to show the extent of this complication, I first need to say something about de Finetti's celebrated *general representation theorem*. This theorem will then allow me to make the final connection to the importance, and difficulty, of addressing the probativist criticism in section 2.6.

[46] After Doob (1949). An excellent discussion of the scope and significance of the results is given by Schwartz (1965). See Huttegger (2017, §5.5) for an introduction and discussion of the philosophical significance. Compare with Earman (1992, Ch. 6). See Baldi (2017) for an advanced but still accessible treatment.

2.5 The General Representation Theorem and Belot's criticism

Exchangeability is a subjective judgment of symmetry about a probability measure on observables X_i, which could be finite or infinite.[47] For $i = 1, \ldots, n, X_i$ are judged to be *finitely exchangeable* to me if their joint probability distribution function is invariant under a permutation of its arguments. An infinite sequence of X_i is *exchangeable* to me if every finite subsequence is exchangeable. Exchangeability and *partial exchangeability* (which I have not defined) operationalize, respectively, random sampling and stratified sampling techniques conditional on parameters.[48]

de Finetti's representation theorem for 0-1 random variables

If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random variables with probability measure P, then there exists a distribution function Q such that the joint mass function $f(X_1, \ldots, X_n)$ for $X_1, \ldots, X_n$ is given by

$$f(X_1, \ldots, X_n) = \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \right\} dQ(\theta) \tag{2.7}$$

where

$$Q(t) = \lim_{n\to\infty} P\left(\frac{\sum_i^n X_i}{n} \leq t \right)$$

and θ is defined as the strong law limit of $\frac{\sum_i^n X_i}{n}$.

[47] All observables are random variables.

[48] See Bernardo and Smith (2000, 168 – 171).

Here are three takeaways from de Finetti's 0-1 representation theorem. First, it justifies the existence and the use of priors on observables (specifically events) by Bayesians. Second, this theorem can be interpreted as a generalization of the Law of Small Numbers. Recall that this Law says that average probability is always the prevision of frequencies. The left hand side of (2.7) is the average probability distribution of the X_i, or as we would say today the *joint marginal distribution* of the $X_i\,(i = 1, \ldots, n)$ by marginalizing out θ with $dQ(\theta)$. The right hand side of (2.7) is the prevision of the joint frequency $\left\{ \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \right\}$ conditional on θ with weights $dQ(\theta)$. The fact that the left hand side is a marginal distribution highlights that θ and the $X_i\,(i = 1, \ldots, n)$ are on par as observables and the representation theorem is de Finetti's way of providing an operational definition of θ in place of the oft-repeated "fixed but non-random parameters" in non-Bayesian frameworks. There is no dualism in de Finetti's framework — everything modeled statistically is an observable. Third, since sometimes θ for 0-1 variables is described as the objective chance of "success", de Finetti's 0-1 representation theorem shows that even this notion can be accommodated within a subjective framework. Parameters are strong law limits of sufficient statistics on observables.[49] These takeaways are genuine unifications without dualism — which is why de Finetti is relevant for the possibility of probative foundations for Bayesian statistics.

de Finetti's General Representation Theorem

If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of variables with probability measure P, then there exists a distribution function Q over $\mathcal{F}$, the set of all distribution functions on $\mathbb{R}$, such that the joint distribution of $(X_1, X_2, \ldots, X_n)$ has the form

$$f(X_1, X_2, \ldots, X_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(X_i) dQ(F)$$

[49] In Appendix 1 of de Finetti (1974b) compares parameters with densities of points in continuous solid bodies. In the case of continuous bodies the density at a point is also a limiting quantity since bodies, in fact, have atomic structure, i.e., have a discrete micro-structure.

where

$$Q(F_{\boldsymbol{X}}(t)) = \lim_{n\to\infty} P(F_n(t))$$

and $F_n(t) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$ is the empirical distribution for $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$.

It is the extension of de Finetti's theorem from the case of 0-1 random variables to the general case that complicates the story about the statistical consistency of Bayesian Estimators. Here's how. Recall that in the 0-1 case, θ in $dQ(\theta)$ is defined as the strong law limit of $\bar{X}$, which is a sufficient statistic. The existence of θ is thus a consequence of the strong law of large numbers (assuming X_i are exchangeable). So in the 0-1 case we have calibration in the sense of statistical consistency. The case of 0-1 random variables and other cases where a sufficient statistic exists are, loosely speaking, the "easy cases" for obtaining statistical consistency.

To see this, ask: what is F in $dQ(F)$? It is tempting to think of F analogously as a "parameter" and the X_i as identically and independently distributed conditional on F assuming they are exchangeable. But the general representation theorem does not operationalize the definition of F as a strong law limit. Notice that in general $F_n(t)$ is a function into a set of probability measures $\mathcal{F}$, which is a topological space — the set of Borel sigma algebras on $\mathcal{F}$ with the topology of weak convergence.[50] Even if we let F be a "parameter", this would imply that Q is a prior distribution from an infinite dimensional family of distributions on $\mathcal{F}$. The family is infinite dimensional because for *each* distribution F (whose *form* we don't know) there is a set of parameters (in the strong limit law sense) that identifies it. Infinite dimensional function spaces lead to *Bayesian Non-parametric* methods because

[50] Let F_n and F be distribution functions, then F_n converges weakly to F if $\lim_{n\to\infty} F_n(x) = F(x)$ for each x at which F is continuous. See Theorem 25.8 in Billingsley (2012, 358) for equivalent formulations of weak convergence for measures.

no assumption regarding the form of the F's, and consequently what parameters they can have, is made. Venturing into Bayesian non-parametrics is beyond the scope of my paper.[51]

However, I wish to draw some connections between the interpretation and implications of de Finetti's general representation theorem to issues within philosophy of science that make the probativist issue salient. Let me start with Dawid then move on to Belot. Both of them charge Bayesians with being too confident, but their points are really entirely different. What Dawid (1982) showed is that a Bayesian who engages in sequential forecasts with feedback will asymptotically be empirically calibrated with the truth almost surely, i.e., with probability 1 in their own beliefs.[52] The talk of *forecasts* is distracting because one can understand Dawid's point without necessarily tying it to forecast calibration. This means, in particular, that objecting to Dawid's results because they concern assessments of forecasters is beside the point. My discussion so far can be used to see this; for it is a corollary to the general representation theorem that predictive distributions are previsions.[53]

Predictive Distributions

If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of real valued observables admitting a density $f(\cdot \,|\, \theta)$ with respect to $dQ(\theta)$, then

$$f(X_{m+1}, \ldots, X_n \,|\, X_1, \ldots, X_m) = \int_\Theta \left\{ \prod_{i=m+1}^{n} f(x_i \,|\, \theta) \right\} dQ(\theta \,|\, x_1, \ldots, x_m)$$

[51] Diaconis and Freedman (1986) is a good place to start for discussions on choices of prior measures $dQ(F)$ in non-parametric settings. See Ghosh and Ramamoorthi (2003), Walker (2005) and Ghosal and van der Vaart (2017) for additional references.

[52] See Kadane and Lichtenstein (1982) for an illuminating discussion.

[53] See Corollary 2 in Bernardo and Smith (2000, 180).

where

$$dQ(\theta|x_1,\ldots,x_m) = \frac{\prod_{i=1}^{m} f(x_i|\theta)dQ(\theta)}{\int_\Theta \prod_{i=1}^{m} f(x_i|\theta)dQ(\theta|x_1,\ldots,x_m)}$$

Here $f(X_{m+1},\ldots,X_n \,|\, X_1,\ldots,X_m)$ is the joint predictive distribution for future observables conditional on past observables (i.e., "feedback" or data). In other words, the predictive distribution of future observables is the prevision of their joint distribution weighted by the posterior distribution.

As a coherent Bayesian, Dawid accepts Cromwell's rule according to which only tautologies receive subjective probability 1 on pain of incoherence. But since contingent matters of fact are not tautologies, Dawid thinks that he cannot assign them subjective probability 1 *even though* his convergence theorem for sequential forecasts with feedback says that he should. He writes:

> We have a paradox: an event can be distinguished (easily, and indeed in many ways) that is given subjective probability one and yet is not regarded as "morally certain." How can the theory of coherence, which is founded on assumptions of rationality, allow such an irrational conclusion?

The complaint here is that Bayesians are always sure (probability one in their own beliefs) that they will converge, although this need not translate to probability one in the real chances.

Dawid's puzzle, in my view, is only apparent. The reason why his convergence result puzzles him is that he maintains a dualism between coherent subjective probabilities and objective

chances. The unification we've seen from de Finetti shows predictive distributions on future observables are previsions, i.e., estimators, not *assertions* about objective chances.[54]

However, there is another way to think about Dawid's result in relation to the de Finetti general representation theorem. It is this. By supposing in section 4.3 "that θ is consistently estimable", Dawid is not, in fact, begging the question. He is restricting the set of distribution functions on $\mathcal{F}$ to finite dimensional parametric models in which the data structure is such that consistency is possible in principle (i.e., parameters can be identified via data). It is quite common, actually, for Bayesian statistics to proceed like this. The idea is to select forms of F (or more technically subclasses of $\mathcal{F}$ called *finite dimensional parametric models*) that allow consistent estimation of parameters. The specification of the prior measure $dQ(\theta)$ implied by the general representation problem is then more tractable for these forms.[55]

This brings us to Belot. Belot has argued, using topology, that events which are measure-theoretically negligible (i.e., the null events, which receive subjective probability zero) can nevertheless be topologically huge (i.e., residual in the space). Therefore, the failure set, even for a coherent Bayesian, is "typical." Some have pointed out that Belot's argument leaves everyone, not just Bayesians, in a quandary. For his complaint can sometimes be read as a complaint regarding using probability at all since the complaint is directed at the strong law of large numbers. So if the argument is successful, then it is successful against both Bayesians and non-Bayesians.

I read Belot differently. The challenge he raises for Bayesians is a genuine challenge. To see this, let's revisit de Finetti's general representation theorem. As mentioned, $\mathcal{F}$ is a

[54]Compare van Fraassen (1984).

[55]Bernardo and Smith (2000, Ch. 4) is an excellent discussion of how this restriction or specification can be accomplished based on principles like *invariance*, *sufficient statistics*, *partial exchangeability and hierarchical models*. See also Fortini, Ladelli, and Regazzini (2000) and Diaconis and Skyrms (2018, Ch. 7) for other principles. Vanpaemel (2010) is an excellent discussion on principles that can guide model specification in social sciences like psychology, although I believe the results here are applicable to Bayesian parametric modeling more generally.

topological space. So $dQ(\cdot)$ is a measure on a topological space.[56] It is not obvious to me that the right way to respond to Belot is to say that he is switching from measure theory to topology, thereby undermining the whole field of statistics by calling into question the strong law of large numbers. In fact, within Bayesian non-parametrics measures on topological spaces are routinely used. My own tentative view is that Belot is drawing attention to a *limitation* on the scope of Bayesian statistical consistency results. His point, I take it, is that the a priori restriction of measures on $\mathcal{F}$ to subclasses of parametric models that guarantee statistical consistency by assigning all other classes $dQ(\cdot)$-measure 0 is a *huge restriction*. The complement of the class of parametric models that results from the a priori restriction is residual in $\mathcal{F}$, although the a priori restriction declares it $dQ(\cdot)$-negligible. Perhaps Belot is correct.[57]

But it does not follow that Bayesians are not humble. The message I take away is different. It is the message that it is difficult to show, as a coherent Bayesian, that one is statistically consistent because the modeling assumptions in specifying Q for a given problem are, where justified, pragmatic and context-dependent.[58] This is humbling and it is what I take to underscore the difficulty of addressing the probativist criticism within a coherent Bayesian framework.

2.6 Is there hope?

The answer to the question in the title of this section depends on one's takeaway from the connection I have drawn between Belot results and the probativist criticism. One takeaway could be to say that the challenge raised by the probativist criticism is to find prior measures

[56] For a rigorous discussion of measures on topological spaces see Bogachev (1998, Ch. 5).

[57] See Ghosal and van der Vaart (2017, 5 – 9, and §§6.1 and 6.3) and compare with the discussion in Cosma Shalizi's blog post from 2023 available here: `http://bactra.org/notebooks/bayesian-consistency.html`.

[58] Compare with Morey, Romeijn, and Rouder (2016, 9 – 10).

dQ on $\mathcal{F}$ that are consistent except on the dQ-negligible sets. In this non-parametric setting, as discussed by Diaconis and Freedman (1986), the class of tail-free and Dirichlet priors can be used. In parametric settings, some subjective Bayesians have argued intersubjective merging of opinions based on weak-star convergence is equivalent to statistical consistency. This is theorem 3 in Diaconis and Freedman (1986).[59]

Another takeaway could be to say that the challenge raised by the probativist criticism, especially in parametric settings, is to exercise *prudence* with regard to the possibility of error. I believe that this is how Mayo would want us to understand the minimal requirement for severity. Recall the question from Mayo (2018, 228):

> How do I criticize your prior degrees of belief? [I]f we are not trying to model opinions, but instead insist on meeting requirements for objective scrutiny? For these goals, inner coherence or consistency among your beliefs is not enough. One can be consistently wrong, as everyone knows (or should know).

On this way of seeing things, Box (1980) recommended using *marginal p-values* to assess the prior distribution. Rubin (1984) and Gelman, Meng, and Stern (1996) introduced *posterior predictive p-values* that could be used for model assessment as well. The main drawback of these checks on "error" is that prima facie these measures "violate" the Likelihood Principle. For this reason, coherent Bayesians like Haaf, Klaasen, and Rouder (2021) advocate the *specification-first principle* in their comparative study of Bayes Factors and posterior predictive assessments.[60] Whether the checks based on Bayesian p-values are genuine violations or simply a failure to apply the Likelihood Principle to certain problems is part of the topic in Mwakima (2024c). Adjudicating on these proposals would be a fruitful way to drive forward the conversation.

[59] See also Huttegger (2015a), Huttegger (2015b) and Huttegger (2017, Ch. 8) for discussions on merging of opinions in variational distance, which underly the Blackwell-Dubins merging results.

[60] Compare with Draper (1995), Draper's discussion in the Gelman, Meng, and Stern (1996) paper, Draper (2006) and Vanpaemel (2010).

2.7 Conclusion

Many within the applied Bayesian statistical community have proposed eclectic unifications between the best of two worlds — fundamentally sound Bayesian methods in statistics and sampling theory methods in statistics (sometimes referred to as classical methods in statistics, which include Mayo's error-statistics based on severity) — in order to address the debate around the possibility of probative foundations for Bayesian statistics. In this paper, my main contribution has been to bring de Finetti into this debate in order to illuminate some of the issues. After all, his work had a formative influence on "the received view" that is being rejected by this community. I have argued that doing so is fruitful for two reasons. First, from de Finetti we get unifications without the dualism between subjective probabilities and objective probabilities (understood as limits of relative frequencies). Second, from de Finetti we have principles (that a coherent Bayesian would accept), which show that the probativist criticism is a genuinely humbling problem.

Chapter 3

On the Scope of the Likelihood Principle

3.1 Introduction

The protracted controversies in statistical practice depend, for the most part, on what one accepts as fundamental principles of correct statistical inference.[1] One such principle is the Likelihood Principle, which was first isolated by Barnard and Feber (1947) and Barnard (1949) from ideas that trace back to Fisher's pioneering work from the 1920s and 30s on parameter estimation.[2] This principle states that parametric inference within a statistical model should be based on the equivalence class of functions of the parameter in which the data is fixed (this class of functions is also called *the likelihood function*). In the next section I say more precisely what I mean by a statistical model and what the relevant equivalence relation that determines the equivalence class is. Call two or more experiments with fixed

[1] See Cox and Hinkley (1974, 36 – 58), Box (1982), Robins and Wasserman (2000) and Mayo (2018, §1.5 and §4.4).

[2] See especially Fisher (1922), Fisher (1925) and Fisher (1934).

data *likelihood equivalent experiments* just in case they have likelihood functions in the same equivalence class. The rough content of the Likelihood Principle is that the *same* parametric inference should be made from likelihood equivalent experiments.

To illustrate the idea, consider the following data from a coin tossing experiment:

$$\boldsymbol{x} = \langle T,\ H,\ T,\ T,\ H,\ H,\ T,\ H,\ H,\ H \rangle$$

The general problem of parametric inference is to model the data-generating process suitably and to use the data to make inferences about the parameters of this process, which we denote as $\boldsymbol{\theta}$. Asked the question, "Will your parametric inference *after* seeing $\boldsymbol{x}$ depend on *how* the 10 coin tosses were obtained?", some people will answer, "No, the data is what it is."[3] To see why other people would answer, "Yes"; imagine that the data could have been observed following any one of the following four experiments (we don't know which):

- E_1: Toss the coin exactly 10 times;
- E_2: Continue tossing until 6 heads appear;
- E_3: Continue tossing until 3 consecutive heads appear;
- E_4: Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.

Those who answer "Yes", often think that even though all of these four experiments can lead to the same data (6 heads and 4 tails); they have *different sample spaces*.[4] The sample

[3]Let me disambiguate a few things. Here the "how" question refers to the description of the coin tossing experiment, which determines a statistical model. The "how" question does not just refer to *the stopping rule*, which is a protocol for determining when a sequential experiment should end. The "No" answer is plausible because it can be shown, mathematically, that there are statistical models — those with equivalent minimal sufficient statistics — such that the description of the experimental set up (specifically the sample space) is *irrelevant* for parametric inference once the data has been obtained.

[4]It is sometimes claimed that these experiments are different because they are sequential experiments with different stopping rules. I do not like this way of motivating the Likelihood Principle because it suggests

space in E_1 is bounded, while the sample spaces for E_2, E_3 and E_4 are unbounded. For this reason, the right statistical model for E_1 is a Binomial model, while the right statistical model for E_2 is Negative Binomial. The sample spaces for E_3 and E_4 are more complex than the sample spaces for the previous two experiments because they involve sequences of coin flips until a specific pattern is achieved. If you still think that the sample space does not matter for "correct" parametric inference about $\boldsymbol{\theta}$ for fixed data $\boldsymbol{x}$, then you accept the Likelihood Principle.

What is controversial within statistical circles is that if one accepts the Likelihood Principle, it is alleged that a lot of seemingly reasonable considerations become irrelevant for parametric inference. For example, the following components of some approaches to parametric inference are, according to (Basu, 1975, 16), inconsistent with the Likelihood Principle. The reason is that each of these quantities (which are mathematical expectations) will vary depending on the sample space even for any two or more experiments that are likelihood equivalent.

- Bias and standard error of point estimates
- Probabilities of the two kinds of errors for a test
- The confidence coefficients associated with interval estimates

One controversial consequence of the Likelihood Principle's disregard for the relevance of the sample space is what has come to be known as "The Stopping Rule Principle" (Berger and Wolpert, 1988, 76).

The Stopping Rule Principle

that the issues raised by this principle only have to do with sequential experiments. However, motivating the Likelihood Principle in terms of the irrelevance of sample spaces generalizes the issues to descriptions of experiments *in general*. This is why the Stopping Rule Principle is a consequence, or special case, of the Likelihood Principle (see a statement of the Stopping Rule Principle below).

In a sequential experiment with observed final data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$*, the information from the experiment should not depend on the stopping rule.*[5]

During a famous discussion, Leonard Savage, one of the pioneers of Bayesian statistics, once remarked (see Barnard and Cox (1962, 76)):

> I learned the stopping-rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right.

The resistance lives on.[6] But so does a defense.[7] Much ink has been spilled on the validity and the scope of the Likelihood Principle both formally in peer-reviewed journals and published books; as well as informally on internet blogs and discussion forums hosted by Deborah Mayo and Andrew Gelman.[8] Such opposition to the Likelihood Principle is fueled in large part by the fact that this principle curtails a significant part of the practice of present day statistics known as *sampling theory*.[9]

[5]This the exact statement of this principle by these authors. My own view is that the phrase "information" should be replaced with "inference made" because "information" is puzzling. Further, the statement of this principle assumes that the stopping rule is *noninformative*, which means that the stopping rule statistic is ancillary for the main parameter(s) of interest from the experiment (see section 3 for discussion of ancillarity).

[6]For a recent discussion see Fletcher (2019), de Heide and Grünwald (2021), and Fletcher (2024).

[7]See Rouder (2014), Rouder and Haaf (2019) and compare with and Hendriksen, de Heide, and Grünwald (2021).

[8]See Birnbaum (1962) (with discussion), Barnard and Cox (1962), Barnard, Jenkins, and Winsten (1962), Basu (1964) Durbin (1970) (with discussion), Birnbaum (1972), Kalbfleisch (1975) (with discussion), Basu (1975), Berger (1984a), Hill (1987) and Berger and Wolpert (1988) (with discussion), Bjørnstad (1996), Mayo and Kruse (2001), Mayo (2014) (with discussion) and Gandenberger (2015).

[9]This term should not be confused with the study of experimental designs or sampling techniques for collecting data from a population. The term 'sampling theory' (also used by Box (1980), Box (1982) and Cox and Hinkley (1974)) is simply a useful contrasting term to coherent Bayesian methods in statistics. One reason for my preference is that it does not necessarily assume that *every* sampling theorist is committed to the limit of relative frequencies interpretation of probability (even though many are).

Sampling theory uses both exact and asymptotic *sampling distributions* of statistics to perform parametric inference. Sampling theorists rely on a different principle — the *Repeated Sampling Principle* — to justify their procedures in opposition to the strictures imposed by the Likelihood Principle.[10]

The Repeated Sampling Principle

Statistical procedures are to be assessed by their behavior in hypothetical repetitions under the same conditions; measures of uncertainty are to be interpreted as hypothetical frequencies in long run repetitions; and criteria of optimality are to be formulated in terms of behavior in hypothetical repetitions.

Bracketing the extreme behaviorist/performance language and the commitment to the limit of relative frequencies interpretation of probability, here's how I *charitably* read this principle: the Repeated Sampling Principle uses optimality criteria based on sampling distributions to calibrate the risk of various methods and decision procedures for making statistical inference. For example, Neyman-Pearson hypothesis testing involves probing statistical models to determine fit or compatibility with the data based on optimal values, which are called the *size* and *power* of a test, of suitably selected risk functions. Mayo and Spanos use the severity function to extend the probing capacities of sampling theory. Bootstrapping methods in statistical inference also rely on sampling theory. Finally, Conditional Bayes, Frequentist Bayes and Pragmatic Bayesians involve pragmatic unifications of sampling theory and Bayesian methods in statistics. In fact, if one would want to calibrate Bayes Factors, one would need to address the scope of the Likelihood Principle and to consider supplementing it in some way using ideas from sampling theory.[11]

[10] See Cox and Hinkley (1974, 45) and Reid and Cox (2015).

[11] See section 3.3.1 below for my discussion of why one would want to do this. It is an open question whether one can do this and whether it is appropriate.

What is interesting philosophically is whether the *scope* of the Likelihood Principle extends to these practices of sampling theorists. If one takes the position that the scope of the Likelihood Principle is *wide* enough to include *all* the different kinds of parametric inference (point estimation, interval estimation, model criticism and model comparison), then the practices of sampling theorists constitute violations of the Likelihood Principle.[12] Such a violation is "a bad thing" in the sense that it is decision-theoretically incoherent (or inadmissible) to contradict the Likelihood Principle.[13] However, one can take a different position and say that the scope of the Likelihood Principle is much *narrower*. In the narrower scope case, some of these practices of sampling theorists are not in violation of the Likelihood Principle — the Likelihood Principle simply doesn't *apply*. Trying to restrict the scope of the Likelihood Principle along these lines has been suggested by several prominent statisticians.[14] However, what we lack from these practitioners are *arguments* and a clear *analysis* for: (1) why imposing some kind of restriction is desirable; and (2) how cases of "violations" can be distinguished from cases of "failures to apply". Here's how Mayo (2018, 303) puts it (LP in her text is an abbreviation of the Likelihood Principle):

> Another little puzzle arises in telling what's true about the LP: Is the LP violated or simply inapplicable in secondary testing of model assumptions[?] For them [Casella and Berger], it appears, the LP is full out violated in model checking. I'm not sure how much turns on whether the LP is regarded as violated or merely inapplicable in testing assumptions; a question arises in either case.

[12]I distinguish these kinds of parametric inference in the next section. My discussion of the Likelihood Principle is in the context of just parametric inference (as opposed to non-parametric inference) because in parametric inference, the concept of a likelihood function is well-understood — the parameter space is finite-dimensional, for example. Moreover, the issues that animate the debate regarding the scope and applicability of the Likelihood Principle arise mainly in parametric contexts.

[13]See Berger (1983). In the ensuing discussion of this paper, Berger agreed with Bernardo that there are contexts in which the principle doesn't apply. So even then, the scope of Berger's result required some qualification. See section 3.3.1 below for more on this.

[14]See Fisher (1956, 49), Barnard and McLaren in the comments of Kalbfleisch (1975), Box (1982), Bernardo and Berger in the discussion of Berger (1983) and Casella and Berger (2002), who are also discussed by Mayo (2018, 302ff.).

My main contribution in this paper is to remove the sort of puzzle or uncertainty Mayo is alluding to and to provide the missing arguments and clarifying philosophical analysis. I will argue that the Likelihood Principle is essentially a dimension reduction principle that only applies to problems of *point estimation*. If likelihood equivalent experiments lead to different values of point estimates, this is a violation of the Likelihood Principle. The Likelihood Principle does not apply to parametric inferences that involve evaluating procedures in accordance with the Repeated Sampling Principle.

Here's how I have organized the rest of my paper. In the next section I give a preliminary overview of the key mathematical concepts that underlie the formulation of "the" Likelihood Principle.[15] Originally introduced by Fisher, these are the concepts of *likelihood*, *sufficiency*, *ancillarity* and *efficiency*. In section 3.3, I discuss what reasons there are for a coherent Bayesian and a sampling theorist to want the kind of scope restriction I am arguing for. In section 3.4, I present my two arguments for restricting the scope: one argument is against formalism; the other argument is based on reinterpreting the *Conditionality Principle* in a way that I believe is consistent with Fisher's original goal for introducing the concept of ancillarity.[16] The goal can be put this way: knowledge of the experiment that was actually performed is an ancillary statistic, which, conditioning upon, can improve the efficiency of point estimators. In section 3.5 I discuss the advantages of the kind of restriction I advocate for. I conclude in section 3.7 after considering a number of objections and replying to them in section 3.6.

[15]The word 'the' is in scare-quotes because it is hard to find a consensus view on, or a univocal statement of, the scope of the Likelihood Principle (see some of the statements I select below). Given this difficulty, the overview I provide in the next section can be read as my way of getting at the precise mathematical *content* of this principle (and what is entailed by it). The discussion here will allow me to contrast my account with the *Formal* Likelihood Principle, which I blame for the conflicts within statistics and its philosophy.

[16]The Conditionality Principle states that parametric inference (in a mixed experiment, for example) should be conditional on the actual experiment that was performed. This statement of this principle comes from Cox and Hinkley (1974, 38)'s influential text, where it is a normative principle rather than an evidential equivalence principle.

3.2 Preliminaries

3.2.1 The Likelihood Function

Parametric inference starts with a statistical model $\mathcal{M} = \{f(\cdot\,;\boldsymbol{\theta}) \,|\, \boldsymbol{\theta} \in \Theta\}$, which is a class of probability distributions (each of which is identified by the value of $\boldsymbol{\theta}$ it takes) for random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. $\boldsymbol{\theta}$ is called a *parameter* and takes values in $\boldsymbol{\Theta} \subset \mathbb{R}^d$, the *parameter space*. The X_i $(i = 1, \ldots, n)$ take values in a sample space $\boldsymbol{\mathcal{X}}$, which can be discrete (e.g., non-negative integers (finite or infinite)) or continuous (e.g., real valued). For a realization $\boldsymbol{X} = \boldsymbol{x}$ (called *the data*), the goal of parametric inference is to use the data to identify the value of $\boldsymbol{\theta}$ such that $f(\boldsymbol{x};\boldsymbol{\theta})$ is a probabilistic description of the process that led to the data $\boldsymbol{x}$. With this goal in mind, let me distinguish the following problems of parametric inference:

1. Given a model $\mathcal{M}$, find an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ using a point estimator $T(\boldsymbol{x})$ according to some optimality criteria. (This is the problem of *point estimation*.)

2. Given a model $\mathcal{M}$ and $\{\theta^0\} \subset \boldsymbol{\theta}$ as the one-dimensional parameter of interest, find an interval estimator $(L(\boldsymbol{x}), U(\boldsymbol{x}))$ that according to some optimality criteria probably covers θ^0.[17](This is the problem of *interval estimation*.)

3. Given a model $\mathcal{M}$, use a statistic $T(\boldsymbol{x})$ to probe the adequacy of a distribution $f(T(\boldsymbol{x});\boldsymbol{\theta})$ in $\mathcal{M}$ according to some optimality criteria. (This is the problem of *model criticism* or *model assessment*.)

4. Given two or more specified models, use statistics to make comparisons among them according to some comparative measure of evidence. (This is the problem of *model comparison*.)

[17]Here $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ are, respectively, the lower bound and upper bound of the interval estimator for θ^0 and are functions of the data. The one-dimensional dimensional case, like the one I have described for estimating θ^0, generalizes to *regional/set estimators* in higher dimensions.

For some model $\mathcal{M}$ and for a fixed value of $\boldsymbol{x}$, progress on solving these problems can be made by conceptualizing $f(\boldsymbol{x};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ alone and by analysing how $f(\boldsymbol{x};\boldsymbol{\theta})$ varies within $\boldsymbol{\Theta}$. This conceptualization leads to the concept of what Fisher (1922) called the *likelihood* or *Likelihood Function*, which I will denote by $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})$ to emphasize that it is a function of $\boldsymbol{\theta}$ for data $\boldsymbol{x}$ that is fixed. If $\boldsymbol{X}$ is a sequence of (conditionally) independent and identically distributed (i.i.d.) random variables,

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) \tag{3.1}$$

Equation (3.1) says that the Likelihood Function for (conditionally) i.i.d. data returns the value of the joint probability distribution of the fixed data $\boldsymbol{x}$ as a function of $\boldsymbol{\theta}$. The *log likelihood function* $\ell(\boldsymbol{\theta};\boldsymbol{x})$ is

$$\ell(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{i=1}^{n} \mathsf{log}(f(x_i;\boldsymbol{\theta})) \tag{3.2}$$

.

One optimality criterion for solving the problem of point estimation is to find an $\hat{\boldsymbol{\theta}}$ such that for fixed $\boldsymbol{x}$, the value of $\ell(\boldsymbol{\theta};\boldsymbol{x})$ is at least locally the highest. We know from optimization problems in calculus that the first and second order derivatives of a function are important for finding and for characterizing the *local extrema* (i.e., local maximum and local minimum values) of many functions. So, in the case of problems of point estimation, one way of finding optimal values (called *maximum likelihood estimates* by Fisher) is to consider the gradient of $\ell(\boldsymbol{\theta};\boldsymbol{x})$ (or *first order* derivatives of $\ell(\boldsymbol{\theta};\boldsymbol{x})$) and the Hessian of $\ell(\boldsymbol{\theta};\boldsymbol{x})$ (or *second order* derivatives of $\ell(\boldsymbol{\theta};\boldsymbol{x})$). This suggests that in problems of point estimation, any part of $\ell(\boldsymbol{\theta};\boldsymbol{x})$ that is free of $\boldsymbol{\theta}$ can be treated as a constant function when taking partial derivatives, which implies that its first and second order derivatives are zero. It is for this reason that

we are interested in $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})$ up to a multiplicative constant function that is free of $\boldsymbol{\theta}$.[18] Let $g(\boldsymbol{y}; \boldsymbol{\theta})$ from a model $\mathcal{M}'$ with the same parameter space as the model $\mathcal{M}$ be the probabilistic description of a process that led to data $\boldsymbol{y}$. Define the equivalence relation $\sim$ by:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) \sim \mathcal{L}'(\boldsymbol{\theta}; \boldsymbol{y}) \Longleftrightarrow f(\boldsymbol{x}; \boldsymbol{\theta}) = c(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{y}; \boldsymbol{\theta}) \tag{3.3}$$

Equation (3.3) says that two likelihood functions (one with fixed data $\boldsymbol{x}$ and the other with fixed data $\boldsymbol{y}$), from possibly *different* models but with the same parameter space, are equivalent if and only if their ratio is a constant function of $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., free of $\boldsymbol{\theta}$. This is what it means to say that in the context of parametric inference the likelihood function is an equivalence class of functions of $\boldsymbol{\theta}$ with fixed data.

3.2.2 Sufficiency, Efficiency and Ancillarity

The discussion I give in this section of some of the technical notions is aimed at a philosophical audience. Specialists who consult Casella and Berger (2002) may find that my discussion deviates a bit from this standard reference in mathematical statistics. Let me say something about this. First, I have used an equally influential text — Cox and Hinkley (1974) — and compared the definitions of these concepts there to those found in Casella and Berger (2002) and Fisher (1925, §§2, 9). Second, what I am doing here constitutes a significant improvement in philosophical analysis of these concepts (especially on the relation between sufficiency, ancillarity and efficiency) compared to the analysis of Casella and Berger (2002, §§6.2.3 – 6.2.4). They write, "paradoxically, an ancillary statistic, when used in conjunction

[18] Some readers will recognize that this is a consequence of the elementary fact that the extrema of a function are exactly the extrema of its (positive) scalar multiples. I am going through the calculational aspects for the following reasons. First, it is often not stated by many authors *why* the likelihood function for $\boldsymbol{\theta}$ is defined up to multiplication by a constant function that is free of $\boldsymbol{\theta}$. Second, the calculational aspects show how or why this definition arises naturally within the context of optimization problems involving point estimation.

with other statistics, sometimes does contain valuable information for inferences about θ." Why they think this is "paradoxical" is unclear to me. My analysis brings out the *intuition* behind Fisher's choices of these specific labels for the concepts he introduced in his papers from the 1920s and 30s — there is nothing paradoxical about them. In fact, the labels are quite *suggestive* of the ideas they capture.

To begin with, these concepts emerged for Fisher in the context is *finite sampling*, rather than *asymptotics*, where the sample size is allowed to go to infinity. This is *small sample* parametric inference. Small sample parametric inference (even with big data) is what finitely resourced individuals and machines are realistically capable of. It proceeds via *dimension reduction* and *optimization*.[19] The sense of "reduction" implied here is reducing the number of dimensions *of points* in the sample space by partitioning the sample space into equivalence classes of lower dimension than the original points using a statistic $T(\boldsymbol{x})$.

For example, suppose the goal is to model the data-generating process of a sequence of three coin tosses. Here the Binomial family of probability distributions $f(\cdot\,;\boldsymbol{\theta}) = \text{Bin}(n = 3, \theta)$ is a suitable model. Parametric inference is about $\boldsymbol{\theta} = \theta$ with $\boldsymbol{\Theta} = [0, 1]$. The sample space is the set of all possible 3-dimensional sequences of heads and tails. The statistic $T(\boldsymbol{x})$, which counts the number of heads, generates one-dimensional equivalence classes.

Here's another example. Suppose that the goal is to model the data-generating process of a sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ (say measurements) of a continuous random variable $\boldsymbol{X}$. Here the Normal family of probability distributions $f(\cdot\,;\boldsymbol{\theta}) = N(\mu, \sigma^2)$ is a suitable model. Parametric inference is about $\boldsymbol{\theta} = (\mu, \sigma^2)$ with $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$. The sample space is the set of all possible n-dimensional sequences in $\mathbb{R}^n$. But the statistic $T(\boldsymbol{x}) = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}$, the mean, generates equivalence classes in $\mathbb{R}$, which are one-dimensional statistics or estimators for μ. At the same time, $T'(\boldsymbol{x}) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1} = s^2$, the estimated unbiased sample variance, generates equivalence classes of one-dimensional estimators for σ^2.

[19] Dimension reduction is sometimes called data reduction or summarization.

Since there are infinitely many statistics that one can use as estimators for $\boldsymbol{\theta}$, are there optimality criteria for selecting these estimators? Fisher in his seminal papers from the 1920s and 30s considered some criteria which included *sufficiency*, *ancillarity*, *consistency*, *efficiency*, *invariance* and so on as optimal properties that an ideal estimator should possess. For the proof of the Likelihood Principle that Birnbaum gave, the key properties are *sufficiency* and *ancillarity*. Although, as we shall see in section 3.4.2. below, *efficiency* is important for restricting the scope of the Likelihood Principle to problems of point estimation.

For ease of notation and for the sake of illustrating the main ideas I will switch from boldface $\boldsymbol{\theta}$ to θ to indicate that I am looking at estimating just one parameter in $\boldsymbol{\theta}$ — the parameter of interest. If $\boldsymbol{\theta}$ is multi-dimensional, the rest of the parameters are *nuisance parameters*. A statistic $T(\boldsymbol{x})$ is *sufficient* for estimating θ if provided you know $T(\boldsymbol{x}) = t$ you *don't need* any other statistic $T'(\boldsymbol{x})$ to estimate θ accurately or with precision.

Prima facie, this qualitative rendering of sufficiency depends on the epistemic position of the agent considering it, which is not a notion that plays a role in sampling theory. So, why am I talking in terms of "accuracy" and "precision"? The answer has to do with the relationship between the variance of $T(\boldsymbol{x})$, its precision and *Fisher Information*, $I(\theta)$. Under suitable regularity conditions (see Miscellanea 10.6.2 in Casella and Berger (2002, 516)), $I(\theta)$ from one observation $X = x$ is defined by:

$$I(\theta) =_{\text{def}} \text{Var}\left(\frac{\partial \ell(\theta; x)}{\partial \theta}\right) = \mathbb{E}\left[\left(\frac{\partial \ell(\theta; x)}{\partial \theta}\right)^2\right]$$

and is computed using:

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2 \ell(\theta; x)}{\partial \theta^2}\right].$$

For a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of size n, the *observed Fisher Information* is $nI(\theta)$. Under suitable regularity conditions, the Cramér-Rao Inequality says that:

$$\mathrm{Var}(T(\boldsymbol{x})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}[T(\boldsymbol{x})]\right)^2}{nI(\theta)}.$$

Let $T(\boldsymbol{x})$ be an unbiased estimator for θ (in this case $\frac{d}{d\theta}\mathbb{E}[T(\boldsymbol{x})] = 1$), then $T(\boldsymbol{x})$ is an *efficient* estimator of θ if and only if

$$\mathrm{Var}(T(\boldsymbol{x})) = \frac{1}{nI(\theta)}.$$

But $\frac{1}{\mathrm{Var}(T(\boldsymbol{x}))}$ is called the *precision*, especially within Bayesian statistics. This means that the accuracy of an estimate $\hat{\theta}$ improves with the precision of an unbiased estimator $T(\boldsymbol{x})$ of θ. This is why I am discussing the concept of sufficiency in terms of accurate point estimation.

In order to tie further the discussion in the previous paragraph to sufficient statistics I need the following facts:

(1) The whole sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a sufficient statistic.

(2) A one-to-one function of a sufficient statistic is a sufficient statistic.

(3) The set of order statistics of a sample is sufficient.[20]

These facts imply that many other sufficient statistics exist provided one exists. Within the class of sufficient statistics, a *minimal sufficient statistic* is a sufficient statistic such that no further dimensional reduction is possible from it without loss of precision. Some facts about minimal sufficient statistics are:

[20]The set of ordered statistics is indeed the whole sample. But the whole sample need not be ordered.

(1) $T(\boldsymbol{x})$ is a minimal sufficient statistic if for any sufficient statistic $S(\boldsymbol{x})$, $T(\boldsymbol{x})$ is a function of $S(\boldsymbol{x})$.

(2) The likelihood function is minimal sufficient.

Fact (2) about minimal sufficient statistics underlies what is known as the *Sufficiency Principle* in Birnbaum's proof. Mayo (2014) identifies the Sufficiency Principle with what Cox and Hinkley (1974) call the *Weak Likelihood Principle*. This principle is generally accepted by many statisticians, hence the non-threatening moniker. Minimal sufficient statistics can be found for the (linear) exponential models, which include most of the models that one encounters in practice, e.g., Normal models (for continuous random variables), Binomial models and Poisson models (for discrete random variables); and the models they are derived from (as special cases) or which are derived from them. They are a small class of models but which, in practice, we can get by with in most applications. For example, the number of heads is a sufficient statistic for θ in the Binomial family of probability distributions in the coin tossing example. The sample mean $\overline{x}$ in the Normal family of probability distributions is a sufficient statistic for μ.

Let $T(\boldsymbol{x})$ be a sufficient statistic based on a sample $\boldsymbol{x}$, then the following statements are equivalent:

(1) $f(\boldsymbol{x} \,|\, T(\boldsymbol{x}), \theta)$ and does not depend on θ.

(2) $\mathcal{L}(\theta; \boldsymbol{x}) = f(\boldsymbol{x} \,|\, \theta)$ can be expressed as $c(\boldsymbol{x})g(T(\boldsymbol{x}), \theta)$.

(3) $\ell(\theta; T(\boldsymbol{x})) = \log \mathcal{L}(\theta; T(\boldsymbol{x}))$ is a minimal sufficient statistic.

(4) $T(\boldsymbol{x})$ and $\boldsymbol{x}$ have the same value of the observed Fisher Information.

The utility of sufficient statistics is underscored by statement (4), which explains Fisher's development of sufficient statistics, and his argument for using them, before the *factorization*

theorem method (statement (2) above) for finding sufficient statistics attributed to either Neyman or Halmos and Savage (1949). What Fisher realized is that with a minimal sufficient statistic you don't need (in particular) the whole sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ as an additional statistic to make an accurate estimate of θ.[21] The sufficient statistic is just as good as the whole sample, in terms of accurate estimation. The sense of "import" or "inference" or "computation" we shall see in subsequent parts of my paper have a precise mathematical meaning — *accurate point estimation*; or as Fisher would say, with efficiency or "without loss of information" (I will use this as a point to criticize the formal statements of the Likelihood Principle). In summary, no information or accuracy is lost in our estimate of θ if we use a sufficient statistic for θ in the process of dimensionally reducing the data.

Related to sufficiency and efficiency is the concept of *ancillarity*. Suppose that the unknown parameter Ω is partitioned into two parts $\Omega = (\theta, \lambda)$, where λ is a nuisance parameter. Let T be the minimal sufficient statistic for Ω and suppose that $T = (S, A)$ where:

(1) The marginal distribution of A depends on λ but is free of θ;

(2) The conditional distribution of S given $A = a$ depends on θ but not on λ for every a;

then A is *ancillary* for θ and S is said to be conditionally sufficient in the presence of nuisance parameter λ. For example, suppose $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. $T = (\overline{x}, s^2)$ is minimal sufficient for $\Omega = (\mu, \sigma^2)$, the marginal distribution of s^2 is free of μ; and for every value of s^2, the conditional distribution of $\overline{x}$ depends on μ alone. For another example, suppose that X is equally likely to be distributed as $N(\mu, \sigma_1^2)$ or $N(\mu, \sigma_2^2)$ where σ_1^2 and σ_2^2 are different but known. Let C be an indicator variable for the variance that takes the value 1 or 2 depending on whether X follows the first or second distribution, respectively. Then $T = (x, c)$ is minimal sufficient for (μ, σ_c^2). The marginal distribution of C, i.e., P(C = 1) = P(C = 2) = 0.5, is free of μ and for every c the conditional distribution of X given

[21] See especially Fisher (1925, 713).

$C = c$ depends on μ alone provided σ_1^2 and σ_2^2 are known. This last example illustrates that the role of ancillary statistics is that they encourage a conditional approach to parametric inference.[22] I return to this in section 3.4.2.

3.3 Why Restrict the Scope?

With these preliminaries, I can now take up the question that marks the title of this section. The reasons for restricting the scope will be different for coherent Bayesians and sampling theorists. So I discuss them in turn, starting with coherent Bayesians.

3.3.1 For a coherent Bayesian

You would think that a coherent Bayesian would have no reason to restrict the scope of the Likelihood Principle to problems of point estimation. After all, if $\mathcal{L}(\theta; \boldsymbol{x}) \sim \mathcal{L}'(\theta; \boldsymbol{y})$, then

$$\begin{aligned} p(\theta \mid \boldsymbol{x}) &= \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)} \\ &= \frac{c(\boldsymbol{x}, \boldsymbol{y})g(\boldsymbol{y} \mid \theta)\pi(\theta)}{c(\boldsymbol{x}, \boldsymbol{y}) \int g(\boldsymbol{y} \mid \theta)\pi(\theta)} \\ &= \frac{g(\boldsymbol{y} \mid \theta)\pi(\theta)}{\int g(\boldsymbol{y} \mid \theta)\pi(\theta)} \\ &= p(\theta \mid \boldsymbol{y}). \end{aligned}$$

The above identity shows that likelihood equivalence implies identical posterior distributions (provided the priors on the parameters match). For example, $f(\boldsymbol{x}; \boldsymbol{\theta})$ could be a sampling model for E_1 ("Toss the coin exactly 10 times.") and $g(\boldsymbol{y}; \boldsymbol{\theta})$ could be a sampling model for E_4 ("Continue tossing until the accumulated number of heads exceeds that of tails by

[22] For excellent discussions of conditional inferences in statistics and their relation to ancillary statistics see Reid (1995), Sundberg (2003) and Ghosh, Reid, and Fraser (2010).

exactly 2."). Since the experiments are likelihood equivalent, it would make no difference for a coherent Bayesian with the same priors on $\boldsymbol{\theta}$ in both experiments. The identity shows that the reason for this is that for a Bayesian *every* kind of parametric inference proceeds via the posterior distribution. There is therefore nothing to be gained from restricting the scope of the Likelihood Principle to problems of point estimation.[23] It is a virtue of the Bayesian framework that no such restriction is required. So, why am I arguing for a restriction?

The first reason has to do with instrumentally calibrating Bayes Factors.[24] Granted that Bayes Factors are *interpretable* as the relative predictive accuracy of our models; and that there are *guidelines* based a scale given by Jeffreys (1961) for how to *interpret* the magnitude of a Bayes Factor, many scientists and Federal Agencies do not find Bayes Factors useful for science because Bayes Factors are not instrumentally calibrated.[25] In order to instrumentally calibrate Bayes Factors one would need: (1) to address the question of what it means to say that a given Bayes Factor is "large" — this is a question of finding a shared objective *scale*; and (2) to quantify *how often* a given value of Bayes Factor is expected to occur with a given set of statistical models due to *sampling variability* — this is accounting for *error*.[26]

Instrumentally calibrating Bayes Factors by accounting for error is related to the problem raised by Mayo (2018, 30 – 32) and in the discussion following O'Hagan (1995)'s paper that Bayes Factors can be used to find evidence for a "wrong" model. According to Mayo, comparative measures of evidence, such as Likelihood Ratios and Bayes Factors, face (at least from a methodological perspective, if not in practice) the problem of *Gellerized Hypotheses*.[27] One can show that a Gellerized alternative hypothesis H_1 always exists for any H_0 (that is

[23]Robert (2007, 7) articulates a position like this — pointing out the artificiality of the kinds of distinctions I am making. See section 3.6 for my response to the artificiality objection.

[24]The varieties of calibration are discussed in Mwakima (2024a).

[25]See Kass and Raftery (1995) for the view that Jeffrey's scale is not a calibration of Bayes Factors.

[26]For the history behind the approaches to accounting for error due to sampling variability that motivates the techniques for calibration in Neyman-Pearson hypothesis testing based on sampling theory, see Mwakima (2024d).

[27]So named after Uri Geller the illusionist who claimed (among other things) that *his* explanation for bending iPhones using the power of his mind had the highest likelihood!

not itself a Gellerized hypothesis for the data $\boldsymbol{x}$ at hand).[28] In practice, the way to get around a Gellerized Hypothesis, which technically speaking is just a *saturated model* with no dimensional reduction, is to consider alternative measures of evidence such as the *Deviance Information Criterion* (DIC), the *Wanatabe-Akaike Information Criterion* (WAIC) or the *Bayesian Information Criterion* (BIC), which score overfitted models (of the saturated kind) poorly.[29] Such measures, have the additional benefit of having, at least asymptotically, a known sampling distribution (a χ^2 distribution with certain degrees of freedom) that can be used to provide a shared or objective scale — usually a *deviance scale* — and to account for error due to sampling variability. However, a coherent Bayesian is not allowed, as a matter of *principle*, to consider the sampling distribution of a statistic because of the wide scope given to the Likelihood Principle! This is why a coherent Bayesian *may* want to reconsider the scope of the Likelihood Principle.

Other reasons for a coherent Bayesian to restrict the scope of the Likelihood Principle are based on what *works* in practice not what looks good philosophically. Prominent Bayesian statisticians such as Andrew Gelman, recognizing the straightjacket the Likelihood Principle places him and others in, argued, in a blog post titled "It is not necessary that Bayesian methods conform to the Likelihood Principle" that there is more to Bayesian methods in statistics than inference.[30] According to Gelman, Bayesian methods include: (1) model specification, (2) conditional inference based on the models specified and the data (where the Likelihood Principle applies) and more importantly (3) model assessment. Gelman's target in his arguments are those people who confine Bayesian methods to (2), where the Likelihood Principle applies. For Gelman, model assessment is an important part of the practice of using Bayesian methods in statistics. But with the exception of robustness

[28]Witness a H_1 that has been *post*-designated or concocted to fit the data *perfectly* in such a way that likelihood ratio or Bayes Factor in its favor is maximal.

[29]The BIC is particularly attractive because it can be derived as an asymptotic approximation of a Bayes Factor. It also has the advantage of not depending on the prior and in some cases (with nested models, especially), it has "nice frequentist properties", i.e., it is instrumentally calibrated.

[30]See `https://rb.gy/44nvtt` (Retrieved on January 2024).

checks on priors, it is alleged that *all* the proposed model assessment techniques mentioned in the previous paragraph including his preferred *posterior predictive checks* "violate" the Likelihood Principle.[31] It is for this reason that he chose the title of this blog post to be what it is. Other Bayesian statisticians are sympathetic with Gelman. Here's how James Berger put it, in his influential book on statistical decision theory:

> A good Bayesian analysis may sometimes require a slight violation of the Likelihood Principle, in attempting to protect against the uncertainties in the specification of the prior distribution [...] analysis compatible with the Likelihood Principle is an ideal towards which we should strive, but an ideal which is not always completely attainable. (Berger, 1985, 33)

Recognizing that the Likelihood Principle is an ideal on paper but not in practice is one way of providing some flexibility to a Bayesian statistician. But it is *safer* and *better* to say that the Likelihood Principle doesn't apply in model assessment than it is to say that the Likelihood Principle is violated, for the simple reason that a principle is not violated where it doesn't apply. Also calling something "a violation" sounds sanctimonious. This is why I am arguing that even for a coherent Bayesian it makes sense to reconsider the scope of the Likelihood Principle — limiting the application of this principle to parametric inference involving point estimation. A coherent Bayesian has nothing to lose and something to gain, namely, the peace of mind of knowing they are not allowing even "slight violations".

[31] See Christensen, Johnson, Branscum, and Hanson (2011, 83) and compare with Kadane and Lazar (2004) and Vanpaemel (2010). Box (1980)'s *marginal p-values* to assess the prior distribution, Rubin (1984) and Gelman, Meng, and Stern (1996)'s *posterior predictive p-values* are mathematical expectations over a sample space. They violate the wide scope of the Likelihood Principle for this reason.

3.3.2 For a sampling theorist

There are at least three reasons why a sampling theorist would want to restrict the scope of the Likelihood Principle to problems of point estimation. A restriction:

(1) Avoids attributing to them an incoherent notion of evidence.

(2) Resolves a conflict with the Repeated Sampling Principle.

(3) Explains why a sampling theorist can accept the Sufficiency Principle and the Conditionality Principle, yet reject the wide scope reading of the Likelihood Principle (even though the Likelihood Principle is supposed to follow from the two former principles by Birnbaum (1962)'s proof).

First, it avoids attributing an incoherent notion of evidence to them. The p-value, which quantifies the attained level of significance, is the classical measure of evidence within a sampling theory framework. However, because its value depends on the sample space of the experiment, scenarios such as those captured by the following set of experiments are possible.

Scenario 1

1. **Negative Binomial Model**: Conduct an experiment with Bernoulli random variables with a protocol to stop after 16 favorable outcomes.

2. Suppose the 16th favorable outcome occurs on the 24th Bernoulli trial.

3. $P(n = 24) = \binom{24-1}{16-1}(0.5)^{16}(0.5)^{8}$

4. Attained significance level (p-value) is 0.077.

Scenario 2

1. **Binomial Model**: Conduct an experiment with Bernoulli random variables with 24 trials fixed in advance.

2. Suppose you obtain 16 favorable outcomes.

3. $P(s = 16) = \binom{24}{16}(0.5)^{16}(0.5)^{8}$

4. Attained significance level (p-value) is 0.032.

In each scenario, what is observed is a sequence of 16 favorable outcomes and 8 unfavorable outcomes. Although the experiments are likelihood equivalent, the inference that can be made in scenario 1 is different from the inference in scenario 2 because the p-values are different. In scenario 1, the p-value of 0.077 is evidence that is consistent with $\theta = 0.5$ at the $\alpha = 0.05$ level of significance. While in scenario 2, the p-value of 0.032 is evidence that is incompatible with $\theta = 0.5$ at the $\alpha = 0.05$ level of significance.

Faced with such outcomes, there are two things one might say. On the one hand, one might say that the p-value is simply an incoherent measure of evidence. It is incoherent in the sense that it supports two different beliefs about $\theta = 0.5$ even with the same set of observations: 16 favorable outcomes and 8 unfavorable outcomes. On the other hand, one might say that there is nothing incoherent about the p-value; instead, the Likelihood Principle is the culprit. For by declaring sample spaces irrelevant, the principle leaves out salient information that would impact our evaluation of evidence even in cases where the experiments are likelihood equivalent.[32] In other words, rather than saying that sampling theorists' measures of evidence are incoherent, a sampling theorist would want to say that the Likelihood Principle should not apply to quantifying evidence.

The second reason why a sampling theorist may wish to restrict the scope of the Likelihood Principle to problems of point estimation is that the wide scope reading conflicts with the Repeated Sampling Principle. Consider the following puzzles.

[32]Mayo, for example, would say something like this. Compare with Fletcher (2024).

	1	2	3
$f_1(x_1; \theta = 0)$	0.9	0.05	0.05
$f_1(x_1; \theta = 1)$	0.09	0.055	0.855
LR	10	0.909	17.1

Table 3.1: Probability mass distributions under Model 1

	1	2	3
$f_2(x_2; \theta = 0)$	0.26	0.73	0.01
$f_2(x_2; \theta = 1)$	0.026	0.803	0.171
LR	10	0.909	17.1

Table 3.2: Probability mass distributions under Model 2

Puzzle 1

This is a variation of an example in Berger and Wolpert (1988), which I find illuminating. Assume $\boldsymbol{\mathcal{X}} = \{1, 2, 3\}$ and $\Theta = \{0, 1\}$ and consider Model 1 and Model 2 which lead to the probability mass distributions in Table (3.1) and Table (3.2), respectively. Since for a fixed value of x and for all values of θ, the distributions in both models are proportional (the proportionality factor equal to the Likelihood Ratio (LR)), the Likelihood Principle implies that both models should lead to the same parametric inference about θ.

Now for Model 1, the uniformly most powerful test of size 0.1 is the test that rejects $f_1(x_1; \theta = 0)$ when $x_1 \in \{2, 3\}$. The power of this test is 0.91. This seems like a reasonably optimal procedure based on the Repeated Sampling Principle. But if Model 1 and Model 2 are evidentially equivalent (by the Likelihood Principle), then we should use critical values $x_2 \in \{2, 3\}$ to reject $f_2(x_2; \theta = 0)$ as well. However, this would imply using a test with an unreasonable size of 0.74. Therefore, since the same fixed observations would imply different optimal test criteria, Model 1 and Model 2 cannot be said to lead to the "same parametric inference" within the Neyman-Pearson hypothesis testing framework.

Puzzle 2

This is a variation of an example in Cox and Hinkley (1974, 38). Fix a Normal model $N(\mu, \sigma^2)$ where σ^2 is known. Construct an ancillary statistic A using σ^2 and the outcome of the toss of a fair coin as follows:

$$A = \begin{cases} \frac{\sigma^2}{n} & \text{if heads} \\ \frac{\sigma^2}{kn} & \text{for an integer } k > 1 \text{ if tails} \end{cases}$$

Here n is the sample size. Since $S = \sum_i^n x_i$ is sufficient for μ and A is ancillary for μ, the Likelihood Principle implies that the parametric inference regarding μ from the mixed experiment should be the same as the parametric inference from the component of it that was actually performed conditional on A. You might say, "Of course. Why should parametric inference about μ depend on what could have possibly been observed but wasn't? If I condition on the ancillary statistic, my inference should depend entirely on that." It can be shown, however, that since the *standard errors* for the different components of the mixed experiment differ, the conclusions arrived at using optimal sampling theory Neyman-Pearson hypothesis testing procedures and confidence interval estimators (based on the Repeated Sampling Principle) will be different for the unconditional mixed experiment compared to any of its components conditional on the value of A.

I believe these puzzles are not serious indictments of sampling theory approaches to statistical inference. For when one looks at these examples, the puzzles only arise when one tries to *apply* the Likelihood Principle to every kind of parametric inference. Writers like Berger and Wolpert (1988) argue that since there are cases where sampling theory gives inequivalent evidential conclusions where the Likelihood Principle prescribes an equivalence, the problem must be with sampling theory. But a sampling theorist can hold on to the Repeated Sampling

Principle by restricting the scope of the Likelihood Principle to problems of point estimation to avoid these sorts of puzzles.

The final reason for restricting the scope of the Likelihood Principle is that doing so can make sense of why many sampling theorists accept the Sufficiency Principle and the Conditionality Principle while rejecting the Likelihood Principle on the wide scope reading. In his comments on Lindley (2000, 330), Brad Efron, a sampling theorist writes:

> The Likelihood Principle seems to be one of those ideas that is rigorously verifiable and yet wrong.

The rigorous verification alluded to here is the original proof by Birnbaum (1962) (which was later improved upon in Birnbaum (1972)) that shows that the Likelihood Principle is a consequence of two principles (The Sufficiency Principle and the Conditionality Principle), which statisticians (both coherent Bayesians and sampling theorists) would consider plausible. There are those (like Mayo and Kruse (2001) and Mayo (2014)) who seek to resist Birnbaum's proof by exposing a fallacy somewhere within it. Although I do not like the formalism (see section 3.4.1 below) with which it is presented, I do not believe there is anything wrong with the proof itself.[33] Sampling theorists, on the account I am proposing, can say that they accept the Sufficiency and Conditionality Principles because these are principles that apply to dimension reduction in the context of point estimation. The Sufficiency Principle prescribes using a minimal sufficient statistic if it exists for accurate point estimation. The Conditionality Principle says conditioning on an ancillary statistic can improve the efficiency of your estimator (I return to say more about this in section 3.4.2). So a sampling theorist can give an account of why she accepts these principles, and consequently a restricted scope of the Likelihood Principle to problems of point estimation.

[33]Compare Gandenberger (2015).

3.4 Arguments for Restricting the Scope

Given that there are reasons why a coherent Bayesian and a sampling theorist may wish to restrict the scope of the Likelihood Principle, there are two independent arguments that I want to give for why a restriction should be adopted. The first argument is against formalism (in section 3.4.1). Using a series of examples, I argue that formal statements of the Likelihood Principle fail to disambiguate between *same inference*, *same information*, *same decisions* and *same evidence*. While the abstract formalism with which the Likelihood Principle is sometimes stated may have the advantage of generality, it: (1) suffers from a lack of precision; and (2) entails more than is warranted by the mathematical equivalence in Equation (3.3). In contrast, this equivalence and what is entailed by it is well-defined in the context of point estimation.

The second argument is based on Fisher's notion of efficiency (in section 3.4.2). Discussions of the Likelihood Principle (except perhaps Barnard, Jenkins, and Winsten (1962)) emphasize the concepts of sufficiency and ancillarity. But for Fisher, there is an inseparable connection between all three in the context of point estimation. I will argue that this kind of connection, which I conjecture doesn't arise in other kinds of parametric inference, offers compelling reasons for restricting the scope of the Likelihood Principle to problems of point estimation.

3.4.1 Part 1: Against Formalism

Here are some statements of the Likelihood Principle which one finds in both the statistical and philosophy of statistics literature. Edwards, Lindman, and Savage (1963, 237), who attribute the principle to Barnard and Feber (1947) and Fisher (1956), write:

> Two possible experimental outcomes D and D' — not necessarily of the same experiment — have the same import if D and D' have the same likelihood [function].

Edwards, Lindman, and Savage (1963) do not define what "same import" means here. Although what they say about "experimental outcomes" having the same likelihood strongly suggests that they are thinking about same parametric inference, it is unclear whether they take a wide or narrow view of what is included within parametric inference. What they say elsewhere in this paper indicates that "same import" can be read as "same (evidential) conclusions."

Now consider Lindley (1972):

> If x_1 and x_2 are two data sets with the same likelihood function apart from a multiplicative constant (that is $p(x_1|\theta) = kp(x_2|\theta)$ for all $\theta \in \Theta$, where k does not depend on θ), then inferences and decisions should be identical for x_1 and x_2.

Here what "an identical inference and decision" is has not been defined.

While stating that the Likelihood Principle "inevitably appears to be rather obvious, and certainly not worth getting exited about," de Finetti (1974b, Ch. 12, 210f.) says of this principle:

> It simply states that the information available from any set of observations is entirely contained in the corresponding likelihood function.

Although de Finetti downplays the significance of the Likelihood Principle, it is far from obvious (to me at least) what "the information available" and what "is entirely contained

in" means on his view.[34] Further, the Likelihood Principle has certainly raised a furor despite de Finetti's brusque admonition against undue excitement. Here's what the *Encyclopedia for Statistical Sciences* entry on the Likelihood Principle says:

> The Likelihood Principle asserts that for a given experiment E, the evidential meaning of any outcome x, for inference regarding θ is contained in the likelihood function determined by x. Hence all other features of the experiment, as, e.g., the sample space, are irrelevant.

This entry adds that the principle is incomplete since it does not say how the likelihood function "determines the evidential meaning." I concur with this and believe this to be an important sticking point within foundational debates in statistics. I would also go further than Joshi (who wrote this entry) and add that prominent statisticians and philosophers of statistics disagree on precisely how to decide what is relevant or irrelevant for statistical inference.

Following Birnbaum's notation, James Berger at one time (Berger, 1984b) stated the Likelihood Principle as:

> $\mathsf{Ev}(E, X)$ should depend on E and X only through the likelihood function. Two likelihood functions for (the same unknown) θ yield identical evidence about θ if they are proportional (as functions of θ).[35]

[34] To be sure, de Finetti does issue some warnings about the principle and takes it, rightly in my view, to lead to the discussion of sufficient statistics. He must be assuming that "the information available" is Fisher information.

[35] Compare with the statement in the later book Berger and Wolpert (1988, 19) where in the statement of the Likelihood Principle "all the information" replaces '$\mathsf{Ev}(E, X)$' and "same information" replaces "identical evidence". See also Berger and Wolpert (1988, 21.1) where "same conclusion about θ" is used instead. Berger and Wolpert (1988, 19) see the connection to Fisher's idea of a minimal sufficient statistic. But they insist, without argument, that the principle goes further than this.

Birnbaum introduced $\mathsf{Ev}(E, X)$ to talk at an abstract level and *very generally* about the "evidence" or "information" about θ that is obtained (or should be reported) upon observing a random variable X in an experiment E. $\mathsf{Ev}(E, X)$ is a *formal symbol* not a function.[36] $\mathsf{Ev}(E, X)$ is not a function because a function requires a specified domain and a *unique* range. Because different people will disagree about what the *range* of measures of evidence $\mathsf{Ev}(E, X)$ maps to, $\mathsf{Ev}(E, X)$ is not a function. Consider a sampling theorist in hypothesis testing, for example. The range of $\mathsf{Ev}(E, X)$ will be a set of p-values. For a coherent Bayesian the range will be a set of Bayes Factors. For Mayo, the range will be a set of values of the severity function.

Mayo (2014) also states the Likelihood Principle in a formal way. She connects Birnbaum's $\mathsf{Ev}(E, X)$ with her own $\mathsf{Infr}(\cdot)$, which she gives a different interpretation as follows:

> $\mathsf{Infr}_E(\boldsymbol{z})$: the parametric statistical inference from a given or known $(E, \boldsymbol{z})$[37]

Using this interpretation, Mayo defines the formal symbol "$\Rightarrow$" to mean:

> $(E, \boldsymbol{z}) \Rightarrow \mathsf{Infr}_E(\boldsymbol{z})$: an informative parametric inference about θ from given $(E, \boldsymbol{z})$ is to be computed by means of $\mathsf{Infr}_E(\boldsymbol{z})$.

and states what she calls the *Strong Likelihood Principle* (SLP).[38] Her statement of the SLP is similar to the statements of the Likelihood Principle by Edwards, Lindman, and Savage (1963) and Lindley (1972) except that what corresponds to "same import" and the undefined "identical inference and decision" in these earlier writers, is $\mathsf{Infr}_{E_1}(\boldsymbol{x_1}) = \mathsf{Infr}_{E_2}(\boldsymbol{x_2})$ for her. Mayo (2014, 229) does this with the intention of:

[36] Compare with Casella and Berger (2002, 292) who talk about "an evidence function" despite Birnbaum (1972)'s resistance to thinking of $\mathsf{Ev}(E, X)$ as a function.

[37] $\boldsymbol{z}$ replaces the X in Birnbaum's notation.

[38] The qualifier 'strong' allows Mayo to identify the *Weak Likelihood Principle* with what in Birnbaum's proof and Cox and Hinkley (1974) is called the *Sufficiency Principle* (the principle that states that the likelihood function is a minimal sufficient statistic).

> [R]eflecting the principles of evidence that arise in Birnbaum's argument, whether mathematical or based on intuitive, philosophical considerations about evidence.

To begin with, while this formal definition enjoys generality and wide scope it leaves "informative parametric inference" undefined (e.g., what does 'informative' mean?). Secondly, what "computation" is being done by means of by means of $\mathsf{Infr}_E(\boldsymbol{z})$? These are unanswered questions on Mayo's formal presentation. Finally, by failing to distinguish the different kinds of parametric inference, Mayo, we saw earlier, is puzzled by the following question: "Is the LP violated or simply inapplicable in secondary testing of model assumptions?"

It is this preoccupation with an abstract formalism that I suspect led an influential textbook in statistics, Casella and Berger (2002, 293 – 294), to distinguish what the authors call the *Formal Likelihood Principle* from the Likelihood Principle simpliciter. The statement of the Formal Likelihood Principle in this textbook is the same as Lindley's statement, except that (like James Berger's) it uses Birnbaum's formalism, which I have criticized for introducing a formal symbol $\mathsf{Ev}(E, X)$, whose range is ambiguous.

What is common to all the examples I have mentioned is that generality is attained at the cost of rigor. My argument, therefore, is that by restricting the scope of the Likelihood Principle to problems of point estimation, "same import", "same parametric inference", "identical decisions" can take on a precise mathematical meaning. It is *the value* of point estimates that is similar or identical whenever two experiments are likelihood equivalent. Sufficient statistics lead to the *same value* of Fisher information as the whole sample. Fisher information is well-defined in the context of point estimation, where it motivates the concepts of sufficient statistics and efficient estimators. This last point provides a nice segue into the second argument I want to give for restricting the scope of the Likelihood Principle to problems of point estimation — what Fisher had to say on efficiency.

3.4.2 Part 2: Fisher on Efficiency

With his definitional introduction of the formal symbol $\mathsf{Ev}(\cdot, \cdot)$ for the "evidential meaning" of an experiment, Birnbaum wished to give the Likelihood Principle a wide scope to include all kinds of parametric inference. In Birnbaum (1962, 286f), he makes a curiously ironical statement:

> The principal writers supporting the use of just the likelihood function for informative inference have not elaborated in very precise and systematic detail the nature of evidential interpretations of the likelihood function.

Moreover, Birnbaum (1962, 284) claimed that Fisher accepted something like the Likelihood Principle as self-evident (note the parentheses):

> Fisher and Barnard have been the principal authors supporting the Likelihood Principle on grounds independent of Bayes' principle. (The principle of maximum likelihood, which is directed to the problem of point-estimation, is not to be identified with the Likelihood Principle. Some connections between the distinct problems of point-estimation and informative inference are discussed below.) Self-evidence seems to be essential ground on which these writers support [The Likelihood Principle].

Here Birnbaum is exaggerating. Neither Fisher nor Barnard considered the Likelihood Principle as "self-evident." Here's how Barnard, Jenkins, and Winsten (1962, 324) put it (my emphasis):

> The term likelihood, as used here, and the recognition of the importance of the concept for statistical inference are due, of course, to Fisher. In his classical series

> of papers (1922, and especially 1925 and 1934) he showed how useful it was to consider the behaviour of the likelihood function in connection with problems of estimation. The *arguments* which we shall present for the Likelihood Principle owe a great deal to Fisher, as will be seen.

The fact that Fisher and Barnard gave arguments shows that they did not think this principle was self-evident.

In fact, a lot of what they *did say* recommends restricting scope of the Likelihood Principle in some way:

> [O]ur position is that two experiments giving the same likelihood function give the same inference, unless one or other has a special feature which justifies the application of a mode of reasoning not applicable to the other. A mere change of sample space does not by itself justify a change of mode. (Barnard, Jenkins, and Winsten, 1962, 334)

This passage is very important.[39] For it shows that the key question is this: what are the "special features" that would place a restriction on the scope of the Likelihood Principle? I will argue that for Fisher, these special features had to do with the connection between conditioning on ancillary statistics and efficiency. Although Fisher did not give a formal definition of ancillary statistics (like the one from section 3.2.2 above), by combining what I read in Fisher (1935):

> Ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it.

[39] See the rest of the discussion in the same page where this passage is from, especially where these authors declare that they are unconvinced of the general validity of Birnbaum's Conditionality Principle.

and what I read in Fisher (1934) that they are used "to recover the whole of the information available"; the connection emerges *but only within the context of problems of point estimation.* This connection, I will argue, is another reason for restricting the scope of the Likelihood Principle to these problems. Here's how Fisher describes his view on the issues in the prefatory note to Fisher (1925):

> When in 1921 the author put forward in the *Phil. Trans*, a paper [Fisher (1922)[40]] on mathematical statistics he was principally concerned, in respect of problems of estimation, with the practical importance of making estimates of high efficiency, i.e., of using statistics which embody a large proportion of the relevant information available in the data, and which ignore, or reject along with the irrelevant information, only a small proportion of that which is relevant. [...] Further work along the lines of the 1921 paper has, however, cleared up the main outstanding difficulties, and seems to make possible a theory of statistical estimation with some approach to logical completeness.

In section 3.2.2, using the Cramér-Rao lower bound on the variance of an estimator, an unbiased estimator $T(\boldsymbol{x})$ was said to be an efficient estimator of θ if and only if $\text{Var}(T(\boldsymbol{x})) = \frac{1}{nI(\theta)}$, where $nI(\theta)$ is the observed Fisher Information. Under regularity conditions, the *ratio* of $\frac{1}{nI(\theta)}$ to the actual $\text{Var}(T(\boldsymbol{x}))$ where $T(\boldsymbol{x})$ is unbiased is the *efficiency* of $T(\boldsymbol{x})$.[41]

With these ideas, what Fisher is saying in this passage is that he was mainly concerned with the practical problem of *finite* sampling without loss of precision (what I have been saying proceeds by dimension reduction and optimization). The efficiency of an unbiased estimator is a measure of how much we need to sample in order to attain a given level of accuracy or

[40]Let me say something about the dates. The paper, which I cite, was actually published in 1922. But Fisher read it on November 17th 1921.

[41]Compare with Fisher (1925, 714). Efficiency can also be considered as an asymptotic optimality criterion. For estimators $T(\boldsymbol{x})$ that are asymptotically normally distributed, the *Asymptotic Relative Efficiency* (ARE) can be used to compare them. See Casella and Berger (2002, 471 – 477).

precision when estimating parameters. The intuitive idea is that if one unbiased estimator $T(\boldsymbol{x})$ needs $n = 10$ samples to get to within 0.01 of θ and another unbiased estimator $T'(\boldsymbol{x})$ needs $n = 1000$ samples to get to within the same level of precision, then $T(\boldsymbol{x})$ is more efficient than $T'(\boldsymbol{x})$.[42]

What does all of this have to do with sufficiency, ancillarity, the Conditionality Principle and the Likelihood Principle (and why I am arguing that it should restricted to point estimation) based on a close reading of Fisher? The answer is given by Fisher (1956, 49 – 50):

> When the general hypothesis is found to be acceptable, and accepting it as true, we proceed to the next step of discussing the bearing of the observational record upon the problem of discriminating among the various possible values of the parameter, we are discussing the theory of estimation itself. In this theory a case of peculiar simplicity arises when an estimate exists which, perhaps in conjunction with ancillary statistics, subsumes the whole of the information, relevant to the parameter, supplied by the observational record [...] In fact for all purposes of inference an exhaustive [i.e., sufficient] statistic, in association perhaps with certain ancillary values, which themselves are independent of the parametric value, can replace the entire observational record from which it was calculated.

This passage is Fisher's version of "the" Likelihood Principle that influenced further developments by Barnard and Birnbaum.[43] But notice that from the quotation, Fisher restricts the scope of the Likelihood Principle to problems of parameter/point estimation, and says that sufficient statistics are to be used in conjunction with ancillary statistics for the sake of efficient estimation of parameters. This is why I am arguing that the right scope, we might say of *the original* Likelihood Principle, is to problems of point/parameter estimation.

[42] Compare Fisher (1925, 703ff).
[43] Compare Fisher (1956, 171).

Here's how I read Fisher on how this "conjunction" is supposed to work and how the Conditionality Principle arises in connection with efficiency and ancillary statistics. The key reference is Fisher (1934, 297 (bottom) – 303), where an illuminating example of the use of ancillary statistics to "recover information" is used for the first time.[44]

To understand Fisher's reasoning, we need the following ideas. Suppose that X is a continuous random variable, the *median* is the value m that satisfies:

$$\int_{-\infty}^{m} f(x)dx = \int_{m}^{\infty} f(x)dx = \frac{1}{2}.$$

Let $f(x)$ be any probability distribution function. Then the family of probability distribution functions $f(x-\theta)$, indexed by the parameter θ with $\Theta = \mathbb{R}$ is called a *location family* with standard probability distribution $f(x)$. θ is called the *location parameter*. For example, for a known σ, $N(\mu, \sigma^2)$ is a location family with μ, the location parameter. The Cauchy family of distributions given by:

$$\frac{1}{\sigma\pi}\frac{1}{\left(1+(\frac{x-\theta}{\sigma})^2\right)}$$

is a location family with location parameter θ with known σ. Similarly, the Laplace family of distributions given by:

$$\frac{1}{2\sigma}\exp\left(\frac{-|x-\theta|}{\sigma}\right)$$

is a location family provided σ is known. Let $f(x)$ be the probability distribution function for a random variable X and let c be a number such that, for all $\varepsilon > 0$, $f(c+\varepsilon) = f(c-\varepsilon)$. Then $f(x)$ is *symmetric* about the point c. The median turns out to be a useful order

[44] However, see also §§14 – 15 in Fisher (1925).

statistic for a symmetric location family of probability distributions. It can be shown that if $X \sim f(x)$, where $f(x)$ is symmetric, then the median of X is the number c. In fact, if $f(x)$ is any probability distribution function with $\mathcal{X} = \mathbb{R}$ that is symmetric about 0, then θ is the median of the location family $f(x - \theta)$.

In the 1934 paper, Fisher considers estimating the location parameter θ for a random variable X, which follows a Laplace distribution with $\sigma = 1$. It was known that the median T is the maximum likelihood estimate for θ and the Fisher information from one observation is $I(\theta) = 1$.[45] What is interesting about T in this case is that it is not a (minimal) sufficient statistic.[46] For an odd-sized sample $n = 2s+1$ the observed Fisher information $nI(\theta) = 2s+1$. Using the exact distribution of the sample median $X_{(s+1)}$ of $2s + 1$ observations from the Laplace distribution, Fisher found that the "lost information" from using the median as an estimator for θ is approximately

$$4\left(\sqrt{\frac{s}{\pi}} - 1\right)$$

which increases with s.[47] Fisher writes:

> Evidently, the simple and convenient method of relying on a single estimate will have to be abandoned. The loss of information has been traced to the fact that samples yielding the same estimate will have likelihood functions of different forms, and will therefore supply different amounts of information. When these functions are differentiable successive portions of the loss may be recovered by using as ancillary statistics, in addition to the maximum likelihood estimate, the second and higher differential coefficients at the maximum. (Fisher, 1934, 300)

[45] See Hogg, McKean, and Craig (2019, 365).

[46] In fact, for the Laplace distribution the set of order statistics is the only minimal sufficient statistic — so there is no dimensional reduction without loss of accuracy/precision.

[47] See also Fisher (1925, 716 – 717) for the same result.

The 1934 paper shows how this lost information can be recovered using ancillary statistics.[48] Let $X_{(i)}$ be the i-th order statistic. For a sample median $X_{(s+1)}$ of $2s+1$ observations, *the configuration of the sample* is given by the ancillary statistics $a_1, a_2, \ldots, a_s; a'_1, a'_2, \ldots, a'_s$ where, for $i = 1, \ldots, s$,

$$\begin{cases} a_i = X_{(s+1+i)} - T & \text{for } X_{(s+1+i)} > T \\ a'_i = T - X_{(s+1+i)} & \text{for } X_{(s+1+i)} < T \end{cases}$$

Fisher was able to show that, *conditional* on the configuration of the sample, T is sufficient. He concludes:

> The process of taking account of the distribution of our estimate in samples of the particular configuration observed has therefore recovered the whole of the information available. (Fisher, 1934, 303)

The key words from this quotation are "the particular configuration observed". For Fisher is saying here that *knowledge* of the experiment that was actually performed can sometimes be used as an ancillary statistic to improve the efficiency of our point estimators.

To illustrate the idea here, consider an experiment that leads to a sequence of measurements $X_1, X_2, \ldots X_n$ for a continuous parameter θ such that for $i = 1, \ldots, n$

$$X_i = \theta + e_i$$

where $e_i \overset{iid}{\sim} f(x)$ with support in $\mathbb{R}$. $f(x-\theta)$ is a location family of probability distributions for X. Suppose that $f(x-\theta)$ is *in fact* the Laplace distribution with $\sigma = 1$. If one uses

[48] An illuminating discussion of Fisher's reasoning in modern notation is given by Gorroochurn (2016). Compare with Fisher (1925, §10).

the median T as an estimator for θ, we know that T is the maximum likelihood estimate. T is asymptotically normally distributed with mean θ and variance $1/n$. However, using the mean $\overline{X}$ as an estimator for θ is less efficient than using the median. By the Central Limit Theorem $\overline{X}$ is asymptotically normally distributed with mean θ and variance $\frac{\sigma^2}{n}$, where $\sigma^2 = \mathbb{E}(e_i^2) = 2$ is taken with respect to the Laplace distribution. The Asymptotic Relative Efficiency of T and $\overline{X}$ is 2. The sample median T is twice as efficient as the sample mean $\overline{X}$ if $f(x - \theta)$ is Laplace.

Suppose instead that in fact $e_i \overset{iid}{\sim} f(x) = N(0, 1)$ for $i = 1, \ldots, n$. Then in this case using the sample median T as an estimator for θ is inefficient. It can be shown that T is asymptotically normally distributed with mean θ and variance $\frac{\pi}{2n}$.[49] $\overline{X}$ is asymptotically normally distributed with mean θ and variance $1/n$. Therefore, the Asymptotic Relative Efficiency of T and $\overline{X}$ is $\frac{2}{\pi} = 0.636$. This means that the median is less efficient than the sample mean for estimating θ assuming the $e_i \, (i = 1, \ldots, n)$ are standard normal random variables by about 36 percentage points.[50]

What these cases show is that knowledge of the experiment that was actually performed is an ancillary statistic that could improve the efficiency of point estimators. If we condition on the knowledge that the $e_i \, (i = 1, \ldots, n)$ determine a Laplace location model, then the sample median is an optimal point estimator for θ. However, conditioning on the knowledge that the $e_i \, (i = 1, \ldots, n)$ are random N(0, 1) variables, then the sample mean is an optimal estimator for θ. Taken together these cases also show how the Conditionality Principle arises in the context of point estimation where conditioning on ancillary statistics leads to conditionally sufficient statistics that are efficient.

[49] See Theorem 10.2.3 in Hogg, McKean, and Craig (2019).
[50] Compare Fisher (1925, 706).

3.5 Advantages of restricting

In the previous section, I gave two arguments for restricting the scope of the Likelihood Principle to problems of point estimation. I argued against the formalism with which the Likelihood Principle has been stated in favor of the precise mathematical content of what is entailed whenever there are likelihood equivalent experiments. The mathematical content is that likelihood equivalent experiments lead to the same value of point estimates. I also argued against extending the scope of the Likelihood Principle to other kinds of parametric inferences. Such an extension is not faithful to Fisher's original goal of using the Likelihood Principle as a dimension reduction principle in the context of finite sampling using estimators that are sufficient (or conditionally sufficient given an ancillary statistic) and efficient. Since the concepts of sufficiency, ancillarity and efficiency arise naturally in the context of point estimation, this is another reason for not extending the scope of the Likelihood Principle. In this section, I want to consider some of the advantages of the kind of restriction I am arguing for. In section 3.6 I will consider some objections.

3.5.1 Violation vs. Failure to apply

Mayo, we saw earlier, thought it was puzzling to distinguish what constitutes a violation of the Likelihood Principle from what is simply a case where the Likelihood Principle does not apply. By restricting the scope of the Likelihood Principle to problems of point estimation in parametric inference it is possible to say what constitutes a violation of the Likelihood Principle and what is simply a case where the Likelihood Principle does not apply. The Likelihood Principle does not apply outside the context of dimension reduction for efficient point estimation. But it is a violation of the Likelihood Principle if two likelihood equivalent experiments lead to different point estimates. The use of *reference priors* in Bayesian inference, for example, is a violation of the Likelihood Principle. Usually, a reference prior

$\pi(\theta)$ for parametric models is one such that:

$$\pi(\theta) \propto I(\theta)^{1/2}.$$

Consider the sequence of Bernoulli trials from the Introduction. If E_1 is "Toss the coin exactly 10 times", then we'd use a Binomial model. However, with E_2, which is "Continue tossing until 6 heads appear", a Negative Binomial model would be appropriate. These experiments are likelihood equivalent. But with reference priors they lead to different values of point estimates because the posterior distributions are different. On the one hand, for the Binomial model where n is fixed in advance and $Y = \sum_{i=1}^{n} X_i$ is the number of favorable outcomes, it can be shown that

$$\pi_1(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which leads to $\pi_1(\theta \mid \boldsymbol{X}) \propto \theta^{y-1/2}(1-\theta)^{(n-y)-1/2}$. So for E_1 with reference priors, the posterior distribution would be proportional to $\theta^{11/2}(1-\theta)^{7/2}$. On the other hand, for the Negative Binomial model, where the random variable is n — the number of trials until $Y \geq 1$ favorable outcomes occur; it can be shown that

$$\pi_2(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$$

which leads to $\pi_2(\theta \mid \boldsymbol{X}) \propto \theta^{y-1}(1-\theta)^{(n-y)-1/2}$. So modeling E_2 with reference priors, the posterior distribution would be proportional to $\theta^{5}(1-\theta)^{7/2}$. Since the posterior distributions are different, the point estimates will be different. This is a violation of the Likelihood Principle. Proponents of reference priors, like Bernardo, account for this violation as follows:

> It is important to stress that reference distributions are, by definition, function[s] of the entire probability model [...] not only of the observed likelihood. Technically, this is a consequence of the fact that the amount of information which an experiment may be expected to provide is the value of an integral over the entire sample space X, which, therefore has to be specified. (Bernardo and Smith, 2000, 315)

Reference prior analysis, we might say, conditions on the knowledge of the experiment — using it as an ancillary statistic for point estimation. Whether this kind of "slight violation", to use James Berger's words, is tolerable is a separate issue from what I am concerned with in this paper.

Related to this violation of the Likelihood Principle by reference prior analysis is an analogous case within a sampling theory framework involving sampling from a Binomial model and from a Negative Binomial model. An unbiased estimator $T(\boldsymbol{x})$ is called a *minimum variance unbiased estimator* (MVUE) of the parameter θ if for every other unbiased estimator $S(\boldsymbol{x})$ of θ, $\text{Var}(T(\boldsymbol{x})) \leq \text{Var}(S(\boldsymbol{x}))$. Suppose X_i for $i = 1, \ldots n$ are Bernoulli random variables. Let Y denote the number of favorable outcomes in a sample of fixed size n, the estimator $T(\boldsymbol{x}) = \frac{Y}{n}$ for θ in the Binomial model is unbiased. $\hat{\theta} = \frac{Y}{n}$ is also the maximum likelihood estimate for θ. Moreover, since $T(\boldsymbol{x}) = \frac{Y}{n}$ is sufficient, $T(\boldsymbol{x})$ is the unique MVUE of θ by the Lehmann-Scheffé theorem. If instead we choose to model a sequence of Bernoulli random variables with the stopping rule to continue sampling until Y favorable outcomes are observed with W unfavorable outcomes and $n = Y + W$, then the random variable W is Negative Binomial. Since the experiments are likelihood equivalent, $\hat{\theta} = \frac{Y}{n}$ is the maximum likelihood estimate for θ. However, it can be shown that $T(\boldsymbol{x}) = \frac{Y}{n}$ in the Negative Binomial model is *not* unbiased.[51] The unbiased estimator is, in fact, $S(\boldsymbol{x}) = \frac{Y-1}{n-1}$ and it is the MVUE by the Lehmann-Scheffé theorem. It is alleged that what this example shows is a case where

[51] See Kendall, Stuart, and Ord (1987, 316) Example 9.15

the experiments are likelihood equivalent yet the point estimates are different. Is this a case where the Likelihood Principle is violated even for a sampling theorist?

I claim that this is not a violation of the Likelihood Principle. There are two ways I wish to defend this claim. First, there is a difference between an estimator, such as $T(\boldsymbol{x})$ and $S(\boldsymbol{x})$ above, and an estimate, such as $\hat{\theta}$. The Likelihood Principle restricted to point estimation is violated if two likelihood equivalent experiments lead to different point *estimates*. In the Binomial and Negative Binomial the maximum likelihood estimates $\hat{\theta}$ are identical. But unbiasedness and efficiency are properties of estimators. The Likelihood Principle does not apply to *evaluations* of point estimators. The second point to make here is that what this example shows is only that $S(\boldsymbol{x}) = \frac{Y-1}{n-1}$ is more efficient than $T(\boldsymbol{x}) = \frac{Y}{n}$ when sampling from a Negative Binomial model. So conditioning on the knowledge of the stopping rule (or equivalently the experiment that was performed), even in this case, can lead to more efficient estimators.

3.5.2 No Salesmanship

Coherent Bayesians typically charge sampling theorists of being incoherent by violating the Likelihood Principle, while Sampling theorists like Mayo raise the probativist criticism against coherent Bayesians. The Likelihood Principle is a selling point for some Bayesians.[52] But sampling theorists like Mayo are not buying it. Here I believe my analysis indicates that one must be careful not to oversell or undersell. By using formal symbols that give the Likelihood Principle a wide interpretation, some of the proponents of the Likelihood Principle oversell what methods which satisfy the Likelihood Principle can accomplish. We saw in section 3.3.1 that Bayesians like Gelman find the Likelihood Principle unduly limiting — there is more to Bayesian methods in statistics than conditional inference through

[52]Berger has a paper with the literal title "Bayesian Salesmanship". See Berger (1984a) and also the discussion in Berger (1983).

the posterior distribution, where the Likelihood Principle should apply. One way of reading Gelman's "It is not necessary that Bayesian methods conform to the Likelihood Principle" is as a claim about what Bayesians can actually sell about their methods.

At the same time, a sampling theorist can know exactly what they are buying when the Likelihood Principle is recommended to them. Drawing on Fisher's work on efficient parameter estimation, I have shown that conditional inferences and ancillaries (such as the knowledge of the actual experiment) can serve an important function for sampling theorists. This had been obscured when the scope of the Likelihood Principle was taken through a *formalism* to be so wide that it led some sampling theorists to demur. Fisher's work indicates that the only *contentful* sense of "same inference" from likelihood equivalent experiments is similar values of point estimates. The puzzles and the problems I discussed in section 3.3.2 only arise outside the context of point estimation. Therefore, there is no need for partisan salesmanship and red herrings. We can adopt the chutzpah attitude advocated by Mayo (2018, 12) and begin finally telling the truth about statistical inference.

3.6 Objections and Replies

It may be objected that the distinction between problems of estimation and other kinds of problems that I have made for the sake of my argument is artificial or ad hoc.[53] My reply to the artificiality objection is that the distinction is real. In section 3.2 the descriptions of the kinds of problems in parametric inference indicate that the problems have different *goals*, *optimality criteria* and choice of comparative *measures of support*. Moreover, far from making an ad hoc distinction, I have given reasons in section 3.3.1 and 3.3.2 for why a coherent Bayesian and a sampling theorist may find a restriction of the scope of the Likelihood Principle desirable. In section 3.5, I have discussed the upshot for foundational

[53]See Robert (2007, 7), for example.

debates in the philosophy of statistics. So the distinctions I am making that lead me to restrict the scope of the Likelihood Principle are not just intended to avoid the puzzles or problems (which would make them ad hoc). Rather the distinctions, and the use I make of them, can address *other* controversial topics in the literature in a non-partisan way.

Still one may press the point that there are statistical tests derived from Likelihood Ratio Test statistics. For example, within the Neyman-Pearson Hypothesis testing framework some of the common tests are based on Likelihood Ratio statistics comparing the restricted null model H_0 (e.g., $\theta = \theta_0$) to an unrestricted alternative model H_1 (e.g., $\theta \in \Theta - \{\theta_0\}$) on the same parametric space Θ. The null model and the alternative model are, therefore, likelihood equivalent. Why shouldn't the Likelihood Principle apply to the inferences here? Interestingly, Fisher (1934, 296) thought that the small sample parametric inference he had developed based on sufficient statistics subsumed the Neyman-Pearson approach based on Likelihood Ratio Test statistics.

> [I]t is surprising that Neyman and Pearson should lay it down as a preliminary consideration that "the testing of statistical hypotheses cannot be treated as a problem in estimation." When tests are considered only in relation to sets of hypotheses specified by one or more variable parameters, the efficacy of the tests can be treated directly as the problem of estimation of these parameters.

Neyman and Pearson (1936) responded to Fisher's claim by strongly denying it. They showed a case where a uniformly most powerful test exists but no minimal sufficient statistic exists; and conversely a case where a minimal sufficient statistic exists but no uniformly most powerful test exists. Therefore, since sufficient statistics and uniformly most powerful tests can come apart, these cases were supposed to support the preliminary consideration, which had been quoted by Fisher. It would be fruitful work for future research to evaluate

Neyman and Pearson's reasons for rejecting the Fisherian connection. This work would also look more closely at "the efficacy of tests" mentioned by Fisher in the above passage.[54]

Here I will only say that underlying their examples is a distinction between *a test statistic* on the one hand and *a statistical test* on the other. Fisher's pure significance testing is not a statistical test in Neyman and Pearson's sense. For one thing, a pure significance test for a hypothesis does not consider *alternatives*. But for Neyman and Pearson, a statistical test is an optimal critical region based on the power function, which *goes beyond* Likelihood Ratios by computing mathematical expectations on the sample space of the alternative — this is the power of the test.[55] In short, while it is true that a test statistic within the Neyman-Pearson school of hypothesis testing can be derived from a Likelihood Ratio; a statistical test, which is *a choice of a critical region*, is based on other optimality criteria based on the value that certain special risk functions take — these are the size and power of the test. The Likelihood Principle does not apply when one goes beyond likelihood equivalence for point estimation to evaluations of decision procedures.

However, there is the curious case of *AN*alysis *O*f *VA*riance (ANOVA). In one-way ANOVA, there is one nominal variable X (e.g., an indicator variable that picks the species of a given genus; or an indicator variable that identifies the treatment group in a controlled experiment) and one response variable Y (e.g., length of wing-span, measured outcome depending on treatment). The goal of one-way ANOVA is to determine whether the levels determined by the nominal variable X are *necessary* in order to account for the total variation observed in Y. If β_0 is the grand mean of the individual measurements regardless of their group; Fisher's idea for doing ANOVA was to partition the total variation (Total Sum of Squares (SS_T))

[54]In advanced mathematical statistics, there is a concept of an *efficient test*. I do not believe that this undermines my argument for restricting the scope of the Likelihood Principle based on the idea that efficiency is a concept that was originally intended for problems of point estimation. Whether a test is efficient is something that is determined by measures (such as power) that go beyond the likelihood function. See Hogg, McKean, and Craig (2019, §10.2.1).

[55]See Mwakima (2024d) for more discussion on this.

into the variation within the groups (SS_W) and the variation between the groups (SS_B).

$$SS_T = SS_W + SS_B$$

If the groups are not significantly different from each other, then $SS_B \ll SS_W$. In this case, the within group variation accounts for the total variation and there is no need to introduce another β_1 for the difference from the grand mean of the group picked out by $X = 1$ in order to understand the variation in Y. The determination of what is a "significant difference" β_1 is based on the F-statistic:

$$F = \frac{MS_B}{MS_W}$$

where MS_B is the mean variation between groups and MS_W is the mean variation within groups. Large values of the F-statistic based on the F-distribution with certain degrees of freedom imply that there is a statistically significant difference between the groups, i.e., $\beta_1 \neq 0$. This is one way of thinking about the problem of ANOVA. It is a model criticism problem where one is testing: $H_0 : \beta_1 = 0$ (no difference between the groups) against the alternative: $H_1 : \beta_1 \neq 0$ (there is a difference between the groups) using an F-statistic.

But there is another to think of one-way ANOVA. This is as a simple linear regression model for Y based on X.

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$$

Here the goal is to estimate β_1 by finding the ordinary least square estimates for β_0 and β_1. It can be shown that the simple linear regression model and one-way ANOVA are equivalent. So isn't ANOVA a case where the problem of parameter estimation is equivalent to a problem

of model criticism? If so, how tenable is the distinction I am making between problems of parameter estimation and problems of model criticism? ANOVA seems like a genuine counter-example.

Here I still appeal to the distinction between a test statistic and a test. While the F-test statistic is a function of sufficient statistics, the test still has to be selected according to some optimality criteria, which go beyond what is contained in the minimal sufficient statistics for this kind of parametric inference. For instance, one would have to consider the sampling distribution of the F-statistic and to pragmatically select a powerful test subject to the size constraint $\alpha = 0.05$ or 0.01, for example.

Another possible objection here is this. Doesn't interval estimation degenerate into point estimation in the limit as $n \to \infty$. Let $\hat{\theta}$ be an estimate for θ and suppose that $T(\boldsymbol{x})$ is an efficient estimator for θ. Then:

$$\hat{\theta} \pm \frac{1}{\sqrt{nI(\hat{\theta})}}$$

is an interval estimator for θ, which would get closer and closer to a point estimate for θ as $n \to \infty$. My response here is to grant the point that *asymptotically*, point estimation and interval estimation coincide in a case like this. However, in the context of small sample parametric inference (with finite sampling), the point estimate and the interval estimate are different. Moreover, when statisticians talk about interval estimation, they usually have in mind $(1 - \alpha)\%$ *confidence* intervals for θ. The α here is an optimality parameter based on the Repeated Sampling Principle for calibrating the interval estimator in the sense of guaranteeing *coverage*.

According to the *Encyclopedia for Statistical Sciences* entry on the Likelihood Principle by Joshi, which I quote here for convenience:

> The Likelihood Principle asserts that for a given experiment E, the evidential meaning of any outcome x, for inference regarding θ is contained in the likelihood function determined by x. Hence all other features of the experiment, as, e.g., the sample space, are irrelevant.

This suggests that the Likelihood Principle is primarily about the evidence and what evidential conclusions can be drawn from likelihood equivalent experiments. Let us call this the *Evidential Likelihood Principle*. Birnbaum, in fact, was concerned with a Likelihood Principle that was intended to state the necessary and sufficient conditions for evidential equivalence. Prima facie, point estimation has nothing to do with evidence. So by restricting the scope to problems of point estimation, what does my account say about the Evidential Likelihood Principle? In what sense does point estimation allow us to make evidential conclusions?

Now for a Bayesian or a Likelihoodist with comparative measures of evidence, inferences based on point estimation will determine the evidential conclusions. The Bayes Factor, which is the ratio of the marginal likelihood functions comparing two models, can be thought of as a point estimator which averages the point estimates under the competing models then returns the quotient.[56] This averaging is possible because of the priors on the parameters. So, for a Bayesian, there is nothing lost by restricting the scope of the Likelihood Principle to point estimation. The conclusions for a coherent Bayesian from likelihood equivalent experiments will be the same conclusions from both the Evidential Likelihood Principle and the Likelihood Principle *simpliciter*. For a Likelihoodist (e.g., Royall (1997) and Forster and Sober (2004)), the Likelihood Ratio will compare simple hypotheses based on the maximum likelihood estimates under each hypothesis, say. So any evidential conclusions will coincide with the outcome of the point estimation process. If the likelihood of the maximum likelihood estimate under one hypothesis is higher than the likelihood of the maximum likelihood

[56] See Mwakima (2024b, §4).

estimate under the alternative, then this could constitute evidence for the hypothesis with a higher relative likelihood.[57]

For the sampling theorist, the restriction of the Likelihood Principle to problems of point estimation, as discussed in section 3.3.2, was meant to avoid attributing to them an incoherent notion of evidence. As mentioned following the entry by Joshi, the Evidential Likelihood Principle is "incomplete" because the principle does not say what determines evidential meaning. I would say that it is unilluminating, especially because the formal symbol $\mathsf{Ev}(\cdot, \cdot)$ has no unambiguous intended interpretation for *every practicing statistician*. In fact, in a posthumous paper (Birnbaum (1977)), Birnbaum analyzes what constitutes an adequate representation of statistical evidence in his criticism of the Lindley-Savage argument for Bayesian methods of statistics. He introduced what he called the *confidence concept* of statistical evidence. This means that even for the originator of the Evidential Likelihood Principle, the adoption and application of this principle will depend on what one believes statistical evidence is. Since there is no agreement, my goal in this paper was to seek common ground on the scope of the Likelihood Principle and to draw out the consequences of the restriction I am advocating for. Seeking common ground may involve giving up on the Evidential Likelihood Principle.

3.7 Conclusion

In conclusion, let me recap the main points of my paper. There has been a long-standing debate within the foundation of statistics regarding the scope of the Likelihood Principle. One issue within this debate is this: what constitutes a violation of the Likelihood Principle as distinguished from a failure to apply it? I have argued that restricting the scope of the Likelihood Principle to problems of point estimation is a way to resolve this issue. My

[57] Compare Berger (1985, 25).

overall argument had two independent parts. In the first part I argued against the formalism with which the Likelihood Principle has been stated in the literature, which leaves "same inference", "same import" and "same decision" up to anyone's interpretation. In the second part I argued that the naturalness with which the connection between sufficiency, ancillarity and efficiency emerges in the context of problems of point estimation in Fisher's work, is a compelling reason to restrict the scope of the Likelihood Principle to problems of point estimation. Such a restriction ought to be welcome to both coherent Bayesians and sampling theorists. Coherent Bayesians have nothing to lose, but something to gain. It is clear from my analysis what constitutes a violation. Further, my analysis rationalizes some of the existing techniques or measures currently in use for model assessment within a Bayesian framework. These techniques and measures are being used in a context where the Likelihood Principle does not apply. Lastly, sampling theorists have a way of shielding their measures of evidence from charges of incoherency. The way consists of saying that the Likelihood Principle does not apply to contexts outside of dimension reduction for point estimation. The purported conflicts with the Repeated Sampling Principle can, consequently, be defused.

Bibliography

Achinstein, Peter (2001). *The Book of Evidence*. Oxford University Press.

Achinstein, Peter (2002). "Is There a Valid Experimental Argument for Scientific Realism?" *The Journal of Philosophy*, 99, no. 9:470–495.

Arfken, George B. and Weber, Hans J. (2001). *Mathematical Methods for Physicists (5th Edition)*. Harcourt Academic Press.

Baldi, Paolo (2017). *Stochastic Calculus*. Springer.

Bandyopadhya, Prasanta S. and Forster, Malcom R. (Eds.) (2011). *Handbook of Philosophy of Science Vol. 7 Philosophy of Statistics*. Elsevier.

Barnard, George A (1949). "Statistical Inference". *Journal of the Royal Statistical Society. Series B (Methodological)*, 11, no. 2:115–149.

Barnard, George A. and Feber, Robert (1947). "Review of Sequential Analysis by Abraham Wald". *Journal of American Statistical Association*, 42, no. 240:658–665.

Barnard, George A, Jenkins, G. M., and Winsten, C. B. (1962). "Likelihood Inference and Time series". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 125, no. 3:321–352.

Barnard, George Alfred and Cox, David Roxbee (1962). *The Foundations of Statistical Inference: A Discussion*. Methuen.

Barnett, Vic (1999). *Comparative Statistical Inference*. John Wiley & Sons, Ltd. Third Edition.

Basu, D. (1964). "Recovery of Ancillary Information". *Sankhyā: The Indian Journal of Statistics, Series A*, pages 3–16.

Basu, D. (1975). "Statistical Information and Likelihood [with discussion]". *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–71.

Belot, Gordon (2013a). "Bayesian Orgulity". *Philosophy of Science*, 80, no. 4:483–503.

Belot, Gordon (2013b). "Failure of Calibration is Typical". *Statistics & Probability Letters*, 83, no. 10:2316–2318.

Berger, James (1984a). "Bayesian Salesmanship". In *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno de Finetti* (edited by P. K. Goel and A. Zellner), pages 473–488. North-Holland Amsterdam.

Berger, James (1984b). *The Frequentist Viewpoint and Conditioning*. Purdue University. Technical Report 83-48. Later version appeared as: Berger, J. (1985) in the Proceedings of the Berkeley Conference in honor of Jack Kiefer and Jerzy Neyman' (L. Le Cam and R. Olshen Eds.), Wadsworth, Belmont.

Berger, James O. (1983). "In Defense of the Likelihood Principle: Axiomatics and Coherency (with Discussion)". In *Bayesian Statistics 2* (edited by J.M. Bernardo, D.V. DeGroot, and A.F.M. Smith), pages 33–65. Elsevier Science. Published in the Proceedings of the Second Valencia International Meeting in 1985. Available online here `https://www.stat.purdue.edu/docs/research/tech-reports/1983/tr83-18.pdf` Retrieved 06/11/2024.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition.

Berger, J.O. and Wolpert, R.L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics. Lecture notes : Monographs Series. Institute of Mathematical Statistics.

Bernardo, José M. and Smith, Adrian F. M. (2000). *Bayesian Theory*. John Wiley & Sons.

Billingsley, Patrick (2012). *Probability and Measure: Anniversary Edition*. Wiley.

Birnbaum, Allan (1962). "On the Foundations of Statistical Inference". *Journal of the American Statistical Association*, 57, no. 298:269–306.

Birnbaum, Allan (1972). "More on Concepts of Statistical Evidence". *Journal of the American Statistical Association*, 67, no. 340:858–861.

Birnbaum, Allan (1977). "The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; with a Criticism of the Lindley-Savage Argument for Bayesian Theory". *Synthese*, pages 19–49.

Bjørnstad, Jan F. (1996). "On the Generalization of the Likelihood Function and the Likelihood Principle". *Journal of the American Statistical Association*, 91, no. 434:791–806.

Bogachev, Vladimir I (1998). "Measures on Topological Spaces". *Journal of Mathematical Sciences*, 91, no. 4:3033–3156.

Bogen, James and Woodward, James (1988). "Saving the Phenomena". *The Philosophical Review*, 97, no. 3:303–352.

Box, George E.P. (1979). "Robustness in the Strategy of Scientific Model Building". In *Robustness in Statistics*, pages 201–236. Elsevier.

Box, George E.P. (1980). "Sampling and Bayes Inference in Scientific Modeling and Robustness". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 143, no. 4:383–404.

Box, George E.P. (1982). "An Apology for Ecumenism in Statistics". *University of Wisconsin-Madison Mathematics Research Center*. Technical Summary Report July 1982.

Box, George E.P. and Jenkins, Gwilym M (1976). *Time Series Analysis. Forecasting and Control*. Holden-Day Series in Time Series Analysis.

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability". *Monthly Weather Review*, 78, no. 1:1–3.

Brown, Jessica (2015). "Evidence and Epistemic Evaluation". In *Oxford Studies in Epistemology Vol. 5* (edited by Tamar Szabó Gendler and John Hawthorne). Oxford University Press.

Burnham, Kenneth P. and Anderson, David R. (2002). *Model selection and Multimodel Inference: A practical Information-theoretic Approach*. Springer New York.

Carnap, Rudolf (1962). *Logical Foundations of Probability*. The University of Chicago Press.

Casella, George and Berger, Roger L. (2002). *Statistical Inference (2nd Edition)*. Brooks/Cole Cengage Learning.

Christensen, Ronald, Johnson, Wesley, Branscum, Adam, and Hanson, Timothy E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press.

Coffey, William, Kalmykov, Yu P., and Waldron, J. T. (2004). *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry, and Electrical Engineering*, volume 14. World Scientific.

Cox, David Roxbee (2006). *Principles of Statistical Inference*. Cambridge University Press.

Cox, David Roxbee and Hinkley, David Victor (1974). *Theoretical Statistics*. Chapman and Hall Press.

Dawid, A Philip (1982). "The Well-Calibrated Bayesian". *Journal of the American Statistical Association*, 77, no. 379:605–610.

de Finetti, Bruno (1937). "Foresight: Its Logical Laws, Its Subjective Sources". In *Studies in Subjective Probability* (edited by H.E. Kyburg and H.E. Smokler). Krieger Publishing. The article first appeared in the *Annales de l'Institut Henri Poincaré, Vol. 7 (1937)*.

de Finetti, Bruno (1962). "Does it Make Sense to Speak of 'Good Probability Appraisers'?" In *The Scientist Speculates: An Anthology of Partly-Baked Ideas* (edited by Irving John Good). Heinemann.

de Finetti, Bruno (1965). "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item". *British Journal of Mathematical and Statistical Psychology*, 18, no. 1:87–123.

de Finetti, Bruno (1969). "Initial Probabilities: A Prerequisite for Any Valid Induction". *Synthese*, 20:2–16.

de Finetti, Bruno (1970). "Logical Foundations and Measurement of Subjective Probability". *Acta Psychologica*, 34:129–145.

de Finetti, Bruno (1972). *Probability, Induction and Statistics: The Art of Guessing*. John Wiley & Sons Ltd.

de Finetti, Bruno (1974a). "Bayesianism: Its Unifying Role for Both the Foundations and Applications of Statistics". *International Statistical Review/Revue Internationale de Statistique*, pages 117–130.

de Finetti, Bruno (1974b). *Theory of Probability: A Critical Introductory Treatment (2 Vols.)*. John Wiley & Sons Ltd. 1990 Edition of the English translation by Antonio Machi and Adrian Smith of the 1970 original.

de Finetti, Bruno (1981). "The Role of 'Dutch Books' and of 'Proper Scoring Rules?" *The British Journal for the Philosophy of Science*, 32, no. 1:55–56.

de Finetti, Bruno (2008). *Philosophical Lectures on Probability*. Springer.

de Finetti, Bruno and Savage, L. J. (1962). "Sul modo di scegliere le probabilitá". *Biblioteca del Metron*, 1:81–147.

de Heide, Rianne and Grünwald, Peter D (2021). "Why Optional Stopping Can Be A Problem For Bayesians". *Psychonomic Bulletin & Review*, 28:795–812.

DeGroot, Morris H. (1970). *Optimal Statistical Decisions*. John Wiley & Sons. 2004 Wiley Classics Library Edition.

Diaconis, P. and Freedman, D. (1983). "Frequency Properties of Bayes Rule". In *Scientific Inference, Data Analysis, and Robustness*, pages 105–115. Academic Press.

Diaconis, Persi and Freedman, David (1980). "Finite Exchangeable Sequences". *The Annals of Probability*, pages 745–764.

Diaconis, Persi and Freedman, David (1986). "On the Consistency of Bayes Estimates". *The Annals of Statistics*, pages 1–26.

Diaconis, Persi and Skyrms, Brian (2018). *Ten Great Ideas About Chance*. Princeton University Press.

Doob, Joseph L. (1949). "Application of the Theory of Martingales". *Le Calcul des Probabilites et ses Applications*, pages 23–27.

Draper, David (1995). "Assessment and Propagation of Model Uncertainty". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57, no. 1:45–70.

Draper, David (2006). "Coherence and Calibration: Comments on subjectivity and "objectivity" in Bayesian Analysis (Comment on Articles by Berger and by Goldstein)". *Bayesian Analysis*, 1, no. 3:423–428.

Durbin, James (1970). "On Birnbaum's Theorem on the Relation between Sufficiency, Conditionality and Likelihood". *Journal of the American Statistical Association*, 65, no. 329:395–398.

Earman, John (1992). *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Edwards, A. W. F. (1992). *Likelihood: Expanded Edition*. The John Hopkins University Press.

Edwards, Ward, Lindman, Harold, and Savage, Leonard J (1963). "Bayesian Statistical Inference for Psychological Research." *Psychological Review*, 70, no. 3:193.

Efron, Bradley (1986). "Why Isn't Everyone a Bayesian?" *The American Statistician*, 40, no. 1:1–5.

Efron, Bradley and Hastie, Trevor (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Einstein, Albert (1956). *Investigations on the Theory of the Brownian Movement (Translated by A.D. Cowper)*. Dover.

Ferguson, Thomas S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*, volume 1. Academic Press.

Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics". *Philosophical Transactions of the Royal Society of London. Series A*, 222, no. 594-604:309–368. Read on November 17th, 1921.

Fisher, R. A. (1925). "Theory of Statistical Estimation". In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725.

Fisher, R. A. (1934). "Two New Properties of Mathematical Likelihood". *Proceedings of the Royal Society of London. Series A*, 144, no. 852:285–307.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh. 2nd Edition of 1959.

Fitelson, Branden (2007). "Likelihoodism, Bayesianism, and Relational Confirmation". *Synthese*, 156, no. 3:473–489.

Fletcher, Samuel C. (2019). "Stopping Rules as Experimental Design". *European Journal for Philosophy of Science*, 9, no. 2:1–20.

Fletcher, Samuel C. (2020). "Of War or Peace? Essay Review of Statistical Inference as Severe Testing". *Philosophy of Science*, 87, no. 4.

Fletcher, Samuel C. (2024). "The Stopping Rule Principle and Confirmational Reliability". *Journal for General Philosophy of Science*, 55:1–28.

Fletcher, Samuel C. and Mayo-Wilson, Conor (forthcoming). "Evidence in Classical Statistics". In *Routledge Handbook of the Philosophy of Evidence* (edited by Maria Lasonen-Aarnio and Clayton Littlejohn). Routledge. Preprint available online at `http://philsci-archive.pitt.edu/16191/1/Evidence_in_Classical_Statistics%20%286%29.pdf` Retrieved on 10/25/2021.

Forster, Malcom and Sober, Elliott (2004). "Why Likelihood?" In *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations* (edited by Mark L. Taper and Subhash R. Lele). The University of Chicago Press.

Fortini, Sandra, Ladelli, Lucia, and Regazzini, Eugenio (2000). "Exchangeability, Predictive Distributions and Parametric Models". *Sankhyā: The Indian Journal of Statistics, Series A*, pages 86–109.

Galavotti, Maria Carla (2019). "Pragmatism and the Birth of Subjective Probability". *European Journal of Pragmatism and American Philosophy*, 11, no. XI-1.

Gandenberger, Greg (2015). "A New Proof of the Likelihood Principle". *The British Journal for the Philosophy of Science*.

Gardiner, C. W. (1983). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer.

Gelman, Andrew (2011). "Bayesian Statistical Pragmatism". *Statistical Science*, 26, no. 1:10–11.

Gelman, Andrew and Hennig, Christian (2017). "Beyond Subjective and Objective in Statistics". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, no. 4:967–1033.

Gelman, Andrew, Meng, Xiao-Li, and Stern, Hal (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies". *Statistica Sinica*, pages 733–760.

Gelman, Andrew and Shalizi, Cosma Rohilla (2013). "Philosophy and the Practice of Bayesian Statistics". *British Journal of Mathematical and Statistical Psychology*, 66, no. 1:8–38.

Gelman, Andrew and Yao, Yuling (2020). "Holes in Bayesian statistics". *Journal of Physics G: Nuclear and Particle Physics*, 48, no. 1:014002.

Ghosal, Subhashis and van der Vaart, Aad (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer.

Ghosh, Malay, Reid, Nancy, and Fraser, D.A.S (2010). "Ancillary Statistics: A Review". *Statistica Sinica*, pages 1309–1332.

Glymour, Clark (1980). *Theory and Evidence*. Princeton University Press.

Gneiting, Tilmann, Balabdaoui, Fadoua, and Raftery, Adrian E (2007). "Probabilistic Forecasts, Calibration and Sharpness". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, no. 2:243–268.

Gneiting, Tilmann and Raftery, Adrian E (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". *Journal of the American statistical Association*, 102, no. 477:359–378.

Gorroochurn, Prakash (2016). *Classic Topics on The History of Modern Mathematical Statistics: From Laplace to More Recent Times*. Wiley.

Haaf, Julia M., Klaasen, Fayette, and Rouder, Jeffrey N. (2021). "Bayes Factor vs. Posterior-Predictive Model Assessment: Insights from Ordinal Constraints". *PsyArXiv Preprints*. Latest version March 18th, 2021. Retrieved on 11th July 2024.

Hájek, Alan (2012). "A Puzzle about Belief". Unpublished manuscript. Available online here `https://fitelson.org/topics/hajek_puzzle.pdf` Retrieved on 06/11/2024.

Halmos, Paul R and Savage, Leonard J (1949). "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics". *The Annals of Mathematical Statistics*, 20, no. 2:225–241.

Hendriksen, Allard, de Heide, Rianne, and Grünwald, Peter (2021). "Optional Stopping with Bayes factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations". *Bayesian Analysis*, 16, no. 3:961–989.

Hill, Bruce M. (1987). "The Validity of the Likelihood Principle". *The American Statistician*, 41, no. 2:95–100.

Hoefer, Carl (2012). "Calibration: Being in Tune with Frequencies". *Dialéctica*, 66, no. 3:435–452.

Hogg, Robert V., McKean, Joseph W., and Craig, Allen T. (2019). *Introduction to Mathematical Statistics (8th Edition)*. Pearson.

Hoijtink, Herbert, van Kooten, Pascal, and Hulsker, Koenraad (2016a). "Bayes Factors Have Frequency Properties This Should Not Be Ignored: A Rejoinder to Morey, Wagenmakers, and Rouder". *Multivariate Behavioral Research*, 51, no. 1:20–22.

Hoijtink, Herbert, van Kooten, Pascal, and Hulsker, Koenraad (2016b). "Why Bayesian Psychologists Should Change the Way They Use the Bayes factor". *Multivariate Behavioral Research*, 51, no. 1:2–10.

Huttegger, Simon M (2015a). "Bayesian Convergence to the Truth and the Metaphysics of Possible Worlds". *Philosophy of Science*, 82, no. 4:587–601.

Huttegger, Simon M (2015b). "Merging of Opinions and Probability Kinematics". *The Review of Symbolic Logic*, 8, no. 4:611–648.

Huttegger, Simon M (2017). *The Probabilistic Foundations of Rational Learning*. Cambridge University Press.

Jeffrey, Richard (1984). "De Finetti's Probabilism". In *Foundations: Logic, Language, and Mathematics*, pages 73–90. Springer.

Jeffreys, Harold (1961). *The Theory of Probability*. Oxford University Press.

Joyce, James M. (1998). "A Nonpragmatic Vindication of Probabilism". *Philosophy of Science*, 65, no. 4:575–603.

Joyce, James M. (2009). "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief". In *Degrees of Belief*, pages 263–297. Springer.

Joyce, Jim (2004). "Williamson on Evidence and Knowledge". *Philosophical Books*, 45, no. 4.

Kadane, Joseph B and Lazar, Nicole A (2004). "Methods and Criteria for Model Selection". *Journal of the American statistical Association*, 99, no. 465:279–290.

Kadane, Joseph B and Lichtenstein, Sarah (1982). "A Subjectivist View of Calibration". *Decision Research Report*.

Kalbfleisch, John D. (1975). "Sufficiency and Conditionality". *Biometrika*, 62, no. 2:251–259.

Kass, Robert E. (2006). "Kinds of Bayesians (Comment on Articles by Berger and by Goldstein)". *Bayesian Analysis*, 1, no. 3:437–440.

Kass, Robert E. (2011). "Statistical Inference: The Big Picture". *Statistical Science*, 26, no. 1:1.

Kass, Robert E. and Raftery, Adrian E. (1995). "Bayes Factors". *Journal of the American Statistical Association*, 90, no. 430:773–795.

Kass, Robert E. and Wasserman, Larry (1996). "The Selection of Prior Distributions by Formal Rules". *Journal of the American Statistical Association*, 91, no. 435:1343–1370.

Katopodes, Nikolaos D (2018). *Free-Surface Flow:: Shallow Water Dynamics*. Butterworth-Heinemann.

Kelley, Mikayla (2023). "On Accuracy and Coherence with Infinite Opinion Sets". *Philosophy of Science*, 90, no. 1:92–128.

Kendall, Maurice, Stuart, Alan, and Ord, J. Keith (1987). *Kendall's Advanced Theory of Statistics: Vol 1 Distribution Theory*. Charles Griffin & Company Ltd.

Kyburg, H.E. and Smokler, H.E. Eds. (1964). *Studies in Subjective Probability*. Krieger Publishing. 2nd Edition 1980.

Lad, Frank (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. John Wiley & Sons, Inc.

Lange, Mark (1999). "Calibration and The Epistemological Role of Bayesian Conditionalization". *The Journal of Philosophy*, 96, no. 6:294–324.

Langevin, Paul (1997). "On the Theory of Brownian Motion (Translated by Don S. Lemons and Anthony Gythiel from "Sur la Théorie du Mouvement Brownien," Fr. Acad. Sci. (Paris) 146, 530–533 (1908))". *American Journal of Physics*, 65, no. 11:1079–1081.

Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D (1977). "Calibration of Probabilities: The State of the Art". In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer.

Lindley, D. V. (1972). *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics.

Lindley, Dennis (1986). "Comments on "Why Isn't Everyone a Bayesian?"". *The American Statistician*, 40, no. 1:6–7.

Lindley, Dennis (2000). "The Philosophy of Statistics". *Journal of the Royal Statistical Society Series D: The Statistician*, 49, no. 3:293–337.

Little, Roderick J (2006). "Calibrated Bayes: A Bayes/Frequentist Roadmap". *The American Statistician*, 60, no. 3:213–223.

Little, Roderick J. (2011). "Calibrated Bayes, for Statistics in General, and Missing Data in Particular". *Statistical Science*, 26, no. 2:162–174.

Maddy, Penelope (1997). *Naturalism in Mathematics*. Oxford University Press.

Maddy, Penelope (2007). *Second Philosophy: A Naturalistic Method*. Oxford University Press.

Magnus, Paul D. and Callender, Craig (2004). "Realist Ennui and the Base Rate Fallacy". *Philosophy of Science*, 71:320 – 338.

Maranda, Guillaume R. (2023). "Sampling Error: The Fundamental Flaw of the Severity Measure of Evidence". *Unpublished Manuscript dated March 26th 2023*.

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

Mayo, Deborah G. (2000). "Experimental Practice and an Error Statistical Account of Evidence". *Philosophy of Science*, 67:S193–S207.

Mayo, Deborah G (2011). "Statistical science and Philosophy of Science: Where do/should they meet in 2011 (and Beyond)?" *RMM*, 2:79–102. Available online here `https://vtechworks.lib.vt.edu/items/55188ef8-b8f8-4a86-b21e-f85bd356f215` Retrieved on 06/11/2024.

Mayo, Deborah G. (2013). "Discussion: Bayesian methods: Applied? Yes. Philosophical defense? In flux". *The American Statistician*, 67, no. 1:11–15.

Mayo, Deborah G. (2014). "On the Birnbaum Argument for the Strong Likelihood Principle". *Statistical Science*, 29, no. 2:227–239.

Mayo, Deborah G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Mayo, Deborah G. and Kruse, Michael (2001). "Principles of Inference and their Consequences". In *Foundations of Bayesianism* (edited by David Corfield and Jon Williamson). Springer.

Mayo, Deborah G. and Spanos, Aris (2011). "Error Statistics". In *Handbook of Philosophy of Science Vol. 7 Philosophy of Statistics* (edited by Prasanta S. Bandyopadhya and Malcom R. Forster), pages 153–198. Elsevier.

Mazo, Robert M. (2002). *Brownian Motion: Fluctuations, Dynamics, and Applications*. Oxford University Press.

Morey, Richard D., Romeijn, Jan-Willem, and Rouder, Jeffrey N. (2013). "The Humble Bayesian: Model Checking from a Fully Bayesian Perspective". *British Journal of Mathematical and Statistical Psychology*, 66, no. 1:68–75.

Morey, Richard D., Romeijn, Jan-Willem, and Rouder, Jeffrey N. (2016). "The Philosophy of Bayes Factors and the Quantification of Statistical Evidence". *Journal of Mathematical Psychology*, 72:6–18.

Morey, Richard D., Wagenmakers, Eric-Jan, and Rouder, Jeffrey N. (2016). "Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker". *Multivariate Behavioral Research*, 51, no. 1:11–19.

Munson, Bruce R., Young, Donald F., Okiishi, Theodore H., and Huebsch, Wade W. (2009). *Fundamentals of Fluid Mechanics (6th Edition)*. John Wiley & Sons, Inc.

Mwakima, David (2024a). "Coherence, Calibration and Severity". Unpublished Manuscript.

Mwakima, David (2024b). "On the Quality of Perrin's Evidence". Unpublished Manuscript.

Mwakima, David (2024c). "On the Scope of the Likelihood Principle". Unpublished Manuscript.

Mwakima, David (2024d). "Statistical Tests, Severity and Error". Unpublished Manuscript.

Neyman, Jerzy and Pearson, Egon S (1928). "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I". *Biometrika*, pages 175–240.

Neyman, Jerzy and Pearson, Egon S (1936). "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses". In *Joint Statistical Papers*, pages 240–264. University of California Press.

Neyman, Jerzy and Pearson, Egon Sharpe (1933). "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses". *Philosophical Transactions of the Royal Society of London. Series A*, 231, no. 694-706:289–337.

Norton, John D (2003). "A Material Theory of Induction". *Philosophy of Science*, 70, no. 4:647–670.

Norton, John D (2021). *The Material Theory of Induction*. University of Calgary Press.

Nye, Mary Jo (1972). *Molecular reality: A perspective on the scientific work of Jean Perrin*. American Elsevier Inc: New York.

O'Hagan, Anthony (1995). "Fractional Bayes factors for Model Comparison". *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, no. 1:99–118.

Perrin, Jean (1910). *Brownian Movement and Molecular Reality (Translated from the 1909 8th Series of Annales De Chimie et de Physique by F. Soddy)*. Taylor and Francis.

Pettigrew, Richard (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Psillos, Stathis (2011). "Making Contact with Molecules: On Perrin and Achinstein". In *Philosophy of Science Matters: The Philosophy of Peter Achinstein* (edited by Gregory J. Morgan). Oxford University Press.

Psillos, Stathis (2014). "The View from Within and the View from Above: Looking at van Fraassen's Perrin". In *Bas van Fraassen's Approach to Representation and Models in Science* (edited by Wenceslao J. Gonzalez), pages 143–166. Springer.

Psillos, Stathis (2018). "Realism and Theory Change in Science". In *The Stanford Encyclopedia of Philosophy* (edited by Edward N. Zalta). The Metaphysics Research Lab, Stanford University.

Psillos, Stathis (2021). "From the Evidence of History to the History of Evidence: Descartes, Newton, and Beyond". In *Contemporary Scientific Realism: Challenges from the History of Science* (edited by Timothy Lyons and Peter Vickers). Oxford University Press.

Quine, W. V. O. (1976). "Posits and Reality". In *The Ways of Paradox and Other Essays*, pages 233–241. Random House Inc. Originally published in 1955.

Rapp, Bastian E (2022). *Microfluidics: Modeling, Mechanics and Mathematics*. Elsevier.

Reid, Nancy (1995). "The Roles of Conditioning in Inference". *Statistical Science*, 10, no. 2:138–157.

Reid, Nancy and Cox, David R. (2015). "On Some Principles of Statistical Inference". *International Statistical Review*, 83, no. 2:293–308.

Robert, Christian P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2 edition.

Robins, James and Wasserman, Larry (2000). "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts". *Journal of the American Statistical Association*, 95, no. 452:1340–1346.

Ross, Sheldon M (1996). *Stochastic Processes (Second Edition)*. John Wiley & Sons.

Rouder, Jeffrey N. (2014). "Optional Stopping: No problem for Bayesians". *Psychonomic Bulletin & Review*, 21:301–308.

Rouder, Jeffrey N. and Haaf, Julia M. (2019). "Optional Stopping and the Interpretation of Bayes Factor". *PsyArXiv Preprints*. Latest version April 14th, 2023. Retrieved on 11th July 2024.

Rouder, Jeffrey N. and Morey, Richard D. (2019). "Teaching Bayes' theorem: Strength of Evidence as Predictive Accuracy". *The American Statistician*, 73:186–190.

Rouder, Jeffrey N., Morey, Richard D., and Wagenmakers, Eric-Jan (2016). "The Interplay Between Subjectivity, Statistical Practice, and Psychological Science". *Collabra*, 2, no. 1.

Roush, Sherrilyn (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press.

Royall, Richard (1997). *Statistical Evidence: A Likelihood Paradigm*, volume 71. Chapman & Hall.

Royall, Richard (2004). "The Likelihood Paradigm for Statistical Evidence". In *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations* (edited by Mark L. Taper and Subhash R. Lele). The University of Chicago Press.

Rubin, Donald B (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". *The Annals of Statistics*, pages 1151–1172.

Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Savage, Leonard J. (1971). "Elicitation of Personal Probabilities and Expectations". *Journal of the American Statistical Association*, 66:783–801.

Schwartz, Lorraine (1965). "On Bayes Procedures". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, no. 1:10–26.

Seidenfeld, Teddy (1985). "Calibration, Coherence, and Scoring Rules". *Philosophy of Science*, 52, no. 2:274–294.

Skyrms, Brian (1984). *Pragmatics and Empiricism*. Yale University Press.

Skyrms, Brian (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Smith, George E. and Seth, Raghav (2020). *Brownian Motion and Molecular Reality: A Study in Theory-Mediated Measurement*. Oxford University Press.

Sober, Elliott (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press.

Sprenger, Jan and Hartmann, Stephan (2019). *Bayesian Philosophy of Science*. Oxford University Press.

Stanford, P. Kyle (2009). "Scientific Realism, The Atomic Theory, and The Catch-all Hypothesis: Can We Test Fundamental Theories Against All Serious Alternatives?" *The British Journal for the Philosophy of Science*, 60, no. 2:253–269.

Stanford, P. Kyle (2021). "Realism, Instrumentalism, Particularism: A Middle Path Forward in the Scientific Realism Debate". In *Contemporary Scientific Realism: Challenges from the History of Science* (edited by Timothy Lyons and Peter Vickers). Oxford University Press.

Stein, Howard (2021). "Physics and Philosophy Meet: The Strange Case of Poincaré". *Foundations of Physics*, 51, no. 3:1–24.

Stern, Hal S. (2016). "A Test By Any Other Name: P values, Bayes Factors, and Statistical Inference". *Multivariate Behavioral Research*, 51, no. 1:23–29.

Sudderth, William D. (1995). "Coherent Inference and Prediction in Statistics". In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 833–844. Elsevier.

Sundberg, Rolf (2003). "Conditional Statistical Inference and Quantification of Relevance". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65, no. 1:299–315.

US Food and Drug Administration (2010). "Guidance for the use of Bayesian Statistics in Medical Device Clinical Trials". *Maryland: US Food and Drug Administration*. Available online at `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials` Retrieved October 2023.

van Dantzig, David (1957). "Statistical Priesthood:(Savage on Personal probabilities)". *Statistica Neerlandica*, 11, no. 1:1–16.

van Dongen, Noah, Sprenger, Jan, and Wagenmakers, Eric-Jan (2023). "A Bayesian Perspective on Severity: Risky Predictions and Specific Hypotheses". *Psychonomic Bulletin & Review*, 30, no. 2:516–533.

van Fraassen, Bas C. (1983). "Calibration: A Frequency Justification for Personal Probability". In *Physics, Philosophy and Psychoanalysis: Essays in Honor of Adolf Grünbaum* (edited by R.S. Cohen and L. Laudan), pages 295–319. D. Reidler Publishing Company.

van Fraassen, Bas C. (1984). "Belief and the Will". *The Journal of Philosophy*, 81, no. 5:235–256.

Vanpaemel, Wolf (2010). "Prior Sensitivity in Theory Testing: An Apologia for the Bayes factor". *Journal of Mathematical Psychology*, 54, no. 6:491–498.

von Mises, Richard (1957). *Probability, Statistics and Truth*. Dover.

von Mises, Richard (1964). *Mathematical Theory of Probability and Statistics*. Academic Press.

Wald, Abraham (1949). "Statistical Decision Functions". *The Annals of Mathematical Statistics*, pages 165–205.

Wald, Abraham (1950). *Statistical Decision Functions*. John Wiley & Sons. 5th Edition of 1964.

Walker, Stephen (2005). "Bayesian Nonparametric Inference". In *Handbook of Statistics, Vol. 25* (edited by Dipak K. Dey and C.R. Rao), pages 339–373. Elsevier B.V.

Wasserman, Larry (2006). "Frequentist Bayes is Objective (Comment on Articles by Berger and by Goldstein)". *Bayesian Analysis*, 1, no. 3:451–456.

Weirich, Paul (2011). "Calibration". In *EPSA Philosophy of Science: Amsterdam 2009*, pages 415–425. Springer.

Williamson, Timothy (2000). *Knowledge and Its Limits*. Oxford University Press.

Woodward, James (1989). "Data and Phenomena". *Synthese*, 79, no. 3:393–472.

Woodward, James (2011). "Data and Phenomena: A Restatement and Defense". *Synthese*, 182, no. 1:165–179.