TWO PRINCIPLES OF EVIDENCE AND THEIR IMPLICATIONS FOR THE PHILOSOPHY OF SCIENTIFIC METHOD

by

Gregory Stephen Gandenberger

BA, Philosophy, Washington University in St. Louis, 2009

MA, Statistics, University of Pittsburgh, 2014

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Gregory Stephen Gandenberger

It was defended on

April 14, 2015

and approved by

Edouard Machery, Pittsburgh, Dietrich School of Arts and Sciences

Satish Iyengar, Pittsburgh, Dietrich School of Arts and Sciences

John Norton, Pittsburgh, Dietrich School of Arts and Sciences

Teddy Seidenfeld, Carnegie Mellon University, Dietrich College of Humanities & Social Sciences

James Woodward, Pittsburgh, Dietrich School of Arts and Sciences

Dissertation Director: Edouard Machery, Pittsburgh, Dietrich School of Arts and Sciences

TWO PRINCIPLES OF EVIDENCE AND THEIR IMPLICATIONS FOR THE PHILOSOPHY OF SCIENTIFIC METHOD

Gregory Stephen Gandenberger, PhD

University of Pittsburgh, 2015

The notion of evidence is of great importance, but there are substantial disagreements about how it should be understood. One major locus of disagreement is the Likelihood Principle, which says roughly that an observation supports a hypothesis to the extent that the hypothesis predicts it. The Likelihood Principle is supported by axiomatic arguments, but the frequentist methods that are most commonly used in science violate it.

This dissertation advances debates about the Likelihood Principle, its near-corollary the Law of Likelihood, and related questions about statistical practice. Chapter 2 provides a new axiomatic proof of the Likelihood Principle that avoids influential responses to previous proofs. Chapter 3 exhibits the close connection between the Likelihood Principle and the Law of Likelihood and responds to three purported counterexamples to them. Chapter 4 presents a new counterexample that is more difficult to avoid but argues that it does not speak against those principles in typical applications.

The next two chapters turn to implications. It is motivated by tension among three desiderata for a method of evaluating hypotheses in light of data. We would like such a method to (1) respect the evidential meaning of data, (2) provide direct guidance for belief or action, and (3) avoid inputs that are not grounded in evidence. Unfortunately, frequentist methods violate (1) by violating the Likelihood Principle, likelihoodist methods violate (2) by directly addressing only questions about the evidential meaning of data, and Bayesian methods violate (3) by using prior probabilities. Chapter 5 sharpens this tension by arguing that no method that satisfies likelihoodist strictures can provide a genuine rival to

frequentist and Bayesian methodologies. Chapter 6 argues that many frequentist violations of the Likelihood Principle are not required by basic frequentist commitments. Those that are may be permissible for the sake of enabling progress in the presence of highly indefinite prior beliefs, but they are not mandatory or even strongly motivated. These considerations support more widespread use of Bayesian methods despite difficulties in specifying prior probability distributions.

TABLE OF CONTENTS

PREFACE xiv

1.0 INTRODUCTION 1

1.1 Evidence and the Norms of Scientific Research 1

1.2 Statistical Inference and Scientific Inference 5

1.3 The Likelihood Principle and Its Bearing on Methodology 10

1.4 The Contributions Made by This Dissertation 11

2.0 A NEW PROOF OF THE LIKELIHOOD PRINCIPLE 13

2.1 Introduction 13

2.2 The New Proof 17

2.3 How the New Proof Addresses Proposals to Restrict Birnbaum's Premises . 30

2.4 A Response to Arguments that the Proofs are Fallacious 36

2.5 Conclusion 39

3.0 NEW RESPONSES TO THREE COUNTEREXAMPLES TO THE LIKELIHOOD PRINCIPLE 41

3.1 Introduction 41

3.2 Why a counterexample to the Law of Likelihood is a *prima facie* counterexample to the Likelihood Principle 42

3.3 Response to Fitelson's counterexample 44

3.3.1 Objection 1: Response conflicts with constraints on evidential favoring 46

3.3.2 Objection 2: Response fails to address tacking paradox 48

3.3.2.1 Response 1: Bite the bullet 49

3.3.2.2 Response 2: Regard Law as explicating "r-favoring" 50

3.3.2.3 Response 3: Restrict Law of Likelihood to structurally identical alternatives . . . 51
3.3.3 Objection 3: Response excludes cases of scientific interest . . . 52
3.3.3.1 Cases involving competing causal claims . . . 52
3.3.3.2 Cases involving nested models . . . 54
3.4 Response to Armitage's counterexample . . . 57
3.4.1 Two problems for attempts to use the Weak Repeated Sampling Principle as an objection to the Law of Likelihood . . . 58
3.4.1.1 Problem 1: The Weak Repeated Sampling Principle is unreasonably strong . . . 58
3.4.1.2 Problem 2: The Law of Likelihood itself cannot conflict with the Weak Repeated Sampling Principle . . . 60
3.4.2 Why Armitage's example is no threat to the Law of Likelihood . . . 61
3.5 Response to Stein's counterexample . . . 63
3.6 Conclusion . . . 68
4.0 A COUNTEREXAMPLE TO THE LAW OF LIKELIHOOD FOR PROBABILITY-ZERO HYPOTHESES . . . 69
4.1 Introduction . . . 69
4.2 A Counterexample to the Law of Likelihood for Probability-Zero Hypotheses 70
4.3 Responses to Worries About the Counterexample . . . 71
4.3.1 Worry 1: Each hypothesis omits two points on the relevant great circle 73
4.3.2 Worry 2: The likelihood ratio is small . . . 73
4.3.3 Worry 3: Real measurement techniques have finite precision . . . 74
4.3.4 Worry 4: The hypotheses belong to different models . . . 74
4.3.5 Worry 5: No limiting operation is specified . . . 78
4.4 A No-Go Result for Addressing the Counterexample Through an Alternative Theory of Conditional Probability . . . 79
4.5 Addressing the Counterexample Within Existing Theories of Conditional Probability . . . 84
4.5.1 Restricting the Law of Likelihood . . . 85

4.5.2 Regarding Evidential Favoring as Either Relative or Indeterminate . 89
4.6 Conclusion . . . 92
5.0 WHY I AM NOT A METHODOLOGICAL LIKELIHOODIST 93
5.1 Introduction . . . 93
5.2 Methodological Likelihoodism . . . 94
5.3 An Overview of My Argument Against Methodological Likelihoodism . . . 96
5.4 Claim 1: An adequate methodology provides a good norm of commitment . 97
5.5 Claim 2: A good purely likelihood-based norm of commitment would have a particular form . . . 99
5.5.1 Objection: A good norm of commitment could allow withholding judgment . . . 101
5.5.2 Reply: Allowing withholding judgment does not help . . . 102
5.6 Claim 3: A good norm of commitment would allow one to prefer chance hypotheses to maximally likely alternatives . . . 103
5.7 Claim 4: A good norm of commitment would treat alike pairs consisting of a hypothesis and its negation that are logically equivalent given one's evidence 106
5.7.1 Objection 1: H_1 and H_2 Are Not Mutually Exclusive . . . 107
5.7.2 Reply to Objection 1 . . . 108
5.7.3 Objection 2: H_1 and H_2 are not statistical hypotheses . . . 108
5.7.4 Reply to Objection 2 . . . 108
5.8 Claim 5: A good norm of commitment would allow one both to obey a highly plausible disjunction rule and to treat pairs of logically equivalent hypotheses alike . . . 109
5.8.1 Objection: This argument relies on a strict notion of neutrality that is not clearly appropriate . . . 112
5.8.2 Reply: Making the notion of neutrality less strict does not result in a successful theory . . . 112
5.9 Conclusion . . . 113
6.0 WHY FREQUENTIST VIOLATIONS OF THE LIKELIHOOD PRINCIPLE ARE PERMISSIBLE AT BEST . . . 115

6.1 Introduction . . . 115
6.2 How Standard Frequentist Practice Conforms to Likelihoodist Principles Within Component Experiments . . . 117
6.3 How Standard Frequentist Practice Violates Likelihoodist Principles Across Component Experiments . . . 120
6.3.1 Frequentists generally use a common Type I error rate across experiments . . . 120
6.3.2 Frequentists Generally Adjust Their Standards on Individual Tests to Control "Familywise" Error Rates . . . 124
6.3.3 Frequentists Generally Require That Hypotheses Be Predesignated . . . 126
6.4 Why Violating the Likelihoodist Principles Is Incompatible With Maximizing Expected Utility . . . 127
6.5 How Thinking of Frequentist Considerations as Tie-Breakers Allows One to Argue That They Are Permissible, but Not That They Are Preferable . . . 132
6.6 Conclusion . . . 134
7.0 CONCLUSION . . . 135
APPENDIX A. PROOF THAT $\mathrm{EV}(E_2, Y_0) = \mathrm{EV}(E^B, G)$ . . . 137
APPENDIX B. PROOF THAT (CE) AND (†) ARE JOINTLY INCOMPATIBLE WITH SIX LEADING MEASURES OF CONFIRMATION 139
B.1 (d) $c(H,E) = \Pr(H|E) - \Pr(H)$ . . . 140
B.2 (s) $c(H,E) = \Pr(H|E) - \Pr(H|\bar{E})$ . . . 140
B.3 (c) $c(H,E) = \Pr(H\&E) - \Pr(H)\Pr(E)$ . . . 140
B.4 (n) $c(H,E) = \Pr(E|H) - \Pr(E|\bar{H})$ . . . 141
B.5 (l) $c(H,E) = \log\left[\frac{\Pr(E|H)}{\Pr E|\bar{H})}\right]$ . . . 142
B.6 (r) $c(H,E) = \log\left[\frac{\Pr(H|E)}{\Pr(H)}\right]$ . . . 142
APPENDIX C. PROOF THAT ARMITAGE'S EXPERIMENT MERELY TRADES OFF ONE KIND OF MISLEADINGNESS AGAINST ANOTHER . . . 143
C.1 Step 1 . . . 143
C.2 Step 2 . . . 144

C.3 Step 3 145
C.4 Step 4 146
APPENDIX D. TECHNICAL DETAILS FROM STEIN'S EXAMPLE . . 147
APPENDIX E. DERIVING THE ANOMALOUS LIKELIHOOD RATIO . 149
APPENDIX F. CONSTRUCTION FOR SECTION 5.7 152
APPENDIX G. DESCRIPTION OF EXPERIMENT FOR SECTION 5.8 . 155
BIBLIOGRAPHY 157

LIST OF TABLES

2.1 Sampling distribution of hypothetical experiment used to motivate Weak Ancillary Realizability 21

2.2 Sampling distribution of E_1^{CE} 29

6.1 Sampling distributions for Example 6.1 118

6.2 Likelihood function for the binomial experiment 121

6.3 Likelihood function for the negative binomial experiment 121

6.4 A pair of hypothetical sampling distributions 128

6.5 A trio of possible tests from Table 6.4 129

A1 Sampling Distribution of E_2^{CE} ($c^* = \frac{1}{1+c}$) 137

F1 A probability assignment that yields the desired result 152

G1 Probabilities of possible die roll outcomes as a function of r' 156

LIST OF FIGURES

2.1 Evidentially equivalent outcomes from the pair of procedures described in Example 2.2.[1] 24

2.2 Evidentially equivalent outcomes from the pair of procedures described in Example 2.3. The bias of the coin for heads p is the same as the fraction of the first spinner that the R regions occupy. 26

2.3 Evidentially equivalent outcomes from the pair of procedures described in Example 2.4. The bias of the coin for heads p is the same as the fraction of the first spinner that the R and S regions occupy in total, which is the same as the fraction that the R and T regions occupy. 27

2.4 Graphical depiction of the series of equivalences used to establish Theorem 1. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (ECP = Experimental Conditionality Principle, WARP = Weak Ancillary Realizability Principle, LP = Likelihood Principle). 30

2.5 Birnbaum's proof of the Likelihood Principle. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (WCP = Weak Conditionality Principle, SP = Sufficiency Principle, LP = Likelihood Principle). 33

4.1 An illustration of Example 4.1 70

4.2 Equator and prime meridional circle are treated differently despite the arbitrariness of the distinction. 72

4.3 The application of Weak Conglomerability to A. 83
4.4 The application of Weak Conglomerability to A'. 83
6.1 α and β for the three tests shown in Table 6.5. 129

PREFACE

Special thanks to my wife for her love and support.

Thanks to my teachers and mentors, particularly Eric Brown, Carl Craver, Lindley Darden, Kenny Easwaran, Branden Fitelson, Satish Iyengar, Edouard Machery, Deborah Mayo, John Norton, Teddy Seidenfeld, Kent Staley, and Jim Woodward.

Thanks to the following people for discussions that contributed to the development of the chapter indicated.

Chapter 1: James Berger, Andrew Gelman, Leon Gleser, Jason Grossman, Nicole Jinn, James Joyce, Kevin Kelly, Jonathan Livengood, Edouard Machery, Deborah Mayo, Conor Mayo-Wilson, John Norton, Jonah Schupbach, Teddy Seidenfeld, Elizabeth Silver, Jan Sprenger, Paul Weirich, James Woodward, John Worrall, and two anonymous referees.

Chapter 2: Jacob Chandler, Kenny Easwaran, Branden Fitelson, Satish Iyengar, Michael Lew, and Mike Titelbaum.

Chapter 3: Adam Brodie, Kenny Easwaran, Branden Fitelson, Satish Iyengar, Michael Lew, Dan Malinsky, Wayne Myrvold, Alexander Pruss, Bryan Roberts, Teddy Seidenfeld, and Nathan VanHoudnos.

Chapter 4: Marcus Adams, Gar Allen, Jake Chandler, Carl Craver, Foad Diazadji-Bahmani, Branden Fitelson, Samuel Fletcher, Clark Glymour, Christopher Hitchcock, Satish Iyengar, Michael Lew, Edouard Machery, Conor Mayo-Wilson, John Norton, Anya Plutynski, Elay Schech, Teddy Seidenfeld, Elliott Sober, Jan Sprenger, and Jim Woodward.

Chapter 5: James Joyce, Kevin Kelly, Michael Lew, Edouard Machery, Deborah Mayo, Teddy Seidenfeld, and Jan Sprenger.

My apologies to anyone whose contributions I have inadvertently failed to acknowledge. Any remaining errors are my own.

1.0 INTRODUCTION

1.1 EVIDENCE AND THE NORMS OF SCIENTIFIC RESEARCH

In 1985, a team of researchers led by Robert Bartlett at the University of Michigan Medical Center reported promising results from the first hundred uses of a technique called Extracorporeal Membrane Oxygenation (ECMO) for the treatment of respiratory failure in infants. The patients in the first fifty cases were newborns who had failed to respond to conventional treatment and were not expected to survive. With ECMO, 54% did survive. The next fifty were moderate to high-risk patients considered to have no more than a 20% chance of survival under conventional therapy. With ECMO, 90% survived.

Those results seem to provide good evidence for the superiority of ECMO to conventional therapy at least for moderate- to high-risk patients. However, they did not satisfy the conventional standard for establishing the efficacy of a medical treatment: a p-value of no more than .05 in a randomized clinical trial published in a peer-reviewed journal. The study was not randomized, meaning that it did not involve using a random process to decide whether to assign a given patient who met the trial's entrance criteria to ECMO or to conventional therapy. In the presence of randomization, standard statistical methods can be used to control the probability that differences in the baseline characteristics of the two groups will lead to incorrect conclusions.

After those promising early results, Bartlett et al. attempted to satisfy the conventional standard for establishing ECMO's efficacy by running a randomized trial with infants who were considered to have a 20% chance or less of surviving on conventional therapy. They were reluctant to use the standard approach to randomization in which each patient is assigned to a treatment group with equal probability because their previous results seemed

to favor ECMO so strongly. Thus, they used an unconventional "randomized play-the-winner" approach. That approach can be understood in terms of an urn model. The first patient's treatment is chosen by a process that has the same stochastic properties as selecting a ball at random (with equal probabilities) from an urn that contains one black ball and one red ball, with the black ball leading to ECMO and the red ball leading to conventional therapy. If a patient survives on a given treatment, then the randomization probabilities are adjusted in a way that is equivalent to adding a ball to the urn with the color that leads to that treatment. If the patient does not survive, then a ball of the other color is added. This approach is continued until the urn contains ten balls of a given color. After that event, randomization ends and subsequent patients are given the "winning" treatment. This approach provided a compromise between the needs of future patients, who would benefit from a careful comparison between ECMO and conventional therapy, and those of the patients enrolled in the trial, who would benefit from using the treatment with the best expected outcome on the basis of the evidence that was already available.

In this study, the first neonate was assigned to ECMO and survived. The next was assigned to conventional therapy and died. The following eight were all assigned to ECMO, and all survived. For reasons not clearly specified in the trial report, two more patients were randomly assigned despite the fact that the hypothetical urn already contained ten ECMO balls. Both received ECMO and survived. Subsequent patients were then given ECMO without randomization. Of the next ten patients, eight received ECMO and survived, and two did not receive it and died.[1] In total, all nineteen patients who received ECMO survived, while all five who received conventional therapy died.

Again, these results seem to provide good evidence for the superiority of ECMO but failed to satisfy the conventional standard of a p-value of .05 or less in a randomized clinical trial published in a peer-reviewed journal. It was fortunate for the patients in the randomized phase of the trial that only one received conventional therapy, but the result was that the p-value of the randomized phase of the trial was .083.

Ware attempted to address this perceived shortcoming in Bartlett et al.'s results by

[1]It is again unclear why two patients were given conventional therapy after randomization had ceased. One possibility is a shortage of resources, such as a lack of availability of ECMO machines.

running a second randomized trial comparing ECMO to conventional therapy. He too used an unconventional design in an attempt to balance the needs of present and future patients. His approach was to use standard 50-50 randomization until four patients died under either treatment, and then to switch to the treatment that did not produce the four deaths. In the randomized phase of the trial, four of ten patients given conventional therapy died, compared to zero of nine given ECMO. Twenty more were given ECMO in the nonrandomized phase, and all but one survived.

Yet again, these results seem to provide good evidence for the efficacy of ECMO but failed to satisfy the conventional standard of a p-value of .05 or less in a randomized clinical trial published in a peer-reviewed journal: the randomized portion had a p-value of .054 [Begg, 1989].

A third trial with conventional randomization was carried out between 1988 and 1992. It resulted in one death among fifteen patients treated with ECMO and six deaths among thirteen given conventional therapy [Gross et al., 1994]. This trial seems to have been at least somewhat justified in that it focused on infants born in small hospitals who would need to be transferred to large medical centers for ECMO but not for conventional therapy.

This third trial did yield a p-value of less than .05 in a randomized clinical trial, but it was published only in conference proceedings. As a result, yet another conventionally randomized trial was carried out, by the UK Collaborative ECMO Group [1996]. The trial protocol called for a sample size of three hundred patients but allowed for a data monitoring committee to end the trial for early evidence of risk or harm. Not surprisingly, the trial was terminated before the planned sample size of three hundred was reached because of strong early evidence of ECMO's superiority. In the end, thirty of ninety-three infants treated with ECMO died, compared to fifty-four out of ninety-two treated with conventional therapy.

This trial too failed to reach the frequentist gold standard! Because the recommendations of the data-monitoring committee and decisions of the trial steering committee are not governed by a strict protocol, the probability of getting a result more extreme than the one observed if in fact both treatments are equally effective—the p-value—lacks an objectively well-specified value. Thankfully, it was nevertheless accepted as conclusive.

One lesson to draw from this example is that how we think about evidence matters.

Each trial was arguably justified on the view that telling evidence requires a result with a p-value of .05 or less from a study that has withstood the scrutiny necessary for publication in a peer-reviewed journal. According to prominent Bayesian and likelihoodist commenters, at most the first randomized trial was needed [Royall, 1989; Berry, 1989]. In all of these trials together, thirty-one out of 156 infants treated with ECMO died (20%), compared to sixty-nine out of 120 given conventional therapy (58%). If the 120 given conventional therapy had been given ECMO instead and 20% of them had survived, then forty-five infants' lives would have been saved—not to mention the many infants outside the trials who died under conventional therapy despite the favorability of the available data toward ECMO's superiority.

Another lesson is that the conventional standard for telling evidence in medical research (as well as many other areas of science) has some features that are at best highly questionable. It ignores trials that are not randomized or are not published in peer-reviewed journals. It does not allow for evidence to accumulate across studies.[2] It has an arbitrary cutoff that is insensitive to the stakes. And it uses p-values, which are sensitive to some considerations that are arguably irrelevant to the evidential import of the data, such as the intentions of the experimenters about when to end data collection.

The conventional standard is a frequentist one, but it goes beyond basic frequentist commitments. It is exactly the kind of mindless, recipe-like approach that frequentist pioneers Neyman, Pearson, and Fisher all rejected [Gigerenzer, 2004, 321–2], using the same cutoff for statistical significance in all cases and refusing to draw any inferences not licensed by an explicit statistical procedure (such as inferences that depend on informally considering the cumulative force of multiple suggestive but non-conclusive results). Thus, we should not be too quick to pin this debacle on frequentism itself. However, it is worth asking whether or not basic frequentist commitments were part of the problem and if so what can be done to improve statistical thinking.

[2]The "gold standard" of a p-value of .05 or less in a randomized trial published in a peer-reviewed journal has arguably been superseded by the "platinum standard" of approval by meta-analysis of randomized trials. Meta-analysis does allow evidence to accumulate across studies and would have been valuable in the ECMO case. However, it faces challenges of its own; see [Stegenga, 2011] for a critical appraisal.

1.2 STATISTICAL INFERENCE AND SCIENTIFIC INFERENCE

Probability theory and statistics are both concerned with the relationship between the stochastic properties of data-generating mechanisms[3] and the data those mechanisms produce. Probability theory addresses *direct inference problems*, which involve using information about the stochastic properties of a data-generating mechanism to determine the probability that it will generate a certain kind of data in a particular case. For instance, it can be used to determine the probability that a coin that has equal probabilities of landing heads or tails on a given toss will land heads five times in a row at least once in a series of fifty independent tosses. Probability theory also addresses questions about what kinds of properties a data-generating mechanism must have in order for its probability of producing a particular kind of data to have a particular stochastic property. A famous example is the birthday problem: assuming that birthdays are uniformly distributed over the three-hundred sixty-five days of the calendar (ignoring leap years), how many people must there be in a group for the probability that two of them share a birthday to be at least one half?

Statistics, by contrast, is concerned with *inverse inference problems*: given some data and information about the circumstances in which it was produced, estimate or infer stochastic properties of the data-generating mechanism. For instance, it addresses questions like, "given that this coin landed heads sixty times in one hundred independent flips, should we reject the hypothesis that it has probability .5 of landing heads in a given toss?" and "What percentage of the votes should we expect the Republican candidate for president to receive in the next election, given that 538 out of 1000 said that they planned to vote for her in a recent telephone survey?"

Within probability theory, there is debate about how probabilistic statements should be understood but little debate about how direct inference should proceed: it should proceed deductively from Kolmogorov's axioms. The only disagreements about calculations that arise concern cases involving infinite disjunctions and probability-zero conditioning events

[3]Some subjective Bayesians de Finetti, 1975 would object to these characterizations, maintaining that "stochastic properties" are a fiction and that probabilistic statements should be interpreted exclusively in terms of subjective degrees of belief. Nevertheless, the notion of a stochastic property is sufficiently well-entrenched to be useful for fixing ideas.

[Seidenfeld, 2001], both of which can generally be avoided in practice. Within statistics, by contrast, there are constantly raging debates about how inverse inference should proceed.

The two dominant approaches to statistical inference are Bayesian and frequentist. Bayesians assimilate statistics to probability theory by assigning probabilities to all propositions, updating those probabilities in light of new data through conditionalization, and using decision theory to decide what inferences to draw and what estimates to give. For instance, suppose that one is asked whether or not one rejects the hypothesis that a particular coin is fair given that it landed heads sixty times in one hundred tosses. A Bayesian would typically[4] proceed by assigning a prior probability Pr(fair) to the hypothesis that the coin is fair—not yet taking into account the sixty heads in one hundred tosses—and then updating that probability on the datum $H = 60$ according to Bayes's theorem:

$$\Pr(\text{fair}|H = 60) = \frac{\Pr(H = 60|\text{fair})\Pr(\text{fair})}{\Pr(H = 60)} \tag{1.1}$$

Calculating $\Pr(H = 60|\text{fair})$ is a straightforward exercise in direct inference. Calculating $\Pr(H = 60)$ is not. One way to proceed is to assign a prior probability distribution f over the whole set of possible biases the coin might have and then using the formula $\Pr(H = 60) = \int \Pr(H = 60|\text{bias} = p)f(\text{bias} = p)dp$. Having calculated $\Pr(\text{fair}|H = 60)$ in this way, one could then use the tools of decision theory to decide whether or not to reject the hypothesis that the coin is fair, for instance by calculating the expected utility of that action.

Frequentists reject this approach because they reject assigning probabilities to hypotheses indiscriminately, at least in scientific contexts. Historically, they have insisted that probability statements have an objective interpretation in terms of either propensities or long-run frequencies. However, one could be a frequentist in the relevant sense while maintaining that probability statements should be interpreted in terms of personal degrees of belief by maintaining that only personal degrees of belief that are objectively well-grounded are scientifically legitimate. The word "frequentist" is less than ideal in this context because it refers both to a body or views and practices having to do with inverse inference and to a view

[4]I write "typically" because Bayesians need not reason in this sequential way, although they typically do. A Bayesian could in principle have direct introspective access to both his or her prior and posterior probabilities without the need to calculate the latter from the former, for instance. Bayesian norms require only that prior and posterior probabilities be related as Bayes's theorem dictates.

about the proper interpretation of probability statements. My concern in this dissertation is with inverse inference rather than with interpretations of probability, so I use "frequentist" in the former sense throughout, unless otherwise noted.

Frequentists deny the legitimacy of assigning probabilities to hypotheses that are not themselves about the outcomes of a data-generating process with given stochastic properties, as most of the hypotheses that are considered in science are not. For that reason, they cannot assimilate inverse inference to direct inference as Bayesians do. They must find some other way to proceed. To a first approximation, they look for methods that would perform well in repeated applications in the long run no matter what the truth may be.

Consider again the question whether or not one should reject the hypothesis that a particular coin is fair on the basis of the fact that it produced sixty heads in one hundred flips. Suppose that the alternative hypothesis of interest is that the probability of heads on a given flip is greater than .5. Then a frequentist would typically proceed by designating $p \leq .5$ the "null hypothesis" H_0 and $p > .5$ the "alternative hypothesis" H_a. He or she would then go through the following steps:

1. **Choose the maximum probability of Type I error that one is willing to accept.** To commit a Type I error is to reject the null hypothesis when it is true. The maximum probability of Type I error that one should be willing to accept depends on how costly a Type I error would be. Frequentists can choose any value here that they want, but 1% and especially 5% are conventional.

2. **Find the test (it it exists) that has the lowest rate of Type II error for any departure from H_0 among all tests that have no more than the desired rate of Type I error.** To commit a Type II error is to fail to reject the null hypothesis when it is false. Such a "uniformly most powerful" test does exist in typical cases when testing a "one-sided" hypothesis of the form $\phi \leq a$ (or $\phi \geq a$) against a hypothesis of the form $\phi > a$ (or $\phi < a$), but not when testing a "point hypothesis" of the form $\phi = a$ against the "two-sided" alternative $\phi \neq a$. Frequentists appeal to further, somewhat *ad hoc* criteria (such as symmetry considerations) to chose among tests when no uniformly most powerful test exists.

Frequentist procedures are often described in terms of p-values. The p-value of a result is the probability that the experiment that produced it would produce a result "at least as extreme" as it under the null hypothesis. The phrase "at least as extreme as" here has multiple interpretations; happily, those interpretations coincide in many standard cases. It turns out that in those cases, a test that rejects the null hypothesis if and only if the p-value of the observed result is α or less has Type I error rate α. Thus, the conventional approach of requiring a p-value of .05 or less to reject a null hypothesis is tantamount to requiring a 5% Type I error rate.

The concepts of alternative hypotheses and Type I and Type II error rates were introduced by Neyman and Pearson [1928], [1933]. Their primary aim was to provide a rationale for techniques that had already been develop by R. A. Fisher. Fisher seemed initially to accept their work [Fisher, 1934] but later became hostile [Fisher, 1955] toward it. He thought of p-values as quantifying the evidence in the data against the null hypothesis rather than as providing the basis for a procedure justified by its long-run operating characteristics. He considered Neyman and Pearson's focus on long-run operating characteristics appropriate perhaps for quality-control procedures in an industrial setting, but not for scientific research.

Fisher later developed the "fiducial approach," which ascribes probabilities to hypotheses in light of data without the use of prior probabilities. He never developed a full-fledged theory of this approach, and there is currently a fair degree of consensus among statisticians and philosophers of statistics that it was misconceived [Seidenfeld, 1992].

Some of Fisher's ideas live on in the *likelihoodist* approach. Followers of this approach attempt to provide a "third way" to do statistics that combines the most attractive features of the frequentist and Bayesian approaches. Like frequentist methods, likelihoodist approaches avoid the use of prior probabilities. Like Bayesian methods, they conform to the *Likelihood Principle*. This principle says that the evidential import of a datum with respect to a set of hypotheses depends only on the probabilities that those hypotheses ascribe to *that* datum. More precisely, it says that the evidential import of the body of data E with respect to he set of hypotheses $\mathbf{Y}$ depends only on $\Pr(E|H)$ as H varies over $\mathbf{Y}$. This principle rules out methods based on p-values, which are sensitive to the probabilities that the hypotheses being considered ascribe to hypotheses *more extreme than* the one that was observed.

The primary tool of the likelihoodist approach is the *Law of Likelihood*, which says that body of data E favors hypothesis H_1 over hypothesis H_2 if and only if the likelihood ratio $\mathcal{L} = \Pr(E|H_1)/\Pr(E|H_2)$ is greater than one, with $\mathcal{L}$ measuring the degree of favoring. Likelihoodists report likelihood ratios and likelihood functions (that is, $\Pr(E|H)$ as a function of H) as a way of quantifying the evidence in the data.

The primary obstacle to widespread acceptance of the likelihoodist approach is that it is not a theory of inference or decision, but only of evidence. It is not clear what one is to do with the output of likelihoodist methods if not use them for Bayesian updating or frequentist testing.

Bayesian, frequentist, and likelihoodist approaches are most easily used with a set of simple statistical hypotheses—that is, a set of hypotheses each of which entails a definite probability distribution over the possible outcomes of a given experiment. They can be extended, with some difficulties, to *complex* statistical hypotheses—that is, disjunctions of simple statistical hypotheses. Simple and complex statistical hypotheses are important in science, but they do not exhaust its content. For instance, the theory that all life on earth evolved from a common ancestor is relevant to the observation that small islands generally have more endemic species than large continents but does not confer a definite probability on that observation.

The Bayesian approach is nicely unified in that it involves using the same basic approach for statistical and non-statistical hypotheses: assign prior probabilities and update using Bayes's theorem. The difference between statistical and non-statistical hypotheses is simply that the likelihood function $\Pr(E|H)$ is entailed by H when H is a simple statistical hypothesis and by H together with a probability distribution over H and simple alternatives to it when H is a complex statistical hypotheses, but is not entailed by H when H is a substantive hypothesis.

There is no consensus on how to think about substantive hypotheses from a frequentist perspective. Perhaps the most well-articulated theory is Deborah Mayo's, according to which the same general principle applies to all hypotheses: good evidence for them comes from their passing severe tests, which means that a data-generating procedure produces results fit them to well to a degree that would be unlikely if they were false. On this account, severe

testing of "high-level" scientific theories somehow arises out of severe testing of its lower-level consequences [2009b].

Likelihoodists are also divided on how to handle substantive hypotheses. Edwards, for instance, does not accept probability statements that do not have frequency interpretations, and so is generally unwilling to contemplate probabilities conditional on substantive hypotheses [1972, xv]. Royall and Sober, on the other hand, are more liberal in the kinds of probability statements they are willing to consider (see e.g. Royall 1997, 13, Sober 2008, 26). Sober in particular requires only that probabilities be "empirically well-grounded," and he considers in some detail how the available evidence bears on the substantive theories of evolution and of intelligent design according to the Law of Likelihood.

In summary, statistical inference is about inferring the stochastic properties of a data-generating mechanism from the data it generates. Frequentists and Bayesians provide different approaches to statistical inference, and likelihoodists provide a third approach that eschews inference in favor of merely characterizing data as evidence. These approaches can be extended to provide more general theories of inductive inference for science, but those extensions involve some difficulties.

1.3 THE LIKELIHOOD PRINCIPLE AND ITS BEARING ON METHODOLOGY

Much of the debate among frequentists, Bayesians, and likelihoodists has centered on the correct interpretation of probability and the legitimacy of using personal probabilities in science. However, these debates can be approached from a different direction. Frequentist methods violate the Likelihood Principle, while likelihoodist methods and Bayesian updating do not. Thus, frequentists seem to be committed either to different ideas about evidence than likelihoodists and Bayesians or to the preeminence of other kinds of considerations (such as the control of Type I and Type II error rates) over evidential ones.

The Likelihood Principle has been much debated among statisticians. It came to prominence in 1962 when Allan Birnbaum showed that it follows from two principles to which

frequentist statisticians appear to be committed. Many responses to Birnbaum's proof have been given since then. Durbin [1970] and Kalbfleisch [1975] argued for restricting its premises in ways that suffice to block the proof while purportedly preserving what is compelling about them from frequentist and pre-theoretic perspectives. Joshi [1990] and more recently Mayo [2014] have argued that the proof is actually invalid. Birnbaum himself thought that a concept of evidence should both conform to the Likelihood Principle and provide frequentist guarantees about long-run performance. He took the fact that those desiderata cannot both be satisfied by the same concept to show that our concept of evidence is "anomalous" [1964].

Frequentists have attempted not only to undermine purported proofs of the Likelihood Principle, but also to defeat it by counterexamples. Most of those counterexamples are in fact directed more immediately against the Law of Likelihood than the Likelihood Principle. However, they are at least *prima facie* counterexamples to the Likelihood Principle because are fairly persuasive arguments from the Likelihood Principle to the Law of Likelihood.

Settling the debate about the Likelihood Principle is only a first step toward settling debates about statistical methods. There are at least two further issues. First, likelihoodist and Bayesian methods both conform to the Likelihood Principle. The choice between them depends on one's willingness to assign prior probabilities to hypotheses and one's view about the adequacy of an approach that merely quantifies degrees of evidential favoring. Second, Frequentist methods violate the Likelihood Principle, but one could support their use while accepting the Likelihood Principle by maintaining that what matters is not respecting evidential equivalence but using techniques that control long-run rates of Type I and Type II error.

1.4 THE CONTRIBUTIONS MADE BY THIS DISSERTATION

This dissertation addresses the issues just sketched. Chapters 1–3 argue for the Likelihood Principle while Chapters 4 and 5 argue for its pro-Bayesian implications.

Chapter 1 provides a new proof of the Likelihood Principle that avoids analogues of Durbin and Kalbfleisch's responses to Birnbaum's proof. It also argues that Joshi and

Mayo's claim that Birnbaum's proof is invalid rests on a misreading of his premises. Chapter 2 addresses three purported counterexamples to the Likelihood Principle and/or the Law of Likelihood that are not adequately addressed in the current literature. Chapter 3 discusses a new counterexample to the Law of Likelihood that exploits Borel's paradox and provides possible responses.

Chapter 4 develops the argument that the likelihoodist approach is inadequate because it merely characterizes data as evidence without thereby providing guidance about what one should believe or do. Chapter 5 argues against the idea that appeals to long-run performance favor the use of frequentist methods even though evidential considerations do not.

In the end, I argue that the evidential import of a body of data for a set of hypotheses depends only on the associated likelihood function, but that it is not possible to provide an adequate methodology for the post-data evaluation of scientific hypotheses on the basis of likelihood functions alone. It follows that it is not possible to provide an adequate methodology for the post-data evaluation of scientific hypotheses on the basis of evidential import alone. I argue further that considerations of long-run performance actually favor the use of methods that respect evidential equivalence understood in this way. Of course, the leading contender for a methodology of this kind that is not based on the evidential import of data alone is the Bayesian approach, which uses prior probabilities as well as likelihood functions. Thus, this dissertation provides support for the use of Bayesian methods in science.

I will now proceed to offer a new axiomatic proof of the Likelihood Principle that avoids influential objections to previous such proofs.

2.0 A NEW PROOF OF THE LIKELIHOOD PRINCIPLE

2.1 INTRODUCTION

Allan Birnbaum showed in his [1962] that the Likelihood Principle follows from the conjunction of the Sufficiency Principle and the Weak Conditionality Principle.[1,2] The Sufficiency Principle and the Weak Conditionality Principle are intuitively appealing, and some common frequentist practices appear to presuppose them. Yet frequentist methods violate the Likelihood Principle, while likelihoodist methods and Bayesian conditioning do not. As a result, many statisticians and philosophers have regarded Birnbaum's proof as a serious challenge to frequentist methods and a promising "sales tactic" for Bayesian methods.[3]

[1]Birnbaum calls the principles he uses simply the Sufficiency Principle and the Conditionality Principle. Dawid [1977] distinguishes between weak and strong versions of the Sufficiency Principle, but this distinction is of little interest: to my knowledge, no one accepts the Weak Sufficiency Principle but not the Strong Sufficiency Principle. The distinction between weak and strong versions of the Conditionality Principle (due to Basu [1975]) is of much greater interest: the Weak Conditionality Principle is much more intuitively obvious, and Kalbfleisch's influential response to Birnbaum's proof (discussed in Section 2.3) involves rejecting the Strong Conditionality Principle but not the Weak Conditionality Principle.

[2]The Conditionality Principle Birnbaum states in his [1962] is actually the Strong Conditionality Principle, but the proof he gives requires only the weak version. Birnbaum strengthens his proof in a later paper [1972] by showing that the logically weaker Mathematical Equivalence Principle can take the place of the Weak Sufficiency Principle. I present Birnbaum's proof using the Weak Sufficiency Principle rather than the Mathematical Equivalence Principle because the former is easier to understand and replacing it with the latter does not address any important objections to the proof.

[3]See for instance Birnbaum [1962], p. 272; Savage [1962], p. 307; Berger and Wolpert [1988], pp. 65–6; Mayo [1996], p. 391, fn. 17; and Grossman [2011a], p. 8, among many others. Birnbaum himself continued to favour frequentist methods even as he refined his proof of the Likelihood Principle [1970b]. He claims that the fact that frequentist principles conflict with the Likelihood Principle indicates that our concept of evidence is "anomalous" [1964]. I regard the frequentist principles that conflict with the Likelihood Principle (such the Weak Repeated Sampling Principle from Cox and Hinkley [1974], pp. 45–6) not as plausible constraints on the notion of evidence, but rather as articulating potential reasons to use frequentist methods despite the fact that they fail to track evidential import in accordance with the intuitions that lead to the Likelihood Principle. This view differs from Birnbaum's only in how it treats words like "evidence" and "evidential import," but this verbal change helps to clarify matters that Birnbaum obscures.

Most frequentist responses to Birnbaum's proof fall into one of three categories: (1) proposed restrictions on its premises, (2) allegations that it is fallacious, and (3) objections to the framework within which it is expressed. In this chapter I respond to objections in categories (1) and (2). In Section 2.2, I give a new proof of the Likelihood Principle that avoids responses in category (1). In Section 2.3, I explain that analogues of the minimal restrictions on Birnbaum's premises that are adequate to block his proof are not adequate to block the new proof. The responses in category (2) apply to this new proof as well, but I argue in Section 2.4 that those responses are mistaken. Arguments in category (3) have been less influential because standard frequentist theories presuppose the same framework that Birnbaum uses. For objections to theories that use a different framework, see Berger and Wolpert [1988], pp. 47–64.

I see little hope for a frequentist in trying to show that Birnbaum's proof and the new proof given here are unsound, but there is room to question the use of these proofs as objections to frequentist methods. The Likelihood Principle as Birnbaum (e.g. [1962], p. 271) and I formulate it says roughly that two experimental outcomes are evidentially equivalent if they have proportional likelihood functions. It does not say that one should use methods that would draw the same conclusions from either of two bodies of data that have proportional likelihood functions. This further claim requires a further argument, such as the one I provide in Chapter 6. The assumption that a method of inference should not be used if it can produce different outputs given evidentially equivalent inputs is intuitively plausible, but not trivially obvious. It is at best a slight oversimplification: frequentist methods are useful even if Bayesian methods are normative because frequentist methods often provide adequate, computationally efficient approximations to Bayesian methods. In addition, although idealized Bayesian conditioning conforms to the Likelihood Principle, the Bayesian methods that statisticians actually use in fact violate the Likelihood Principle as well, albeit in subtle ways that generally have little or no effect on the results of their analyses.[4] More importantly, the assumption presupposes an "evidentialist" epistemology

[4]I say that a method of inference *violates* a sufficient condition for evidential equivalence such as the Likelihood Principle if in some possible situation it would produce different outputs depending on which datum it receives among a set of data that the principle implies are evidentially equivalent without a difference in utilities, prior opinions, or background knowledge. I say that a method that does not violate a given condition of this kind *conforms to* it. In theory, subjective Bayesians conform to the Likelihood Principle

that some statisticians and philosophers reject. For instance, frequentist pioneers Jerzy Neyman and Egon Pearson claim that frequentist methods should be interpreted not in evidential terms but simply as decision rules warranted by their long-run performance (e.g. [1933], pp. 290–1).[5] The use of the Likelihood Principle as an objection to frequentist methods simply begs the question against this view. Many frequentists regard the Neyman-Pearson approach as too "behavioristic" for use in science (e.g. Fisher [1955]), but there are "conditional frequentist" approaches (initiated by Kiefer [1977]) that attempt to address this problem by ensuring that the measure of long-run performance used is relevant to the particular application in question. I do not claim that evidentialism is false or that conditional frequentism is viable, but only that those who would use the Likelihood Principle as an argument against frequentist methods need to account for such views.

Proofs of the Likelihood Principle have implications for the philosophy of statistics and perhaps for statistical practice even if they do not warrant the claim that frequentist methods should not be used. The Likelihood Principle does imply that evidential frequentism—the view that frequentist methods track the evidential import of data—is false.[6] This conclusion is relevant to debates internal to frequentism that plausibly hinge on whether frequentist methods should be understood as tracking evidential import, as decision rules justified by

by updating their belief in H upon learning E by the formula $P_{new}(H) = P_{old}(H|E) = \frac{P_{old}(H)P(E|H)}{\sum_i P(E|H_i)P_{old}(H_i)}$, where i ranges over an index set of the set of hypotheses under consideration. (The sum becomes an integral in the case of continuous hypothesis spaces.) Thus, $P_{new}(H)$ depends on E only through the likelihoods $P(E|H_i)$, as conforming to the Likelihood Principle requires. In practice, subjective Bayesians typically use methods that depend on the sampling distribution to estimate an expert's $P_{old}(H)$, such as methods that involve fitting a prior distribution that is conjugate to the sampling distribution. Objective Bayesians use priors that depend on the sampling distribution in order to achieve some aim such as maximizing a measure of the degree to which the posterior distribution depends on the data rather than the prior, as in the reference Bayesian approach (Berger [2006] p. 394). Some contemporary Bayesians (e.g. the authors of Gelman et al. [2003], pp. 157–96) also endorse model-checking procedures that violate the Likelihood Principle more drastically. It is worth noting that neither subjective nor objective Bayesians violate the Likelihood Principle in a different sense of "violates" than the one used here, even when checking their models: they do not allow information not contained in the likelihood of the observed data to influence the inferences they draw conditional on a model (Gelman [2012]). But they generally do allow the sampling distribution of the experiment (for instance, whether the experiment is binomial or negative binomial) to influence their choice of a model, and thereby potentially influence the conclusions they reach.

[5]Incidentally, there is evidence that Pearson was never fully committed to this view (Mayo [1992]).

[6]I take it that what makes a method frequentist is that it would provide some kind of guarantee about long-run performance in repeated applications to the same experiment with varying data, which requires that its outputs depend on the probabilities of unobserved sample points in violation of the Likelihood Principle. I also take it that a method that violates a true sufficient condition for evidential equivalence thereby fails to track evidential import.

their operating characteristics, or in some other way (Mayo [1985]). Topics of such debates include the use of accept/reject rules rather than p values, predesignation rules, stopping rules, randomized tests, and the use of conditional procedures.

A few technical notes are in order before proceeding. I assume that inferences are being performed in the context of a statistical model of the form $(\mathcal{X}, \Theta, \mathbf{P})$, where $\mathcal{X}$ is a finite sample space of (possibly vector-valued) points $\{\mathrm{x}\}$, Θ a possibly uncountable parameter space of (possibly vector-valued) points $\{\theta\}$, and $\mathbf{P}$ a family of probability distributions P_θ over $\mathcal{X}$ indexed by θ.[7] Against standard practice in the philosophy of science, I follow Birnbaum (e.g. [1962], pp. 269–70) and others writers in the literature about the Likelihood Principle in using the term "experiment" to refer to any data-generating process with such a model, even when that process is not manipulated.[8] I assume that $\mathbf{P}$ contains all of the hypotheses of interest. The model may also include a prior probability distribution over Θ and/or a utility/loss function defined on the Cartesian product of Θ and a set of possible outputs. Following Grossman [2011b], p. 561, I assume that the choice of experiments is not informative about θ.[9]

The assumption that sample spaces are finite restricts the scope of the proof given here, but this fact is not a serious problem for at least two reasons. First, infinite sample spaces are merely convenient idealizations. Real measuring devices have finite precision, and real measurable quantities are bounded even if our knowledge of their bounds is vague.[10] Second, the view that the Likelihood Principle holds for experiments with finite sample spaces but not for infinite sample spaces is implausible, unappealing, and insufficient to save evidential frequentism. Thus, a proof of the Likelihood Principle for experiments with finite sample

[7] Although I refer to Θ as a parameter space, it is really nothing more than an index set. $\mathbf{P}$ need not be a parametric family; that is, it need not be possible to characterize the elements of $\mathbf{P}$ by means of a relatively simple equation written in terms of θ.

[8] This broad use of the term "experiment" is not ideal, but there is no alternative that is obviously better. A nontechnical term such as "observational situation" fails to convey the presence of a statistical model. Grossman's term "merriment" ([2011a], p. 63) is not widely recognized.

[9] I do not follow Grossman ([2011b], p. 561) in assuming that utilities are either independent of the observation or unimportant. This restriction is needed when the Likelihood Principle is formulated as a claim about what kinds of methods should be used, but not when it is formulated as a sufficient condition for two outcomes to be evidentially equivalent.

[10] For instance, we might model the entry of customers into a bank as a Poisson process, but we would not take seriously the implication of this model that with positive probability ten billion customers will enter the bank in one second. The model neglects constraints such as the sizes of the bank and of the world population that become relevant only far out in the tails of the distribution.

spaces would be sufficient for present purposes even if there were actual experiments with truly infinite sample spaces. Moreover, it seems likely that the proof given here could be extended to experiments with continuous sample spaces, as Berger and Wolpert extend Birnbaum's proof ([1988], pp. 32–6). Attempting to extend the proof given here in this way would be an interesting technical exercise, but for the reasons just discussed it would not be either philosophically or practically illuminating.

Birnbaum's proof is more elegant than the proof given here, and its premises are easier to grasp. On the other hand, the new proof is safe against natural responses to Birnbaum's proof that have been influential. Thus, while Birnbaum's proof may have more initial persuasive appeal than the proof given here, the new proof is better able to withstand critical scrutiny.

2.2 THE NEW PROOF

I show in this section that the Likelihood Principle follows from the conjunction of the Experimental Conditionality Principle and what I call the Weak Ancillary Realizability Principle. In the next section I display the advantages this proof has over Birnbaum's.

Both the Experimental Conditionality Principle and the Weak Ancillary Realizability Principle appeal to the notion of a *mixture experiment*. A mixture experiment consists of using a random process to select one of a set of component experiments and then performing the selected experiment, where the component experiments share a common index set Θ that is independent of the selection process. For instance, one might flip a coin to decide which of two thermometers to use for a measurement that will be used to test a particular hypothesis, and then perform that measurement and the associated test. A non-mixture experiment is called *minimal*. A mixture experiment with two minimal, equiprobable components is called *simple*.

Roughly speaking, the Experimental Conditionality Principle says that the outcome of a mixture experiment is evidentially equivalent to the corresponding outcome of the component experiment actually performed. The Weak Ancillary Realizability Principle says that the outcome of a minimal experiment is evidentially equivalent to the corresponding outcome of

a two-stage mixture experiment with an isomorphic sampling distribution.

The Experimental Conditionality Principle can be expressed more formally as follows:

> **The Experimental Conditionality Principle.** For any outcome x of any component E' of any mixture experiment E, $\mathrm{Ev}(E, (E', x)) = \mathrm{Ev}(E', x)$.

where "$\mathrm{Ev}(E, x)$" refers to the "evidential import" of outcome x of experiment E, and "$(E, (E', x))$" refers to outcome x of component E' of mixture experiment E. "Evidential import" is an undefined primitive notion that principles like those discussed in this chapter are intended partially to explicate. In words, the Experimental Conditionality Principle says that the evidential import of the outcome of an experiment does not depend on whether that experiment is performed by itself or as part of a mixture.

The Experimental Conditionality Principle has considerable intuitive appeal: denying it means accepting that the appropriate evidential interpretation of an experiment can depend on whether another experiment that was not performed had a chance of being performed. If this claim does not seem odd in the abstract, then consider it in the case of the following example:

> **Example 2.1.** Suppose you work in a laboratory that contains three thermometers, T_1, T_2, and T_3. All three thermometers produce measurements that are normally distributed about the true temperature being measured. The variance of T_1's measurements is equal to that of T_2's but much smaller than that of T_3's. T_1 belongs to your colleague John, so he always gets to use it. T_2 and T_3 are common lab property, so there are frequent disputes over the use of T_2. One day, you and another colleague both want to use T_2, so you toss a fair coin to decide who gets it. You win the toss and take T_2. That day, you and John happen to be performing identical experiments that involve testing whether the temperature of your respective indistinguishable samples of some substance is greater than 0°C or not. John uses T_1 to measure his sample and finds that his result is just statistically significantly different from 0°. John celebrates and begins making plans to publish his result. You use T_2 to measure your sample and happen to measure exactly the same value as John. You celebrate as well and begin to think about how you can beat John to publication. "Not so fast," John says. "Your experiment was different from mine. I was bound to use T_1 all along, whereas you had only a 50% chance of using T_2. You need to include that fact in your calculations. When you do, you'll find that your result is no longer significant."

According to radically "behavioristic" forms of frequentism, John may be correct. You performed a mixture experiment by flipping a coin to decide which of two thermometers to

use, and thus which of two component experiments to perform. The uniformly most powerful level α test[11] for that mixture experiment does *not* consist of performing the uniformly most powerful level α test for whichever component experiment is actually performed. Instead, it involves accepting probability of Type I error greater than α when T_3 is used in exchange for a probability of Type I error less than α when T_2 is used, in such a way that the probability of Type I error for the mixture experiment as a whole remains α (see Cox [1958], p. 360).

Most statisticians, including most frequentists, reject this line of reasoning. It seems suspicious for at least three reasons. First, the claim that your measurement warrants different conclusions from John's seems bizarre. They are numerically identical measurements from indistinguishable samples of the same substance made using measuring instruments with the same stochastic properties. The only difference between your procedures is that John was "bound" to use the thermometer he used, whereas you had a 50% chance of using a less precise thermometer. It seems odd to claim that the fact that you could have used a instrument other than the one you actually used is relevant to the interpretation of the measurement you actually got using the instrument you actually used. Second, the claim that John was "bound" to use T_1 warrants scrutiny. Suppose that he had won that thermometer on a bet he made ten years ago that he had a 50% chance of winning, and that if he hadn't won that bet, he would have been using T_3 for his measurements. According to his own reasoning, this fact would mean that his result is not statistically significant after all.[12] The implication that one might have to take into account a bet made ten years ago that has nothing to do with the system of interest to analyze John's experiment is hard to swallow. In fact, this problem is much deeper than the fanciful example of John winning the thermometer in a bet would suggest. If John's use of T_1 as opposed to some other thermometer with different stochastic properties was a nontrivial result of *any* random process at any point in the past

[11] A uniformly most powerful test of significance level α is a test that maximises the probability of rejecting the null hypothesis under each simple component of the alternative hypothesis among all tests that would reject the null hypothesis with probability no greater than α if the null hypothesis were true.

[12] One could argue that events like outcomes of coin tosses are determined by the laws of physics and relevant initial conditions anyway, so that both you and John were bound to use the thermometer you actually did use. However, applying the same argument to any appeal to randomness that does not arise from genuine indeterminism would undermine frequentist justifications based on sampling distributions in most cases. In addition, this argument could be avoided by replacing the coin toss in the example with a truly indeterministic process such as (let us suppose) radioactive decay.

that was independent of the temperature being measured, then the denial of Weak Conditionality Principle as applied to this example implies that John analyzed his data using a procedure that fails to track evidential import.[13] Third, at the time of your analysis you *know* which thermometer you received. How could it be better epistemically to fail to take that knowledge into account?

It is sometimes said (e.g. Wasserman [2012]) that an argument for the Weak Conditionality Principle like the one just given involves a hasty generalization from a single example. However, the purpose of the example is merely to make vivid the intuition that features of experiments that could have been but were not performed are irrelevant to the evidential import of the outcome of the experiment that actually was performed. The intuition the example evokes, rather than the example itself, justifies the principle.

The Experimental Conditionality Principle does go beyond the example just discussed in that it applies to mixture experiments with arbitrary probability distributions over arbitrarily (finitely) many components. The intuition the example evokes has nothing to do with the number of component experiments in the mixture or the probability distribution over those experiments, so this extension is innocuous. Also, the Experimental Conditionality Principle does not require that the component experiment actually performed be minimal. Thus, it can be applied to a nested mixture experiment, and by mathematical induction it implies that the outcome of a nested mixture experiment with any finite number of "stages" is evidentially equivalent to the corresponding outcome of the minimal experiment actually performed. There does not seem to be any reason to baulk at applying the principle repeatedly to nested mixture experiments in this way. Thus, the Experimental Conditionality Principle is not a hasty generalization because insofar as it generalizes beyond the example just discussed it does so in an unobjectionable way.

It is sometimes useful to express conditionality principles in terms of ancillary statistics. An ancillary statistic for an experiment is a statistic that has the same distribution under each of the hypotheses under consideration. A statistic that indexes the outcomes of a random

[13]This kind of reasoning makes plausible the claim that most if not all real experiments are components of mixture experiments that we cannot hope to identify, much less model, and thus that assuming something like the Experimental Conditionality Principle is necessary for performing any data analysis at all (Kalbfleisch [1975], p. 254).

process used to decide which component of a mixture experiment to perform is ancillary for that mixture experiment. Other ancillary statistics arise not from the mixture structure of the experiment but from the set of hypotheses under consideration. Kalbfleisch [1975] calls a statistic that indexes the outcomes of an overt random process used to decide which component experiment to perform *experimental* ancillaries, and a statistic that is ancillary only because of features of the set of hypotheses under consideration a *mathematical* ancillary. The Experimental Conditionality Principle permits conditioning on experimental ancillaries but does not address conditioning on mathematical ancillaries. The distinction between experimental and mathematical ancillaries is extra-mathematical in the sense that it is not given by the model $(\mathcal{X}, \Theta, \mathbf{P})$. As a result, it is difficult if not impossible to formulate that distinction precisely. This fact is not an objection to my proof in the relevant dialectical context because I use the distinction between experimental and mathematical ancillaries only to address a response to Birnbaum's proof due to Kalbfleisch [1975] that requires it. (See Section 2.3).

The Weak Ancillary Realizability Principle essentially says that one may replace one binary mathematical ancillary in a minimal experiment with an experimental ancillary without changing the evidential imports of the outcomes. The following reasoning helps motivate this principle. Consider a hypothetical experiment with the sampling distribution given by Table 2.2 below, where the cells of the table correspond to a partition of the sample space.

Table 2.1: Sampling distribution of hypothetical experiment used to motivate Weak Ancillary Realizability

	v_1	v_2	v_3	v_4
h	$P_\theta(x_1)$	$P_\theta(x_2)$	$P_\theta(x_3)$	$P_\theta(x_4)$
t	$P_\theta(x_5)$	$P_\theta(x_6)$	$P_\theta(x_7)$	$P_\theta(x_8)$

For all $i = 1, \ldots, 8$, $P_\theta(x_i)$ is known only as a function of θ. However, it is known that $P_\theta(h) = P_\theta(v_1) = \frac{1}{2}$. Such a sampling distribution could arise by many processes, such as (1) the roll of an appropriately weighted eight-sided die with faces labelled $x_1, \ldots, x_8$; (2) the flip of a fair coin with sides labelled h and t followed by the roll of either an appropriately weighted

four-sided die with faces labelled $x_1, \ldots, x_4$ (if the coin lands heads) or an appropriately weighted four-sided die with faces labelled $x_5, \ldots, x_8$ (if the coin lands tails); or (3) the flip of a fair coin with sides labelled v_1, $\sim v_1$ followed by either the flip of an appropriately biased coin with sides labelled x_1 and x_5 (if the first coin lands on side v_1) or the roll of an appropriately weighted six-sided die with faces labelled $x_2, x_3, x_4, x_6, x_7, x_8$ (if the first coin lands tails). The intuition that the Weak Ancillary Realizability Principle aims to capture is that the outcome with likelihood function $P_\theta(x_1)$ has the same evidential import with respect to θ regardless of which of these kinds of process produces it: it makes no difference evidentially whether such an outcome arises from a one-stage process, a two-stage process in which the row is selected and then the column, or a two-stage process in which either the first column or its complement is selected and then the cell, provided that the overall sampling distribution is the same in the three cases. It is worth taking a moment to be sure that one has grasped this somewhat complicated example and satisfied oneself that the intuition I claim it evokes is indeed rather compelling. The key step in the proof of the Likelihood Principle given below is the construction of a hypothetical experiment that has the same essential features as the experiment just described.

The formal statement of the Weak Ancillary Realizability Principle uses the following terminology and notation. For a given set of probability distributions $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ with support $\mathcal{X}$ and a given set $A \subset \mathcal{X}$, let ${}^A\mathbf{P}$ refer to the set of distributions obtained by replacing $P_\theta(X) \in \mathbf{P}$ with ${}^AP_\theta(X) = P_\theta(X|X \in A)$ for each $\theta \in \Theta$ and $X \in \mathcal{X}$, and let $A^C = \mathcal{X} \setminus A$. Call experimental models $(\mathcal{X}, \Theta, \{P^1_\theta\})$ and $(\mathcal{Y}, \Theta, \{P^2_\theta\})$ isomorphic under the one-to-one mapping $f : \mathcal{X} \to \mathcal{Y}$ if and only if $P^1_\theta(x) = P^2_\theta(f(x))$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta$. The Weak Ancillary Realizability Principle can now be stated as follows.

> **The Weak Ancillary Realizability Principle.** Let E be a minimal experiment with model $(\mathcal{X}, \Theta, \mathbf{P})$ such that there is a set $A \subset \mathcal{X}$ such that for all $\theta \in \Theta$, $\Pr_\theta(X \in A) = p$ for some known $0 \leq p \leq 1$. Let E' consist of flipping a coin with known bias p for heads to decide between performing component experiment E_1 if the coin lands heads and E_2 if the coin lands tails, where E_1 is a minimal experiment with model isomorphic to $(A, \Theta, {}^A\mathbf{P})$ under f_1, and E_2 is a minimal experiment with model isomorphic to $(A^C, \Theta, {}^{A^C}\mathbf{P})$ under f_2. Under these conditions $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', (E_1, f_1(x)))$ for all $x \in A$, and $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', (E_2, f_2(x)))$ for all $x \in A^C$.

In words, if a minimal experiment's sample space can be partitioned into two sets of outcomes

A and A^C such that the probability that the observed outcome is in A is known to be p, then the outcome of that experiment has the same evidential import as the corresponding outcome of an experiment that consists of first flipping a coin with bias p for heads and then performing a minimal experiment that is isomorphic to a minimal experiment over A if the coin lands heads and a minimal experiment that is isomorphic to a minimal experiment over A^C if it lands tails. Roughly speaking, this principle allows one to break a minimal experiment into two stages by turning a mathematical ancillary into an experimental ancillary.

The Weak Ancillary Realizability Principle does not require that the outcomes of the components E_1 and E_2 of E' be literally the same as the corresponding outcomes of E. An alternative approach to the one I have taken here would be to include this requirement and to adopt in addition a weak version of Dawid's Distribution Principle ([1977], p. 247) which says that corresponding outcomes of minimal experiments with isomorphic models are evidentially equivalent. I have chosen not to take this approach in order to make it easier to compare my proof to Birnbaum's, but considering it is instructive. The Distribution Principle this approach requires is weaker than Dawid's in that it only applies to minimal experiments and thus is compatible with Kalbfleisch's response to Birnbaum's proof that is discussed in the next section. Adding to the Weak Ancillary Realizability Principle the requirement that the outcomes of the components E_1 and E_2 of E' be literally the same as the corresponding outcomes of E makes it possible to argue for the Weak Ancillary Realizability Principle as follows. It is always possible to form an appropriate E' for a given E provided that E can be repeated indefinitely many times. Simply flip a coin with probability p for heads. If the coin lands heads, repeat E until an outcome in A occurs, and then report that outcome. If the coin lands tails, repeat E until an outcome in A^C occurs, and then report that outcome. The claim that corresponding outcomes of some E and the E' formed from it in this way are evidentially equivalent is highly intuitive. The fact that the E' outcome occurred after some unspecified number of unspecified outcomes in the sample space of the component experiment not selected is uninformative because one already knows the probability of such outcomes (p or $1 - p$).

The Weak Ancillary Realizability Principle is different from what one might call the Strong Ancillary Realizability Principle, which says across the board that the distinction

between experimental and mathematical ancillaries is irrelevant to evidential import. In contrast, the Weak Ancillary Realizability Principle permits conditioning only on a *single*, *binary* ancillary (the indicator for A) from a *minimal* experiment. The proof given here would be redundant otherwise: the conjunction of the Strong Ancillary Realizability Principle and the Experimental Conditionality Principle obviously implies a Strong Conditionality Principle that permits conditioning on any ancillary, which has already been shown to imply the Likelihood Principle (Evans et al. [1986]).

The Strong Ancillary Realizability Principle may be true and does have some intuitive appeal, but the Weak Ancillary Realizability Principle is far easier to defend because one can construct a single simple illustration in which the principle seems obviously compelling that essentially covers all of the cases to which the principle applies. To demonstrate this point, I will start with an illustration that is a bit too simple and then add the necessary additional structure (see Figure 2.1). These illustrations make the same point as the example given in conjunction with Table 2.2 above, but their concreteness makes them more vivid and thus perhaps more convincing.

Figure 2.1: Evidentially equivalent outcomes from the pair of procedures described in Example 2.2.[15]

Example 2.2. Consider an ideal spinner divided into regions R_1, R_2, S_1, and S_2 such that one knows that R_1 and R_2 together occupy half of the spinner's area. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following two procedures:

One-Stage Procedure. Spin the spinner and report the result.

Two-Stage Procedure. Flip a coin with bias $\frac{1}{2}$ for heads. If the coin lands heads, replace the spinner with one divided into two regions R_1^* and R_2^* such that for $i = 1$ or 2, R_i^* occupies the same fraction of the new spinner that R_i occupies of the half of the original spinner that R_1 and R_2 occupy together. If this spinner lands on R_i^*, report R_i as the result. If the coin lands tails, do likewise with the S regions.

Intuitively, knowing whether the one-stage or the two-stage procedure was performed would not help in drawing inferences about the relative sizes of the spinner regions from one's data. The difference between these two procedures does not matter for such inferences. Each procedure generates the same sampling distribution; the fact that one does so in one step while the other does so in two steps is irrelevant.

In Example 2.2, the fraction of the spinner that R_1 and R_2 are known to occupy and the bias of the coin used in the two-stage procedure are $\frac{1}{2}$. But the intuition that the example evokes has nothing to do with the fact that this number is $\frac{1}{2}$. Nor does it have anything to do with the fact that there are two R regions and two S regions. Thus, we can safely extend this intuition to the following more general example (see Figure 2.2).

Example 2.3. Consider an ideal spinner divided into regions R_1, R_2, ..., R_n and S_1, S_2, ..., S_m such that one knows that R_1, R_2, ..., R_n together occupy proportion p of the spinner for some particular $0 \leq p \leq 1$. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following two procedures:

One-Stage Procedure. Spin the spinner and report the result.

Two-Stage Procedure. Flip a coin with bias p for heads. If the coin lands heads, replace the spinner with one divided into n regions R_1^*, R_2^*, ..., R_n^* such that for $i = 1, \ldots, n$, R_i^* occupies the same fraction of the new spinner that R_i occupies of the

[15] Quarter clipart courtesy FCIT: Portrait on a Quarter, retrieved February 2, 2013 from <etc.usf.edu/clipart/40200/40232/quart_front_40232.htm>.

Figure 2.2: Evidentially equivalent outcomes from the pair of procedures described in Example 2.3. The bias of the coin for heads p is the same as the fraction of the first spinner that the R regions occupy.

> percentage p of the original spinner that $R_1, \ldots, R_n$ occupy together. If this spinner lands on R_i^*, report R_i as the result. If the coin lands tails, do likewise with the S regions.

Again, it seems obvious that knowing whether the one-stage or the two-stage procedure was performed would not help in drawing inferences about the relative sizes of the spinner regions from one's data. As long as the sampling distribution remains unchanged, whether an experiment is performed in one step or two is irrelevant to the evidential imports of its outcomes.

The Weak Ancillary Realizability Principle does extend the intuition Example 2.3 evokes to experiments involving data-generating mechanisms that are not spinners. There is nothing special about spinners driving that intuition, so this extension is innocuous.

There is one fact about the Weak Ancillary Realizability Principle that is important to the proof of the Likelihood Principle which Example 2.2 does not illustrate, namely that it can apply to a single experiment in more than one way. Call a partition of a sample space that is indexed by an ancillary statistic (such as $\{A, A^c\}$ in the statement of the Weak Ancillary Realizability Principle) an *ancillary partition*. An experiment can have multiple

ancillary partitions to which the Weak Ancillary Realizability Principle applies. The proof of the Likelihood Principle presented below involves applying the Weak Ancillary Realizability Principle to two ancillary partitions of the same experiment, so it is worth considering an example of an experiment of this kind in order to confirm that it does not violate our intuitions (See Figure 2.3).

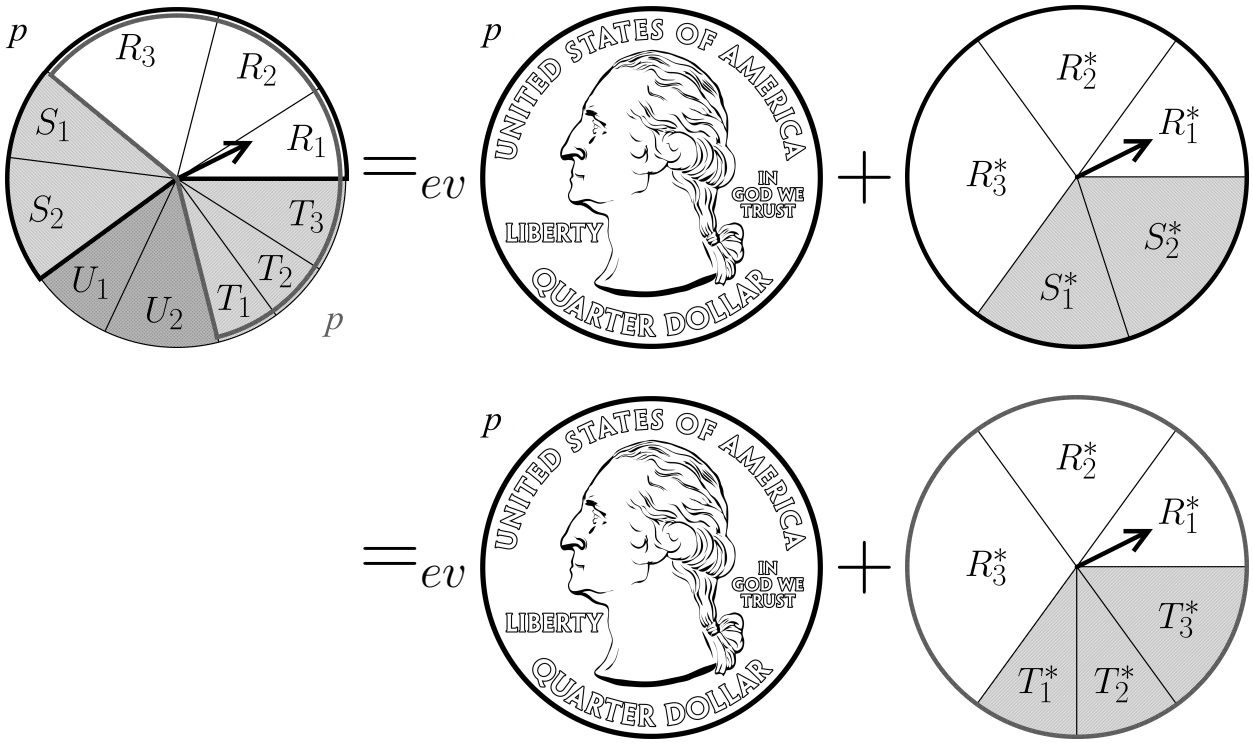

Figure 2.3: Evidentially equivalent outcomes from the pair of procedures described in Example 2.4. The bias of the coin for heads p is the same as the fraction of the first spinner that the R and S regions occupy in total, which is the same as the fraction that the R and T regions occupy.

Example 2.4. Consider an ideal spinner divided into regions R_1, R_2, ..., R_n; S_1, S_2, ..., S_m; T_1, T_2, ..., T_l; and U_1, U_2, ..., U_k, such that for some $0 \leq p \leq 1$ one knows both that the R regions and the S regions together occupy proportion p of the spinner and that the R regions and the T regions together occupy proportion p of the spinner. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following three procedures:

One-Stage Procedure. Spin the spinner and report the result.

Two-Stage Procedure A. Flip a coin with bias p for heads. If the coin lands heads, replace the spinner with one divided into $n + m$ regions R_1^*, R_2^*, ..., R_n^*, S_1^*, S_2^*, ..., S_m^* such that for all $i = 1, \ldots, n$, R_i^* occupies the same fraction of the new spinner that R_i occupies of the percentage p of the original spinner that $R_1, \ldots, R_n$ and $S_1, \ldots, S_m$ occupy together, and likewise for S_i^* and S_i for all $i = 1, \ldots, m$. If this spinner lands on R_i^* or S_i^*, report R_i or S_i as the result, respectively. If the coin lands tails, do likewise with the T and U regions.

Two-Stage Procedure B. Perform Two-Stage Procedure A but reverse the roles of S and T.

The intuition that corresponding outcomes of the One-Stage Procedure and of Two-Stage Procedure A are evidentially equivalent seems to be completely unaffected by the fact that one could also perform Two-Stage Procedure B, and vice versa. Thus, there is no reason not to apply the Weak Ancillary Realizability Principle twice to experiments with two ancillary partitions.

The Likelihood Principle can be expressed formally as follows:

The Likelihood Principle. Let E_1 and E_2 be experiments with a common parameter space Θ, and let x and y be outcomes of E_1 and E_2, respectively, such that $P_\theta(x) = cP_\theta(y)$ for all $\theta \in \Theta$ and some positive c that is constant in θ. Then $\mathrm{Ev}(E_1, x) = \mathrm{Ev}(E_2, y)$.

In words, two experimental outcomes with proportional likelihood functions for the same parameter are evidentially equivalent.

I can now prove the following result:

Theorem 1. *The Experimental Conditionality Principle and the Weak Ancillary Realizability Principle jointly entail the Likelihood Principle.*

Proof. Consider an arbitrary pair of experiments E_1 and E_2 with respective models $(\mathcal{X}, \Theta, \{P_\theta^1\})$ and $(\mathcal{Y}, \Theta, \{P_\theta^2\})$ such that $\mathcal{X} = \{x_0, x_1, \ldots, x_n\}$, $\mathcal{Y} = \{y_0, y_1, \ldots, y_m\}$, and $P_\theta^1(x_0) = cP_\theta^2(y_0)$ for all $\theta \in \Theta$ and some $c \geq 1$ that is constant in θ. There is no loss of generality in the assumption $c \geq 1$ because one can simply swap the labels of E_1, E_2, and their outcomes if $P_\theta^1(x_0) < P_\theta^2(y_0)$. x_0 is the outcome $x_0^\dagger$ of some unique minimal

experiment $E_1^\dagger$ with sample space $\mathcal{X}^\dagger$ that is performed with some known probability q when E_1 is performed.[16] $E_1^\dagger$ is either E_1 itself or a proper component of E_1.[17] x_0 just is $(E_1^\dagger, x_0^\dagger)$, so by the reflexivity of the evidential equivalence relation $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E_1, (E_1^\dagger, x_0^\dagger))$.[18] By the Experimental Conditionality Principle, $\mathrm{Ev}(E_1, (E_1^\dagger, x_0^\dagger)) = \mathrm{Ev}(E_1^\dagger, x_0^\dagger)$.[19]

Construct a hypothetical minimal experiment E_1^{CE} with sample space $\mathcal{X}^{CE}$ and sampling distribution given by Table 2.2[20]. Although E_1^{CE} is minimal, I trust that no confusion will result from the use of expressions of the form (d, z_i) and (e, z_i) to refer to points in $\mathcal{X}^{CE}$ in accordance with Table 2.2. The arrangement of sample points into rows and columns in Table 2.2 only serves to display the relevant (mathematical) ancillary partitions of $\mathcal{X}^{CE}$. The outcomes in the first row that correspond to outcomes of $E_1^\dagger$ (that is, $\{(d, z_i) : \exists(x^\dagger \in \mathcal{X}^\dagger)(x_i = (E_1^\dagger, x^\dagger)\}$) constitute a set $A \subset \mathcal{X}$ such that $\Pr(X \in A) = p$ for all $X \in \mathcal{X}$ and some known $0 \leq p \leq 1$, namely $\frac{q}{2}$. Likewise, the outcomes in the first column (that is, $\{(d, z_0), (e, z_0)\}$ constitute a set $A \subset \mathcal{X}$ such that $\Pr(X \in A) = p$ for all $X \in \mathcal{X}$ and some known $0 \leq p \leq 1$, namely $\frac{1}{2}$.

Table 2.2: Sampling distribution of E_1^{CE}

	z_0	z_1	z_2	z_3	$\ldots$	z_n
d	$\frac{1}{2}P_\theta^1(x_0)$	$\frac{1}{2}P_\theta^1(x_1)$	$\frac{1}{2}P_\theta^1(x_2)$	$\frac{1}{2}P_\theta^1(x_3)$	$\ldots$	$\frac{1}{2}P_\theta^1(x_n)$
e	$\frac{1}{2} - \frac{1}{2}P_\theta^1(x_0)$	$\frac{1}{2}P_\theta^1(x_0) - \frac{1}{2}\min_\theta P_\theta^1(x_0)$	$\frac{1}{2}\min_\theta P_\theta^1(x_0)$	0	$\ldots$	0

Let E_1^M be an experiment that consists of flipping a coin with bias $\frac{q}{2}$ for heads to choose between performing $E_1^\dagger$ if the coin lands heads and performing a minimal experiment with sampling distribution given by the distribution of E_1^{CE} conditional on the complement of the set of outcomes that correspond to outcomes of $E_1^\dagger$ if the coin lands tails. By the Weak Ancillary Realizability Principle, $\mathrm{Ev}(E_1^{CE}, (d, z_0)) = \mathrm{Ev}(E_1^M, (E_1^\dagger, x_0^\dagger))$. By the Experimental Conditionality Principle, $\mathrm{Ev}(E_1^M, (E_1^\dagger, x_0^\dagger)) = \mathrm{Ev}(E_1^\dagger, x_0^\dagger)$. From all of the equivalences established so far it follows that $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E_1^{CE}, (d, z_0))$.

Next construct a hypothetical Bernoulli experiment E^B with sample space (g, h) and sampling distribution given by $P_\theta^B(g) = P_\theta^1(x_0)$. Finally, construct a mixture experiment $E_1^{M_B}$ that consists of first flipping a coin with bias $\frac{1}{2}$ for heads to decide between performing E^B and performing a minimal experiment with the known sampling distribution given by the distribution of E_1^{CE} conditional on the complement of the first-column outcomes $\{(d, z_0), (e, z_0)\}$. By the Weak Ancillary Realizability Principle, $\mathrm{Ev}(E_1^{CE}, (d, z_0)) =$

[16] $E_1^\dagger$ is unique because experimental ancillaries involve overt randomisation. Thus, the problem of the nonuniqueness of maximal ancillaries that plagues frequentist attempts to incorporate conditioning on ancillary statistics in general (see Basu [1964] and subsequent discussion) does not arise. That q is known follows from the fact that the model specifies P_θ^1 for each $\theta \in \Theta\}$ and the stipulation that the process by which the component of a mixture experiment to be performed is selected is independent of θ.

[17] I am treating the "component" relation for experiments as transitive.

[18] Evidential equivalence is assumed to be an equivalence relation.

[19] When E_1 is minimal, $E_1^\dagger = E_1$ and $x_0^\dagger = x_0$, so $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E_1^\dagger, x_0^\dagger)$ by reflexivity alone.

[20] The construction used in this table is a modified version of the construction Evans et al. use to show that the Likelihood Principle follows from the Strong Conditionality Principle alone ([1986], p. 188).

$\mathrm{Ev}(E_1^{M_B}, (E^B, g))$. By the Experimental Conditionality Principle, $\mathrm{Ev}(E_1^{M_B}, (E^B, g))$= $\mathrm{Ev}(E^B, g)$. It follows that $\mathrm{Ev}(E_1^{CE}, (d, z_0))$= $\mathrm{Ev}(E^B, g)$.

From $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E_1^{CE}, (d, z_0))$ and $\mathrm{Ev}(E_1^{CE}, (d, z_0)) = \mathrm{Ev}(E^B, g)$, it follows that $\mathrm{Ev}(E_1, x_0)$= $\mathrm{Ev}(E^B, g)$. An analogous construction establishes $\mathrm{Ev}(E_2, y_0) = \mathrm{Ev}(E^B, g)$, and thus $\mathrm{Ev}(E_1, x_0)$= $\mathrm{Ev}(E_2, y_0)$. (See the appendix for the details of this construction.) The only restriction we placed on (E_1, x_0) and (E_2, y_0) in establishing this result is that they have proportional likelihood functions, so the Likelihood Principle follows by universal generalization. □

Figure 2.4 provides a graphical depiction of this proof.

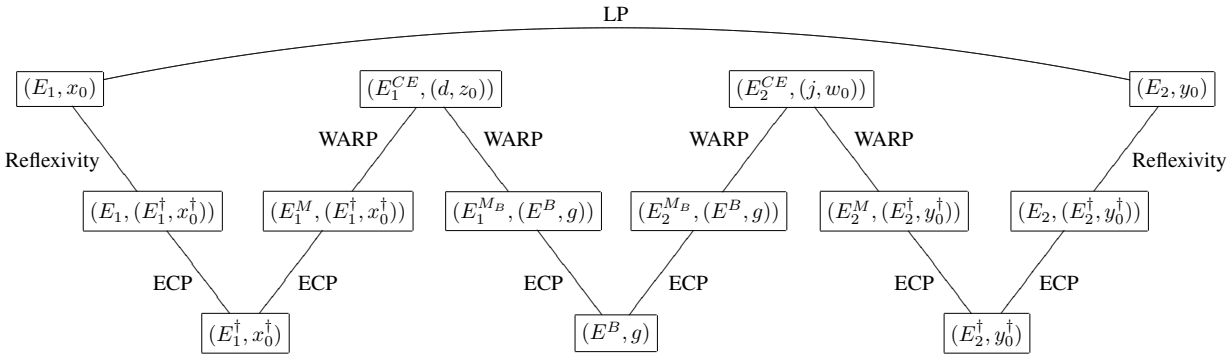

Figure 2.4: Graphical depiction of the series of equivalences used to establish Theorem 1. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (ECP = Experimental Conditionality Principle, WARP = Weak Ancillary Realizability Principle, LP = Likelihood Principle).

2.3 HOW THE NEW PROOF ADDRESSES PROPOSALS TO RESTRICT BIRNBAUM'S PREMISES

Birnbaum [1962] shows that the Likelihood Principle follows from the conjunction of the Sufficiency Principle and the Weak Conditionality Principle. Durbin [1970] responds to Birnbaum's proof by restricting the Weak Conditionality Principle, while Kalbfleisch [1975]

responds by restricting the the Sufficiency Principle.[21] Analogous restrictions on the premises of the proof given in the previous section do not suffice to save evidential frequentism.

I will briefly present Birnbaum's proof, and then explain how the new proof addresses Durbin's and Kalbfleisch's responses to it. The Weak Conditionality Principle is the Experimental Conditionality Principle restricted to simple mixture experiments. It can be stated as follows:

> **The Weak Conditionality Principle**: For any outcome x of any component E' of any simple mixture experiment E, $\mathrm{Ev}(E, (E', x)) = \mathrm{Ev}(E', x)$.

This principle is logically weaker than the Experimental Conditionality Principle, but there does not seem to be any reason to accept the former but not the latter.

The Sufficiency Principle says that two experimental outcomes that give the same value of a sufficient statistic are evidentially equivalent. The notion of a *sufficient statistic* is intended to explicate the informal idea of a statistic that simplifies the full data without losing any of the information it provides about the model. Formally, a statistic is called sufficient with respect to the hypotheses of interest if and only if it takes the same value for a set of outcomes only if the probability distribution over those outcomes given that one of them occurs does not depend on which of those hypotheses is true. The Sufficiency Principle seems eminently plausible: if the probability distribution over a set of outcomes given that one of them occurs does not depend on which hypothesis is true, then it is not clear how those outcomes could support different conclusions about those hypotheses. The Weak Sufficiency Principle can be stated formally as follows:

> **The Sufficiency Principle (S)**: Consider an experiment $E = (\mathcal{X}, \{\theta\}, \mathbf{P})$ where T(X) is a sufficient statistic for θ. For any $x_1, x_2 \in \mathcal{X}$, if $T(x_1) = T(x_2)$ then $\mathrm{Ev}(E, x_1) = \mathrm{Ev}(E, x_2)$.

The Sufficiency Principle would underwrite the practice, common among frequentists as well as advocates of other statistical paradigms, of "reducing to a sufficient statistic," that is, reporting only the value of a sufficient statistic rather than reporting the full data. For instance, from a sequence of a fixed number of coin tosses that are assumed to be independent

[21] Cox and Hinkley suggest a response similar to Kalbfleisch's in their earlier [1974], but they do not develop the idea as fully as Kalbfleisch.

and identically distributed with probability p of heads, a frequentist would typically report only the number of heads in the sequence (a sufficient statistic for p) rather than the sequence itself.

One can also formalize the notion of a *minimal sufficient* statistic, which retains all of the information about the model that is in the full data but cannot be simplified further without discarding some such information. Formally, a statistic is minimal sufficient for θ in a given experiment if and only if it is sufficient for θ and is a function of every sufficient statistic for θ in that experiment. A minimal sufficient statistic is more coarse-grained than any non-minimal sufficient statistic for the same parameter and experiment, so it provides the greatest simplification of the data that the Sufficiency Principle underwrites.

Minimal sufficient statistics are unique up to one-to-one transformation, and a statistic that assigns the same value to a pair of outcomes if and only if they have the same likelihood function is minimal sufficient (Cox and Hinkley [1974], p. 24). Thus, the Sufficiency Principle implies that two outcomes of the same experiment are evidentially equivalent if they have the same likelihood function. This consequence of the Sufficiency Principle is sometimes called the Weak Likelihood Principle (Cox and Hinkley [1974], p. 24); it differs from the Likelihood Principle only in that the latter applies also to outcomes from different experiments. This difference may seem slight, but the Likelihood Principle has implications that are radical from a frequentist perspective, such as the evidential irrelevance of stopping rules, that the Weak Likelihood Principle lacks.

Proving the Likelihood Principle from the Sufficiency Principle requires an additional principle that allows one to "bridge" different experiments. The Weak Conditionality Principle plays this role in Birnbaum's proof: one simply constructs a hypothetical mixture of the two experiments in question. The proof proceeds as follows. Take an arbitrary pair of experimental outcomes (E_1, x) and (E_2, y) that have the same likelihood function for the same parameter θ. Construct a simple mixture E^M of E_1 and E_2. $(E^M, (E_1, x_0))$ and $(E^M, (E_2, y_0))$ are two outcomes of the same experiment that have the same likelihood function, so a minimal sufficient statistic for E^M has the same value for those two outcomes. By the Sufficiency Principle, then, $\mathrm{Ev}(E^M, (E_1, x_0)) = \mathrm{Ev}(E^M, (E_2, y_0))$. By the Conditionality Principle, $\mathrm{Ev}(E^M, (E_1, x_0)) = \mathrm{Ev}(E_1, x_0)$ and $\mathrm{Ev}(E^M, (E_2, y_0)) = \mathrm{Ev}(E_2, y_0)$. It follows

that $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E_2, y_0)$. (E_1, x_0) and (E_2, y_0) are arbitrary except for the fact that they have the same likelihood function, so the Likelihood Principle follows by a universal generalization. Figure 2.5 displays the steps of this proof in a graphical format.

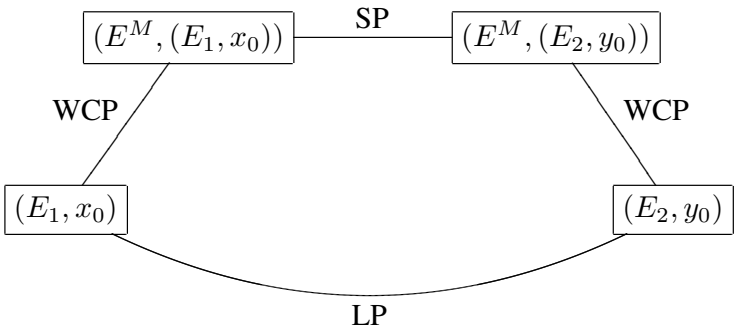

Figure 2.5: Birnbaum's proof of the Likelihood Principle. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (WCP = Weak Conditionality Principle, SP = Sufficiency Principle, LP = Likelihood Principle).

Having presented Birnbaum's proof, I will now discuss Durbin's and Kalbfleisch's responses to it in turn. Durbin [1970] proposes restrict the Weak Conditionality Principle to experimental ancillaries that are functions of a minimal sufficient statistic. This restriction suffices to block Birnbaum's proof because the outcomes (E_1, x_0) and (E_2, y_0) of Birnbaum's mixture experiment E^M share the same value of a minimal sufficient statistic. Thus, one value of a minimal sufficient statistic for E^M corresponds to two different values of the experimental ancillary that indexes the outcomes of the process used to select which of E_1 and E_2 to perform. It follows that the ancillary statistic that indexes the outcome of the coin flip used to decide which component experiment to perform is not a function of a minimal sufficient statistic, and thus that Durbin's restricted Weak Conditionality Principle does not underwrite conditioning on it.

Durbin's proposal faces many strong objections (see e.g. Birnbaum [1970a]; Savage [1970]; and Berger and Wolpert [1988], pp. 45–6). However, influential authorities continue to cite it as a reason to reject Birnbaum's proof (e.g. Casella and Berger [2002], p. 296). Regardless of how those objections to Durbin's approach fare, the proof presented in

the previous section makes it effectively moot because the analogue of Durbin's response—restricting the Experimental Conditionality Principle to experimental ancillaries that are functions of a minimal sufficient statistic—allows the proof to go through in all but a restricted class of cases. The view that the Likelihood Principle holds only outside that class of cases is implausible, unattractive, and insufficient to satisfy an evidential frequentist, so for present purposes a proof of the Likelihood Principle outside that class is as good as a proof of the Likelihood Principle in general.

That the proof in the previous section goes through in all but a restricted class of cases when one restricts the Experimental Conditionality Principle as Durbin proposes to restrict the Weak Conditionality Principle can be seen as follows. It suffices to consider the applications of the Experimental Conditionality Principle to $(E_1, (E_1^\dagger, x_0^\dagger))$, $(E_1^M, (E_1^\dagger, x_0^\dagger))$, and $(E_1^{M_B}, (E^B, d))$ because its applications to $(E_2, (E_2^\dagger, y_0^\dagger))$, $(E_2^M, (E_2^\dagger, y_0^\dagger))$, and $(E_2^{M_B}, (E^B, d))$, respectively, are analogous. With Durbin's restriction the Experimental Conditionality Principle can be applied to $(E_1, (E_1^\dagger, x_0^\dagger))$ unless at some stage in the possibly nested mixture experiment E_1, two experimental outcomes from different component experiments have the same likelihood function. E_1 may have this property, but actual experiments typically do not, and the view that the Likelihood Principle holds only for those that do not is implausible, unattractive, and insufficient to satisfy an evidential frequentist. The principle can be applied to $(E_1^M, (E_1^\dagger, x_0^\dagger))$ and $(E_1^{M_B}, (E^B, d))$ unless outcomes in the two components of one of E_1^M or $E_1^{M_B}$, respectively, have the same likelihood function. The sampling distribution of these experiments, given by Table 2.2, has been chosen so that it is not the case that outcomes in the two components of either of those experiments have the same likelihood function *by construction*. Some such pair of outcomes may have the same likelihood function because of incidental features of the sampling distribution of E_1, but such cases are of no special interest. Again, the claim that the Likelihood Principle applies only outside such cases is implausible, unattractive, and insufficient to satisfy an evidential frequentist. Thus, the analogue of Durbin's restriction allows the proof given in the previous section to go through for a large class of cases that is sufficient for present purposes.

A second proposal to weaken the premises of Birnbaum's proof is Kalbfleisch's proposal to

restrict the Sufficiency Principle to outcomes of minimal experiments [1975].[22] Kalbfleisch's proposal, like Durbin's, faces many serious objections but continues to be cited influential authorities such as Casella and Berger ([2002], p. 296).[23] The proof given in the previous section neatly sidesteps the analogue of Kalbfleisch's proposal because the Weak Ancillary Realizability already applies only to minimal experiments.

Kalbfleisch's proposal is a bit stronger than it needs to be to block Birnbaum's proof: Kalbfleisch prohibits applying the Sufficiency Principle to any mixture experiment, but Birnbaum's proof involves applying it only to a simple mixture experiment. However, a weakened version of Kalbfleisch's proposal that applied only to simple mixtures, call it "Kalbfleisch*," would be inadequate because one could easily avoid it by modifying Birnbaum's proof slightly, for instance by adding a third component to E^M or by giving E_1 and E_2 unequal probabilities within E^M.

One might wonder if there is a set of restrictions on the Weak Conditionality Principle and the Sufficiency Principle not stronger than Durbin's or Kalbfleisch* that would suffice to block Birnbaum's proof and the analogue of which would suffice to block the new proof. In fact, any such set of restrictions would have to appeal to possible incidental features of the arbitrary experiments E_1 and E_2 between outcomes of which evidential equivalence is to be established, rather than to features of other experiments in the proofs that are present by construction. Thus, it would only suffice to restrict the scope of the new proof in a way that one suspects would be implausible, unattractive, and insufficient to satisfy an evidential frequentist. The only experiment Birnbaum constructs in his proof is the simple mixture E^M of E_1 and E_2, which are arbitrary except for the fact that they have a pair of respective outcomes with proportional likelihood functions. Thus, the weakest restriction on the Weak Conditionality Principle that appeals only to features of the proof that are

[22] Kalbfleisch also proposes to restrict the Strong Conditionality Principle to allow conditioning on mathematical ancillaries only after reducing by minimal sufficiency, but this change irrelevant to Birnbaum's proof.

[23] For objections to Kalbfleisch's proposal, see (Birnbaum [1975]) and (Berger and Wolpert [1988], pp. 46–67). Savage's objection to Durbin's approach also applies to Kalbfleisch's approach with slight modifications [1970]. Savage's objection to Durbin's approach is that it could lead statisticians to draw very different conclusions from experiments that differ "only microscopically" when the microscopic difference makes a minimal sufficient statistic no longer quite sufficient. The same point applies to Kalbfleisch's approach within a minimal component experiment.

present by construction which suffices to block Birnbaum's proof is to prohibit applying it to mixture experiments outcomes from different components of which have proportional likelihood functions. This restriction is equivalent to Durbin's. And the weakest restriction on the Sufficiency Principle that appeals only to features of experiments in the proof that are present by construction which suffices to block the proof is to prohibit applying it to simple mixture experiments, which is exactly Kalbfleisch*.

Thus, Durbin and Kalbfleisch's responses are in a sense the only options for those who would like to block Birnbaum's proof by restricting its premises. Of course, there are stronger responses that would suffice to block both Birnbaum's proof and the new proof given here, but such a strong response would require a proportionally strong argument.

2.4 A RESPONSE TO ARGUMENTS THAT THE PROOFS ARE FALLACIOUS

In addition to Durbin's and Kalbfleisch's proposals to block Birnbaum's proof by restricting its premises, there are also arguments due to Joshi [1990] and Mayo ([2009a], [2011], [2012]) that Birnbaum's proof is fallacious. The objections Joshi and Mayo present would also apply (*mutatis mutandis*) to the proof presented here, but I will argue that they are mistaken.

The arguments that Birnbaum's proof is fallacious mistake Birnbaum's premises for what we might call their operational counterparts. The Weak Conditionality Principle and the Sufficiency Principle each posit sufficient conditions for experimental outcomes to be evidentially equivalent. They are different, respectively, from what we might call the Operational Weak Conditionality Principle and the Operational Sufficiency Principle. The Operational Weak Conditionality Principle says that in drawing inferences from the outcome of a simple mixture experiment, one ought to use the sampling distribution of the component experiment actually performed, rather than the sampling distribution of the mixture experiment as a whole. The Operational Sufficiency Principle says that one ought to "reduce by sufficiency" as far as possible—that is, to use for inference the sampling distribution of a minimal sufficient statistic rather than the sampling distribution of the original sample space. Even

on the assumption that one ought to use methods that conform to the Weak Conditionality Principle and Sufficiency Principle if they are true, it does not follow that one ought to change the sample space as their operational counterparts prescribe: Bayesian conditioning and likelihoodist methods are insensitive to sample spaces, so they conform to the Weak Conditionality Principle and the Sufficiency Principle regardless of whether one conditions on ancillaries and/or reduces by sufficiency or not.

The Weak Conditionality Principle and the Sufficiency Principle are logically consistent: each of those principles merely asserts that certain sets of experimental outcomes are evidentially equivalent, and the assertion of any set of equivalences is logically consistent with the assertion of any other set of equivalences. However, their operational counterparts can conflict: conditioning on an ancillary statistic can preclude reducing by a particular sufficient statistic and vice versa. A conflict of this kind does arise in Birnbaum's proof: a minimal sufficient statistic of the mixture experiment E^M assigns the same value to the outcomes (E_1, x_0) and (E_2, y_0), so conditioning on which component experiment is performed precludes reducing to that statistic because only one of (E_1, x_0) and (E_2, y_0) is in the resulting sample space, and reducing to that statistic precludes conditioning on which component experiment is performed when either (E_1, x_0) or (E_2, y_0) occurs because those outcomes come from different component experiments but are indistinguishable in the reduced sample space.

It is easy to see how this conflict between the Operational Weak Conditionality Principle and the Operational Sufficiency Principle could give rise to the claim that Birnbaum's premises are inconsistent. For a frequentist, the only way to conform to each of Birnbaum's premises is to follow the corresponding operational principle. The two operational principles come into conflict, so from a frequentist perspective the original premises seem inconsistent. But this apparent inconsistency is not a legitimate objection to Birnbaum's proof because it presupposes a frequentist use of sampling distributions. Whether or not sampling distributions are relevant to evidential import is exactly what is at issue in the debate about the Likelihood Principle, so this presupposition begs the question.

The following passage shows that Joshi does in fact mistake Birnbaum's premises for their operational counterparts:

> For the assumed set-up [in Birnbaum's proof], the conditionality principle essentially

> means that only the experiment actually performed (E_1 or E_2) is relevant[...] But the same relevancy must hold good when applying the sufficiency principle[...] [A minimal sufficient statistic for E^M] is not a statistic—and hence not a sufficient statistic—under the probability distribution relevant for the inference in the assumed set-up. Hence the sufficiency principle cannot yield [$\mathrm{Ev}(E^M, (E_1, x_0)) = \mathrm{Ev}(E^M, (E_2, y_0))$] [...]so the proof fails. ([1990], pp. 111–2)

Joshi assumes that the Weak Conditionality Principle implies that one must condition on which component of E^M is actually performed before applying the sufficiency principle. But the Weak Conditionality Principle is not a directive to condition at all. Joshi appears to have in mind the Operational Weak Conditionality Principle and a version of the Operational Sufficiency Principle that is restricted along the lines of Kalbfleisch's approach to require conditioning on an experimental ancillary before reducing by sufficiency. Conforming to Birnbaum's premises requires following their operational counterparts only within a frequentist approach, so Joshi's objection begs the question.

Birnbaum himself insisted that his premises were to be understood as "equivalence relations" rather than as "substitution rules" and recognized that his proof is valid only when they are understood in this way. As he put, "It was the adoption of an unqualified equivalence formulation of conditionality, and related concepts, which led, in my 1972 paper, to the monster of the [Likelihood Principle]" ([1975], 263).

Mayo's objection to Birnbaum's proof is more elaborate than Joshi's but rests on the same error.[24] Mayo reconstructs Birnbaum's argument as having two premises. She writes the following about the first of those premises ([2012], p. 19, notation changed for consistency):

> Suppose we have observed (E_1, x_0) [such that some (E_2, y_0) has the same likelihood function]. Then we are to view (E_1, x_0) as having resulted from getting heads on the toss of a fair coin, where tails would have meant performing E_2[...] Inference based on [a minimal sufficient statistic for E^M] is to be computed averaging over the performed and unperformed experiments E_1 and E_2. This is the *unconditional formulation* of [E^M].

Mayo's comments here are true of the Operational Sufficiency Principle, but not of Birnbaum's Sufficiency Principle. Birnbaum's Sufficiency Principle does not say anything about how inference is to be performed—in particular, it does not say that inference is to be computed by averaging over E_1 and E_2. It is compatible with the possibility that inference is to

[24] I consider Mayo's most recent presentation of the objection here ([2012]). See also Cox and Mayo ([2011]) and Mayo ([2009a]).

be performed in that way, but it is also compatible with the possibility that inference is to be performed in a way that does not take sampling distributions into account at all, such as by a likelihoodist method or Bayesian conditioning.

Mayo then states her second premise as follows:

> Once it is known that E_1 produced the outcome x_0, the inference should be computed just as if it were known all along that E_1 was going to be performed, i.e. one should use the conditional formulation, ignoring any mixture structure.

She then claims that Birnbaum's argument is unsound because her two premises are incompatible: premise one says that one should use the unconditional formulation, while premise two says that one should use the conditional formulation. It is true that the argument Mayo has constructed is unsound, but that argument is not Birnbaum's. In reconstructing Birnbaum's argument Mayo has assumed that one must choose between a "conditional" and an "unconditional" formulation of the mixture experiment E^M. Frequentists need to make that choice because they use sampling distributions for inference, but likelihoodists and Bayesians do not. Thus, Mayo's response to Birnbaum's proof begs the question by presupposing a frequentist approach.

In personal communication, Mayo has responded that Birnbaum's proof is irrelevant to a sampling theorist if it requires assuming that sampling distributions are irrelevant to evidential import. But his proof does not require that assumption. Each of Birnbaum's premises is compatible with the claim that sampling distributions are relevant to evidential import. The fact that they are not *jointly* compatible with that assumption is not an *objection* to Birnbaum's proof—it is the whole *point* of Birnbaum's proof!

2.5 CONCLUSION

I have shown that the Likelihood Principle follows from the conjunction of the Experimental Conditionality Principle and the Weak Ancillary Realizability Principle. My proof of this result addresses responses to Birnbaum's proof that involve restricting its premises. Joshi's and Mayo's arguments for the claim that Birnbaum's proof is logically flawed would with

appropriate modifications apply to the new proof given here as well, but those arguments are in error.

The case for the Likelihood Principle seems quite strong. However, the Likelihood Principle as formulated here only implies that one ought to use methods that conform to the Likelihood Principle on the assumption that one ought to use methods that track evidential import. This assumption is not mandatory. Frequentists claim that their methods have many virtues, including objectivity and good long-run operating characteristics, that are not characterized in terms of evidential import. Tracking evidential import is intuitively desirable, but one could maintain that it is less important than securing one or more of those putative virtues.

3.0 NEW RESPONSES TO THREE COUNTEREXAMPLES TO THE LIKELIHOOD PRINCIPLE

3.1 INTRODUCTION

In the previous chapter, I presented a new proof of the Likelihood Principle and responded to frequentist attempts to undermine such proofs. In this chapter, I respond to three frequentist attempts to defeat the Likelihood Principle by counterexample.

Two of the counterexamples I consider are directed in the first instance against the Law of Likelihood rather than the Likelihood Principle. The Likelihood Principle does not entail the Law of Likelihood, but there are arguments from the former to the latter that are compelling on the assumption that there is such a thing as the degree to which a datum favors one hypothesis over another where each of those hypotheses assigns to that datum a determinate probability. This assumption is not obviously true, but it is sufficiently plausible to warrant regarding a counterexample to the Law of Likelihood is a *prima facie* counterexample to the Likelihood Principle as well.

There are many purported counterexamples to the Likelihood Principle.[1] I have chosen to respond to three that to my mind have no been adequately addressed elsewhere. After presenting an argument from the Likelihood Principle to the Law of Likelihood in Section 3.2, I respond to examples from Fitelson [2007], Armitage [1961], and Stein [1962] in Sections

[1]For additional counterexamples and the responses to them that I endorse, see [Birnbaum, 1964, 12–3] and [Royall, 1997]; [Stone, 1976], [Fraser et al., 1985], [Evans et al., 1986], [Berger and Wolpert, 1988, 127–36], and [Hill, 1988, 161–74]; [Sober, 1983, 354–6]; [Leeds, 2004], and [Chandler, 2013, 133–4]; [Sober, 2005, 128–9] and [Fitelson, 2007, 4–5]; and [Sober, 2008, 37–8]). The counterexamples in [Forster, 2006] are not in either of these categories because they lie outside the scope of the Likelihood Principle as I formulate it: they concern only the broader principle Forster calls the Likelihood Theory of Evidence, which extends the Likelihood Principle to composite as well as simple statistical hypotheses.

3.3, 3.4, and 3.5, respectively.

3.2 WHY A COUNTEREXAMPLE TO THE LAW OF LIKELIHOOD IS A *PRIMA FACIE* COUNTEREXAMPLE TO THE LIKELIHOOD PRINCIPLE

The Likelihood Principle does not entail the Law of Likelihood. Thus, a proponent of the Likelihood Principle could respond to counterexamples to the Law of Likelihood (including those presented in Sections 3.3 and 3.4, but not the one presented in Section 3.5) simply by denying the Law of Likelihood. I argue in this section that this response has a nontrivial cost: it requires denying that there is an objective fact of the matter regarding the degree to which data favors one hypothesis over another (even up to ordinal equivalence) in the typical case in science in which likelihoods are objectively well-defined while prior probabilities are not. Some Bayesians would be willing to pay this cost, in either of two ways. First, one could accept that evidential favoring is subjective in typical scientific cases. Second, one could give up the notion of evidential favoring entirely, in the same way that some advocates of Bayesian conditioning deny the significance of notions of confirmation (e.g. Brössel and Huber, 2014). It is not clear that the first of these approaches is adequate; in order to escape from counterexamples by adopting an alternative subjective account of evidential favoring, one would need to show that the subjective account in question does in fact avoid the counterexamples in question without generating new ones. The second approach is an option for those who are firmly committed to Bayesian ideas, but it precludes the likelihoodist strategy of giving an objective account of evidential favoring as a "fallback option" for cases in which prior probabilities are not objectively well-grounded [Sober, 2008]. I criticize this strategy in Chapter 5, but I maintain that to reject it on the basis of the purported counterexamples discussed in this chapter would be a mistake.

Let me now say why advocates of the Likelihood Principle are committed to the Law of Likelihood on pain of denying that there are objective facts about degrees of evidential favoring (taken to be defined up to ordinal equivalence) in typical scientific cases. The

Likelihood Principle says that the evidential import of datum E with respect to a set of hypothesis $\mathbf{H}$ depends only on $\Pr(E|H)$ as a function of H on $\mathbf{H}$, up to a constant of proportionality. The Law of Likelihood is usually taken to be[2] the claim that E favors H_1 over H_2 if and only if the likelihood ratio $\mathcal{L} = \Pr(E|H_1)/\Pr(E|H_2)$ is greater than one, with $\mathcal{L}$ measuriing the degree of that favoring, at least in the kinds of cases that are at issue, in which $\mathcal{L}$ is objectively well-defined. I take it that evidential favoring is defined only up to strictly monotone increasing transformations, so that the Law of Likelihood is compatible with the use of a given measure of evidential favoring if and only if that measure is strictly monotone increasing in $\mathcal{L}$.[3]

One can argue from the Likelihood Principle to the Law of Likelihood as follows. Suppose that for any H_1 and H_2 each of which has an objectively well-defined likelihood on datum E, it is objectively the case either that E is neutral between H_1 to H_2; that E favors H_1 over H_2 to a determinate, real-valued degree (relative to a scale that is defined up to ordinal equivalence); or vice versa. The Likelihood Principle says that the evidential import of E with respect to the pair of hypotheses $\{H_1, H_2\}$ depends only on $\Pr(E|H_1)$ and $\Pr(E|H_2)$ up to a constant of proportionality—equivalently, it depends only on the likelihood ratio $\mathcal{L}$. A rule for assessing E as evidence with respect to H_1 and H_2 should obviously be symmetric with respect to interchange of the labels H_1 and H_2. It follows that E must be evidentially neutral between H_1 and H_2 when $\mathcal{L} = 1$. Presumably, E is not neutral between H_1 and H_2 when $\Pr(E|H_1) \neq \Pr(E|H_2)$ and does not favor H_2 over H_1 when $\Pr(E|H_1) > \Pr(E|H_2)$. It follows that E favors H_1 over H_2 if and only if $\mathcal{L} > 1$, and that the degree to which it does so depends only on $\mathcal{L}$.

It remains to be shown only that the degree to which E favors H_1 over H_2 is monotone

[2] I write "is usually taken to be..." rather than "is..." because I argue in Section 3.3 that the Law of Likelihood is appropriate only for mutually exclusive hypotheses, contrary to this formulation.

[3] The analogous assumption is standard in the literature on confirmation theory (see e.g. Fitelson, 2001). An advocate of the Law of Likelihood might wish to require in addition to being monotone increasing in $\mathcal{L}$ that a measure of evidential favoring allow for some nice way of calculating the degree to which the conjunction of multiple pieces of independent evidence favors one hypotheses over another from the degrees to which the individual pieces of evidence do so. When one uses $\mathcal{L}$ itself as a measure, for instance, multiplying the degrees of evidential favoring from individual, independent pieces of evidence gives the degree of evidential favoring from their conjunction. $\log(\mathcal{L})$ is arguably even nicer in this respect because it allows one to aggregate by summing rather than multiplying. Insofar as such arguments are well-motivated, they can merely be adding on to the argument given here to yield an argument from the Likelihood Principle to a more constrained version of the Law of Likelihood.

increasing in $\mathcal{L}$. Suppose this claim were false. Then for an observation O of a string of heads as long as one likes on a sequence of independent and identically distributed coin tosses, it would be possible to construct pairs of hypotheses H and H' such that O favors H' over the hypothesis H_F that the coin is fair at least as strongly as it favors H over H_F even though H posits a higher probability of heads on each toss than H'.[4] I take it that this conclusion is unacceptable, and thus that the degree to which E favors H_1 over H_2 must be monotone increasing in $\mathcal{L}$.

Again, an advocate of the Likelihood Principle can avoid this argument by denying the assumption that there are facts about evidential favoring of the kind described. Those who are not willing to pay this cost seem to be committed to the Law of Likelihood. This fact warrants treating counterexamples to the Law of Likelihood as at least *prima facie* counterexamples to the Likelihood Principle. Fortunately for the Likelihood Principle, I will argue, the purported counterexamples to the Law of Likelihood discussed below are unsuccessful.

3.3 RESPONSE TO FITELSON'S COUNTEREXAMPLE

Fitelson presents the following as a counterexample to the Law of Likelihood [2007, 5, original emphasis, my notation]:

> **Example 3.1.** ...we're going to draw a single card from a standard (well-shuffled) deck.... E = the card is a spade, H_1 = the card is the ace of spades, and H_2 = the card is black. In this example...$\Pr(E|H_1) = 1 > \Pr(E|H_2) = 1/2$, but it seems absurd to claim that E favors H_1 over H_2, as is implied by the [Law of Likelihood]. After all, E *guarantees the truth of* H_2, but E provides only non-conclusive evidence for the truth of H_1.

[4] If the degree to which E favors H_1 over H_2 is not monotone increasing in $\mathcal{L}$, then there are an l_1 and an l_2 such that $l_1 \geq l_2$, yet for any $H_1, \ldots, H_4$ such that $l_1 = \Pr(E|H_1)/\Pr(E|H_2)$ and $l_2 = \Pr(E|H_3)/\Pr(E|H_4)$, E favors H_3 over H_4 at least as strongly as it favors H_1 over H_2. Simply choose H and H' so that $\Pr(O;H)/\Pr(O;H_F) = l_1$ and $\Pr(O;H')/\Pr(O;H_F) = l_2$ to get the result that O favors H' over H_F at least as strongly as it favors H over H_F.

I propose responding to this example by restricting the Law of Likelihood to mutually exclusive hypotheses. This response blocks not just this specific counterexample, but any potential counterexample in which a datum provides conclusive evidence for one hypothesis and non-conclusive evidence for another: if a pair of hypotheses is mutually exclusive, then E provides conclusive evidence for one only if it conclusively refutes the other.

Chandler [2013] and Steel [2007] also propose restricting the Law of Likelihood to mutually exclusive hypotheses, but the main advocates of the Law state it without that restriction (Edwards 1972, 31; Royall 1997, 3; Sober 2008, 32). There are at least two plausible explanations for the fact that the need for this restriction often goes unnoticed. First, one might simply take it for granted that the "favors over" relation holds only between mutually exclusive hypotheses. As Sober puts it, the Law of Likelihood addresses questions about "what the evidence says about the competition between two hypotheses" [2008, 34]. There is no genuine competition between two hypotheses if they are not mutually exclusive and thus could both be true. For instance, there is no genuine competition between the claim that a baseball team in the National League will win the next World Series and the claim that a baseball team based in Chicago will win the next World Series, because it is possible that a National League baseball team based in Chicago will win the next World Series.

Second, one might have in mind statistical hypotheses each of which posits a probability distribution for the same random variable. Hypotheses of this kind cannot be distinct without being mutually exclusive.[5] However, the Law of Likelihood is generally taken to apply to substantive hypotheses that imply statistical hypotheses yet can be distinct without being mutually exclusive, such as the hypothesis that the card is black and the hypothesis that it is the ace of spades.[6] For applications of this kind, restricting the Law of Likelihood to

[5]Two probability density functions can be distinct in that they differ on sets of measure zero yet compatible in the sense that they imply all the same probabilities for observable events. However, probability density functions that differ only on sets of measure zero are not distinct in the sense that is relevant to statistical practice, precisely because they imply all the same probabilities for observable events. From the standpoint of statistical practice, the fact that one can alter a probability density function on a set of measure zero without changing its implications for the probabilities of observable events is merely an artifact of the measure-theoretic formalism of probability theory.

[6]Compatible substantive hypotheses can imply incompatible statistical hypotheses because the manner in which a substantive hypothesis implies a statistical hypotheses is non-monotonic. For instance, the information that the card is black implies that the probability that it is a spade is 1/2, but the statement that the probability that the card is a spade is 1/2 is no longer correct given the additional information that the card is the ace of spades.

mutually exclusive hypotheses is indeed necessary to avoid counterexamples.

Restricting the Law of Likelihood to mutually exclusive hypotheses might seem to be *ad hoc*, particularly in light of the preceding arguments for the Likelihood Principle and from the Likelihood Principle to the Law of Likelihood that never mentioned mutual exclusivity. However, it is easy to see where the arguments break down and why the breakdown does not pose a threat to the principles in general. The argument from the Likelihood Principle to the Law of Likelihood starts with the supposition that E either favors H_1 over H_2 or vice versa or is neutral between them. None of these concepts apply when H_1 and H_2 are compatible and thus not genuine competitors. This fact does nothing to undermine the argument in other cases, including the standard case in science in which the hypotheses in question are the simple elements of a statistical model.

Restricting the Law of Likelihood to mutually exclusive hypotheses seems natural and suffices to block Fitelson's counterexample, but it faces at least three significant objections. In Subsection 3.3.1, I respond to the objection that restricting the Law of Likelihood to mutually exclusive hypotheses violates plausible constraints on the notion of evidential favoring. In Subsection 3.3.2, I respond to the objection that it fails to solve the "tacking paradox." In Subsection 3.3.3, I respond to the objection that it excludes cases of genuine scientific interest.

3.3.1 Objection 1: Response conflicts with constraints on evidential favoring

Fitelson claims in his [2013] that restricting the Law of Likelihood to mutually exclusive hypotheses makes it "too easy to refute" the following "bridge principle" that connects the notion of evidential favoring to that of incremental confirmation (67):[7]

> (†) Evidence E favors hypothesis H_1 over hypothesis H_2 if and only if E confirms H_1 more than H_2.[8]

[7] Evidential favoring is a three-place relation between a bit of data and a pair of hypotheses. Confirmation is a two-place relation between a bit of evidence a single hypothesis. Incremental confirmation concerns the "change in firmness" of a hypothesis as a result of some datum. It is contrasted with absolute confirmation, which concerns the not the change in firmness but rather the terminal degree of firmness of the hypothesis after receiving the datum.

[8] Fitelson relativizes this principle to a measure of confirmation. I have implicitly stated it in terms of the "true" measure of confirmation because differences between measures of confirmation are not relevant to the issue at hand.

Fitelson also claims that the following "should be a desideratum for any adequate explication of favoring" (69):

> (CE) If E constitutes conclusive evidence for H_1, but E constitutes less than conclusive evidence for H_2 (where it is assumed that E, H_1, and H_2 are all contingent), then E favors H_1 over H_2.

Restricting the Law of Likelihood to mutually exclusive hypotheses seems to be incompatible with accepting either (CE) or (†). For (CE) implies that the "favors over" relation can hold between compatible hypotheses on the extremely mild assumption that E can be conclusive evidence for an hypothesis H_1 and less than conclusive evidence for a hypothesis H_2 that is compatible with H_1. (†) does likewise on the even milder assumption that E can confirm a compatible pair of hypotheses H_1 and H_2 to different degrees.

One could respond to this objection by claiming that the Law of Likelihood explicates the "favors over" relation for mutually exclusive hypotheses, while a different principle explicates it for compatible hypotheses. But while (†) and (CE) both have intuitive appeal, we can see that at least one of them must be false without even considering the Law of Likelihood and Fitelson's counterexample. For they jointly entail the following, which is not plausible:

> (CE′) If E constitutes conclusive evidence for H_1, but E constitutes less than conclusive evidence for H_2 (where it is assumed that E, H_1, and H_2 are all contingent), then E confirms H_1 more than it confirms H_2.

Contrary to (CE′), E seems to confirm H_2 more than it confirms H_1 when it constitutes conclusive evidence for an H_1 that was already nearly certainly true and near-conclusive evidence for an H_2 that was previously nearly certainly false. For instance, let $\Pr(H_1) = .99, \Pr(H_1|E) = 1, \Pr(H_2) = .01$, and $\Pr(H_2|E) = .99$. In this case, E confirms H_2 more than it confirms H_1 both intuitively and according to what Chandler [2013, 130] identifies as the six most popular measures of confirmation currently on offer, given the additional stipulation that H_1 and H_2 are exhaustive. (See Appendix B for a proof of this claim.) Thus (CE′) is false, which implies that either (CE) or (†) is false.

This problem is easy to resolve by restricting the "favors over" relation itself to mutually exclusive hypotheses, thereby restricting (CE), (†), and (CE′) in the same way. This restriction is quite natural and does not seem to have any substantial bad consequences. It

is compatible with the highly plausible "left-to-right" direction of (†): if E favors H_1 over H_2, then E confirms H_1 more than H_2. It does require giving up the "right-to-left" claim that E favors H_1 over H_2 if it confirms H_1 more than H_2 for the special case in which H_1 and H_2 are compatible, but that consequence is perhaps not undesirable. Restricting (†) in this way is not "too easy," but rather is well motivated by a natural interpretation of what it means for evidence to favor one hypothesis over another and the fact that (CE) and (†) in their original formulations cannot both be true.

Similarly, restricting the "favors over" relation to mutually exclusive hypotheses is compatible with (CE) for mutually exclusive H_1 and H_2, where it is obviously true because E constitutes conclusive evidence for H_1 but less than conclusive evidence for H_2 only if it refutes H_2. (CE′) is obviously true for essentially the same reason. We must accept that (CE) and (CE′) are false when H_1 and H_2 are compatible, but only because the notion "favors over" simply does not apply in those cases.

Restricting the "favors over" relation in general to mutually exclusive hypotheses resolves Fitelson's concern that so restricting the Law of Likelihood is incompatible with plausible constraints (†) and (CE) on the notion of evidential favoring. Moreover, this maneuver allows one to preserve the obviously true parts of (†) and (CE) without having to accept their clearly false consequence (CE′).

3.3.2 Objection 2: Response fails to address tacking paradox

Fitelson also objects that restricting the Law of Likelihood to mutually exclusive hypotheses fails to address examples similar to his Example 3.1 that give rise to a version of the tacking paradox.[9] One such example is as follows [Fitelson, 2013, 77]:

> **Example 3.2.** Let E be the proposition that the card is black. Let X be the hypothesis that the card is an ace, let H_1 be the hypothesis that the card is a spade, and let H_2 be the hypothesis that the card is a club.

In this case the Law of Likelihood implies that E is neutral between H_1 and the conjunction of H_2 and X, because $\Pr(E|H_1) = \Pr(E|H_2 \ \& \ X) = 1$. But this claim is implausible because

[9] The tacking paradox is also known as the "problem of irrelevant conjunction."

because X is an "irrelevant conjunct" that has been "tacked on" to H_2. Appealing to mutual exclusivity does not help because H_1 and H_2 & X are mutually exclusive.

More familiar versions of the tacking paradox arise in qualitative confirmation theory, where the problematic conclusion is that an E that confirms some hypothesis H also confirms H & X, where X is irrelevant; and in quantitative confirmation theory, where the problematic conclusion is that an E that confirms H also confirms H & X to the same degree.

I offer three possible responses to the tacking paradox. Which response one should accept depends on one's intuitions and other commitments.

3.3.2.1 Response 1: Bite the bullet My favorite response to the tacking paradoxes is to bite the bullet and accept their supposedly problematic conclusions, in this case that E is neutral between H_1 and H_2 & X. I know of no argument against those conclusions claim beyond a bare appeal to intuition. For what it is worth, I lack the relevant intuitions. It seems right to me, for instance, that if E is neutral between H_1 and H_2, then it is also neutral between H_1 and H_2 & X for some irrelevant X. This claim would be clearly problematic if it implied that E confirms or disconfirms X itself, but it does not.

Milne takes the same line in the context of quantitative confirmation theory. He claims that "prediction and confirmation are two sides of the same coin" and thus that "evidence equally to be expected in the light of two theories cannot differentially confirm them" [1996, 23]. If we substitute "is evidentially neutral between" for "does not differentially confirm," then this statement is just a weak version of the Law of Likelihood and thus cannot be regarded as an argument for accepting that E is neutral between H_1 and the H_2 & X rather than restricting the Law of Likelihood. However, Milne also addresses a possible objection to the claim that E is neutral between H_1 and H_2 & X by pointing out that this claim is compatible with believing that there are powerful methodological arguments against theories that are "merely tacked together." It is only incompatible with the claim that those methodological arguments have to do with the evidential favoring relation between a pair of theories and a datum—as opposed to, for instance, the way in which considerations of unification and simplicity affect prior probabilities.

While the fact that I am unmoved by the supposedly unintuitive nature of the examples that give rise to the tacking paradox does not make me unique, it does seem to place me in the minority among philosophers of science. For those who differ from me in this respect, I offer two additional responses to the tacking paradox.

3.3.2.2 Response 2: Regard Law as explicating "r-favoring" Those who cannot accept that E is neutral between H_1 and H_2 & X may prefer the analogue of a response Maher gives to the tacking paradox in the context of qualitative Bayesian confirmation theory. Qualitative Bayesian confirmation theory provides the predicate C as an explicatum:

$$C(H, E, K) \equiv_{df} \Pr(H|E\&K) > \Pr(H|K)$$

What gives rise to the tacking paradox within qualitative Bayesian confirmation theory, Maher claims, is a lack of clarity about its explicandum. The tacking paradox (among other arguments) shows that C is a poor explicatum for the "everyday notion" of confirmation. Thus, a Bayesian confirmation theorist must identify some other explicandum. Maher suggests the notion he calls "r-confirmation:"

> **Definition.** E r-confirms H given K iff the inductive probability of H given E and K is greater than the inductive probability of H given K alone.

That E r-confirms H & X if it r-confirms H and is irrelevant to X is a consequence of the fundamental Bayesian assumption that inductive probability obeys the probability calculus. This conclusion does not give rise to paradox once one realizes that r-confirmation is not the everyday notion of confirmation.

One could avoid the tacking paradox for the Law of Likelihood in a similar manner by claiming that the Law of Likelihood explicates a notion of "r-favoring:"

> **Definition.** E r-favors H_1 over H_2 given K iff the ratio of the inductive probability of H_1 given E and K to that of H_2 given E and K is greater than the ratio of the inductive probability of H_1 given K alone to that of H_2 given K alone.

The primary difficulty this maneuver faces is that it appeals to the ill-understood and arguably misguided notion of "inductive probability." Nevertheless, it does indicate a pos-

sible line of response to the tacking paradox for those who consider such a response to be necessary. However, this line of response is not available to conventional likelihoodists, who reject the assignment of probabilities (inductive or otherwise) to typical scientific hypotheses. They must take a third line of response if they are no twilling to bite the bullet.

3.3.2.3 Response 3: Restrict Law of Likelihood to structurally identical alternatives A third possible response to the tacking paradox for the Law of Likelihood is to require that H_1 and H_2 be not only mutually exclusive but also *structurally identical*, meaning that they assign values to the same set of random variables. This response is sufficient to prevent the paradox from arising: the fact that X assigns a value to a different variable from H_1 and H_2 is what allows E to be relevant to H_1 and H_2 but not to X.

Steel suggests this approach in his [2007, 68] but then rejects it because he claims that it excludes some cases of genuine scientific interest. Newton's theory of gravity and Einstein's general theory of relativity, for instance, certainly differ over more than the values of some common set of variables [Steel, 2007, 70].

This objection to restricting the Law of Likelihood to structurally identical hypotheses is far from conclusive. While the Law of Likelihood restricted to structurally identical hypotheses does not apply to comparisons between high-level theories such as Newton's and Einstein's theories of gravity directly, it does apply to the kinds of low-level consequences of those theories that individual experiments actually probe. For instance, it applies to the Eddington eclipse observations, which were directly concerned with consequences of Newton's and Einstein's theories for a measure of the deflection of starlight as it passed by the sun. Thus, the Law of Likelihood is relevant to the comparison between Newton's and Einstein's theories even if it does not address that comparison directly.

Fully developing this idea would require providing an account of the relationship between tests of low-level consequences of high-level theories and the evaluation of those theories themselves. Such an account lies beyond the scope of this dissertation; see [Suppes, 1962]; [Mayo, 1996, Ch. 5]; and [Mayo and Spanos, 2009, Ch. 2] for possible avenues to pursue. In any case, Steel's objection shows at most that the Law of Likelihood does not apply directly to the evaluation of high-level theories if it restricted in the way he considers. It

might nevertheless be a useful epistemic tool. The tacking paradox is even less problematic for those who are willing to accept that the Law of Likelihood explicates the notion of r-favoring rather than the everyday notion of evidential favoring and for those who do not find the claim that E is neutral between H_1 and H_2 & X problematic in the first place. Thus, the fact that restricting the Law of Likelihood to mutually exclusive hypotheses leaves it unaddressed is not a persuasive objection to that maneuver.

3.3.3 Objection 3: Response excludes cases of scientific interest

The final possible objection I will consider to restricting the Law of Likelihood to mutually exclusive hypotheses is that it unduly restricts the scope of the Law. In particular, it seems to exclude cases involving competing causal claims and cases involving nested models. However, in cases of these two kinds there are many possible interpretations of the hypotheses being tested on which they are mutually exclusive. I conjecture that some such interpretation is appropriate whenever claims of one of these kinds can truly be said to be tested against one another. The burden is on those who doubt this claim to provide a counterexample. Moreover, like Steel's objection to restricting the Law of Likelihood to structurally identical hypotheses, this objection shows at most that the scope of the Law of Likelihood is somewhat smaller than one might have thought.

3.3.3.1 Cases involving competing causal claims Machery [2014] argues that the psychologists Greene et al. [2001] are best understood as using a likelihoodist methodology to test the following hypotheses against one another:

Example 3.3.

H_1: People respond differently to moral-personal and moral-impersonal dilemmas because the former elicit more emotional processing than the latter.

H_2: People respond differently to moral-personal and moral-impersonal dilemmas because the single moral rule that is applied to both kinds of dilemmas (for example, the doctrine of double effect) yields different permissibility judgments.

H_1 and H_2 are ambiguous, but on the surface they do not seem to be mutually exclusive. They certainly are not mutually exclusive if they merely assert, respectively, that eliciting more emotional processing and yielding different permissibility judgments under certain moral rules are each *a* cause of responding differently to moral-personal and moral-impersonal dilemmas, where causation is understood in the manipulationist sense according to which (roughly) A causes B if and only if it is possible to change B by manipulating A (see Woodward 2003). Moreover, it seems quite plausible that a true general theory of human moral judgment would have to account for both emotional processing and the use of moral rules, thus providing a synthesis of H_1 and H_2 that would not be possible if H_1 and H_2 were mutually exclusive. On the other hand, it does seem to be possible to test H_1 and H_2 against one another.

These statements can be generalized to any case involving a pair of claims of the form "X causes Z" and "Y causes Z." Such claims are plausibly regarded as compatible, but it does seem possible to test them against one another. Because such claims are ubiquitous in science, it might seem that restricting the Law of Likelihood to mutually exclusive hypotheses is a considerable concession.

The problem with this argument is that while there may be both interpretations of claims of the form "X causes Z" and "Y causes Z" on which they are compatible and interpretations of such claims on which they can be tested against one another, it is not clear that there are any interpretations that have both of these characteristics. It does not seem plausible to me that scientists can truly be said to test claims of the form "X is *a* cause of Z" and "Y is *a* cause Z" against one another, precisely because those claims could both be true. However, they can truly be said to test claims of the form "X causes Z" and "Y causes Z" against one another under many possible interpretations of those claims on which they are mutually exclusive. For instance, in simple cases such hypotheses could be understood as asserting, respectively, "X causes Z and Y does not" and "Y causes Z and X does not." In cases involving complex systems, they will generally be more plausibly understood as asserting something like, respectively, "X is the most important cause of Z" and "Y is the most important cause of Z" (or perhaps "X is a more important cause of Y than Z" and vice versa), where the notion of importance could generally be operationalized as, for instance,

the percent of variance in Z accounted for in appropriate experiments.

Another possibility for a given pair of claims of the form "X causes Y" and "Z causes Y" is that they are best understood not as hypotheses properly speaking, but rather as loose expressions of the stances of two competing research programs. Particular experiments in the relevant domain do not test even disambiguations of "X causes Z" and "Y causes Z" against one another directly, but rather a more specific claim "in the spirit of 'X causes Z' " against a more specific claim "in the spirit of 'Y causes Z.' " Rather than testing the research programs themselves against one another directly, scientists use outcomes from tests of more specific claims to inform judgments about those research programs in light of their fruitfulness, empirical adequacy, and so forth.

This last interpretation seems quite plausible in the case of Example 3.3. It is compatible with the plausible conjecture that the view that ultimately prevails will be a synthesis of the research programs that H_1 and H_2 represent. Individual experiments that test mutually exclusive hypotheses "drawn from" those respective research programs against one another simply provide evidence for phenomena for which a successful synthesis of this kind would have to account.

In summary, there are plausible ways of understanding Greene et al.'s work on which the hypotheses they can properly be said to testing against one another in particular experiments are mutually exclusive. Thus, that work does not speak against my conjecture that some suitable interpretation of the claims "X causes Z" and "Y causes Z" can be found that make them mutually exclusive whenever scientists can truly be said to test such claims against one another.

3.3.3.2 Cases involving nested models

Chandler [2013] discusses a different kind of case in which scientists appear to test compatible hypotheses against one another, namely cases in which they choose among nested models. (A *model* in the sense that is relevant here is a composite statistical hypothesis, that is, a disjunction of hypotheses each of which ascribes a probability distribution to some variable.) Scientists often test, for instance, the hypothesis (LIN) that some variable Y is a quadratic function of X plus a normally distributed error term against the hypothesis (QUAD) that Y is a linear function of X plus

a normally distributed error term. But (LIN) is simply a special case of (QUAD) with the coefficient of the quadratic (X^2) term of (QUAD) set to zero. Thus, (LIN) and (QUAD) are compatible. Model selection is an important part of science, so this kind of example seems to indicate that restricting the Law of Likelihood to mutually exclusive hypotheses excludes cases of genuine scientific interest.

As in the case of competing causal claims, so too in the case of nested models, there are several possible interpretations according to which the scientists actually are testing mutually exclusive hypotheses against one another. Once again, I conjecture that some such interpretation can always be found and contend that the burden of proof is on those who doubt this claim to provide a counterexample.

Chandler provides one possible interpretation when he points out that "as Forster and Sober use the term, 'E favors model M_1 over model M_2' is actually shorthand for 'E favors the likeliest (in the technical sense) disjunct $L(M_1)$ of model M_1 over the likeliest disjunct $L(M_2)$ of model M_2,' with $L(M_1) \cap L(M_2) = \emptyset$" [2013, 133]. $L(M_1)$ and $L(M_2)$ are simple statistical hypotheses, so they are either mutually exclusive or identical. We can allow the Law of Likelihood to apply to a simple statistical hypothesis and itself, with the intuitively obvious result that any piece of evidence is evidentially neutral between them. Thus, under this interpretation statements of the form "E favors model M_1 over model M_2" fall within the scope of the Law of Likelihood even when it is restricted to mutually exclusive hypotheses.

This response is fine as far as it goes, but scientists seem to use phrases of the form "E favors M_1 over M_2" in other ways as well. For instance, they might say "E favors (QUAD) over (LIN)" when they mean that their data favors the claim that the element of (QUAD) that is by some measure closest to the true model has a nonzero coefficient for the X^2 term over the negation of that claim. However, in many applications it is implausible that the coefficient of the X^2 term for the element of (QUAD) that is closest to the true model would be exactly zero. What a scientist might mean instead in such cases is that E favors over its negation the hypothesis that a nonzero coefficient for the X^2 term is necessary for producing a statistically adequate curve, meaning (roughly) a curve the residuals of which look like white noise to a degree that is adequate for his or her aims. On both of these interpretations, "E favors (QUAD) over (LIN)" actually means "E favors (QUAD)\(LIN) over (LIN)," where

(QUAD)\(LIN) is the set of elements of (QUAD) that are not also elements of (LIN). Thus, scientists who have one of these interpretations in mind are in fact considering mutually exclusive hypotheses.

Fitelson [2013] claims that interpreting what look like comparisons between nested models as comparisons between mutually exclusive hypotheses is undesirable from a likelihoodist perspective because the fact that one can apply the Law of Likelihood to nested hypotheses such as (LIN) and (QUAD) is supposed to be an advantage for likelihoodism over Bayesianism. It is difficult to see how one could give a Bayesian account of the widespread preference for the simple model (LIN) over the complex model (QUAD) given that the fact that (LIN) $\subset$ (QUAD) entails $\Pr(\text{LIN}) \leq \Pr(\text{QUAD})$. Likelihoodism seems to be in a better position in this respect because it is concerned with evidential favoring rather than with posterior probability.

This argument is problematic in at least two respects. First, likelihoodism fares little better than Bayesianism in accounting for the widespread preference for (LIN) over (QUAD). The likelihood ratio for (LIN) and (QUAD) themselves is defined only relative to a prior probability distribution over the simple components of these models, and thus is typically unavailable from a likelihoodist perspective. A standard way to address this problem is to use the likelihood ratio of the best-fitting element of (LIN) to the best-fitting element of (QUAD), but this comparison can never favor (LIN) over (QUAD) because the best-fitting element of (LIN) is also an element of (QUAD). In fact, data from typical problems is all but guaranteed to favor (QUAD) over (LIN) according to the Law of Likelihood even if (LIN) is true.[10]

The Law of Likelihood does not vindicate the widespread preference for simple models. Some other account is needed, such as perhaps Forster and Sober's appeal to overfitting [1994] or an extension to Kevin Kelly's theory of Ockham's razor (see e.g. Kelly 2007) to statistical problems. Many model selection criteria designed to address overfitting, such as the Akaike Information Criterion that Forster and Sober [1994] promote, do include a likelihood ratio term, but it is the likelihood ratio of the best-fitting member of one model to

[10]Formally, three or more data points will favor (QUAD) over (LIN) with probability one even if (LIN) is true in the standard case in which the observations are affected by continuously distributed random noise.

the best-fitting member of the other rather than a likelihood ratio for the models themselves. Moreover, standard arguments for the use of such criteria such as the argument given by Forster and Sober [1994] do not require that the likelihood ratio term be interpreted as a measure of evidential favoring. Thus, the use of such criteria is quite compatible with restricting the Law of Likelihood to mutually exclusive hypotheses.

Second, even if likelihoodism did provide an account of the widespread preference for simple models, this fact would not be an advantage for likelihoodism over Bayesianism because that account would be available to Bayesians as well. A Bayesian can say anything that a likelihoodist can say. Likelihoodism and Bayesianism come into conflict only over the legitimacy of distinctively Bayesian statements that involve assignments of probabilities to hypotheses.

The upshot of Fitelson's counterexample is that the Law of Likelihood applies only to mutually exclusive hypotheses. This restriction is natural and adequate to block the counterexample and can withstand all of the objections that have been raised against it.

3.4 RESPONSE TO ARMITAGE'S COUNTEREXAMPLE

The example discussed in this section has been claimed to illustrate a conflict between the Law of Likelihood and what Cox and Hinkley [1974, 45–46] call *the Weak Repeated Sampling Principle*:[11]

> **The Weak Repeated Sampling Principle.** We should not follow procedures which for some possible parameter values would give, in hypothetical repetitions, misleading conclusions most of the time.[12]

[11] Similar principles include Birnbaum's (Conf) [1977, 24] and Berger and Wolpert's "Formal Confidence Principle" [1988, 71–72], among others. Such principles attempt to elaborate Neyman's conception of statistics as being concerned with the appropriate regulation of one's "inductive behavior" (e.g. Neyman and Pearson 1933, 291).

[12] Cox and Hinkley's *Strong Repeated Sampling Principle* says simply that "statistical procedures are to be assessed by their behavior in hypothetical repetitions under the same conditions" (45). This claim is not logically stronger than the Weak Repeated Sampling Principle because, unlike the latter, it says nothing about *how* behavior in hypothetical repetitions is to be used in choosing statistical procedures. However, it is stronger in one respect if it is taken to mean that procedures are to be assessed *only* on their behavior in hypothetical repetitions.

In other words, we should not use a procedure when it's possible that it's probable that it will mislead us.

Any attempt to argue that the Law of Likelihood is unacceptable because it conflicts with the Weak Repeated Sampling Principle faces two immediate problems: the Weak Repeated Sampling Principle is unreasonably strong, and the Law of Likelihood itself is not the sort of claim that could conflict with it. I will discuss these difficulties in turn, and then use the lessons drawn from that discussion to explain why Armitage's purported counterexample is no threat to the Law of Likelihood.

3.4.1 Two problems for attempts to use the Weak Repeated Sampling Principle as an objection to the Law of Likelihood

3.4.1.1 Problem 1: The Weak Repeated Sampling Principle is unreasonably strong Suppose that for some inductive inference problem one must choose between Procedure A and Procedure B. For all but one possible parameter value, Procedure A would outperform Procedure B by a wide margin in repeated applications. But for that one possible but highly implausible value, Procedure A would yield misleading conclusions most of the time, whereas there are no possible parameter values on which Procedure B would yield misleading conclusions most of the time. Surely there are cases of this kind in which Procedure A is by far the more reasonable choice. Nevertheless, the Weak Repeated Sampling Principle requires choosing Procedure B. Thus, the Weak Repeated Sampling Principle goes too far by requiring that one avoid the bare possibility of a high probability of a misleading result regardless of the expected consequences.

Moreover, when taken at face value the Weak Repeated Sampling Principle is so strong that even frequentists universally violate it even in the most basic kinds of textbook examples. Take the case of using a predesignated number of observations from a normal distribution with unknown mean μ and known variance to test the null hypothesis $H_0 : \mu \leq \mu_0$ against $\mu > \mu_0$, where one's options are either to reject or fail to reject the null hypothesis. The obvious and standard procedure on a problem of this kind is to reject H_0 if and only if the mean of a set of observations drawn from the distribution in question is greater than some

specific cutoff value c. (A test of this kind is a uniformly most powerful test for its Type I error rate.) It is standard to designate as the null hypothesis the one that would be more costly to reject if it were true, so c should be greater than μ_0. Fixing c determines the procedure's Type I error rate: the probability of rejecting H_0 if it is true, maximized[13] over the set of possibilities that are consistent with H_0. Fixing c also determines for each simple component of the alternative hypothesis a corresponding Type II error rate, that is, for each hypothesis H_a of the form $\mu = \mu_a$ for some $\mu_a > \mu_0$, it determines the probability of failing to reject the null hypothesis if H_a is true.

The problem for the Weak Repeated Sampling Principle is that the sum of the Type I error rate and the Type II error rate goes to one as the value μ_a of the component of the alternative hypothesis one considers decreases toward μ_0. Thus, for any procedure of the kind described there is some hypothesis such that the procedure is as close as one likes to being at least as likely as not to deliver the wrong answer if that hypothesis is true. Using a different kind of test procedure will not address this problem because any such procedure has a larger Type II error rate than the test of the form in question that has the same Type I error rate.

One could respond that the Weak Repeated Sampling Principle only requires that a procedure not give a misleading conclusion *most* of the time, and merely "as close as one likes to as often as not" is not most of the time. But this response rules out the standard frequentist practice of choosing a cutoff so that the Type I error rate has some small value, most often .05, because then the Type II error rate would rise to .95, which presumably would qualify as "most of the time."

A more promising response is that failing to reject the null hypothesis when the true mean is just a tiny bit larger than μ_0 is not a "misleading conclusion" in the intended sense. One is concerned only to detect discrepancies from the null hypothesis that are large enough to be of practical importance. The sum of the Type I error rate and the limit of the Type II error rate as μ_a decreases to $\mu_0 + d$ for some positive d does not go to one and can be made as small as one likes in principle by increasing the sample size while holding c and d

[13]Technically, the frequentist error rates are defined in terms of *suprema* rather than maxima, but this distinction is unimportant for our purposes.

fixed. Thus, allowing one to use a procedure that has a high probability of yielding a result that is only slightly misleading makes it possible to conform to the Weak Repeated Sampling Principle on this problem. It thereby allows frequentists to avoid what would otherwise be a fatal objection to the Weak Repeated Sampling Principle. However, we will see in Subsection 3.4.2 that it also defuses Armitage's purported counterexample to the Law of Likelihood.

3.4.1.2 Problem 2: The Law of Likelihood itself cannot conflict with the Weak Repeated Sampling Principle

A second immediate problem for attempts to argue that the Law of Likelihood is unacceptable because it conflicts with the Weak Repeated Sampling Principle is that the Law of Likelihood is not even the sort of claim that could conflict with the Weak Repeated Sampling Principle. The Weak Repeated Sampling Principle prohibits procedures that yield misleading conclusions most of the time, but the Law of Likelihood is not a procedure and does not yield conclusions. What can conflict with the Weak Repeated Sampling Principle is not the Law of Likelihood itself, but rather various procedures that the Law of Likelihood vaguely suggests. For instance, when there are multiple hypotheses of interest of which one must choose exactly one to accept, the Law of Likelihood suggests the maximum likelihood estimation procedure of accepting the hypothesis that ascribes the highest probability to the data. But the Law of Likelihood does not require the use of maximum likelihood estimation. Indeed, a Bayesian could very well accept the Law of Likelihood but would not typically use maximum likelihood estimation.

One type of procedure that the Law of Likelihood suggests is simply to "announce" through some appropriate channel for some H_1 and H_2 of particular interest that one's data favor H_1 over H_2 (or vice versa) to the degree given by the associated likelihood ratio. An output of this procedure is misleading when the hypothesis that is said to be disfavored is true, with the likelihood ratio a measure of the degree of misleadingness for fixed H_1 and H_2. Armitage's example appears to provide a recipe for causing this procedure to generate an announcement that is by this standard as misleading as one likes with probability one, in violation of the Weak Repeated Sampling Principle. I will show, however, that this appearance is misleading.

3.4.2 Why Armitage's example is no threat to the Law of Likelihood

Armitage's example involves taking observations from a normal distribution with unknown mean and known variance. The key to the example is that the number of observations is not fixed in advance. Instead, one keeps taking observations until the sample mean $\bar{x}_n$ is a certain distance away from zero. This distance decreases as the number of observations increases, at a rate that is fast enough to ensure that the experiment ends eventually even if the true mean μ is zero.[14]

Armitage's example seems problematic because by making the value of $|\bar{x}_n|$ required to end the experiment sufficiently large, one can guarantee that the experiment generates a likelihood ratio for the hypothesis that $\mu = \bar{x}_n \neq 0$ against that hypothesis that $\mu = 0$ that is as large as one likes,[15] even when the latter is true. Thus, the practice of announcing that the observations taken from this experiment favor $\mu = \bar{x}_n$ over $\mu = 0$ to the degree given by the likelihood ratio of the former to the latter appears to violate the Weak Repeated Sampling Principle in a dramatic fashion. At the end of the previous subsection, I wrote that such a procedure is misleading when $\mu = 0$ is true, with the likelihood ratio a measure of the degree of misleadingness for fixed H_1 and H_2. It follows that the result that the Armitage example is bound generate is indeed misleading. However, it does not follow that this result can be made as misleading as one likes. The hypothesis that $\mu = \bar{x}_n$ is not fixed, but random, so the claim that the likelihood ratio of $\mu = \bar{x}_n$ to $\mu = 0$ is a measure of misleadingness for fixed H_1 and H_2 does not apply.

One might think that the fact that the Armitage example provides a recipe for producing a result that is misleading at all is enough to refute the Law of Likelihood. But the Armitage example is not special in this regard. A fixed number of observations from a normal distri-

[14] Observations are taken until the first time the sample mean is at least k standard deviations away from 0 for some k ($|\bar{x}_n| > k\sigma_0/\sqrt{n}$). With probability one, sampling stops after some finite n. When the experiment ends, the likelihood at $\mu = \bar{x}_n$ is more than $e^{\frac{1}{2}k^2}$ times that at $\mu = 0$, so by choice of k the likelihood ratio for $\mu = \bar{x}_n$ against $\mu = 0$ can be made arbitrarily large. See [Cox and Hinkley, 1974, 50–1].

[15] It is important to avoid misinterpreting Armitage's example as providing a recipe for generating a result which according to the Law of Likelihood disfavors the hypothesis that $\mu = 0$ is true *relative to its negation* $\mu \neq 0$ even when $\mu = 0$. The likelihood ratio for this pair of hypotheses is defined only relative to a prior probability distribution for μ. Given such a distribution, that likelihood ratio need not be less than one. For instance, Berger and Wolpert [1988, 81–2] consider a prior that yields for one possible outcome of the experiment a likelihood ratio of 3.5 for the hypothesis that $\mu = 0$ against $\mu \neq 0$, which a likelihoodist would conventionally interpret as weakly favoring $\mu = 0$ over $\mu \neq 0$.

bution with unknown mean also produces a maximum likelihood estimate of that mean that is inevitably different from its true value, simply because the data are noisy. The fact that the Law of Likelihood says that the data favor the maximum likelihood estimate over the true value in such simple cases is not an objection to the Law of Likelihood. The evidence itself is misleading. The Law of Likelihood characterizes it correctly.

The Armitage example differs from the fixed-sample-size case in that it allows one to put a lower bound on the likelihood ratio for the maximum likelihood estimate against the true hypothesis. It thus provides a recipe for producing a result that is more misleading than would be expected in the fixed-sample-size case in one respect. However, making the result more misleading in that respect requires making it more accurate in expectation in a different respect: when $\mu = 0$, increasing the lower bound on the misleading likelihood ratio the experiment produces requires decreasing the expected distance from the truth of the maximum likelihood estimate (that is, the expected value of $|\bar{x}_n|$). (See Appendix C.) If a frequentist can say that he or she is not worried about a large probability of failing to reject the null hypothesis when it is not far from the truth—as he or she must in order to avoid violating the Weak Repeated Sampling Principle—then so too a likelihoodist can claim that he or she is not worried about a large probability of finding strong evidential support for $\mu = \bar{x}_n$ against $\mu = 0$ even though $\mu = 0$ is true when one can secure a given degree of strength for that support only by allowing $\bar{x}_n$ to be very close to the truth with high probability.

The Armitage example merely allows one to trade one kind of misleadingness for another. This tradeoff is what justifies regarding the likelihood ratio as a measure of degree of misleadingness only for a pair of hypotheses that is fixed in advance. How the two kinds of misleadingness should be traded off against one another depends on the relevant utility function. Thus, there is no general, principled argument for the claim that the Armitage example is any more problematic for the Law of Likelihood than the standard fixed-sample-size case, and the latter is not problematic at all.

3.5 RESPONSE TO STEIN'S COUNTEREXAMPLE

In his [1962], Stein presents what he takes to be a counterexample to the Likelihood Principle. Advocates of the Likelihood Principle such as Berger and Wolpert [1988, 133–5] and Grossman [2011a, 311–3] accept that Stein's example illustrates a conflict between the Likelihood Principle and frequentist reasoning but argue that the problem lies with the frequentist reasoning. I argue that the purportedly frequentist reasoning used to generate the conflict is not correct by any reasonable frequentist lights, and thus that Stein's example fails to illustrate any real conflict at all.

My response to Stein's counterexample turns on an elementary point about frequentist methods. Consider the basic textbook example of using a single observation from a random variable X that is normally distributed with unknown mean θ and known variance σ^2 to produce an interval estimate of θ. A standard frequentist solution to this problem would be to use the random interval $(X - 1.96\sigma, X + 1.96\sigma)$. For instance, for the observation $X = 50$ and known variance $\sigma^2 = 1$, a frequentist would give the interval $(48.04, 51.96)$ as an estimate of θ. He or she could justify this procedure by citing the fact that $(X - 1.96\sigma, X + 1.96\sigma)$ contains the true value of θ 95% of the time in the limit of indefinitely many repeated applications with varying data, regardless of what that value may be. Moreover, this interval is the shortest among all random intervals in this problem with that property.

These properties of $(X - 1.96\sigma, X + 1.96\sigma)$ provide a kind of justification for the *method* of using that interval to estimate θ. Whether or not a *particular instance* of that interval such as $(48.04, 51.96)$ somehow "inherits" that justification is controversial among frequentists. Frequentists in the "evidentialist" (Fisherian) tradition contend that their methods do allow for epistemic appraisal of hypotheses in light of particular outcomes, while frequentists in the "reliabilist" (Neyman-Pearson) tradition do not.

Even Fisherian frequentists have to acknowledge that the coverage probability of a random variable is a good measure of the warrant for one of its particular instances only under special circumstances. They have to contend with the fact that every particular interval is an instance of countably infinitely many random intervals for the observed value of the relevant random variable, and the coverage probabilities of those random intervals vary widely. For

instance, the particular interval $(48.04, 51.96)$ is an instance not only of the standard frequentist 95% confidence interval $(X-1.96\sigma, X+1.96\sigma)$, but also of the interval $(X-1.96X/50, X+1.96X/50)$, which has coverage probability that varies with θ and is zero in the worst case of $\theta = 0$. It is also an instance of the interval $(X - 1.96\sigma, X + 1.96\sigma)I(X = 50)$, which has coverage probability 0 for all θ, and $(X - 1.96\sigma, X + 1.96\sigma)I(X = 50) + (-\infty, \infty)I(X \neq 50)$, which has coverage probability 1 for all θ. The claim that the right measure of the warrant for asserting that θ is in $(48.04, 51.96)$ is the coverage probability of $(X - 1.96\sigma, X + 1.96\sigma)$ rather than that of of some other random interval that would also generate $(48.04, 51.96)$ given $X = 50$ requires an argument.

Stein's error is to use the coverage probability of a random interval of which a given particular interval is an instance as a measure of the warrant for that particular interval when the choice of that random interval for that purpose has not been justified and at least appears to be unjustifiable.

Stein begins[16] in his example with an observed value x_0 of a random variable X that is normally distributed with unknown mean $\theta > 0$ and known variance σ_0^2. He then constructs a random variable Y such that the likelihood function of $Y = x_0$ is very nearly proportional to that of $X = x_0$. He shows that the particular interval $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ is both the instance of the random interval $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ that the observation $X = x_0$ generates and the instance of the random interval $(Y + 1.96Y/d, Y - 1.96Y/d)$ that the observation $Y = x_0$ generates, where $d = x_0/\sigma_0$. The first of these random intervals $(X-1.96\sigma_0, X+1.96\sigma_0)$ is the standard frequentist confidence interval with (approximately[17]) 95% coverage probability, while the second random interval $(Y - 1.96Y/d, Y + 1.96Y/d)$ has terrible coverage probability (less than 10^{-100}) [Berger, 1980, 154, 400–1]. It supposedly follows by frequentist lights that $X = x_0$ warrants $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ as an estimate of θ, while $Y = x_0$ does not, in violation of what we might call the Extended Likelihood Principle, according to which (roughly speaking) outcomes with nearly proportional likelihood functions are nearly evidentially equivalent.

[16] The version of Stein's argument I present actually incorporates minor refinements from Berger and Wolpert [1988, 133–4] that do not affect its substance.

[17] Because of the restriction $\theta > 0$, this interval does not have exact 95% coverage. This restriction is insignificant provided $x_0 > 3\sigma_0$ or so, which we can simply stipulate.

The assumption that the Likelihood Principle implies a version of the Extended Likelihood Principle of the kind that Stein's example requires is open to question, but I will grant it for the sake of argument.

Unfortunately for Stein, the same kind of reasoning he uses can also be used to "show" that $X = x_0$ both warrants and does not warrant $(x_0 + 1.96\sigma_0, x_0 + 1.96\sigma_0)$ as an estimate of θ. For $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ is the instance of both $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ and $(X - 1.96X/d, X + 1.96X/d)$ that $X = x_0$ generates, where $d = x_0/\sigma_0$. The former has 95% coverage probability, while the latter has coverage probability that varies with θ all the way down to 0% in the worst case ($\theta = 0$). Worse, $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ is the instance of $((X - 1.96\sigma_0)I(X = x_0), (X + 1.96\sigma_0)I(X = x_0))$ that $X = x_0$ generates, where that random interval has 0% coverage probability for all θ. It follows by the kind of purportedly frequentist reasoning Stein uses that $X = x_0$ both warrants and does not warrant $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ as an estimate of θ.

The problem with both this argument and Stein's argument is that the coverage probability of a random interval cannot be used indiscriminately as a measure of the warrant for asserting the instance of that random variable that the observed value of the random variable at issue generates. On the other hand, it does seem that frequentists must use the coverage probability of *some* random interval that generates the particular interval in question as a measure of the warrant for that interval if they are to give such a measure at all.[18] Perhaps there is a way to pick out the right random interval in each case. If $(Y - 1.96Y/d, Y + 1.96Y/d)$ were the right random interval in Stein's example, then his argument would illustrate a genuine conflict between the Likelihood Principle and frequentist reasoning.

Standard presentations of Stein's example (e.g. Berger and Wolpert 1988, 133–5) simply assume without argument that the coverage probability of $(Y - 1.96Y/d, Y + 1.96Y/d)$ is the relevant one for a frequentist assessment of the degree to which observing $Y = x_0$ warrants asserting that θ is in the interval $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0) = (x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$. But why that interval rather than $(Y - 1.96\sigma_0, Y + 1.96\sigma_0)$? For that matter, why use

[18] Of course, frequentists in the Neyman-Pearson tradition deny that there is such a thing as the degree to which an observation warrants an hypothesis. I am considering Fisherian interpretations here because they are the interpretations that apparently conflict with the Likelihood Principle here.

$(X - 1.96\sigma_0, X + 1.96\sigma_0)$ to assess the degree to which $X = x_0$ warrants asserting that θ is in the interval $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0) = (x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$, rather than $(X - 1.96X/d, X + 1.96X/d)$?

These questions are difficult to answer. Frequentism is most easily understood as a "forward-looking" theory about how to design experiments in a way that controls the probability that one will accept an erroneous conclusion, rather than as a "backward-looking" theory for assessing particular data as evidence. The most well-developed frequentist theory for interpreting data as evidence, Deborah Mayo's theory of error statistics, does not help in this case. Mayo claims that data x_0 provide good evidence for hypothesis H just in case H passes a severe test T with data x_0, which means that the following two conditions are satisfied [Mayo and Spanos, 2011, 164]:

> (S-1) x_0 accords with H (for a suitable notion of accordance) and
> (S-2) with very high probability, test T would have produced a result that accords less well with H than x_0 does, if H were false or incorrect.

This account does not help because the interval $(x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$ was chosen to fit the datum $Y = x_0$, rather than the datum being taken from a genuine test of that interval. (S-1) and (S-2) are satisfied in this case, but it is clear that the taking of the datum $Y = x_0$ should not count as a severe test of the hypothesis that θ is in the interval $(x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$ that was designated after the fact. Either Mayo's theory is false, or this case lies outside its scope because of it egregiously violates the standard frequentist requirement to predesignate hypotheses for testing.[19]

Perhaps, from a frequentist perspective, the right measure of the warrant for asserting that θ is within a particular interval is the coverage probability of the random interval from which that particular interval came or would have come. For instance, it seems to make sense to use 95% as a measure of the warrant for asserting that θ is within $(x_0 - 1.96\sigma_0, x_0 + 1.96\sigma_0)$ upon observing $X = x_0$ because a frequentist actually would use a random interval of the form $(X - k\sigma_0, X + k\sigma_0)$ to estimate θ from X, with $k = 1.96$ giving coverage probability of 95%. If this proposal is correct, then an advocate of Stein's argument needs to argue that

[19] Mayo denies that predesignation is always strictly necessary for any frequentist evaluation [Mayo, 1996, Ch. 9] [Mayo and Spanos, 2011, 164], but not that absence of predesignation is ever problematic, as it clearly would be for her in this case.

a frequentist actually would use an interval of the form $(Y - kY/d, Y + kY/d)$ in order to establish that from a frequentist perspective it is appropriate to use the coverage probability of $(Y - 1.96Y/d, Y + 1.96Y/d)$ as a measure of the warrant for asserting that θ is within $(x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$ upon observing $Y = x_0$. Now, $(Y - 1.96Y/d, Y + 1.96Y/d)$ is of the form (aY, bY), which is convenient because intervals of that form have coverage probability that does not depend on θ. However, it is also of the more specific form $((1 - 1.96a)Y, (1+1.96a)Y)$, which is far from optimal in terms of maximizing coverage probability for a given width because Y is almost certain to be at least ten times larger than θ [Basu, 1975, 52]. For that reason, a frequentist would not use an interval of that form: it would be better to use an interval of the form (aY, bY) with $b > a \gg 1$. Thus, the proposal that the right measure of the warrant for asserting that θ is within $(x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$ upon observing $Y = x_0$ is the coverage probability of the random interval from which that particular interval came or would have come implies that the coverage probability of $(Y - 1.96Y/d, Y + 1.96Y/d)$ is not the right measure.

One might think that Stein showed that an advocate of the Likelihood Principle who is committed to using $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ as an estimator of θ for all values of X is thereby also committed to using $(Y - 1.96Y/d, Y + 1.96Y/d)$ for all values of Y. The terrible coverage probability of $(Y - 1.96Y/d, Y + 1.96Y/d)$ is a good reason not to be committed to using it for all values of Y, so if Stein had demonstrated this result, then an advocate of the Likelihood Principle would have to argue that it would be a mistake be to committed to using $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ for all values of X. Berger and Wolpert [1988, 133–5] and Grossman [2011a, 311–3] make strong arguments for this claim,[20] but those arguments are not needed because Stein did not establish the result that would make its conclusion necessary. His argument involves constructing a *different* random variable Y for each observed value x_0 of X. Thus, while an advocate of the Likelihood Principle is committed to saying that $X = x_0$ warrants the claim that θ is within $(x_0 - 1.96x_0/d, x_0 + 1.96x_0/d)$ to the same degree as observing $Y = x_0$, he or she is not committed to that claim for all values x_0 of X and a

[20] No proper prior probability distribution gives $(x - 1.96\sigma_0, x + 1.96\sigma_0)$ as the Bayesian highest posterior density region for all possible values x of X. The improper uniform prior over all $\theta > 0$ does give approximately $(x - 1.96\sigma_0, x + 1.96\sigma_0)$ for all values x of X that are not too close to zero, but the use of this improper prior is highly suspect in this case for reasons Berger [1980, 154–5] explains.

single random variable Y. (See Appendix D for details.)

Advocates of the Likelihood Principle have argued that Stein's example is not fatal even though it shows that the Likelihood Principle conflicts with frequentist reasoning. If the argument presented here is correct, then their arguments are unnecessary because Stein's example does not in fact illustrate any conflict between likelihoodism and any reasonable form of frequentism.

3.6 CONCLUSION

I have provided new responses to three purported counterexamples to the Likelihood Principle. Together with the proof of the Likelihood Principle presented in the previous chapter, these responses strengthen the case for a likelihoodist or Bayesian approach rather than a frequentist approach to statistical inference. In the next chapter, I present a new counterexample to the Law of Likelihood that I argue does require modifying or reinterpreting it. However, the blame for the problem it illustrates lies with the theory of conditional probability rather than the connections between conditional probabilities and evidence the the Likelihood Principle and Law of Likelihood posit.

4.0 A COUNTEREXAMPLE TO THE LAW OF LIKELIHOOD FOR PROBABILITY-ZERO HYPOTHESES

4.1 INTRODUCTION

The Law of Likelihood says that a datum E evidentially favors a hypothesis H_1 over an incompatible[1] hypothesis H_2 if and only if the likelihood ratio $k = \Pr(E|H_1)/\Pr(E|H_2)$ is greater than one, with k measuring the degree of favoring. This claim is problematic when either H_1 or H_2 has probability zero. When H has probability zero, $\Pr(E|H)$ is defined in Kolmogorov's theory of regular conditional distributions only relative to a sub-σ field in which H is embedded. One can embed a pair of hypotheses H_1 and H_2 in a sub-σ-field that treats them differently even though the setup of the problem is symmetric with respect to those hypotheses. Doing so gives rise to instances of Borel's paradox,[2] in which probabilities conditional on probability-zero hypotheses that intuitively should be equal turn out to be unequal. In this way, one can produce counterexamples to the Law of Likelihood in which $\Pr(E|H_1)/\Pr(E|H_2) > 1$ even though the setup of the problem is symmetric with respect to H_1 and H_2.

I present a counterexample of that kind in Section 4.2. In Section 4.3, I respond to several attempts to dismiss the counterexample. I then present a no-go result in Section 4.4 which shows that avoiding such counterexamples by adopting an alternative theory of conditional probability requires giving up either even a weak form of conglomerability or

[1] Standard formulations of the Law of Likelihood do not require that H_1 and H_2 be incompatible. I argue for this requirement in Chapter 2, Section 2.

[2] "Borel's paradox" has also been called "Bertrand's Paradox" because Joseph Bertrand seems to have been the first to discuss it [Bertrand, 1889]. Émile Borel responds to Bertrand's discussion in his [1909, 100–4]. It is also sometimes called "the Borel–Kolmogorov paradox" because Kolmogorov gives the canonical response to it in his [1956, 50–1].

one of two other highly plausible principles. I then present two possible responses to the counterexample within extant theory of conditional probability in Section 4.5. In the end, I conclude that either of those responses is acceptable and that there are no strong grounds on which to decide between them.

It should be noted that Borel's paradox only arises for hypotheses that have probability zero. When H has strictly positive probability, the formula $\Pr(E|H) = \Pr(E \cap H)/\Pr(H)$ holds, and the values of these probabilities are not relative to a sub-σ-field. Thus, Borel's paradox does not create any difficulties for the Law of Likelihood in many ordinary cases.

4.2 A COUNTEREXAMPLE TO THE LAW OF LIKELIHOOD FOR PROBABILITY-ZERO HYPOTHESES

In the following example, standard mathematical techniques yield $\Pr(E|H_1)/\Pr(E|H_2) > 1$, but it is clear that the datum E is evidentially neutral between H_1 and H_2. Thus, it is a counterexample to at least a naïve formulation of the Law of Likelihood.

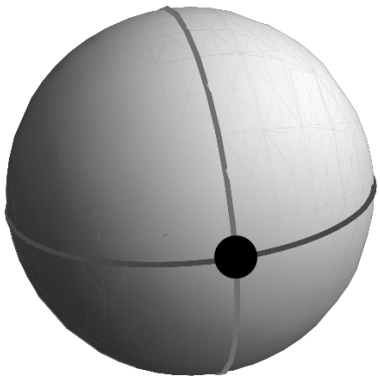

Figure 4.1: An illustration of Example 4.1

Example 4.1. Consider a unit sphere equipped with an arbitrary system of latitudes and longitudes. Let E be the datum that a point P randomly selected from a uniform distribution on the surface of the sphere lies within $\pi/180$ units of the intersection of the equator and the prime meridian along the surface of the sphere. Let H_1 be the

hypothesis that P lies on the "prime meridional circle," i.e. union of the prime meridian and the line of 180° longitude, omitting the points at which that circle intersects the equator. Let H_2 be the hypothesis that P lies on the equator, omitting the same two points.

By stipulation, the system of latitudes and longitudes in this example is arbitrary. Moreover, the datum E is symmetric with respect to the equator and the prime meridional circle. Thus, it is intuitively clear that E is evidentially neutral between H_1 and H_2. But standard mathematical techniques yield the likelihood ratio $\Pr(E|H_1)/\Pr(E|H_2) = 1.57$, which according to the Law of Likelihood indicates that E favors H_1 over H_2. Thus, an unrestricted version of the Law of Likelihood is false when the conditional probabilities to which it appeals are calculated using standard mathematical techniques.

See Appendix E for a derivation of the anomalous likelihood ratio. You can see roughly how this result arises by thinking of the prime meridional circle as the limit of the region between two lines of longitude and thinking of the equator as the limit of the region between two lines of latitude: two lines of longitude are farther apart close to the equator than they are around the poles, while any two lines of latitude are equally spaced all the way around the globe. This difference corresponds to the fact that, using standard mathematical techniques, the conditional probability distribution on the equator is uniform, while that on the prime meridional circle increases as one approaches the equator. (See figure 4.2.)

Treating a great circle as a limit for the purpose of calculating probabilities conditional on it makes sense when one is conditioning on the datum that some point of interest lies on it to within the resolving power of some measurement technique. It does not make sense in Example 4.1, in which the circles correspond to sharp hypotheses and the identification of one as the equator and the other as the prime meridional circle is arbitrary.

4.3 RESPONSES TO WORRIES ABOUT THE COUNTEREXAMPLE

There are several apparent reasons to think that Example 4.1 is somehow illegitimate and thus not a proper counterexample to the Law of Likelihood as standardly formulated. In this

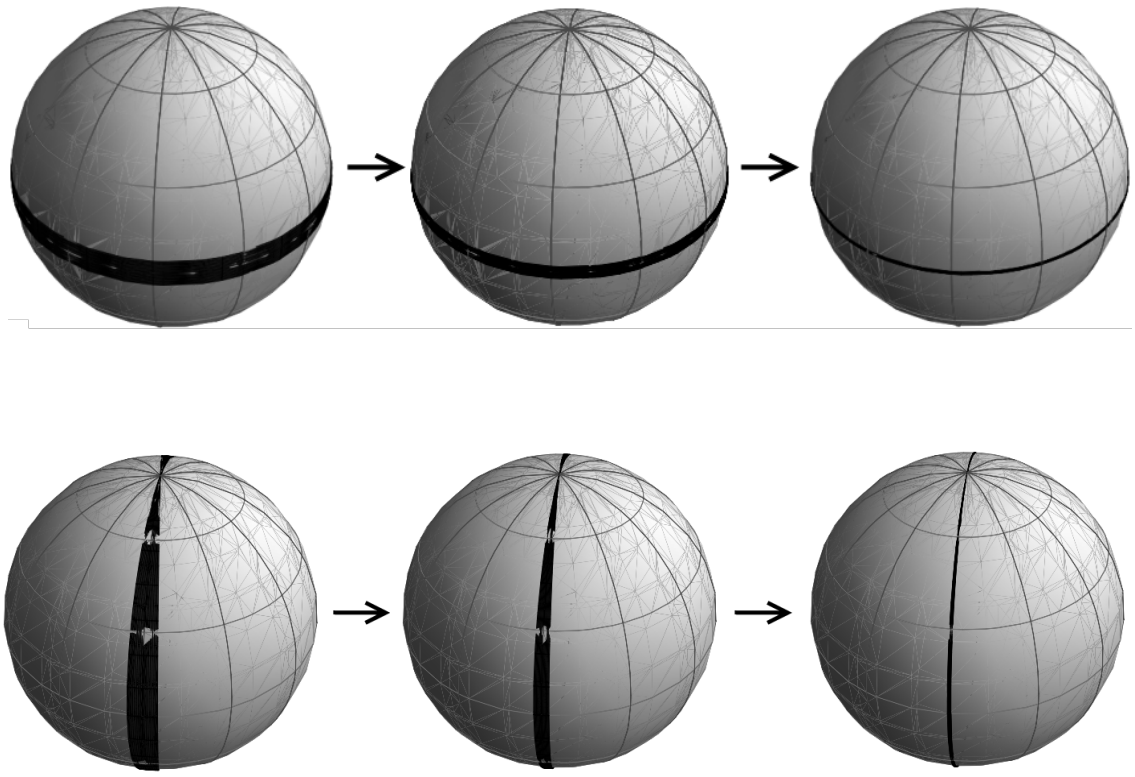

Figure 4.2: Equator and prime meridional circle are treated differently despite the arbitrariness of the distinction.

section I will consider five of these concerns and argue that none of them defuses the counterexample. This is not to say that likelihoodists have no response to the counterexample, but only that they cannot simply dismiss it without providing a proper response.

4.3.1 Worry 1: Each hypothesis omits two points on the relevant great circle

It might seem suspicious that H_1 and H_2 in Example 4.1 each omit two points from the relevant great circle. However, the omission of those points plays no role in generating the anomalous likelihood ratio, as the calculation in Appendix E shows. Because those points have no length, the same likelihood ratio arises regardless of how they are handled.

I exclude the points at which the relevant great circles intersect only so that one cannot avoid the counterexample by restricting the Law of Likelihood to mutually exclusive hypotheses, as I advocate in Chapter 3.

4.3.2 Worry 2: The likelihood ratio is small

The value 1.57 of the anomalous likelihood ratio is significantly smaller than the value of 8 that likelihoodists conventionally require in order to declare a result "fairly strong" evidence [Royall, 2000, 761]. One might think that this fact somehow excuses the Law of Likelihood. But it does not, for two reasons. First, the Law of Likelihood implies the incorrect qualitative claim that the result favors the prime meridional circle hypothesis over the equator hypothesis regardless of what it says about the degree of favoring. Second, one could produce an analogous but more dramatic result by using a sub-σ-field that corresponds to a strange coordinate system. As I explain above, the fact that likelihood ratio in this case is greater than one corresponds to the fact that lines of longitude are farther apart at the equator than near the poles, while lines of latitude are spaced equally all the way around the sphere. One could get a larger likelihood ratio by using a system of "pseudo-longitudes" that exaggerates this effect around the prime meridian and/or a system of "pseudo-latitudes" that are closer together near the prime meridional circle than elsewhere around the equator. It is a plausible conjecture that one could generate any likelihood ratio one likes by using a sufficiently strange coordinate system; see [Arnold and Robertson, 2003] for a similar result

in a related problem.

4.3.3 Worry 3: Real measurement techniques have finite precision

Borel himself points out that actual methods of observation do not allow you to learn that a particular point on a sphere lies on a particular great circle [1909, 102–3]. From a position on the prime meridian, for instance, you might be able to use astronomical observations and a chronometer to determine that your longitude is between 0.1″ East and 0.1″ West, but you would never be able to determine that your longitude is exactly 0.

This point would be relevant if $\Pr(E|H_1)$ were supposed to reflect uncertainty about the the longitude of P given a *measurement* with a high but finite degree of precision indicating that it lay on the equator, and $\Pr(E|H_2)$ were supposed to reflect uncertainty about the the latitude of P given a measurement with a high but finite degree of precision indicating that it lay on the prime meridional circle. But those probabilities are supposed to reflect uncertainty given *hypotheses*, not measurements. The fact that measurement techniques have finite precision does not preclude considering "sharp" hypotheses.

4.3.4 Worry 4: The hypotheses belong to different models

R.A. Fisher [1922, 310] and A.W.F. Edwards [1972, 9], among others, characterize a likelihood function as defined only up to a constant of proportionality. This fact might lead one to think that the Law of Likelihood applies only to hypotheses that belong to a common model. If two hypotheses belong to a common model, then the constant of proportionality associated with their likelihoods will be the same and thus will cancel out when one takes their ratio. If they belong to different models, then there will be no determinate fact about how their constants are related to one another. Thus, there will be no determinate fact about their likelihood ratio. H_1 and H_2 in Example 4.1 seem to belong to different models, one of which carves up the sphere into meridional circles while the other carves it up into zonal circles, so one might think that restricting the Law of Likelihood to hypotheses that belong to a common model is both independently well-motivated and sufficient to avoid the counterexample.

This objection is dubious for two main reasons. First, it is not clear what it means to say that two hypotheses belong to a "common model." In statistics jargon, "model" typically refers to a triple of the form $\{\mathcal{X}, \Theta, \mathbf{P}\}$, where $\mathcal{X}$ is a sample space, Θ is an index set, and $\mathbf{P}$ is a set of probability distributions indexed by the elements of Θ. H_1 and H_2 cannot belong to such a model because they are neither elements of a sample space, nor indices, nor probability distributions. We could thus avoid them by saying that we can apply the Law of Likelihood only to hypotheses that belong to a model of this kind (presumably by positing that the statistical distribution underlying the data belongs to some particular subset of $\mathbf{P}$), but this maneuver excludes too much. It would prevent us from applying the Law of Likelihood to substantive scientific hypotheses such as the theory of evolution, as Sober [2008] does to good effect. We might want to disallow the use of the Law of Likelihood in some such cases because it is difficult at best to assign values to the relevant likelihoods in an objective way, but this point clearly does not apply to all substantive hypotheses. For instance, there is no question that the probability that a card drawn from a standard, well-shuffled deck is a spade given that it is black king is 1/2 while the probability that is is a spade that it is a queen is 1/4, and it seems quite reasonable to say for that reason that the information that the card is a spade favors the hypothesis that it is a black king over the hypothesis that it is a queen.

Even when we leave aside this difficulty by focusing on hypotheses that simply posit probability distributions, it is still not clear under what circumstances a pair of hypotheses should be regarded as belonging to a common model. Consider a paradigm example of hypotheses that come from different models, intuitively speaking—for instance, the hypothesis that an outcome of five heads in ten tosses was produced by a fair coin under binomial sampling—meaning that the number of tosses was fixed and the number of heads random—or under negative binomial sampling—meaning that the number of heads was fixed and the number of tosses random. Binomial and negative binomial sampling are standard examples of distinct models, so if taking hypotheses from distinct models is problematic in general than it should be problematic here.

However, it is unclear what principled distinction one can draw between this pair of hypotheses and, for instance, two hypotheses within the "binomial sampling model" that

posit different biases for the coin. At first it might seem that this pair of hypotheses must be regarded as belonging to different models because they correspond to different sample spaces: the binomial hypothesis says that the possible observations are $\{H = 0, H = 1, \ldots, H = 10\}$, where H is the number of heads, with the number of flips $n = 10$ being a fixed quantity, while the negative binomial hypothesis says that the possible observations are $\{N = 5, N = 6, \ldots\}$, where N is the number of flips, with the number of heads $h = 5$ a fixed quantity. However, we can give these hypotheses a common sample space simply by expressing the sample points in each sample space in terms of both the number of heads and the number of flips and then taking the union of the results. The fact that each element of this sample space is characterized by two numbers is no objection: we can project this two-dimensional characterization onto one dimension if we like, given that the cardinality of $\mathbb{N} \times \mathbb{N}$ is the same as that of $\mathbb{N}$ itself. Each hypothesis assigns probability zero to some elements of this combined sample space, but that is no obstacle to treating it as a sample space.

In general, if hypothesis H_1 says that some random vector Y has probability density function $f_1(Y)$ and another hypothesis H_2 says that Z has probability density function $f_2(X)$, where the union of the sample spaces of X and Y is a partition, then those hypotheses belong to a common model according to which $f(X) = I(b = 1)f_1(X) + I(b = 2)f_2(X)$, where X ranges over the union of the sample spaces, I is an indicator function, and $b = i$ if and only if hypothesis H_i is true. This kind of approach might seem artificial, but it would be appropriate, for instance, in the event that one were given the number of heads and the number of tosses and wished to evaluate the hypothesis that sampling was binomial against the hypothesis that it was negative binomial. I see no reason why the Law of Likelihood could not be applied in this case, despite the fact that the hypotheses come from "different models," intuitively speaking.

That brings me to my second reason for finding dubious the objection that Example 4.1 is illegitimate because H_1 and H_2 belong to different models: there do not seem to be any good arguments for denying that the Law of Likelihood can be applied to hypotheses from different models. I gave an argument for this claim at the start of this section, but it relies on the idea that a likelihood function should is defined only up to a model-specific constant of proportionality. There does not seem to be any argument behind that idea beyond an

appeal to the authority of figures such as Fisher. There are reasons to accept the assertion of the Likelihood Principle that evidential import depends on the likelihood function only up to a constant of proportionality, but this fact does not require regarding likelihood functions as *defined* only up to such a constant. One can simply say that the mapping from likelihood functions to evidential import is many-to-one. At most, the Likelihood Principle makes it *convenient* to regard the likelihood function as defined only up to a constant of proportionality. It would be a mistake to derive methodological conclusions from this fact.

The possibility of using data on coin flips to evaluate hypotheses involving binomial sampling against negative binomial sampling as described above might seem to generate objections to the idea that the Law of Likelihood can be applied to hypotheses drawn from different models.[3] For it might seem that a datum E that a coin landed heads h times in n flips contains no information about whether the number of heads or the number of tosses was fixed in advance. But applying the Law of Likelihood yields the conclusion that E favors the hypothesis H_1 that the bias of the coin is p_0 and the number of heads is fixed at n over the hypothesis H_2 that that the bias of the coin is p_0 and the number of heads is fixed at h to the degree n/h for any p_0: $\frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{\binom{n}{h}p^h(1-p)^{n-h}}{\binom{n-1}{h-1}p^h(1-p)^{n-h}} = \frac{n}{h}$. Because this relationship holds for all p_0, it presumably follows that E favors the more general "binomial" hypothesis H_B that the number of heads is fixed at n over the more general "negative binomial" hypothesis H_{NB} that the number of heads is fixed at h.[4] But then it seems that we are not just treating the uninformative observation E as if it were informative, but doing so in a way that is biased toward the conclusion that sampling is binomial.

These concerns are misplaced. First, while it might seem intuitively plausible that E is uninformative, this claim is not obviously true. Consider an extreme case, such as $n = 1000$ and $h = 1$. There are 1000 ways this result could have arisen if $n = 1000$ was fixed in advance: the head could have occurred in the first toss, the second toss, ..., or the thousandth toss. But there is only one way it could have arisen if $h = 1$ was fixed in advance: the head must have occurred on the thousandth toss. Given these considerations, it seem plausible that

[3]This objection was presented to me by Michael Lew in personal correspondence.

[4]This claim can be shown to follow if there is a probability distribution over the bias of the coin, regardless of what that distribution is, and it seems eminently plausible even in the absence of a definite probability distribution.

the datum $n = 1000, h = 1$ does in fact favor the hypothesis that $n = 1000$ was fixed over the hypothesis that $h = 1$ was fixed to the degree $n/h = 1000$, and this result accords with the fact that the odds of the former to the latter would increase one-thousandfold under Bayesian updating on that datum.

Second, the Law of Likelihood does not say that the fact that a coin landed heads h times in n flips favors the hypothesis of binomial sampling over the hypothesis of negative binomial sampling to the degree n/h; it says that it favors the hypothesis of binomial sampling *with the number of flips fixed at the actually observed number* n over the hypothesis of negative binomial sampling *with the number of heads fixed at the actually observed number* h to that degree. Thus, there is no pair of particular, fixed hypotheses such that the Law of Likelihood says that any set of coin flip outcomes favors the one over the other, as the charge of bias would require. It is somewhat curious and surprising that the particular binomial hypothesis that is compatible with the data is always favored over the particular negative binomial hypothesis that is compatible with the data, but this result is vindicated by Bayesian reasoning and the kind of combinatorial reasoning used above in the case of $n = 1000, h = 1$.

4.3.5 Worry 5: No limiting operation is specified

Jaynes [2003, 469–70] argues that the problem described in Example 4.1 is ill-formed. When we have a probability density on one space and wish to generate from it a density on a subspace of measure zero, we have to think of the subspace as the limit of a sequence of positive-measure subspaces and specify how that limit is to be approached. Jaynes goes on to claim that the term "great circle" is ambiguous until a limiting operation is specified; a great circle approached through the "equatorial limit'"is not the same object as a great circle approached through the "meridional limit."

The claim that "great circle" is ambiguous seems to go too far. After all, one can characterize a great circle in terms of the points that comprise it without reference to a limiting operation. However, the claim that the problem is ill-formed is more plausible and does not obviously require that "great circle" be ambiguous. The key point for the purposes

of this chapter is that Borel's paradox cases create puzzles for the Law of Likelihood even if this claim is granted, because it entails that either the Law of Likelihood does not apply to such cases or that evidential favoring itself is relative to a limiting operation. In this respect, it fares no better than Kolmogorov's theory that conditional probabilities are relative to a sub-σ-field.

4.4 A NO-GO RESULT FOR ADDRESSING THE COUNTEREXAMPLE THROUGH AN ALTERNATIVE THEORY OF CONDITIONAL PROBABILITY

Maintaining that the Law of Likelihood holds and that evidential favoring is absolute and determinate in the kinds of cases we are considering requires departing from Kolmogorov's theory of regular conditional distributions, according to which the relevant conditional probabilities are relative to a sub-σ-field. There are several alternatives to Kolmogorov's approach, including theories due to Popper [1968], Dubins [1975], Jaynes [2003], Rényi [1955], and Rao [2005]. Unfortunately, none of those theories determines the result that $\Pr(E|H_1)/\Pr(E|H_2) = 1$, and thus that E is evidentially neutral between H_1 and H_2 according to the unrestricted Law of Likelihood.

Existing theories of conditional probability fall into three categories with respect to how they handle conditioning on probability-zero hypotheses:

1. **Relativizing.** Theories in this category determine probabilities conditional on hypotheses of probability zero only relative to something beyond the standard $\{\mathcal{X}, \Theta, \mathbf{P}\}$ model. They include Kolmogorov's theory [1956], which appeals to sub-σ-fields, and Jaynes's theory [2003], which appeals to limiting operations.
2. **Underdetermining.** Theories in this category provide axioms that constrain possible conditional probability assignments but do not determine them, even relative to a structure such as a sub-σ-field. Dubins's "coherent conditional distributions" [1975] and Popper's "Popper functions" fall into this category.

3. **Misdetermining.** Theories in this category provide particular, non-relative probability assignments in a way that fails to avoid counterexamples to the Law of Likelihood. Rényi's theory falls into this category.[5]

Given that there is no extant theory of conditional probability that determines the desired result, someone who wishes to maintain that the Law of Likelihood holds and that evidential favoring is absolute can either opt for a theory such as Popper's or Dubins's that at least permits that result or maintain that some new theory of conditional probability that determines it is needed. Unfortunately, any theory that even permits the desired result in a class of cases like Example 4.1 must violate at least one of three highly intuitive principles. It is far from obvious whether it is better to violate one or more of those principles or to permit the desired results and thus avoid counterexamples to an unrestricted and absolute Law of Likelihood, but we can at least say that there appears to be a significant cost to avoid such counterexamples.

Any theory of conditional probability that permits one to avoid counterexamples to the Law of Likelihood like Example 4.1 is consistent with the following principle.

> **Rotation Preservation.** Let R be a region of the sphere discussed in Example 4.1 that is invariant under arbitrary rotations of the sphere about some axis X. Let $\mathbf{G}$ be the set of great circles that passes through X. It is permitted that $\Pr(R|G) = a$ for all great circles $G \in \mathbf{G}$ and some constant a.

This principle says, roughly, that it is permissible to have the same probability that the point P that was randomly selected from a uniform distribution on the surface of the sphere lies on a given great circle for each great circle that passes through a particular point, given that P is in a region that is invariant under rotations about the central axis that passes through that point. Conformity to this principle is not *sufficient* to avoid counterexamples to the Law of Likelihood arising from Borel's paradox: one could replace the circular region P

[5]The core of Rényi's theory does not apply to conditioning on measure-zero sets and thus does not assign values to $\Pr(E|H_1)$ and $\Pr(E|H_2)$. However, Rényi extends that core to produce the non-uniform conditional distributions over meridians. Rao [2005, 118] claims that the extension Rényi gives is "tailored the the problem at hand" and that it is not clear how to generalize it. In any case, it cannot give the result $\Pr(E_m|H_m)/\Pr(E_m|H_2) = 1$ for all hypotheses H_m according to which P lies on a meridian and all data E_m according to which P lies within one degree latitude and one degree longitude of the intersection of that meridian with the equator. On Rényi's approach, $\Pr(E_m|H_m) = \sin \pi/180$ for all such hypotheses and data. $\Pr(E_m|H_2$ cannot equal $\sin \pi/180$ for all such hypotheses, because if it did then by finite additivity it would follow that $\Pr(H_2|H_2) = 360 \sin \pi/180 = 6.28 > 1$.

in Example 4.1, for instance, with a square region which is not invariant under rotations and thus is not addressed by Rotation Preservation. However, conformity to Rotation Preservation is *necessary* for avoiding counterexamples. For suppose that $\Pr(R|G_1) \neq \Pr(R|G_2)$ for some great circles G_1 and G_2 that pass through a point such that the region R is invariant under rotations of the sphere around the central axis that passes through that point. Then, according to the Law of Likelihood, learning that P lies in R favors G_1 over G_2 or vice versa (removing the points at which these circles intersect so that they are disjoint).[6] This result is incorrect by the same kind of symmetry reasoning that indicates that the result that E favors H_1 over H_2 in Example 4.1 is incorrect, so any instance of it counts as a counterexample to the Law of Likelihood understood in accordance with a theory of conditional probability that generates it.

Unfortunately, any theory of conditional probability that is consistent with Rotation Preservation is inconsistent with the conjunction of the following three highly plausible additional principles (with a minor caveat that is described below).

> **Weakening.** If $\Pr(A) > 0$, $\Pr(B) > 0$, and $A \cap B = \emptyset$, then $\Pr(A \cup B) > \max\{\Pr(A), \Pr(B)\}$.

Weakening is a consequence of the axiom of finite additivity, according to which $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ provided that $A \cap B = \emptyset$. (It also follows from the weaker principle of superadditivity according to which $\Pr(A \cup B)$ is at least as great as the sum of $\Pr(A)$ and $\Pr(B)$, given $A \cap B = \emptyset$.) Unlike the stronger principle of countable additivity, finite additivity is essentially uncontested.

> **Intersection Identity.** If $A \cap E = B \cap E$, then $\Pr(A|E) = \Pr(B|E)$.

This principle comes from Easwaran [2008, 81]. It follows from the identity $\Pr(C|E) = \Pr(C \cap E)/\Pr(E)$ in the case in which $\Pr(E) > 0$. It also seems highly intuitive even in the general case: given that A and B are the same "within E," so to speak, they should be assigned the same probability on the assumption that E holds.

[6] I am assuming that $\Pr(R|G_1)$ and $\Pr(R|G_2)$ are both real-valued. If they are not, then the Law of Likelihood does not apply, so we still have a restriction on the scope of the Law of Likelihood of a kind that we are currently trying to avoid.

Weak Conglomerability. If there is a constant a such that $\Pr(A|E) = a$ for all $E \in \mathcal{E}$, where $\mathcal{E}$ is a partition, then $\Pr(A) = a$.

This principle follows from the law of total probability $\Pr(C) = \sum_i \Pr(C|E_i)\Pr(E_i)$ when the relevant partition is countable and is indexed by i. It also seems to follow from any reasonable interpretation of probability in the general case. For instance, if $\Pr(C|E_i)$ corresponds to the degree to which one would (or should) believe C upon learning E_i then it amounts to the statement that if one would (or should) believe C to a particular degree regardless of which element of a partition one learned, then one should already believe C to that degree.

Advocates of Dubins's theory of coherent conditional probability reject Weak Conglomerability despite its intuitive appeal, regarding its rejection as the price to be paid for avoiding Borel's paradox and other difficulties for the Kolmogorov approach. I do not wish to take a side in this debate; I only wish to point out that avoiding Borel's paradox does not come without a cost.

An argument from [Easwaran, 2008, 84–5] can be adapted to show that Rotation Preservation is incompatible with the conjunction of Weakening, Intersection Identity, and Weak Conglomerability, with the minor caveat that at least one of those principles must be very slightly strengthened to accommodate the fact that all of the great circles that pass through a particular point contain that point and its opposite and thus do not form a partition. One simple and intuitively plausible way to address this complication is to maintain (quite plausibly) that Weak Conglomerability holds for "almost-partitions" the elements of which have a finite number of points in common, provided that the probability distribution over the space is continuous (and thus puts no probability mass on those points). Other options are to arbitrarily assign the two points in question to particular great circles or to remove the relevant points from the sphere and make appropriate slight adjustments to Rotation Preservation. How exactly we handle these points is not crucial: intuitively, it should not matter how we handle two extensionless points on a continuous surface in the presence of a continuous probability distribution. For the sake of definiteness, I will opt for the slightly extended version of Weak Conglomerability.

Here is the argument. Returning to Example 4.1, let A be the region of points within d

degrees of the poles in some system of latitudes and longitudes for some $0 < d < 90$. Let $\mathcal{G}_1$ be an "almost-partition" of the sphere into the great circles that pass through the poles. A is invariant under rotations of the sphere about the axis passing through the poles, so Rotation Preservation says that it is permitted that $\Pr(A|G) = a$ for some a and for each hypothesis G corresponding to an element of $\mathcal{G}_1$. By Weak Conglomerability for "almost-partitions," it follows that $\Pr(A) = a$. (See figure 4.4)

Figure 4.3: The application of Weak Conglomerability to A.

Now imagine a rotation of the sphere around some axis perpendicular to the one just discussed, and consider the band A' swept out by A. Let $\mathcal{G}_2$ be the "almost-partition" of the sphere into great circles passing through the endpoints of this second axis. Take the unique great circle G^* that is in both $\mathcal{G}_1$ and $\mathcal{G}_2$. $A \cap G^* = A' \cap G^*$, so Intersection Identity requires $\Pr(A|G^*) = \Pr(A'|G^*)$. By Rotation Preservation, it is permitted that that $\Pr(A'|G) = a$ for all $G \in \mathcal{G}_2$. By Weak Conglomerability for "almost-partitions," it follows that $\Pr(A') = a$. Thus, it is permitted that $\Pr(A) = \Pr(A')$.

But $\Pr(A) = \Pr(A')$ is incompatible with Weakening. The region $A^\dagger = A' \setminus A$ has positive area, so it has positive probability. $A' = A^\dagger \cup A$, and $A^\dagger \cap A = \emptyset$, so $\Pr(A') = \Pr(A^\dagger \cup A) > \Pr(A)$.

This no-go result shows that anyone who wishes to maintain an absolute and unrestricted Law of Likelihood must deny at least one of Weakening, Intersection Identity, and Weak Conglomerability for "almost-partitions." Denying Weak Conglomerability seems to be the most promising option, given that advocates of the Dubins theory of conditional probability

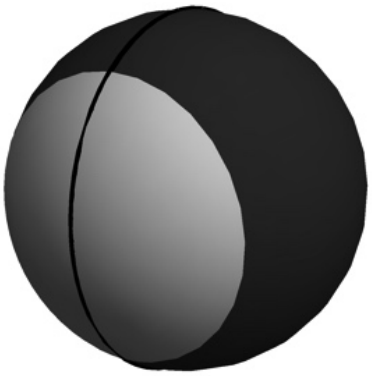

Figure 4.4: The application of Weak Conglomerability to A'.

are willing to deny it on other grounds [Dubins, 1975]. However, it is a fairly radical step. Moreover, it is still not enough to *determine* the intuitively correct result in Example 4.1 that E is evidentially neutral between H_1 and H_2: the Dubins theory, for instance, permits many different conditional probability assignments, including the assignment that yields the intuitively incorrect 1.57 likelihood ratio.

A full discussion of the relative merits of various theories of conditional probability lies beyond the scope of this dissertation; see for instance [Seidenfeld, 2001], [Hájek, 2003], [Easwaran, 2008], and [Rescorla, 2014].

4.5 ADDRESSING THE COUNTEREXAMPLE WITHIN EXISTING THEORIES OF CONDITIONAL PROBABILITY

I showed in the previous section that any attempt to address counterexamples to the Law of Likelihood arising from Borel's paradox by adopting an alternative to Kolmogorov's theory of conditional probability will require violating at least one of three highly plausible principles. One could maintain that we should give up one of those principles rather than accept either restrictions on the Law of Likelihood or the view that facts about evidential favoring are either indeterminate or relative to something not contained in the standard statistical model

in the kinds of cases that give rise to Borel's paradox. However, in this section I argue that the cost of accepting either of the latter for a likelihoodist is slight. I will consider first the possibility of restricting the Law of Likelihood, and then the possibility of regarding evidential favoring as either indeterminate or relative.

4.5.1 Restricting the Law of Likelihood

One possible response to Example 4.1 is to restrict the Law of Likelihood so that it does not apply to the kinds of cases that gives rise to Borel's paradox. This response is obviously sufficient to avoid those examples as counterexamples, but it faces at least two objections: it seems *ad hoc*, and it limits the scope of the Law of Likelihood. I will address these objections in turn.

There are at least two reasons to worry about an apparently *ad hoc* restriction on a principle. First, it might make one wonder where arguments for an unrestricted version of the principle went wrong. Second, it might make one worry that perhaps there are other possible counterexamples that the restriction fails to address.

Both of these worries are mitigated by the fact that counterexamples arising from Borel's paradox have as their source an acknowledged difficulty in the theory of conditional probability. Thus, they do not reveal a problem in the relationship between conditional probability and evidential favoring that the Law of Likelihood posits when the conditional probabilities are unproblematic. This fact mitigates the first worry: arguments for the Law of Likelihood go wrong in ignoring the possibility of anomalous conditional probabilities. In considering the premises of those arguments, one naturally thinks of cases in which the conditional probabilities are unproblematic, and counterexamples arising from Borel's paradox give no reason to think that the arguments are wrong in those cases. It also mitigates the second worry: there may be other anomalies in the theory of conditional probability that could be exploited to produce other counterexamples—cases involving self-locating belief come to mind—but this possibility is no threat to use of the Law of Likelihood in typical cases.

The non-*ad hoc* restriction we really need is that, roughly, the Law of Likelihood applies only when the relevant conditional probabilities are unproblematic. It is unclear how exactly

this restriction should be spelled out. For instance, some likelihoodists such as Edwards [1972] say that the Law of Likelihood should be used only when the relevant conditional probabilities express known long-run frequencies, while others such as Sober [2008] require only that they be empirically well-supported. In any case, saying that it should not be used in the kinds of cases that give rise to Borel's paradox seems reasonable and is non-*ad hoc* in that it can be seen as an instance of a more general and principled restriction, albeit one that is difficult to specify completely.

One might worry a vague restriction to the effect that the Law of Likelihood does not apply when conditional probabilities are problematic gives likelihoodists too much latitude: when faced with a potential counterexample, he or she could always simply claim that the conditional probabilities are at fault rather than the Law itself. However, this worry should not be oversold. We have a fair amount of independent purchase on the question of whether or not a given conditional probability assignment is problematic. As a result, the claim that conditional probabilities are to blame would be a complete non-starter as a response to many purported counterexamples to the Law of Likelihood, including those addressed in Chapter 3. In the case at hand, the conditional probabilities in question have been regarded as paradoxical since Bertrand [1889] first investigated them and have provided some of the impetus for the development of alternatives to Kolmogorov's theory of conditional distributions.

In sum, restricting the Law of Likelihood so that it does not apply to the kinds of cases that give rise to Borel's paradox is not badly *ad hoc* because it can be seen as part of specifying a well-motivated general restriction of the Law to cases in which the relevant conditional probabilities are unproblematic. Moreover, it does not raise the kinds of worries that *ad hoc* restrictions sometimes raise which would make one worry about the correctness of the Law even in ordinary cases.

A second reason to worry about a restriction on the Law of Likelihood is that it might substantially reduce the scope of the principle and thereby limit its usefulness. If the restriction is necessary, then resistance to it on these grounds is futile. However, it does seem to give the likelihoodist a reason at least to search for other options.

I have two responses to this concern. First, it is not clear that the likelihoodist has

any better options. Adopting a different theory of conditional probability generates other apparently serious problems, as I argued in the previous section. In the next subsection, I will argue that the advantages of the other option—regarding evidential favoring as either relative or indeterminate—are superficial.

Second, the restriction in question would have little if any effect on the scope of the Law of Likelihood in practice, both because it is not necessary to rule out all cases involving probability-zero hypotheses and because truly probability-zero hypotheses are rarely if ever of genuine interest.

It is not necessary to rule out all cases involving probability-zero hypotheses, because many such hypotheses have perfectly determinate likelihoods. They have determinate likelihoods because they just are specific hypotheses about the probability distribution that is generating the data. For instance, the hypotheses that a particular coin has probability .5 of landing heads when flipped entails a definite probability for the possibility that the coin will land heads six times in ten independent tosses even if that hypotheses is regarded as having probability zero. (It would be surprising if a real coin were *exactly* fair.) More generally, Borel's paradox does not arise for a *simple statistical hypothesis,* which specifies values for all parameters of the statistical distribution from which the observations are taken to have been drawn.

One can also distinguish between *composite* statistical hypotheses and *substantive* hypotheses, where the former are disjunctions of simple statistical hypotheses while the latter are not immediately about the probability distribution that is generating the data at all. However, this distinction is not clear or helpful in the present context. Consider hypothesis H_1 from Example 4.1, which says that the point selected from the surface of the sphere lies within a particular distance of the intersection of the equator and the prime meridian. On the surface, this hypothesis seems to be substantive rather than statistical. However, you can write down the unconditional probability distribution over the surface of the sphere in terms of parameters for latitude and longitude and write down H_1 as a disjunction of simple statistical hypotheses about those parameters. Borel's paradox arises for this hypothesis whether we think of it as substantive or statistical. The key point is that the conditional probability in question cannot simply be read off of the hypothesis but has to be determined

in accordance with some theory such as those mentioned in Section 4.4.

In at least most of the cases in which scientists seem to be considering a probability-zero hypothesis, Borel's paradox will not arise either because that hypothesis is a simple statistical hypothesis or because it should not in fact be regarded as having probability zero.

Probability-zero hypotheses often appear in science as point-null hypotheses–for instance, the hypothesis that the mean recovery rate is the same in across treatment groups in a randomized clinical trial. Given certain assumptions, this hypothesis is a simple statistical hypothesis. For instance, given common normality assumptions, this hypothesis entails that a particular function of the data is drawn from a t distribution the only free parameter of which is determined by the sample size.

Hypotheses which assign a sharp value to a quantity of interest are not always simple statistical hypotheses, however. Consider, for instance, stellar parallax. One objection to heliocentrism that Copernicus faces is that if the earth revolves around the sun, then we should observe a corresponding shift in the apparent positions of stars at different times of year. Copernicus argues that this objection is not conclusive because that stellar parallax would be too small to observe if the stars were sufficiently far away, which for all anyone knew they could be. In this case, the question whether some parameter is precisely zero or merely very small has great theoretical importance. The point null hypothesis that it is precisely zero can be treated as a simple statistical hypothesis given a sufficiently developed theory about the error characteristics of the method that is used to make the relevant measurements, but one might worry about Borel's paradox arising in the absence of such a theory.

In this case, however, the point-null hypothesis should not be regarded as having probability zero even though it posits a sharp value for the parameter in question. A Ptolemaic geocentric theory was a live possibility at the time, and such a theory entails that there is no stellar parallax, so the hypothesis that there is no stellar parallax would have receive a positive probability. Thus, Borel's paradox could not arise for it. The same point applies in many other cases involving theoretically important point-null hypotheses in science, such as Michelson and Morley's measurements of the "ether wind."

In other cases, scientists formally test a point-null hypothesis that is plausibly regarded as having zero probability (for instance, that some parameter is exactly zero) but are actually

interested in a vague range-null hypothesis according to which the point-null hypothesis is not too far from the truth (for instance, that the value of the parameter in question is within ϵ of zero for some small positive ϵ). Many cases of null hypothesis significance testing in the social sciences are of this kind. For instance, a recent paper in *Psychological Science* reports results from tests of various versions of the point null hypothesis that children's tendency toward altruistic behavior is the same one month before and one month after experiencing a major earthquake [Li et al., 2013]. This point null hypothesis is arguably highly implausible—perhaps even a probability-zero claim: it would be incredible if the traumatic experience of living through an earthquake had *precisely* zero effect on altruistic behavior. But it would not be incredible if the effect were small enough to be negligible. In cases like this one, scientists appear to be testing what could be regarded as a probability-zero point-null hypothesis, but are better regarded as testing the vague range-null hypothesis that any departure from the point-null hypothesis is small. The latter should be regarded as having positive probability and thus does not give rise to Borel's paradox.

This interpretation of null-hypothesis significance testing is somewhat problematic for a likelihoodist because range-null hypotheses generally lack well-defined objective likelihoods. They are, after all, composite statistical hypotheses. However, this problem will generally not be serious. The likelihood on a given datum of a composite hypothesis is (the continuous analog of) a weighted average of the likelihoods on that datum of the simple hypotheses it contains. Thus, it must be within the range of those likelihoods. That range will be small to the extent that the likelihood function does not vary much over the elements of the composite hypothesis. In many cases, the range will be small enough that the likelihood of the point-null hypothesis will provide a good approximation to the likelihood of the range-null hypothesis.

There may be some cases of genuine scientific interest in which Borel's paradox arises, so that restricting the Law of Likelihood has some practical cost. However, many cases in which scientists seem to be considering probability-zero hypotheses are not of this kind, either because the hypotheses have determinate likelihoods despite being probability-zero or because, on a proper interpretation, the hypotheses being considered have positive probability after all. Moreover, the purported fact that the restriction has a practical cost is no argument against its necessity. However, it does provide motivation for looking for alternatives, such

as the approach to be described in the next subsection.

4.5.2 Regarding Evidential Favoring as Either Relative or Indeterminate

An alternative to restricting the Law of Likelihood is to maintain that any indeterminacy or relativity in the relevant likelihoods produces corresponding indeterminacy or relativity in evidential favoring. For instance, if we understand conditional probabilities in accordance with the theory of regular conditional distributions, then we would say in Example 4.1 that E favors H_1 over H_2 to the degree 1.57 not absolutely, but relative to the sub-σ-field that is implicit in the calculation.

This approach is attractive in its unity and simplicity. It involves no restrictions on the Law of Likelihood and falls under the same general principle as the idea that evidential favoring is objective to the extent that the relevant likelihoods are objective: any qualification on the status of the relevant likelihoods ratio entails a corresponding qualification on the status of the associated evidential favoring.

Unfortunately, the claims this approach yields seem intuitively wrong. In Example 4.1, for instance, the pre-theoretically correct result is that E is determinately and absolutely neutral between H_1 and H_2. It seems wrong to say that E favors H_1 over H_2 even relative to the sub-σ-field in question.

This objection is not a knockdown. Intuitions about the infinite and the infinitesimal are notoriously unreliable, so it would not be completely surprising if we had to give up our intuitions about this case. A view that broke the symmetry between H_1 and H_2, for instance by saying that E determinately and absolutely favors H_1 over H_2 or vice versa, would seem quite unacceptable, but the view in question maintains that symmetry: for every sub-σ-field relative to which E favors H_1 over H_2 to a given degree, there is one relative to which E favors H_2 over H_1 to the same degree.

A second objection to this approach is that it is not clear what purpose it serves beyond the aesthetic one of making our account of evidential favoring simpler and more unified. In standard cases, the Law of Likelihood explicates the pre-theoretically important notion of evidential favoring. It is not clear what it is doing in Borel's paradox cases, given the current

approach. We do not seem to have a pre-theoretic notion of evidential favoring relative to a sub-σ-field for it to explicate.

A possible response to this objection is that the likelihoodist's measure of evidential favoring derives its meaning from the fact that it gives the ratio of posterior to prior odds—or in cases with continuous hypothesis spaces, probability density ratios—when that ratio is well determined. That is, Bayes's theorem entails

$$\frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{p(H_1|E)}{p(H_2|E)} \Big/ \frac{p(H_1)}{p(H_2)} \tag{4.1}$$

where $p(\cdot)$ can indicate either a probability or a probability density. This expression holds in the Kolmogorov theory of conditional probability provided that the same sub-σ-field is used for all of the calculations. Thus, the likelihood ratio of 1.57 one gets when one applies standard techniques to Example 4.1 reflects the change in relative probability density that happens under Bayesian conditioning using the same kinds of techniques.

Unfortunately, this response only serves to move the problem back one step: now we face questions about the significance of the relevant probability density ratios. Intuitively, one should be neutral between H_1 and H_2 both before and after conditioning. Standard mathematical techniques do not deliver this result, and the Kolmogorov approach delivers it only relative to a particular sub-σ-field. On the Dubins approach, by contrast, the relevant probability densities are simply chosen by the user to reflect his or her conditional and unconditional beliefs, subject to certain axioms that can be regarded as consistency requirements, without any reference to sub-σ-fields needed. Given this approach, it seems reasonable to say that evidential favoring is indeterminate, or rather determined only by the relevant dispositional beliefs of the user. The cost, again, is giving up Weak Conglomerability.

One final objection to this approach is that there is an important difference between the distinction between objective and subjective likelihoods on the one hand and likelihoods that are and are not relative to a sub-σ-field on the other hand. The first distinction has to do with what grounds the relevant probability assignment—is it a fact about the world, or a fact about some individual's psychology? The second has to do, at least in cases in which there is no "preferred" sub-σ-field, with whether or not the probability assignments have any non-arbitrary grounding at all. Justifications for the Likelihood Principle and Law of

Likelihood presuppose that conditional probabilities can be interpreted in the natural way as measuring the degree to which one should or would expect the observation in question if the relevant hypothesis were true. When conditional probabilities cannot be interpreted in that way, the justifications break down, and it should not surprise us if the principles break down as well.

There are in some cases grounds for preferring one sub-σ-field over others. For instance, one sub-σ-field might properly reflect the operating characteristics of a measurement technique in a case in which one is conditioning on *data*. Thus, although evidential favoring is in the first instance relative to a sub-σ-field, it becomes absolute when a preferred sub-σ-field is picked out. The problem with this argument is that there does not seem to be anything to pick out one sub-σ-field as the relevant one when one is conditioning on hypotheses rather than data.

In sum, regarding evidential favoring as relative or indeterminate in the cases in which Borel's paradox arises may be a tenable alternative to restricting the Law of Likelihood, but its benefits are weak and it suffers from interpretive difficulties.

4.6 CONCLUSION

I have argued that Borel's paradox gives rise to genuine counterexamples to the Law of Likelihood when conditional probabilities are calculated using standard mathematical techniques. It is possible to avoid the counterexamples by appealing to an alternative theory of conditional probability, but at what appears to be a rather high cost. It is also possible to avoid the counterexamples by either restricting the Law of Likelihood so that it does not apply to the kinds of cases in which Borel's paradox arises or regarding evidential favoring as inheriting any relativity or indeterminacy that affects the likelihood ratio. In the end, the view I prefer is that the Law of Likelihood holds only when the conditional probabilities to which it appeals can legitimately be interpreted as measuring the degree to which the data in question would or should be expected if the hypothesis in question were true, which is not the case when their values depend on an arbitrary choice among limiting operations or

sub-σ-fields.

This concludes my investigation of arguments for and against the Likelihood Principle and the Law of Likelihood. I have found that both principles have strong arguments in their favor and can withstand many objections that have been raised against them. I will now turn my attention to their methodological implications.

5.0 WHY I AM NOT A METHODOLOGICAL LIKELIHOODIST

5.1 INTRODUCTION

Methodological likelihoodists go beyond merely accepting the Likelihood Principle and Law of Likelihood: they claim that it is possible to provide an adequate post-data methodology for science on the basis of likelihood functions alone. Methods based on likelihood functions alone are appealing in that they combine major advantages of Bayesian and frequentist methods. Like Bayesian and unlike frequentist methods, they conform to the Likelihood Principle. Like frequentist and unlike Bayesian methods, they avoid the use of prior probabilities.

Many Bayesians reject methodological likelihoodism because they maintain that an adequate post-data methodology for science would provide guidance for belief and action in the form of posterior probability distributions, which requires appealing to prior probability distributions as well as likelihood functions [e.g. Berger and Wolpert, 1988, 124–36]. The impact of this argument is limited by the fact that non-Bayesians reject the claim that providing a posterior probability distribution is the only or even the best way to provide guidance for belief and action. Frequentists, for instance, favor methods of inference or decision that purport to be justified by results concerning their long-run operating characteristics that do not appeal to posterior probabilities.

One could give a stronger argument against methodological likelihoodism by showing that no norm of belief or action based on likelihood functions alone satisfies requirements for such a norm that are acceptable from both Bayesian and non-Bayesian perspectives. I aim to give such an argument in this chapter. I argue that methodological likelihoodism is false by arguing that (1) an adequate post-data methodology for science would provide a good norm of commitment, (2) a norm of commitment based on likelihood functions alone would have a

particular form, and (3) no norm of the specified form satisfies certain minimal requirements. "Commitment" here can be understood as referring to either belief in a psychological sense or the "behavioristic" notion of adopting a disposition to act on the assumption that the claim in question is true.

This argument has potentially far-reaching implications. The Likelihood Principle says that evidential import depends only on likelihood functions. Thus, if the Likelihood Principle is true and methodological likelihoodism is false, as I am claiming, then it is impossible to provide an adequate post-data methodology for science on the basis of evidential import alone. Science requires either an approach such as the frequentist one that does not respect evidential equivalence or an approach such as the Bayesian one that uses other inputs in addition to evidential import.

I characterize the methodological likelihoodist position more precisely in Section 5.2. Section 5.3 provides an overview of my argument against that position, which I then develop in Sections 5.4–5.8.

5.2 METHODOLOGICAL LIKELIHOODISM

My concern for methodological likelihoodism is not that its principles are false, but that they are insufficient as the sole basis of a methodology for science. Methodological likelihoodists agree with Bayesians in accepting the Likelihood Principle but maintain that the use of prior probabilities is inappropriate in typical cases in science. They simply report likelihood ratios and likelihood functions to be interpreted in accordance with the Law of Likelihood.[1] The following example illustrates this approach.

> **Example 5.1.** In the 1980s, a new treatment called ECMO for severe respiratory problems in newborns was tested in a population of patients each of whom was estimated to have a chance of no more than 20% of surviving under conventional therapy. Seventy-two of the first one hundred patients given ECMO survived. [Bartlett et al., 1985]

[1] Methodological likelihoodists also use various techniques to eliminate "nuisance parameters," that is, parameters that appear in the likelihood function but are not of interest in their own right. They generally admit that those techniques are *ad hoc* [see e.g. Royall, 1997, Ch. 7], but problems arising from nuisance parameters are not my concern in this paper. Similar techniques are used for likelihood-based prediction; see [Bayarri et al., 1988, Section 3] for a discussion of the challenges these techniques face.

Assume for the sake of simplicity that each infant in this population has an equal probability p of surviving under ECMO and that their outcomes are independent. Under this assumption, outcomes of this experiment are like independent tosses of a coin with probability p for heads. Let H_p be the hypothesis that a given infant has probability p of surviving under ECMO. The Law of Likelihood says that the observation E of seventy-two survivals in the first one hundred cases favors H_{p_1} over H_{p_2} to the degree $\Pr(E|H_{p_1})/\Pr(E|H_{p_2}) = (p_1/p_2)^{72}[(1-p_1)/(1-p_2)]^{28}$. For instance, it says that E favors $H_{50\%}$ over $H_{20\%}$ to the degree 8.6×10^{22}. This is an enormous degree of favoring: for comparison, Royall [2000] suggests using 8 as a benchmark for "fairly strong" favoring and 32 for "strong" favoring.

A few points of clarification are in order before I proceed. Methodological likelihoodism is the view that it is possible to provide an adequate post-data methodology for science on the basis of likelihood functions alone.[2] The word "adequate" here is used in an ordinary evaluative sense. A *post-data* methodology is a set of procedures for processing and interpreting the results of an experiment, as opposed to a *pre-data* methodology for planning experiments. I use "experiment" in this paper in a broad sense, to refer to any observational situation with a definite hypothesis space and a definite set of possible outcomes, including "observational studies" in which no intervention is being performed on the system or population of interest.

Methodological likelihoodists can be pluralists: they need not maintain that methods based on likelihood functions alone are appropriate for *all* genuine scientific problems, but only that they are *sufficient* for *some* of them. Methodological likelihoodists typically say that methods based on likelihood functions alone should be used when a Bayesian approach would be problematic because well-grounded prior probabilities are not available [e.g. Sober, 2008, 32]. What they cannot say (*qua* methodological likelihoodists) is that scientists should report likelihood ratios and likelihood functions merely so that others can use them to update their subjective prior probability distributions by Bayesian conditioning. That approach is

[2]Prominent methodological likelihoodists such as Edwards, Royall, and Sober do not express their view in this way, but they are committed to methodological likelihoodism as I characterize it because they advocate reporting facts about evidential favoring as explicated by the Law of Likelihood as a genuine alternative to Bayesian and frequentist post-data methodologies for science.

not a methodology based on likelihood functions alone, but a Bayesian methodology in which likelihood functions are used merely for purposes of communication.

Having characterized the methodological likelihoodist position, I will now introduce my argument against it.

5.3 AN OVERVIEW OF MY ARGUMENT AGAINST METHODOLOGICAL LIKELIHOODISM

The Likelihood Principle and Law of Likelihood have many virtues. They are intuitively plausible and have compelling axiomatic bases. They cohere well with Bayesian approaches and are useful for diagnosing what has gone wrong when frequentist approaches yield intuitively unreasonable results [Berger and Wolpert, 1988, 65ff.]. The likelihood functions on which they are based are often objective, and even when they are not objective they are often easier to assess at least in a rough, qualitative way than prior probabilities [Sober, 2008]. The Law of Likelihood is useful for resolving disputes about the significance of data for a pair of theories: Sober, for instance, uses it to good effect in discussing disputes about whether certain pieces of evidence favor the theory of intelligent design over the theory of evolution or vice versa [2008].

What, then, could be wrong with the methodological likelihoodist position? Simply the idea that likelihood functions are sufficient for a methodology that can stand on its own. Using the Law of Likelihood to clarify the evidential import of data is fine, but the aims of science require that we then be able to say something about what we should believe or do in light of the data. I will argue that attempts to answer such questions on the basis of likelihood functions alone cannot succeed.

My argument against methodological likelihoodism begins in Section 5.4 with the claim that an adequate methodology for the post-data analysis of experimental outcomes would provide a good norm of commitment. I then argue in Section 5.5 that if there are good norms of commitment based on likelihood functions alone, then they include a norm that directs one to accept one hypothesis over another to some function of the degree to which

one's total evidence favors the former over the latter, where that function satisfies a few minimal constraints. In each of Sections 5.6–5.8 I show that any norm of that kind is incompatible with compelling requirements for a good norm of commitment. If the central claims of Section 5.4, Section 5.5, and *any of* Sections 5.6–5.8 are true, then methodological likelihoodism is false.

A few points of clarification are in order before I proceed. My term "acceptance," like "commitment," is ambiguous between a psychological attitude and a behavioral disposition. Thus, my argument is relevant both to those who are interested in inductive inference and those who follow Neyman [1957] in seeking instead a theory of "inductive behavior." I assume that degrees of relative acceptance indicate attitudes that span the range from full acceptance in one direction to full acceptance in the other, with a definite point of neutrality in between. I speak of "preferring" one hypothesis to another as shorthand for accepting the former over the latter to a degree greater than the degree that indicates neutrality.

I assume that norms of action are either reducible to or conceptually posterior to norms of commitment, and thus that there is no need to consider them separately. They are reducible to norms of commitment if commitments just are dispositions to act in certain ways. They are conceptually posterior to norms of commitment if commitments are beliefs and a good norm of action is one that integrates one's beliefs and values in an appropriate way (e.g. by maximizing expected utility).

5.4 CLAIM 1: AN ADEQUATE METHODOLOGY PROVIDES A GOOD NORM OF COMMITMENT

In this section, I argue that an adequate post-data methodology for science provides a good norm of commitment. That is, it provides some kind of rule for accepting hypotheses at least as guides for further action.

Scientists evaluate hypotheses in terms of various "theoretical virtues" such as accuracy, consistency, breadth, simplicity, and fruitfulness [Kuhn, 1977]. They collect data in order to evaluate them for accuracy, where "accuracy" can refer to various notions concerning

some kind of correspondence between what the hypothesis says and what the world is like, such as truth, empirical adequacy [van Fraassen, 1980], and predictive accuracy [Forster, 2002]. The purpose of a post-data methodology is to provide principled guidance in making such evaluations. Evaluating a hypothesis for accuracy involves forming a belief about its accuracy, or at least adopting a disposition to act on some assumption about its accuracy. Thus, an adequate post-data methodology for science provides a principled way to form commitments, which requires that it provide a good norm of commitment.

The relevant alternative to this claim is the view that an adequate post-data methodology for science might merely characterize data as evidence. This view gives insufficient regard to the practical and intellectual problems that science is meant to address. Consider again the experiment (Example 5.1) in which seventy-two of the first hundred infants treated with ECMO survived. The Law of Likelihood can tell us to what degree datum favors the hypothesis that the probability of survival given ECMO is 50% over the hypothesis that it is 20%, for instance. However, it does not tell us *anything*, on its own, about the accuracies of those hypotheses. This fact makes it inadequate as the sole basis for a methodology for science: scientists do not collect data merely for the sake of evaluating it as evidence, but for the sake of evaluating hypotheses for accuracy in order to decide what to believe about them or do in light of them. In this case, they are collecting data in order to decide whether to start using ECMO routinely for infants who are at a high risk of dying from severe respiratory problems. Even in "pure science" cases that are not so immediately tied to applications, scientists want to know whether the hypotheses they are considering are accurate, at least in some weak sense such as approximate statistical adequacy [Spanos, 2010].

This is not to say that using the Law of Likelihood to clarify the status of data as evidence with respect to some set of hypotheses is useless. It can be illuminating as a *preliminary step* toward evaluating the accuracies of hypotheses. Sober uses the Law of Likelihood to assess, for instance, the bearing of the imperfect adaptations that organisms exhibit on the hypothesis that they were created by an intelligent designer against the hypothesis that they evolved by natural selection [2008, 107]. Such an assessment could be a useful step toward deciding which (if either) of those hypotheses to accept. But principled guidance in taking that further step requires some further norm or norms. I will now consider the possibility

that one could provide such a norm on the basis of likelihood functions alone. If not, then methodological likelihoodism is false.

5.5 CLAIM 2: A GOOD PURELY LIKELIHOOD-BASED NORM OF COMMITMENT WOULD HAVE A PARTICULAR FORM

Given that an adequate methodology for science would provide a good norm of commitment, it is incumbent on methodological likelihoodists to show that it is possible to provide such a norm on the basis of likelihood functions alone. In this section I argue that if there are good purely likelihood-based norms of commitment, then they include a norm that directs one to accept one hypothesis over another to some function of the degree to which one's total evidence favors the former over the latter, where that function satisfies a few minimal constraints. When I argue in subsequent sections that no such norm is a good one, it follows that there are no good purely likelihood-based norms of commitment, and thus, given Claim 1 from the previous section, that methodological likelihoodism is false.

An attractive starting point for providing a norm of commitment based on a measure of evidence is Hume's dictum that a wise person proportions his or her belief to his or her total evidence [1825, 111, paraphrased]. Methodological likelihoodists provide only a contrastive measure of evidential favoring, and we are interested in a notion of commitment that is broader than belief, so we need a variant of Hume's dictum which says that a wise person proportions his or her *commitment* in one proposition *relative to another* to the degree to which his or her total evidence favors the one over the other.

This contrastive variant of Hume's dictum is attractive but too specific for present purposes. I do not want to rule out, for instance, being more cautious than it prescribes by remaining neutral when one's total evidence is non-neutral but weak. However, it does seem safe to require that a rule relating evidential favoring to relative acceptance be *nondecreasing*, so that an increase in favoring never leads to a decrease in relative acceptance; *neutrality-calibrated*, meaning that neutral evidence leads to neutrality of acceptance; *permutation-consistent*, meaning that it yields the same judgments about H_1 and H_2 regardless of whether

one considers H_1 against H_2 or H_2 against H_1; and *nontrivial*, in the sense that there is some degree of favoring that suffices for it to subscribe preferring one hypothesis to another.

If we require proportioning relative acceptance to a *function of* degree of evidential favoring that satisfies those requirements, then we get the following class of norms.

> **P**roportion **R**elative **A**cceptance to a **F**unction of the **E**vidence (**PRAFE**): Accept H_1 over H_2 to the degree $f(\mathcal{L}_T) = f(\Pr(E_T|H_1)/\Pr(E_T|H_2))$, where E_T is one's total relevant evidence and f is some nondecreasing function such that $f(1) = 1$, $f(1/x) = 1/f(x)$, and $f(a) > 1$ for some a.

On the intended interpretation of these PRAFE-rules, accepting H_1 over H_2 to degree one means being neutral between them, doing so to a degree greater than one means preferring H_1 to H_2, and doing so to a degree less than one means preferring H_2 to H_1. Thus, each PRAFE-rule entails a corresponding qualitative norm that says to prefer one hypothesis to another just in case the likelihood ratio of the former to the latter exceeds some threshold. Accepting H_1 over H_2 to degree a means the same thing as accepting H_2 over H_1 to degree $1/a$. $f(1) = 1$ is necessary and sufficient to ensure that each PRAFE-rule is neutrality-calibrated; $f(1/x) = 1/f(x)$, that it is permutation-consistent; and $f(a) > 1$ for some a, that it is nontrivial.

The class of PRAFE-rules is very large. In addition to the simple variant of Hume's dictum described above, it also includes "aggressive" rules that prescribe believing one hypothesis over another to a degree greater than the relevant degree of favoring; "stingy" rules that prescribe the opposite; and rules that oscillate between aggressive and stingy (while still being non-decreasing) as the degree of favoring increases.

No other purely likelihood-based norm of commitment suggests itself. Methodological likelihoodists are committed to the Principle of Total Evidence and to the claim that $\mathcal{L}_T$ measures the degree to which one's total evidence favors H_1 over H_2. They are not committed to any other interpretive principles. Bayesian confirmation theory provides several non-comparative measures of evidential support, but none of those measures are purely likelihood-based.[3] The restrictions PRAFE places on the function f of the degree of favoring to use as

[3] According to Chandler [2013], the following are currently the six most popular measures of the degree to which E supports (or "incrementally confirms") H: (d) $\Pr(H|E) - \Pr(H)$, (s) $\Pr(H|E) - \Pr(H|\neg E)$, (c) $\Pr(H\&E) - \Pr(H)\Pr(E)$, (n) $\Pr(E|H) - \Pr(E|\neg H)$, (l) $\log\left[\frac{\Pr(E|H)}{\Pr(E|\neg H)}\right]$, and (r) $\log\left[\frac{\Pr(H|E)}{\Pr(H)}\right]$. None

one's degree of relative acceptance are well motivated. Thus, PRAFE seems to include every purely likelihood-based norm of commitment that a methodological likelihoodist might want to consider.

For the sake of generality, I leave the notion of acceptance open to a range of interpretations. All I will assume is that preferring H_1 to H_2 is extensionally equivalent to being disposed to choose a course of action that will yield a good outcome if H_1 is true and a bad outcome otherwise over one that will yield the same good outcome if H_2 is true and the same bad outcome otherwise. I am not introducing overt betting scenarios: the acts in question could be purely cognitive acts (e.g., believing H_1 more strongly than H_2), and the good and bad outcomes could be good and bad epistemically (e.g. believing something true and believing something false, respectively). A Bayesian would interpret preferring H_1 to H_2 as having higher credence in H_1 than H_2, but I am not assuming that one has credences or that credences satisfy the axioms of probabilities.

One might want to use different PRAFE-rule for different pairs of hypotheses, but such an approach would not be purely likelihood-based because it could lead one to have different degrees of acceptance for pairs of hypotheses that have the same respective likelihood functions on one's total evidence. Thus, such an approach is not available to a methodological likelihoodist, who seeks a methodology for science that is based on likelihood functions alone.

5.5.1 Objection: A good norm of commitment could allow withholding judgment

One feature of PRAFE that is open to question is that it does not allow for withholding judgment about a pair of hypotheses, as opposed to being neutral between them. I have assumed that degrees of relative commitment are real numbers on a scale that goes from complete acceptance in one direction to complete acceptance in the other direction. One might think that if one has no evidence bearing on a pair of hypotheses, or only neutral evidence, then one should be in an uncommitted state that is distinct from neutrality and

of these measures are purely likelihood-based. All involve explicit reference to either the prior probability $\Pr(H)$ or the posterior probability $\Pr(H|E)$ except (n) and (l). (n) and (l) both refer to $\Pr(E|\neg H)$, which is typically a disjunction of hypotheses H_i such that $\Pr(E|H_i)$ is well defined by frequentist lights, but $\Pr(E|\neg H) = \sum_i \Pr(E|H_i)\Pr(H_i|\neg H)$ depends on the conditional prior probability distribution $\Pr(H_i|\neg H)$.

does not correspond to any point on this scale [Norton, 2008]. Additionally, one might think that given certain kinds of evidence one should be in a semi-definite state that corresponds to a set or interval of values on this scale rather than a single point.

5.5.2 Reply: Allowing withholding judgment does not help

I am open to the possibility that degrees of relative acceptance should in some cases be set- or interval-valued, but I will not pursue it in this paper. Allowing for it would make methodological likelihoodism harder to refute but hardly any more plausible. Imprecise Bayesians allow probability functions to be interval-valued, but in doing so they are generalizing an arguably successful theory. By contrast, a methodological likelihoodist who found my arguments against PRAFE-rules persuasive would be allowing for interval-valued degrees of relative acceptance in an attempt to rescue a failed theory. That approach does not seem promising.

I am willing to grant the somewhat plausible idea that one should withhold judgment in the complete absence of relevant (non-neutral) evidence. I will not grant that one should withhold judgment given multiple pieces of evidence that are individually non-neutral but collectively neutral. That proposal is unattractive on its face and has very unattractive consequences. For instance, suppose that E_1 favors H_1 over H_2 to some degree $c > 1$, while E_2 favors H_2 over H_1 to the same degree, where E_1 and E_2 are independent given either H_1 or H_2. Then the conjunction of E_1 and E_2 is neutral between H_1 and H_2.[4] It is implausible that if one learns first E_1 and then E_2, with no other relevant evidence, then one should prefer H_1 to H_2 after learning E_1 but have no definite state of commitment with regard to H_1 and H_2 after learning E_2, rather than being neutral between them. After all, if E_2 had been stronger evidence by any degree ϵ as small as one likes, then one would not have withheld judgment, but either would have preferred H_2 to H_1 or would have been neutral between them (given a sufficiently stingy PRAFE-rule). This discontinuity in the prescribed response to E_2 as its strength varies seems unnatural and ill motivated. I will return to this issue at the end of Section 5.8.

[4] $\Pr(E_1 \,\&\, E_2|H_1) = \Pr(E_1|H_1)\Pr(E_2|H_1)$ by independence, which equals $c(1/c) = 1$ by the fact that E_1 favors H_1 over H_2 to the degree c and E_2 favors H_2 over H_1 to the same degree.

I cannot see any other way in which PRAFE might be ruling out a purely likelihood-based norm of commitment that is worth considering. The burden is on anyone who would claim otherwise to provide an example and to argue that it is a good norm of commitment.

I have argued in this section that if there are good, purely likelihood-based norms of commitment, then they include PRAFE-rules. In each of the next three sections, I argue that PRAFE-rules are not good norms of commitment. Given my Claim 1 that a good post-data methodology for science provides a good norm of commitment, it follows that methodological likelihoodism is false.

5.6 CLAIM 3: A GOOD NORM OF COMMITMENT WOULD ALLOW ONE TO PREFER CHANCE HYPOTHESES TO MAXIMALLY LIKELY ALTERNATIVES

The Law of Likelihood says that E favors H_1 over H_2 if H_1 ascribes a higher probability to E than H_2 does. As Barnard [1972, 129] points out, it follows that the outcome of an experiment always favors the "maximally likely hypothesis" that the experiment was bound to produce that outcome over any hypothesis that makes the outcome of the experiment a matter of chance. (That a hypothesis is "maximally likely" does not mean that it is maximally probable, but rather that it *makes the data* maximally probable.) Some regard this fact as a problem for the Law of Likelihood [e.g. Barnard, 1972, 129; Mayo, 1996, 200–3], while others do not [e.g. Royall, 1997, 13–5]. In this section, I argue that it is a problem for PRAFE-rules even if it is not a problem for the Law of Likelihood.

Consider the following example.

Example 5.2. Suppose you observe ten radioactive isotopes labeled i_1 to i_{10} of a particular species for a period equal to their half-life. You observe E_{23469}: i_2, i_3, i_4, i_6, and i_9 decay, and the rest of the isotopes do not. How does that observation bear on the deterministic hypothesis H_{23469} that those isotopes were bound to decay against the chance hypothesis $H_{50\%}$ that each of the ten isotopes had a 50% chance of decaying independently of the others?

H_{23469} says that the observed datum E_{23469} was inevitable, while $H_{50\%}$ says that it had

probability $(1/2)^{10} = 1/1024$. Thus, the Law of Likelihood says that E_{23469} favors H_{23469} over $H_{50\%}$ to the degree 1024. This result looks bad for the Law of Likelihood. Observations about which isotopes decay do not seem to bear on the question of determinism versus indeterminism at all, yet the Law of Likelihood says that E_{23469} favors the relevant deterministic hypothesis H_{23469} over the most likely indeterministic hypothesis $H_{50\%}$ (among those that treat the decay events as independent and equiprobable) to a large degree.

Two considerations seem to make this problem worse. First, the Law of Likelihood would have said that the result of the experiment favored the deterministic hypothesis that exactly the isotopes that did decay were bound to decay over $H_{50\%}$ to the same large degree regardless of how the experiment had turned out. There is nothing special about E_{23469}: the Law of Likelihood seems to have a general bias toward deterministic hypotheses. Second, the degree of favoring would have been greater if the number of isotopes had been greater, increasing by a factor of two for each additional isotope. This fact is particularly significant for PRAFE-rules. Some PRAFE-rules are sufficiently "stingy" that they would not prescribe preferring H_{23469} to $H_{50\%}$ even on a likelihood ratio of 1024. However, every PRAFE-rule has a threshold such that it prescribes preferring one hypothesis to another when the likelihood ratio of the former to the latter exceeds that threshold. For any such threshold, we could produce an example analogous to Example 5.2 that will inevitably generate a datum that exceeds it for some deterministic hypothesis relative to $H_{50\%}$, simply by increasing the number of isotopes observed.

In general, PRAFE-rules always say not to prefer a hypothesis according to which the outcome of an experiment was due to chance to a maximally likely hypothesis, and they always say to prefer a maximally likely hypothesis to a chance hypothesis that makes any particular outcome sufficiently unlikely (such as $H_{50\%}$ when the number of isotopes is large). This kind of behavior is at odds with both common sense and scientific practice. Scientists continue to regard quantum indeterminism as a live possibility despite having observed an enormous number of radioactive decay events, and they seem to be reasonable in doing so. For this reason, PRAFE-rules are not good norms of commitment.

Similar examples have been used to argue not against PRAFE-rules, but against the Law of Likelihood itself. Royall gives two responses to those arguments. Whether those

arguments as sufficient defenses of the Law of Likelihood or not, they cannot be used to rescue PRAFE-rules.

The first response Royall gives is that the Law of Likelihood is about evidential favoring rather than belief or action. Thus, the Law of Likelihood at least does not direct one to believe or act on H_{23469} rather than $H_{50\%}$ upon observing E_{23469}. This response is not available for PRAFE-rules, which are about belief or action.

The second response Royall gives is that while the Law of Likelihood does say that E_{23469} favors H_{23469} over $H_{50\%}$, it need not say that it favors the more generic hypothesis H_d that the radioactive decay process in question is deterministic over $H_{50\%}$.[5] Whether or not it says the latter depends on the likelihood ratio $\Pr(E_{23469}|H_d)/\Pr(E_{23469}|H_{50\%})$. If all of the hypotheses like H_{23469} that say that exactly some particular subset of isotopes i_1 to i_{10} are bound to decay are given the same prior probability, then $\Pr(E_{23469}|H_d)/\Pr(E_{23469}|H_{50\%}) = 1$,[6] and the Law of Likelihood says that E_{23469} is evidentially neutral between H_d and $H_{50\%}$. If no prior probability over those hypotheses is available, as is typical in science according to methodological likelihoodists, then the relevant likelihood ratio is undefined and the Law of Likelihood is silent. In general, the Law of Likelihood always says that the data favors *some* deterministic hypothesis over any chance hypothesis, but it does not always say that it favors a *generic* hypothesis of determinism over a given chance hypothesis.

One might wonder how an observation can favor H_{23469} over $H_{50\%}$ but not favor H_d over $H_{50\%}$, given that H_{23469} entails H_d. A Bayesian can answer this question: an observation can do so in the sense that conditioning on it increases the probability ratio (odds) $\Pr(H_{23469})/\Pr(H_{50\%})$ while leaving $\Pr(H_d)/\Pr(H_{50\%})$ unchanged. Some methodological likelihoodist say accordingly that the Law of Likelihood measures evidential favoring in the sense that it measures the factor by which Bayesian conditioning *would* increase the probability ratio in question if the relevant prior probabilities were available [e.g. Royall, 1997, 13].

An analogous response is not available for PRAFE-rules, whether we interpret acceptance in terms of belief or in terms of behavioral dispositions. On either interpretation, preferring

[5] Royall actually addresses a different but analogous example, but the translation between the two examples is straightforward.

[6] $\Pr(E_{23469}|H_d) = 1/1024$, and thus is equal to $\Pr(E_{23469}|H_{50\%})$, if all deterministic hypotheses have equal prior probability because there are 1024 such distinct hypotheses: one for each of the $2^{10} = 1024$ possible outcomes of the experiment.

H_{23469} to $H_{50\%}$ is extensionally equivalent to being disposed to choose a course of action that yields a good outcome if H_{23469} is true and a bad one otherwise over one that yields the same good action if $H_{50\%}$ is true and the same bad one otherwise. Because H_d is true if H_{23469} is true, it would be irrational to have this disposition toward H_{23469} and $H_{50\%}$ but not toward H_d and $H_{50\%}$.

Thus, neither of Royall's responses on behalf of the Law of Likelihood helps in defending PRAFE-rules. The fact that PRAFE-rules never direct one to prefer chance hypotheses to maximally likely hypotheses and always direct one to prefer maximally likely hypotheses to sufficiently diffuse chance hypotheses indicates that they are not good norms of commitment. Given Claims 1 and 2 from the previous two sections, it follows that methodological likelihoodism is false.

I present an additional problem for PRAFE-rules in each of the next two sections.

5.7 CLAIM 4: A GOOD NORM OF COMMITMENT WOULD TREAT ALIKE PAIRS CONSISTING OF A HYPOTHESIS AND ITS NEGATION THAT ARE LOGICALLY EQUIVALENT GIVEN ONE'S EVIDENCE

In this section I argue that no PRAFE-rule is a good norm of commitment because any such rule can force one to violate the following principle:

> (R1) If H_1 and H_2 are logically equivalent given your evidence and you prefer H_1 to $\neg H_1$, then you should not prefer $\neg H_2$ to H_2.

H_1 and H_2 are logically equivalent given E if and only if E entails the biconditional $H_1 \leftrightarrow H_2$. For instance, "all ravens are either red or black" is logically equivalent to "all ravens are black" given "no ravens are red" because "no ravens are red" entails that "all ravens are either red or black" is true if and only if "all ravens are black" is true.

(R1) is intuitively compelling. After all, if H_1 and H_2 are logically equivalent given your evidence, then your evidence entails that H_1 is true if and only if H_2 is true. To take a particular case, (R1) prohibits someone who knows that no ravens are red from preferring

"all ravens are either black or red" to its negation while dispreferring "all ravens are black" to its negation. For someone who knows that no ravens are red, those hypotheses have the same content and thus should be assessed alike.

Something like (R1) is generally taken for granted in formal theories of commitment. For instance, it is a standard technique in performing conditional probability calculations to replace A in an expression of the form $\Pr(A|C)$ with some proposition that is logically equivalent to A given C but not otherwise. For instance, one might replace $\Pr(X_1 + X_2 = 2|X_1 = 1)$ with $\Pr(X_2 = 1|X_1 = 1)$, because $X_2 = 1$ is logically equivalent to $X_1 + X_2 = 2$ given $X_1 = 1$. Standard textbooks take for granted the permissibility of such substitutions. Thus, if one used a person's odds $\Pr(H_1)/\Pr(H_2)$ to formalize the degree to which he or she accepts H_1 over H_2 (implicitly conditioning on his or her evidence), then (R1) would hold automatically. Likewise for any other formalization of relative acceptance that permits substitution of logical equivalents given one's evidence.

My proof that any PRAFE-rule can force one to violate (R1) goes as follows. Let X_1 and X_2 record the outcomes of two coin flips. If the first flip lands heads, then $X_1 = 1$. Otherwise $X_1 = 0$. Likewise for the second flip and X_2. Let E be the evidence $X_1 = 1$, H_1 the hypothesis $X_1 = X_2 = 1$, and H_2 the hypothesis $X_2 = 1$. Suppose that E is the only information one has about X_1 and X_2. Pick a PRAFE-rule, thereby fixing a threshold a for the likelihood ratio $\Pr(E|H_1)/\Pr(E|H_2)$ beyond which one is required to prefer H_1 to H_2. In Appendix F, I show how to construct a joint distribution over X_1 and X_2 such that both $\Pr(E|H_1)/\Pr(E|\neg H_1)$ and $\Pr(E|\neg H_2)/\Pr(E|H_2)$ are greater than a. From that construction, it follows that the PRAFE-rule in question can force one to prefer H_1 to $\neg H_1$ and $\neg H_2$ to H_2. But given E, H_1 and H_2 are equivalent. Therefore, that PRAFE-rule can force one to violate (R1).

I will now consider possible objections to using this result to argue against PRAFE-rules.

5.7.1 Objection 1: H_1 and H_2 Are Not Mutually Exclusive

One strategy for responding to this argument is to restrict PRAFE-rules or even the Law of Likelihood itself to hypotheses that are not logically related in the way that H_1 and H_2

in my proof are related. In fact, in Chapter 3 I argued on other grounds for restricting the Law of Likelihood to mutually exclusive hypotheses. H_1 and H_2 are not mutually exclusive, so one might think that this restriction would prevent violations of (R1) from arising.

5.7.2 Reply to Objection 1

In fact, restricting the Law of Likelihood to mutually exclusive hypotheses does not prevent violations of (R1) from arising. H_1 and H_2 in my proof are not mutually exclusive, but the Law of Likelihood is not applied to the comparison between H_1 and H_2: it is applied to the comparison between H_1 and its negation and to the comparison between H_2 and its negation.

Restricting a PRAFE-rule so that it does not apply to a hypothesis and its negation even when both have well defined likelihoods would be an *ad hoc* response to the argument presented in this section. An alternative response would be to prohibit certain *combinations* of applications of PRAFE-rules; for instance, one could prohibit applying a PRAFE-rule both to H_1 and its negation and to H_2 and its negation, where H_1 is the conjunction of H_2 with some further claim. However, this move too looks *ad hoc*. It would also be very limiting and practically impossible to enforce, particularly across entire communities of scientists.

5.7.3 Objection 2: H_1 and H_2 are not statistical hypotheses

There is something that distinguishes the case considered in the proof in this section from the kinds of cases that are commonly considered in likelihoodist writings: H_1 and H_2 are not "statistical hypotheses" with respect to E, that is, hypotheses solely about the stochastic properties of the mechanism that produced E. Instead, H_1 says that E occurs and some other outcome occurs, while H_2 says only that the other outcome occurs. Some writers do restrict the Law of Likelihood to statistical hypotheses [e.g. Hacking, 1965, 59 and Edwards, 1972, 57].

5.7.4 Reply to Objection 2

While some methodological likelihoodists restrict the Law of Likelihood to statistical hypotheses, they seem to do so simply because they have statistical applications in mind. There is no obvious principled basis for that restriction, and the authors in question do not provide one. Moreover, that restriction has the unfortunate consequence of restricting the scope of the Law of Likelihood substantially. For instance, it would prevent one from applying the Law of Likelihood to high-level, substantive scientific theories, as Sober does with the theories of evolution and intelligent design [2008]. In addition, it does not address the other problems for PRAFE-rules presented in this paper. Thus, restricting the Law of Likelihood to statistical hypotheses might allow one to avoid violations of (R1), but it does so in a way that is not independently well motivated, has a high cost, and leaves other problems untouched.

In sum, PRAFE-rules are not good norms of commitment because they can lead to violations of (R1). Given Claims 1 and 2, it follows that methodological likelihoodism is false. Restricting PRAFE-rules to mutually exclusive hypotheses does not allow one to avoid this problem. Restricting them to statistical hypotheses lacks independent motivation and would substantially reduce the scope of the methodological likelihoodist approach.

5.8 CLAIM 5: A GOOD NORM OF COMMITMENT WOULD ALLOW ONE BOTH TO OBEY A HIGHLY PLAUSIBLE DISJUNCTION RULE AND TO TREAT PAIRS OF LOGICALLY EQUIVALENT HYPOTHESES ALIKE

In this section I argue that no norm of the form given by (PRAFE) is a good one because there are cases in which any such norm would force one to violate either (R2) or (R3):

(R2) If H_1 is logically equivalent to H_2, H_3 is logically equivalent to H_4, and you prefer H_1 to H_3, then you should not prefer H_4 to H_2.

(R3) Suppose H_1, H_2, and H_3 are mutually exclusive and exhaustive. If you are neutral

between H_1 and H_2 and do not prefer H_1 to H_3, then you should prefer (H_2 or H_3) to H_1.

(R2) is similar to (R1) and just as compelling. It is stronger than (R1) in that it is not restricted to hypotheses and their negations. However, the fact that each pair of hypotheses to which (R1) applies consists of a hypothesis and its negation plays no role in my argument for (R1). Moreover, (R2) is weaker than (R1) in that the relevant pairs of hypotheses need to be logically equivalent full stop, rather than merely logically equivalent given one's evidence.

Violations of (R2) again seem unacceptable. H_1 is true if and only if H_2 is true, and likewise for H_3 and H_4, so one's assessment of H_1 relative to H_3 should be the same as one's assessment of H_2 relative to H_4. For instance, (R2) says that you should not prefer "all ravens are black" to "some ravens are white" while at the same time preferring "some white things are ravens" (logically equivalent to "some ravens are white") to "all non-black things are non-ravens" (logically equivalent to "all ravens are black"). One's preference between a pair of propositions should depend on the content of those propositions rather than on the form in which those propositions are expressed.

(R2), like (R1), would hold automatically in a variety of possible formalizations of the notion of relative acceptance. It is typically assumed in probability theory, for instance, that one can replace a hypothesis with a logically equivalent hypothesis without changing the probability. Thus, if one used a person's odds $\Pr(H_1)/\Pr(H_2)$ to formalize the degree to which he or she accepts H_1 over H_2, then (R2) would hold automatically. Likewise for any formalization that allows substitutions of logical equivalents.

(R3) follows from the Bayesian assumption that one should have definite credences in H_1, H_2, and H_3 that obey the axioms of probability. Those axioms include the axiom of finite additivity, which entails that one's credence in (H_2 or H_3) should be the sum of one's credence in H_2 and one's credence in H_3. The fact that one is neutral between H_1 and H_2 and does not prefer H_1 to H_3 implies that one has equal credence in H_1 and H_2 and that one's credence in H_3 is no less than that common credence. It follows that one's credence in (H_2 or H_3) is greater than one's credence in H_1. Because one is neutral between H_1 and H_2, the only way it could not be greater is if one's credence in H_3 were zero. But then one's credences in H_1 and H_2 would also have to be zero, which is inconsistent with the

requirement that one's credences in the mutually exclusive and exhaustive hypotheses H_1, H_2, and H_3 sum to one. Thus, (R3) is a consequence of basic Bayesian principles.

One can also argue for (R3) without appealing to contentious Bayesian assumptions. Suppose you are neutral between H_1 and H_2 and do not prefer H_1 to H_3. You must be certain that (H_1 or H_2) is true, certain that it is false, or neither certain that it is true nor certain that it is false . You cannot be certain that it is true, because you would then be certain that H_3 is false (given simple single-premise closure), which would mean that you do prefer H_1 to H_3, since you are neutral between H_1 and H_2. If you are certain that it is false, then you must be certain that H_3 is true, and thus you should prefer (H_2 or H_3) to H_1. If you are neither certain that it is true nor certain that it is false, then you must likewise be neither certain that H_3 is true nor certain that it is false. Disjoining H_3 to H_2 should thus be sufficient to tip the scales in favor of (H_2 or H_3) against H_1, given that you are neutral between H_1 and H_2. Therefore, if you are neutral between H_1 and H_2 and do not prefer H_1 to H_3, then you should prefer (H_2 or H_3) to H_1. This reasoning works even if you "do not prefer" H_1 to H_3 only in the sense that you withhold judgment between them.

The following example shows that no PRAFE-rule conforms to both (R2) and (R3).[7]

Example 5.3. Suppose a mad genius has mixed water and wine in a bottle. You know that the ratio r of water to wine is in the interval $(1/2, 2]$. The mad genius knows the value of r but refuses to tell it to you. He does agree to run three trials of an experiment each possible outcome of which yields relevant, non-neutral evidence concerning the hypotheses $H_1^r : r \in (1/2, 1]$, $H_2^r : r \in (1, 3/2]$ $H_3^r : r \in (3/2, 2]$. As it turns out, the data from the three trials is collectively neutral with respect to those hypotheses. (A description of the experiment and its result is given in Appendix G.) PRAFE-rules thus require being neutral among them. (R3) thus says that one should prefer (H_2^r or H_3^r) to H_1^r.

A problem arises when we apply the same kind of reasoning to the ratio r' of wine to water. r' must be in the interval $[1/2, 2)$. The pieces of evidence from the mad genius's experiment are individually non-neutral between $H_1^{r'} : r' \in [1/2, 1)$ and $H_3^{r'} : r' \in [3/2, 2)$, but collectively neutral among those hypotheses and $H_2^{r'} : r' \in [1, 3/2)$. (See Appendix G.) PRAFE-rules thus require one to be neutral between $H_1^{r'}$ and $H_3^{r'}$ and either to be neutral between or to withhold judgment between $H_2^{r'}$ and $H_3^{r'}$. (R3) thus entails that one should prefer ($H_2^{r'}$ or $H_3^{r'}$) to $H_1^{r'}$. However, ($H_2^{r'}$ or $H_3^{r'}$) is logically equivalent to H_1^r, and $H_1^{r'}$ is logically equivalent to (H_2^r or H_3^r). Thus, PRAFE-rules and (R3) require a patern of preferences that (R2) forbids, namely preferring (H_2^r or H_3^r) to H_1^r (which are equivalent to $H_1^{r'}$ and ($H_2^{r'}$ or $H_3^{r'}$), respectively) while also preferring ($H_2^{r'}$ or $H_3^{r'}$) to $H_1^{r'}$.

[7] This example is loosely related to a well-known example due to von Mises [1957].

This example shows that any PRAFE-rule can force one to violate either (R2) or (R3). Thus, no PRAFE-rule is a good norm of commitment. Given my Claims 1 and 2, it follows once again that methodological likelihoodism is false.

5.8.1 Objection: This argument relies on a strict notion of neutrality that is not clearly appropriate

(R3) is only plausible if "being neutral" between a pair of hypotheses is understood in a strong sense, so that adding any "weight" to either of them by disjoining it with a mutually exclusive hypothesis that one is not sure is false is sufficient to "tip the scales." It is not clear that one should be neutral among H_1^r, H_2^r, and H_3^r, for instance, in this strong sense. It is clear that one is not warranted in believing any of those hypotheses over any other, but perhaps the appropriate doxastic attitude to have about them is sufficiently indefinite that it does not warrant the claim that one should believe (H_2^r or H_3^r) over H_1^r. (See Norton, 2008 for arguments that might be used to motivate this response.)

5.8.2 Reply: Making the notion of neutrality less strict does not result in a successful theory

I address the proposal that PRAFE-rules should allow for indefinite or semi-definite doxastic states in subsection 5.5.2. Some versions of this proposal have some plausibility and would make the methodological likelihoodist position harder to refute, but a methodological likelihoodist needs to show that some approach of this kind results in a useful and well-functioning theory. The prospects for this approach seem dim.

I specifically argue in subsection 5.5.2 against the proposal that one should withhold judgment about a pair of hypotheses given collectively neutral but individually non-neutral pieces of evidence about them, which is the situation with regard to H_1^r, H_2^r, and H_3^r. But perhaps the appropriate attitude to have about a pair of these hypotheses in light of the outcome of the mad genuis's experiment lies somewhere in between strict neutrality and complete withholding of judgment. For the sake of definiteness, let us suppose that it is the attitude represented by some set of real numbers, which for the sake of permutation-

consistency must be of the form $(1/a, a)$[8] for some $a \geq 1$, and that finite additivity applies to the endpoints of these intervals. On this proposal, the degree to which one accepts (H_2^r or H_3^r) over H_1^r should be of the form $(2/a, 2a)$. If a is greater than 2, then this interval includes one. Thus, this proposal would not lead one to say that one should determinately prefer (H_2^r or H_3^r) to H_1^r, thereby allowing one to avoid violating (R2).

This approach faces a dilemma that seems to rule out the possibility that it could yield a useful and well-functioning theory. Either a goes to one as the size of the body of relevant but collectively neutral data increases, or it does not. If it does, then one simply needs to modify Example 5.3 by increasing the size of the data set enough to make a less than two to get an example in which any PRAFE-rule still forces one to violate either (R2) or (R3). If it does not, then either that approach yields strange discontinuities like those discussed in subsection 5.5.2, in which tiny differences in the strength of a particular piece of evidence yield great differences in the definiteness of one's doxastic state, or it yields only interval-valued degrees of relative acceptance even for large non-neutral bodies of data. If the latter is the case, then the usefulness of the theory for scientific purposes is doubtful.

This argument generalizes to other possible ways of representing doxastic attitudes that are somewhere in between strict neutrality and complete indefiniteness. Either those attitudes converge to strict neutrality as the amount of relevant but collectively neural evidence increases or they do not. If they do, then the problem discussed in this section will re-emerge. If they do not, then either the approach has strange discontinuities or its usefulness is limited by the fact that it fails to yield definite recommendations even in some situations that involve large, collectively non-neutral bodies of evidence.

5.9 CONCLUSION

Methodological likelihoodism is attractive because methods based on likelihood functions alone combine some of the major advantages of frequentist and Bayesian methods, respectively. Like frequentist methods, they do not require prior probability distributions. Like

[8]The endpoints of this interval could be either closed or open; nothing substantial hangs on this point.

Bayesian methods, they conform to the Likelihood Principle.

The first problem for methodological likelihoodism is that its central principle, the Law of Likelihood, addresses only questions about evidential favoring. An adequate methodology for science needs to do more than that: it needs to provide a good norm of commitment. After all, we do science not for the sake of gathering data and characterizing it as evidence, but for the sake of deciding what to believe and do in principled, empirically well-informed ways. One could attempt to give a norm of commitment based on likelihood functions alone, but PRAFE-rules seem to encompass all the norms of this kind that are worth considering, and I have shown that they have severe problems. They lead one to prefer maximally likely hypotheses to chance alternatives when such preferences are unwarranted, and they can lead to violations of the highly plausible rules (R1)–(R3). Thus, it does not seem to be possible to provide an adequate self-contained methodology for science on the basis of likelihood functions alone. Methodological likelihoodism is false.

The claim that the Likelihood Principle is true and methodological likelihoodism false implies that an adequate post-data methodology for science cannot be based on evidential import alone. A methodology can be based on long-run operating characteristics, as in the frequentist approach, but that approach faces many objections, including problematic consequences of its violations of the Likelihood Principle (see Chapter 6). It can alternatively be based on both evidential import and other inputs, as in the Bayesian approach, but that approach requires explicit use of prior probabilities. We would like to think that science can rest on an assessment of the evidential import of our empirical data uncolored by prejudices such as those represented by prior probabilities, but if my conclusions are correct then it cannot.

6.0 WHY FREQUENTIST VIOLATIONS OF THE LIKELIHOOD PRINCIPLE ARE PERMISSIBLE AT BEST

6.1 INTRODUCTION

The standard frequentist approach to testing simple statistical hypotheses conforms to likelihoodist principles of evidence within but not across component experiments. This pattern of behavior is not required by the frequentist goal of controlling long-run error rates: there are most powerful tests that violate the Likelihood Principle within an experiment, and there are ways to control Type I error while conforming to the Likelihood Principle across experiments, provided that hypotheses are predesignated. It is hard to justify from likelihoodist and Bayesian perspectives because it seems to fail to respect evidential equivalence across experiments and fails to maximize expected utility on any utility function and prior probability distribution.

These objections can be addressed to some extent by thinking of frequentist methods as appealing to objective considerations to "break the tie" among subjective perspectives over which the relevant agent is indifferent. However, there does not seem to be any compelling reason to prefer the frequentist approach to other methods of tie-breaking, some of which are better integrated into an overarching Bayesian methodology.

I will unpack this argument in the following way. In Section 6.2, I describe the standard frequentist approach to testing simple statistical hypotheses, explaining how restrictions on the application of the Neyman-Pearson theory of most powerful tests allow them to avoid violating likelihoodist principles within experiments. In Section 6.3, I explain how frequentists violate the likelihood principle across experiments and argue that the particular approach they have taken is not required by the basic frequentist goal on controlling long-

run error rates. In Section 6.4, I explain how any violation of the Likelihood Principle within or across experiments leads to a failure to maximize expected utility. In Section 6.5, I explain how thinking of frequentist methods as appealing to objective considerations as tie-breakers helps address this objection, but argue that there does not seem to be any reason to prefer this approach to tie-breaking over various other possible approaches. In the end, I conclude that the standard frequentist approach to testing simple statistical hypotheses is permissible in some cases, despite the fact that it violates the Likelihood Principle, but far from compelling.

One possible response to this chapter is to reject likelihoodist principles of evidence in favor of some other account such as the error-statistical account due to Mayo [1996], according to which standard frequentist practice is justified by evidential considerations. Those who wish to pursue that line will need to take up the arguments for those principles that I present in previous chapters.

I focus in this chapter on cases of simple-versus-simple hypothesis testing, in which the question of interest is whether or not one should reject one simple statistical hypothesis in favor of another simple statistical hypothesis. A simple statistical hypothesis relative to a particular experiment is one that entails a definite probability distribution over the "sample space" of possible outcomes of an experiment. The notion of rejecting one hypothesis in favor of another is easiest to understand when the two hypotheses in question are the only relevant possibilities. In that case, it means something like adopting a disposition to act as if the latter were true rather than the former, subject to new information or a substantial change in stakes. It exhibits the same kinds of context-sensitivities as the non-relative notion of full belief. Nevertheless, it is a standard notion in frequentist discourse.

Frequentist theory also addresses testing one-sided hypotheses (hypotheses according to which some parameter lies on a particular side of a particular value) and testing simple hypotheses against complex alternatives (e.g., the hypothesis that the mean of a distribution is zero against the hypothesis that it is non-zero), as well as point and interval estimation. I focus on simple-versus-simple cases because the Neyman-Pearson theory of most powerful tests applies most cleanly in those cases. If it does not work there, then the claim that it works elsewhere is dubious. Estimation is a related but distinct topic that is best addressed

separately.

A "component experiment" is an arm of a mixture experiment. A mixed experiment consists of first using a randomizer the behavior of which is independent of the hypothesis space to select which of a set of component experiments to perform, and then performing that experiment. Frequentists generally endorse the use of the same Type I error rate rather than the same likelihood-ratio cutoff across components of a mixture experiment as well as across completely separate experiments [Cox, 1958].

6.2 HOW STANDARD FREQUENTIST PRACTICE CONFORMS TO LIKELIHOODIST PRINCIPLES WITHIN COMPONENT EXPERIMENTS

Standard frequentist practice can be viewed as a compromise between "behavioristic" and "evidential" impulses. It is based on the theory of most powerful tests that originated with Neyman and Pearson. That theory was originally developed and justified exclusively in terms of long-run error rates on decisions about how to behave with regard to hypotheses, rather than in terms of the degree to which the data support judgments about the alethic or epistemic value of particular hypotheses. As Neyman and Pearson put it, "without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in the following which we insure that, in the long run of experience, we shall not too often be wrong" [Neyman and Pearson, 1933, 291]. In practice, however, that theory is applied in a way that exhibits some sensitivity to evidential considerations. The result of this compromise is that standard frequentist practice conforms to likelihoodist principles within but not across component experiments.

In many cases, the Neyman-Pearson approach of constructing a uniformly[1] most powerful test (which is always possible in simple-versus-simple testing) yields a *likelihood test*, i.e., a test that rejects H_0 in favor of H_a if and only if the likelihood ratio $\mathcal{L} = \Pr(E|H_a)/\Pr(E|H_0)$

[1] In simple-versus-simple testing, the word "uniformly" in the phrase "uniformly most powerful tests" is superfluous because the alternative experiment is simple.

for the alternative hypothesis against the null on the observed outcome E is above some threshold. Such tests have a nice likelihoodist interpretation: they reject H_0 in favor of H_a if and only if the degree to which E favors H_a over H_0 is above some threshold. They also have a nice Bayesian interpretation: they reject H_0 in favor of H_a if and only if the posterior odds of H_a against H_0 are above some threshold.

However, the Neyman-Pearson theory of most powerful tests is not always in such nice agreement with likelihoodist and Bayesian perspectives. By the Neyman-Pearson lemma [Neyman and Pearson, 1933], every likelihood test is a most powerful test. However, not every most powerful test is a likelihood test. Moreover, different likelihood tests with the same Type I error rate can have different likelihood-ratio cutoffs for rejection. The frequentist commitment to controlling Type I error rates also leads them to require that hypotheses be predesignated and to set more stringent standards for rejection when multiple hypotheses are being tested side-by-side. Those practices have no likelihoodist or Bayesian justification.

There are two kinds of most powerful tests that are not likelihood tests, as I will explain by means of the following example.

Example 6.1. Suppose one is able to run an experiment with possible outcomes x_1, x_2, and x_3, the probabilities of which on the hypotheses of interest are given by the following table.

Table 6.1: Sampling distributions for Example 6.1

	x_1	x_2	x_3
$\Pr(X\|H_0)$	.89	.05	.06
$\Pr(X\|H_a)$	.46	.04	.5
$\mathcal{L} = \Pr(X\|H_0)/\Pr(X\|H_a)$	1.9	1.25	.12

One kind of most powerful test that is not a likelihood test is a *randomized* test. A randomized test uses an auxiliary randomizer to decide whether or not to reject H_0 on one or more possible outcomes. For instance, suppose that in the situation described in the example above one is unwilling to tolerate a Type I error rate greater than .01. Then the most powerful experiment one can run would reject H_0 with probability 1/6 when x_3 is

observed—say, if and only if a particular die produces a 6 when rolled.

This procedure is most powerful, but it can lead to violations of the Likelihood Principle by leading one to draw different conclusions in different possible scenarios in which the data are evidentially equivalent: one in which x_3 is observed and the die produces a 6, and one in which x_3 is observed and the die produces some other number. Frequentists generally do not use randomized tests even when they are most powerful, for the same reason that they would not use a procedure that rejects an arbitrary hypothesis if and only if a particular (irrelevant) twenty-sided die produced a twenty when rolled. Such a procedure has a low Type I error rate (.05), and is even a most powerful test based on that die roll with that error rate, but one does not have to be a likelihoodist to think that it is inappropriately sensitive to evidentially irrelevant considerations.

Next suppose that in the above example one is unwilling to tolerate a Type I error rate greater than .05. If randomized tests are impermissible, then the most powerful permissible test with that Type I error rate is to reject H_0 if and only if x_2 occurs. But that approach seems unreasonable because x_3 clearly favors H_a over H_0 more than x_2 does, according to both the Law of Likelihood and pretheoretic intuitive judgments. It is "non-monotonic" in that as one increases the Type I error rate one is willing to tolerate, it adds x_2 to the rejection region before x_3 even though x_3 has a smaller likelihood ratio for H_0 against H_a. Frequentists generally do not use non-monotonic tests even when they are most powerful. In this case, they would either decline to reject H_0 on the basis of any data or raise the Type I error rate they are willing to accept to at least .06.

By disallowing randomized and non-monotonic tests, frequentists avoid violating the Likelihood Principle and the Law of Likelihood within component experiments.[2] However, they do not avoid violating it across component experiments, as I explain in the next section.

[2]As a historical aside, Neyman and Pearson come close to ruling out most powerful tests that are not likelihood tests in their original presentation of the theory. They maintain that tests should be based on ordered contour levels on the sample space such that one's inclination to reject H_0 for H_a would be the same for all points on the same contour level and would increase as one moves from one contour level to the next [Neyman and Pearson, 1933]. They do not require that contour levels be levels of constant likelihood-ratio and that the ordering on levels be the order of decreasing likelihood ratio, but their theory is compatible with that requirement and not with other ways of carving up the space into contour levels that would yield tests that were not most powerful. Given that requirement, one gets likelihood tests by requiring that contour levels be added to the rejection region in order of decreasing inclination to reject H_0 for H_a. This feature of Neyman and Pearson's theory is often forgotten, although Mayo [1996] emphasizes it.

6.3 HOW STANDARD FREQUENTIST PRACTICE VIOLATES LIKELIHOODIST PRINCIPLES ACROSS COMPONENT EXPERIMENTS

Standard frequentist practice violates likelihoodist principles across component experiments in three ways: (1) using a common Type I error rate rather than a common likelihood-ratio cutoff for rejection; (2) requiring adjustments for multiple testing; and (3) requiring that hypotheses be predesignated. I discuss these three practices in Subsections 6.3.1, 6.3.2, and 6.3.3, respectively. I argue that the basic frequentist commitment to controlling Type I and Type II error rates requires (3) but not (1) or (2).

6.3.1 Frequentists generally use a common Type I error rate across experiments

Frequentists generally use a common Type I error rate across component experiments. The following example illustrates one way in which this practice can lead to violations of likelihoodist principles across experiments: it can lead to the use of different likelihood-ratio cutoffs in experiments with different stopping rules (i.e., different procedures for deciding when to stop collecting data), and thus to different outcomes from data that the Likelihood Principle entails are evidentially equivalent.

Example 6.2. Suppose one wishes to test the hypothesis that the bias of a particular coin for heads is .5 against the hypothesis that it is .6. One can run either of the following experiments.

Experiment 1. *Flip the coin twelve times and count the number of heads.*

Experiment 2. *Flip the coin until the third tail and count the number of flips.*

The likelihood functions for the two experiments are as follows, where the column headings indicate the number of heads. The third row gives the likelihood ratios $\Pr(E|H_0)/\Pr(E|H_a)$

Table 6.2: Likelihood function for the binomial experiment

	$N \leq 4$	5	6	7	8	**9**	10	11	12
bias = .5	.19	.19	.23	.19	.12	**.054**	.016	.0029	.00024
bias = .6	.057	.10	.18	.23	.21	**.14**	.064	.017	.0022
likelihood ratio	3.3	1.9	1.3	.85	.57	**.38**	.25	.17	.11

Table 6.3: Likelihood function for the negative binomial experiment

	$N \leq 8$	9	10	11	**12**	13	14	15	≥ 16
bias = .5	.87	.055	.035	.022	**.013**	.0081	.0048	.0028	.0037
bias = .6	.68	.084	.064	.048	**.035**	.026	.018	.013	.027
likelihood ratio	1.3	.65	.55	.45	**.38**	.32	.26	.22	.14

A likelihoodist or Bayesian would reject H_0 on the basis of 9 heads in Experiment 1 if and only if one would reject H_0 on the basis of 12 tosses in Experiment 2. After all, the likelihood ratios for these two outcomes are the same.

On the Neyman-Pearson approach with the tolerated Type I error rate fixed at .05, one would not reject H_0 on the basis of 9 heads in Experiment 1, but would reject it on the basis of 12 tosses in Experiment 2: the most powerful level .05 test in Experiment 1 rejects H_0 if and only if the number of heads is at least ten, while that in Experiment 2 rejects H_0 if and only if the number of tosses is at least twelve.

This example illustrates the fact that the Type I error rate generated by a test with a particular likelihood-ratio cutoff can vary with the *stopping rule* used, i.e., with the criterion used to decide when to stop collecting data. After all, the likelihood-ratio cutoff is a function of the probability of the result just sufficient to reject the null hypothesis on the null hypothesis and on the alternative hypothesis, whereas the Type I error rate is a function of the probabilities of all of the results at least as extreme as that result on the null hypothesis. The set of results at least as extreme as a given result can vary with the stopping rule, so it

should not be surprising that the Type I error rate associated with taking that result to be just sufficient to reject the null hypothesis can also vary with the stopping rule. In the case at hand, the set of results more extreme than nine heads in twelve tosses when the number of tosses is fixed is {ten heads, eleven heads, twelve heads}. The set of results more extreme than nine heads in twelve tosses when the number of tails is fixed is {ten heads, eleven heads, ...}.

This sensitivity to stopping rules strikes many commenters as contrary to intuitive judgments about evidential relevance. The only difference between the two nine-heads-in-twelve-tosses outcomes has to do with *what the experimenter would have done of the data had been different*. How could such *counterfactuals* about the *experimenter's behavior* be relevant to the significance of the *actual data* for the hypotheses in question? Pointing out that the p-values (the probabilities of results at least as extreme as the observed results) are different in the same cases does not help: how could the implications of hypotheses for *merely possible data that were not observed* be relevant to their assessment in light of the datum that *was* observed?

A frequentist might object at this point that ignoring stopping rules leads to losing control over Type I error rates. Consider again the example due to Armitage discussed in Chapter 3. Suppose one sets the likelihood-ratio threshold for rejecting $H_0 : \mu = 0$ against $H_a : \mu = \bar{x}_n$ at 1/3000. Then a nefarious experimenter can cause one to reject H_0 with probability 1 even if it is true by running the experiment just described with $k = 4$. Moreover, he or she can do the same thing for any likelihood-ratio threshold by making k sufficiently large. It seems that by ignoring stopping rules, one loses control over Type I error.

However, losing control over Type I error in this case is a consequence of not only ignoring the stopping rule, but also (essentially) testing $\mu = 0$ against *every* hypothesis of the form $\mu = \mu_0$, where μ_0 is a constant not equal to zero, using the same likelihood-ratio threshold. It is not so surprising that given enough time, one can eventually produce evidence that is highly misleading with respect to H_0 relative to at least one of those (uncountably many!) alternatives.

Frequentists have the resources to keep Type I error rates bounded even if they use fixed likelihood-ratio cutoffs for rejection rather than fixed Type I error rates. They typically

require that the null and alternative hypotheses be *fixed* and *predesignated* and that adjustments be made for *multiple testing*. The hypothesis $\mu = \bar{X}$ is not fixed but *random* because the numerical value it assigns to μ depends on the data. Thus, a frequentist would not allow it as an alternative hypothesis. Nor would he or she allow one to observe the data and then decide which "fixed" hypothesis of the form $\mu = \mu_a$ to use as one's alternative, which is just an informal way of testing a random hypothesis. One could test $\mu = 0$ against many hypotheses of the form $\mu = \mu_a$ simultaneously, but a frequentist would typically require lowering the Type I error rate of each test in order to keep the "familywise" Type I error rate (the probability of making at least one Type I error over all of the experiments being performed side-by-side) adequately low.

In general, a test of a fixed hypothesis against a fixed alternative that is based on a likelihood-ratio cutoff controls long-run error rates automatically. It can be shown that if the fixed hypothesis H_0 is true, then the probability that a given experiment will yield a result with likelihood ratio for H_a against H_0 greater than k is at most $1/k$—regardless of the stopping rule [Robbins, 1970]![3] Thus, for instance, a test with a likelihood ratio cutoff of .05 has Type I error rate at most .05. By virtue of the Neyman-Pearson lemma, such a test is also a most powerful test, and thus controls Type II error rate (the probability of failing to reject H_0 when H_a is true) as far as possible consistent with its Type I error rate.

Given this result, I do not see why a frequentist would oppose using the same likelihood-ratio cutoff across tests of the same pair of hypotheses in the same decision-theoretic context even if the stopping rules are different, provided that the hypotheses are fixed and predesignated and there are no issues with multiple hypotheses being tested simultaneously. Doing so would bring frequentist methods into closer alignment with likelihoodist and Bayesian methods without giving up anything essential to the frequentist approach. Long-run error rates would still be controlled; they would just be determined by the likelihood-ratio cutoff rather than the reverse. I will give a decision-theoretic argument against the use of fixed Type I error rates in the next section. Before that, I need to address predesignation and adjustments for multiple testing.

[3]Tighter bounds are available in some cases—see Royall 2000.

6.3.2 Frequentists Generally Adjust Their Standards on Individual Tests to Control "Familywise" Error Rates

Many frequentists maintain that when one runs multiple hypothesis tests "side by side," so to speak, one should adjust for that fact in order to keep sufficiently low not only the probability of rejecting *each individual* hypothesis if true, but also the probability of rejecting *some hypothesis or other* if they all are true.

Intuitions seem to pull in both directions on this issue. If I test twenty null hypotheses at the .05 level and get one statistically significant result (i.e., one result that leads me to reject the null hypothesis), then it seems that you should be more skeptical that I have discovered a true departure from the null hypothesis than if I had tested that hypothesis at the .05 level in isolation—even if the hypotheses and the data are the same in the two cases. After all, in the first case I had a 64% chance of getting a statistically significant result just by chance even if all of the null hypotheses were true, whereas in the second case I had only a 5% chance.

Standard frequentist practice in this kind of situation is to maintain the probability of rejecting some true null hypothesis at the desired level (often .05) by lowering the probability of Type I error for each individual test. The simplest and most common procedure is a *Bonferroni correction*, which simply sets the Type I error rate for each individual test at the desired "familywise" error rate divided by the number of tests being performed. This procedure leads to a substantial loss in power for each test, but it is sufficient to address what many frequentists consider an unacceptable risk of Type I error.

Although the argument for multiple-testing corrections has some intuitive appeal, there is also something odd about the idea that what I should conclude about a pair of hypotheses from some body of data depends on what other hypotheses I was testing. Consider a case in which the hypotheses have nothing to do with each other—for instance, let them be the hypothesis that the bias of a particular coin is .5, that smoking causes lung cancer, and that the Higgs boson does not exist in a particular energy range. Why on earth should my standard for rejecting the hypothesis about the coin be more stringent because I am also considering the hypothesis about the Higgs?

This example points to a thorny questions for those who believe in multiple corrections: where does it end? A common frequentist practice is to control familywise Type I error for all of the hypotheses tested for a single publication. Why stop there? Why not control familywise Type I error over a scientist's career, or an entire branch of science, or over all of science? If there really any difference between a case in which I test twenty hypotheses over the course of my career and publish the one that gives a statistically significant result and a case in which I test them all on the same day and write them up together in a single paper? If one should think that the result is probably just the product of noise in the second case, then shouldn't one think so in the first case as well?

It seems that if the convention of doing Bonferroni corrections within a paper and only within a paper has any justification, then that justification has something to do with how the publication process in science currently works, and not with basic epistemic issues. The more hypotheses we test, the more likely we are to reject some falsely. But it's also the case that the more hypotheses we test, the more likely we are to fail to reject some falsely. Corrections for multiple comparisons make the former problem better at the expense of making the latter problem worse. I cannot see how that tradeoff can be worthwhile generally, though of course it will be in some cases. There might seem to be a relevant difference in the two kinds of errors in that a rejection of a null hypothesis is generally regarded as more decisive than a failure to reject it. But that asymmetry does not seem to have any epistemic justification, at least in the case of simple-versus-simple hypothesis testing in which both hypotheses are sufficiently plausible that the test is worth doing.

Likelihoodist principles militate against multiple-testing corrections: $\Pr(E|H)$ as a function of H in $\mathbf{H}$ is not influenced by whether or not other another set of hypotheses were tested alongside those in $\mathbf{H}$. Thus, the Likelihood Principle entails that multiple-testing issues are irrelevant to evidential import. From a Bayesian perspective, the posterior probability of E given H is the same.

An exception to these statements occurs when the data from the other tests is also relevant to some extent to the hypotheses in question. For instance, suppose that a drug is being tested for effectiveness in four different racial groups. In most cases, presumably, enough of the relevant biology is shared that results on, say, study participants of Asian

heritage gives some information about the effectiveness of the drug in participants with African heritage. In such a case, one can use "partial pooling" to share information across the experiments. This approach will have the effect of pulling the estimated treatment effects toward each other and will thus reduce the familywise Type I error rate, but as a side effect of using more of the information that is in the data rather than as a direct effect of artificially increasing the stringency of each test [Gelman et al., 2012].

It is not clear that a frequentist cannot budge on multiple testing. A frequentist has a basic commitment to controlling long-run error rates, but that commitment cannot extend to controlling the familywise Type I error rate over all of science at, say, the .05 level. Such a commitment would make science impossible. And if a frequentist is not committed to controlling that familywise Type I error rate, then why should he or she be committed to controlling any familywise Type I error rate? For any particular familywise error rate, we can ask "why that family, and not some other?" The only way to keep this question from arising is to control error rates on individual tests alone and to accept the inevitable fact that the more tests one runs, the more errors one will make—but the more correct inferences one will draw as well.

6.3.3 Frequentists Generally Require That Hypotheses Be Predesignated

The Armitage experiment illustrates the fact that frequentists need predesignation requirements in order to control Type I error rates: if one allows the alternative hypothesis to be specified after the data are in, then one allows an experiment that rejects the null hypothesis relative to some alternative with probability one even if the null hypothesis is true. The control over long-run error rates that frequentists prize requires that the planning and execution of an experiment be "part of a single whole," as Pearson put it [1962]. For that reason, I do not see how a frequentist can budge on predesignation while retaining the basic frequentist commitment to the importance of long-run error rates.

Again, however, likelihoodist theories of evidence entail that whether or not a pair of hypotheses was predesignated is irrelevant to the evidential import of the data with respect to those hypotheses. $\Pr(E|H)$ is the same for a simple statistical hypothesis H whether it was

predesignated or not; thus, the Likelihood Principle entails that the evidential import of E with respect to **H** does not depend on whether or not the hypotheses in **H** were predesignated. Predesignation is also irrelevant to a Bayesian, who uses the likelihood function for **H** on E to update a prior probability distribution on **H**.

6.4 WHY VIOLATING THE LIKELIHOODIST PRINCIPLES IS INCOMPATIBLE WITH MAXIMIZING EXPECTED UTILITY

A combination of tests maximizes expected utility relative to some utility function and prior probability distribution if and only if it conforms to the Law of Likelihood (in a sense to be explained). Of course, frequentists generally do not use utility functions and prior probability distributions because they are not objective, but they might nevertheless be bothered by the fact that their approach to statistics is not optimal from *any* subjective Bayesian perspective.

This result is of course closely related to Bayesian optimality results and should not be surprising given the close ties between the Law of Likelihood and Bayesian updating. However, it does provide a new perspective on those results. Moreover, it addresses the question of what the practical value of conforming to likelihoodist principles of evidence might be, beyond the fact that doing so respects the intuitions about evidential equivalence from which the Likelihood Principle has been proven.

Let me explain. Choosing a rejection region for a hypothesis test (i.e., the set of possible outcomes on which one will reject the null hypothesis in favor of the alternative) determines the Type I error rate α and the Type II error rate β. For any experiment with a finite sample space (which is any experiment that can actually be performed, given the finite precision and range of every measurement device), there is a finite set of possible error-rate pairs (α, β). For concreteness, consider an experiment with the sampling distributions shown in Table 6.4 and the associated tests shown in Table 6.5.

One can create a rule for choosing among possible hypothesis tests by specifying a family of "indifference curves" in the (α, β) plane, where two points a and b lie on the same indifference curve if and only if one does not strictly prefer a test with the error probabilities

Table 6.4: A pair of hypothetical sampling distributions

	x_1	x_2	x_3
$\Pr(X\|H_0)$	.01	.05	.94
$\Pr(X\|H_a)$	.04	.05	.91
$\mathcal{L} = \Pr(X\|H_0)/\Pr(X\|H_a)$	.25	1	1.03

given by a over one with the error probabilities given by b, and vice versa. Every indifference curve lies either wholly above or wholly below every other indifference curve, and a test on an indifference curve I is strictly preferred to every test on an indifference curve that lies above I.

Frequentists typically allow indifference curves to take any shape as long as they do not cross and their tangent lines have negative slope everywhere. But it's easy to show that maximizing expected utility with one's choice of test requires that one's indifference curves be parallel straight lines. With this point in mind, it's easy to see that a test that is not a likelihood test in my sense does not maximize expected utility relative to *any* prior odds on H_0 and H_a

Consider the example plotted above. Assuming that one is aiming to maximize expected utility, one's indifference curves are parallel straight lines that have either smaller slope, larger slope, or the same slope as the line that connects Test 1 and Test 3 in the (α, β) plane. If they have the same slope, then Tests 1 and 3 are both preferred to Test 2. If they have larger (less negative) slope, then Test 3 is preferred to Test 2 (and Test 1). If they have smaller (more negative) slope, then Test 1 is preferred to Test 2 (and Test 3). Thus, violating the Law of Likelihood by preferring Test 2 to both Test 1 and Test 3 is incompatible with maximizing expected utility.

The same argument can be made for any test that is not a likelihood test. One can compare that test to a test that simply omits an "offending" point (one that is in the rejection region but has larger likelihood ratio of H_0 to H_a than some point outside the

Table 6.5: A trio of possible tests from Table 6.4

Test	Rejection region	α	β
1	$\{x_1\}$	.01	.96
2	$\{x_2\}$	.05	.95
3	$\{x_1, x_2\}$	.06	.91

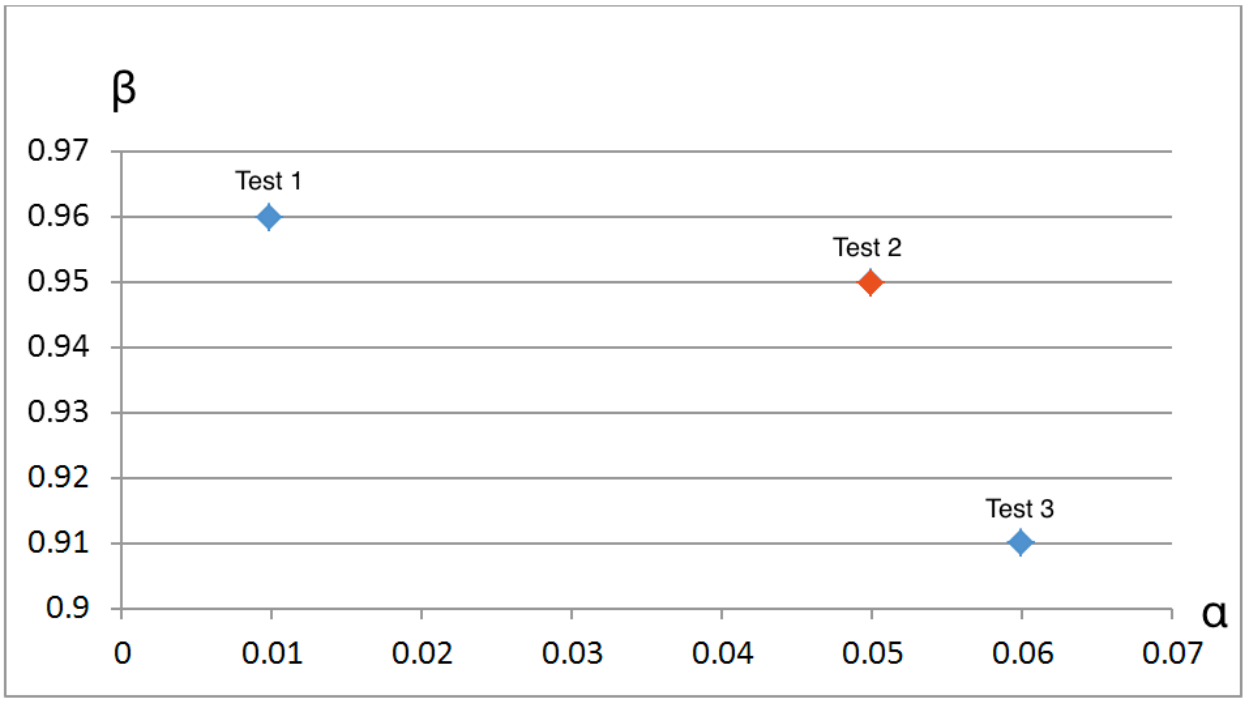

Figure 6.1: α and β for the three tests shown in Table 6.5.

rejection region) from its rejection region and to a test that simply replaces an offending point with an "offended" point (one that is not in the rejection region but has smaller likelihood ratio of H_0 to H_a than some point that is in the rejection region). The slope of the straight line connecting those two tests in the (α, β) plane is the likelihood ratio of H_a to H_0. The slope of the straight line connecting the offending test to the test that omits an offending point is greater (less negative) than this slope, so we have a situation like the one above.

The argument given above shows that any test that violates the Law of Likelihood in the sense of having in its rejection region a point that favors H_a over H_0 more than another point that is outside its rejection region fails to maximize expected utility relative to any utility function and prior probability distribution. But this argument only addresses violations of the Law of Likelihood *within* an experiment. I have argued that frequentists can conform to the Law of Likelihood within an experiment without abandoning anything essential to their position. What about violations of the Law of Likelihood across experiments? I have argued that a frequentist can give up some standard frequentist violations of the Law of Likelihood across experiments without doing violence to his or her basic position—in particular, he or she can give up violations arising from fixing the Type I error rate rather than the likelihood-ratio cutoff in cases differing only in their stopping rules and those arising from the use of multiple-testing corrections. However, it seems that he or she cannot give up predesignation requirements.

Unfortunately for frequentists, the argument given above can be extended to all of these kinds of violations of the Law of Likelihood across experiments. The extension requires only the following mild assumption: if a frequentist prefers test T_1 to all other tests she could perform on Experiment 1 and prefers test T_2 to all other tests she could perform on Experiment 2, then on a simple mixture of Experiments 1 and 2 she prefers performing T_1 if Experiment 1 is performed and T_2 if Experiment 2 is performed to all other tests he or she could perform. (A mixture of two experiments is an experiment that consists of using an auxiliary randomizer the behavior of which is independent of the hypotheses of interest to decide which of those two experiments to perform, and then performing it.) One could come up with strange counterexamples to this principle—for instance, involving an evil demon

that unleash terrible consequences if and only if one performs the combination test in the mixture experiment—but not any, I think, that are relevant to ordinary scientific practice.

Let me illustrate. Return to the coin-flipping case, in which one either flips a coin twelve times and counts the number of heads or flips until three tails appear and counts the number of flips. Suppose that a frequentist prefers the level .05 test in each case. Then he or she violates the Law of Likelihood by rejecting the null hypothesis on nine heads in twelve flips in the second case but not the first. I am assuming that if he or she prefers the level .05 test in each case, then he or she prefers the combination of those tests in a mixture of the two experiments. But then we have a case of him or her violating the Law of Likelihood in a single experiment, so that the argument above applies.

Actually, there is slight subtlety here: the argument given above applies only when one chooses a test that does not reject H_0 in favor of H_a on some datum that favors H_a over E more strongly than another datum on which one does reject H_0 in favor of H_a. In the coin-flipping case just considered, one does not reject H_0 in favor of H_a on some datum that favors H_a over H_0 *more strongly* than another datum on which one does reject H_0 in favor of H_a, but on one that favors H_a over H_0 *to the same degree* as another datum on which one does reject H_0 in favor of H_a. This approach could maximize expected utility if the expected utility of rejecting H_0 on a datum that favors H_0 over H_a to that degree is exactly. But that result requires a rather specific balancing of utilities and probabilities that certainly cannot account for any *general* preference for violating the Law of Likelihood. Indeed, it cannot account for any preference for violating the Law of Likelihood at all, but only for very occasional indifference to whether one violates the Law of Likelihood or not.

Given the assumption I've made about the relationship between preferences on tests in individual experiments and preferences on tests in mixture experiments, the same kind of argument can be run for any violation of the Law of Likelihood across experiments, including violations introduced by predesignation requirements.

I have assumed that the expected utility of a test depends only on the utilities and prior probabilities of the possible right and wrong decisions. One might think that running a test with a particular Type I error rate, or having the same Type I error rate across components of a mixture experiment, has some utility in itself. However, this claim seems quite odd.

Even the most radically behavioristic frequentists think of controlling long-run error rates as a means to the ends of finding truth and of making better decisions, rather than as an end in itself.

6.5 HOW THINKING OF FREQUENTIST CONSIDERATIONS AS TIE-BREAKERS ALLOWS ONE TO ARGUE THAT THEY ARE PERMISSIBLE, BUT NOT THAT THEY ARE PREFERABLE

I have argued that frequentists cannot eliminate all violations of likelihoodist principles across experiments without abandoning their basic commitment to controlling long-run error rates. I have also shown that violations of likelihoodist principles are incompatible with maximizing expected utility. Frequentist methods are supposed to be objective, but what is the value of an objective approach that is not optimal from any particular point of view?

One way to defend frequentist methods against this criticism involves thinking of them from no particular point of view, so to speak. Let me explain. One of the major objections to orthodox subjective Bayesianism is that it presupposes degrees of belief that are far more precise than any real person actually has. It often seems more appropriate to represent doxastic agents (particularly group agents) as having set- or interval-valued degrees of belief rather than sharp, real-numbered degrees of belief. Perhaps a justification for frequentist approaches appears when we move from considering particular possible *opinioned* points of view to considering more realistic points of view that involve imprecise doxastic attitudes.

For instance, suppose that in the coins case my prior probability that the coin is fair is appropriately represented by the entire $[0, 1]$ interval. The coin is flipped 100 times, resulting in 57 heads, and I need somehow to decide whether or not to reject the hypothesis that the coin is fair in favor of the hypothesis that it has bias .6 for heads. My credal state is too indefinite for one decision or the other determinately to maximize expected utility (assuming that my utility function rewards getting the right answer). Yet I need to find some way to proceed. One option is the standard frequentist way: to choose one hypothesis to be the null, choose the probability of Type I error I am willing to accept, and so on.

This approach could lead me to use different likelihood-ratio cutoffs for rejection depending on the experiment's stopping rule. It thus fails to maximize expected utility from the standpoint of any one member of my "credal committee." But my perspective on the situation is not that of any particular member of my credal committee. No matter what I do, some members of my credal committee will be unhappy; and no member of my committee will be happy with everything I *might* do. But I can constrain the set of tests I am considering to ensure that whatever I do, some member of my credal committee will be happy. It is hard to ask for anything more given an imprecise doxastic state.

On the other hand, while thinking about frequentist testing as a tie-breaking procedure allows one to respond to the objection that standard frequentist practice fails to maximize expected utility from any Bayesian agent's point of view, it also does not seem to have any particularly compelling arguments in its favor. It seems like one permissible approach to a situation in which just about any approach is permissible. Choosing a Type I error rate is effectively equivalent to choosing one of the set of utility functions and prior probability distributions that would yield the same likelihood-ratio cutoff for rejection. Type I error rates are objective, but a preference for one error rate over another is not. Prior probability distributions and utility functions are also objectively real entities; what are subjective are the doxastic and axiological states that they represent. If it is permissible to break a tie by appealing to a conventionally accepted Type I error rate or by choosing a Type I error rate arbitrarily, then it should also be permissible to do so by appealing to a conventionally accepted prior probability distribution (as so-called objective Bayesians propose) or by choosing one arbitrarily. It may as a matter of sociological fact be easier currently to get agreement on an appropriate long-run error rate than on appropriate prior probabilities, but this fact seems to have to do with historical contingencies such as the widespread adoption of the arbitrary .05 and .01 levels as standard benchmarks rather than with any epistemologically well-grounded reasons.

6.6 CONCLUSION

Frequentist methods seem torn between behavioristic and evidential impulses. The original Neyman-Pearson theory focused on controlling Type I and Type II error rates, but it is standard practice to avoid randomized and non-monotonic tests that look unreasonable evidentially despite being most powerful tests. Frequentists could go further by using likelihood tests with a common likelihood ratio cutoff across experiments. This approach does not conflict with control over long-run error rates provided that predesignation requirements are preserved. However, essentially all violations of the Likelihood Principle and the Law of Likelihood—including violations arising from predesignation requirements—involve failures to maximize expected utility from any subjective perspective. This failure does not look so bad if we think of use of a frequentist method as a way to proceed in a way that accords with the evidence locally despite not having a definite prior probability distribution. In that kind of situation, however, just about any approach looks just as good as just about any other, so this claim does not provide grounds for a ringing endorsement.

7.0 CONCLUSION

We want a statistical methodology for science to (1) respect the evidential import of the data, (2) avoiding taking into account considerations outside of the evidential import of the data (except for the role that values play in decision-making), and (3) provide good guidance for belief or action. Each of the three leading approaches in statistics violates one of these desiderata.

To review, Chapter 1 provides a proof of the Likelihood Principle. Chapters 2 and 3 reinforce this proof by defending the Likelihood Principle against purported counterexamples. If the Likelihood Principle is true, then desideratum (1) requires conforming to it by drawing the same conclusions from two bodies of data that have the same likelihood function for the same set of hypotheses given no difference in prior beliefs, background assumptions, or utilities. Frequentist methods do not conform to it in this sense, so they violate (1)

The Bayesian approach of assigning probabilities universally and updating those probabilities on new data in accordance with Bayes's theorem violates (2) by taking using prior probabilities that cannot ultimately be based on evidence. One might think that this claim is only true of some prior probabilities—roughly, those that are "merely subjective," as opposed to being derived from either data or a formal rule. However, any method of deriving a probability distribution from data is effectively equivalent to updating some prior probability distribution on that data. One can simply identify probabilities with observed frequencies, but then one is implicitly using a uniform "pre-prior" probability distribution on the hypotheses in question. One can appeal to formal rules such as maximum-entropy to motivate the choice of a "pre-prior" probability distribution, but the general consensus among methodologists is that such rules may provide convenient defaults but do not provide privileged, truly non-informative distributions that aptly represent a state of ignorance (see

Norton 2008).

Chapter 4 argues that no method based on likelihood functions alone can provide a good norm of belief or action. Thus, all possible likelihoodist methods violate (3). Moreover, given the Likelihood Principle, there is no way to satisfy all of (1)–(3). Chapter 4 also argues that satisfying (3) is necessary for providing an adequate methodology for science. Merely characterizing data as evidence is not sufficient for the aims of science.

We are thus left with a choice between violating (1) by using frequentist approaches and violating (2) by using Bayesian approaches. In Chapter 6, I argued that violating (1) is sometimes permissible on the grounds that frequentist considerations provide somewhat plausible ways to break "ties" in the presence of an indefinite doxastic state. However, I saw no reason to prefer this approach to tie-breaking to other approaches that conform to the Likelihood Principle and thus are consistent with maximizing expected utility from a coherent Bayesian point of view. These considerations favor more widespread use of Bayesian methods in science.

APPENDIX A

PROOF THAT $\mathrm{EV}(E_2, Y_0) = \mathrm{EV}(E^B, G)$

What follows is completely analogous to the proof in the main text that $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E^B, g)$. Some expository comments and footnotes given there are not repeated here.

Proof. We have assumed that $P^1_\theta(x_0) = cP^2_\theta(y_0)$ for all $\theta \in \Theta$ and some $c \geq 1$ that is constant in θ. y_0 is the outcome $y_0^\dagger$ of some unique minimal experiment $E_2^\dagger$ with sample space $\mathcal{Y}^\dagger$ that is performed with some known probability r when E_2 is performed. $E_2^\dagger$ is either E_2 itself or a proper component of E_2. y_0 just is $(E_2^\dagger, y_0^\dagger)$, so by the reflexivity of the evidential equivalence relation $\mathrm{Ev}(E_2, y_0) = \mathrm{Ev}(E_2, (E_2^\dagger, y_0^\dagger))$. By the Experimental Conditionality Principle, $\mathrm{Ev}(E_2, (E_2^\dagger, y_0^\dagger)) = \mathrm{Ev}(E_2^\dagger, y_0^\dagger)$.

Construct a hypothetical minimal experiment E_2^{CE} with sample space and sampling distribution given by Table A1. The outcomes in the first row that correspond to outcomes of $E_2^\dagger$ (that is, $\{(j, w_i) : \exists(y^\dagger \in \mathcal{Y}^\dagger)(y_i = (E_2^\dagger, y^\dagger)\})$ constitute a set $A \subset \mathcal{Y}$ such that $\Pr(Y \in A) = p$ for some known $0 \leq p \leq 1$, namely rc^*c where $c^* = \frac{1}{1+c}$. Likewise, the outcomes in the first column (that is, $\{(j, w_0), (k, w_0)\}$ constitute a set $A \subset \mathcal{Y}$ such that $\Pr(Y \in A) = p$ for some known $0 \leq p \leq 1$, namely c^*.

Table A1: Sampling Distribution of E_2^{CE} ($c^* = \frac{1}{1+c}$)

	w_0	w_1	w_2	w_3	...	w_n
j	$c^*cP^2_\theta(y_0)$	$c^*cP^2_\theta(y_1)$	$c^*cP^2_\theta(y_2)$	$c^*cP^2_\theta(y_3)$	...	$c^*cP^2_\theta(y_n)$
k	$c^* - c^*cP^2_\theta(y_0)$	$c^*cP^2_\theta(y_0) - c^*\min_\theta P^2_\theta(y_0)$	$c^*\min_\theta P^2_\theta(y_0)$	0	...	0

Let E_2^M be an experiment that consists of flipping a coin with bias rc^*c for heads to choose between performing $E_2^\dagger$ if the coin lands heads and performing a minimal experiment with sampling distribution given by the distribution of E_2^{CE} conditional on the complement of $\{(j, w_i) : \exists(y^\dagger \in \mathcal{Y}^\dagger)(y_i = (E_2^\dagger, y^\dagger))\}$ if the coin lands tails. By the Weak Ancillary Realizability Principle, $\mathrm{Ev}(E_2^{CE}, (j, w_0)) = \mathrm{Ev}(E_2^M, (E_2^\dagger, y_0^\dagger))$. By the Experimental Conditionality Principle, $\mathrm{Ev}(E_2^M, (E_2^\dagger, y_0^\dagger)) = \mathrm{Ev}(E_2^\dagger, y_0^\dagger)$. It follows that $\mathrm{Ev}(E_2, y_0) = \mathrm{Ev}(E_2^{CE}, (j, w_0))$.

Next take the hypothetical Bernoulli experiment E^B constructed in the proof given in the main text, which has sample space (g, h) and sampling distribution given by $P^B_\theta(g) = P^1_\theta(x_0) = cP^2_\theta(y_0)$. Finally, construct a mixture experiment $E_2^{M_B}$ that consists of first flipping a coin with bias c^* for heads to decide between performing E^B and performing a minimal experiment with the known sampling distribution given by the distribution of E_2^{CE} conditional on the complement of the first-column outcomes $\{(j, w_0), (k, w_0)\}$. By the Weak Ancillary Realizability Principle, $\mathrm{Ev}(E_2^{CE}, (j, w_0)) = \mathrm{Ev}(E_2^{M_B}, (E^B, g))$. By the Experimental Conditionality Principle, $\mathrm{Ev}(E_2^{M_B}, (E^B, g))= \mathrm{Ev}(E^B, g)$. It follows that $\mathrm{Ev}(E_2^{CE}, (j, w_0))= \mathrm{Ev}(E^B, g)$. From $\mathrm{Ev}(E_2, y_0) = \mathrm{Ev}(E_2^{CE}, (j, w_0))$ and $\mathrm{Ev}(E_2^{CE}, (j, w_0))= \mathrm{Ev}(E^B, g)$, it follows that $\mathrm{Ev}(E_2, y_0)= \mathrm{Ev}(E^B, g)$. □

In the main text it was shown that $\mathrm{Ev}(E_1, x_0)= \mathrm{Ev}(E^B, g)$, from which it now follows that $\mathrm{Ev}(E_1, x_0)= \mathrm{Ev}(E_2, y_0)$.

APPENDIX B

PROOF THAT (CE) AND (†) ARE JOINTLY INCOMPATIBLE WITH SIX LEADING MEASURES OF CONFIRMATION

Here I demonstrate the statement made on page 47 that E confirms H_2 more than it confirms H_1 according to what Chandler [2013, 130] identifies as the six most popular measures of confirmation currently on offer when $\Pr(H_1) = .99, \Pr(H_1|E) = 1, \Pr(H_2) = .01$, and $\Pr(H_2|E) = .99$ and H_1 and H_2 are exhaustive. Those six measures are as follows:

1. (d) $c(H, E) = \Pr(H|E) - \Pr(H)$
2. (s) $c(H, E) = \Pr(H|E) - \Pr(H|\bar{E})$
3. (c) $c(H, E) = \Pr(H\&E) - \Pr(H)\Pr(E)$
4. (n) $c(H, E) = \Pr(E|H) - \Pr(E|\bar{H})$
5. (l) $c(H, E) = \log\left[\frac{\Pr(E|H)}{\Pr(E|\bar{H})}\right]$
6. (r) $c(H, E) = \log\left[\frac{\Pr(H|E)}{\Pr(H)}\right]$

I show that $c(H_2, E) > c(H_1, E)$ for each of these measures in turn. I am not aware of any other measures that have been seriously proposed that do not yield the same conclusion. This conclusion is incompatible with (CE′), which follows from the conjunction of (CE) and (†), so it follows from the results presented here that (CE) and (†) are jointly incompatible with any of the measures of confirmation considered. This result strongly suggests that at least one of (CE) and (†) is false, as I argue in Subsection 3.3.1

B.1 (D) $C(H, E) = \mathrm{PR}(H|E) - \mathrm{PR}(H)$

.

The result for this measure is trivial: $c(H_2, E) = \Pr(H_2|E) - \Pr(H_2) = .99 - .01 = .98$, while $c(H_1, E) = \Pr(H_1|E) - \Pr(H_1) = 1 - .99 = .01$, so $c(H_2, E) > c(H_1, E)$.

B.2 (S) $C(H, E) = \mathrm{PR}(H|E) - \mathrm{PR}(H|\bar{E})$

.

$$\Pr(H_2|E) - \Pr(H_2) > \Pr(H_1|E) - \Pr(H_1) \qquad \text{(Shown in item 1 above.)}$$

$$\Pr(H_2|E) - \Pr(H_2) + \Pr(H_2\&E) - \Pr(H_2\&E) >$$
$$\Pr(H_1|E) - \Pr(H_1) + \Pr(H_1\&E) - \Pr(H_1\&E)$$

$$\Pr(H_2|E) - \Pr(H_2) + \Pr(E|H_2)\Pr(H_2) - \Pr(H_2|E)\Pr(E) >$$
$$\Pr(H_1|E) - \Pr(H_1) + \Pr(E|H_1)\Pr(H_1) - \Pr(H_1|E)\Pr(E)$$

$$\Pr(H_2|E)[1 - \Pr(E)] - \Pr(H_2)[1 - \Pr(E|H_2)] >$$
$$\Pr(H_1|E)[1 - \Pr(E)] - \Pr(H_1)[1 - \Pr(E|H_1)]$$

$$\Pr(H_2|E)\Pr(\bar{E}) - \Pr(H_2)\Pr(\bar{E}|H_2) > \Pr(H_1|E)\Pr(\bar{E}) - \Pr(H_1)\Pr(\bar{E}|H_1)$$

$$\Pr(H_2|E) - \frac{\Pr(H_2)\Pr(\bar{E}|H_2)}{\Pr(\bar{E})} > \Pr(H_1|E) - \frac{\Pr(H_1)\Pr(\bar{E}|H_1)}{\Pr(\bar{E})}$$

$$\Pr(H_2|E) - \Pr(H_2|\bar{E}) > \Pr(H_1|E) - \Pr(H_1|\bar{E}) \qquad \text{(Bayes's theorem)}$$

$$c(H_2, E) > c(H_1, E)$$

B.3 (C) $C(H, E) = \mathrm{PR}(H\&E) - \mathrm{PR}(H)\mathrm{PR}(E)$

.

$$\Pr(H_2|E) - \Pr(H_2) > \Pr(H_1|E) - \Pr(H_1) \qquad \text{(Shown in item 1 above.)}$$
$$\Pr(H_2|E)\Pr(E) - \Pr(H_2)\Pr(E) > \Pr(H_1|E)\Pr(E) - \Pr(H_1)\Pr(E)$$
$$\Pr(H_2\&E) - \Pr(H_2)\Pr(E) > \Pr(H_1\&E) - \Pr(H_1)\Pr(E)$$
$$c(H_2, E) > c(H_1, E)$$

B.4 (N) $C(H, E) = \mathrm{PR}(E|H) - \mathrm{PR}(E|\bar{H})$

$$0.9801 > 0.01$$
$$(0.99)^2 > 0.01$$
$$\frac{.99}{.01} > \frac{1}{.99}$$
$$\frac{\Pr(H_2|E)}{\Pr(H_2)} > \frac{\Pr(H_1|E)}{\Pr(H_1)}$$
$$\frac{\Pr(H_2|E)\Pr(E)}{\Pr(H_2)} > \frac{\Pr(H_1|E)\Pr(E)}{\Pr(H_1)}$$
$$\Pr(E|H_2) > \Pr(E|H_1)$$
$$2\Pr(E|H_2) > 2\Pr(E|H_1)$$
$$\Pr(E|H_2) - \Pr(E|H_1) > \Pr(E|H_1) - \Pr(E|H_2)$$
$$\Pr(E|H_2) - \Pr(E|\bar{H}_2) > \Pr(E|H_1) - \Pr(E|\bar{H}_1) \qquad (H_1 \text{ and } H_2 \text{ are exhaustive})$$
$$c(H_2, E) > c(H_1, E)$$

B.5 (L) $C(H,E) = \text{LOG}\left[\frac{\text{PR}_{(E|H)}}{\text{PR}_{E|\bar{H})}}\right]$

$$
\begin{aligned}
\Pr(E|H_2) &> \Pr(E|H_1) && \text{(See proof of item 4 above.)}\\
[\Pr(E|H_2)]^2 &> [\Pr(E|H_1)]^2 &&\\
\frac{\Pr(E|H_2)}{\Pr(E|H_1)} &> \frac{\Pr(E|H_1)}{\Pr(E|H_2)} &&\\
\log\left[\frac{\Pr(E|H_2)}{\Pr(E|H_1)}\right] &> \log\left[\frac{\Pr(E|H_1)}{\Pr(E|H_2)}\right] &&\\
\log\left[\frac{\Pr(E|H_2)}{\Pr(E|\bar{H}_2)}\right] &> \log\left[\frac{\Pr(E|H_1)}{\Pr(E|\bar{H}_1)}\right] && (H_1 \text{ and } H_2 \text{ are exhaustive})\\
c(H_2,E) &> c(H_1,E) &&
\end{aligned}
$$

B.6 (R) $C(H,E) = \text{LOG}\left[\frac{\text{PR}_{(H|E)}}{\text{PR}_{(H)}}\right]$

.

The result for this measure is trivial: $c(H_2,E) = \log\left[\frac{\Pr_{(H_2|E)}}{\Pr_{(H_2)}}\right] = \log\left[\frac{.99}{.01}\right] = 4.60$, while $c(H_1,E) = \log\left[\frac{\Pr_{(H_1|E)}}{\Pr_{(H_1)}}\right] = \log\left[\frac{1}{.99}\right] = .01$, so $c(H_2,E) > c(H_1,E)$.

APPENDIX C

PROOF THAT ARMITAGE'S EXPERIMENT MERELY TRADES OFF ONE KIND OF MISLEADINGNESS AGAINST ANOTHER

$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. $N_k = \min\{n : |\bar{X}_n| \geq k/\sqrt{n}\}$. We want to show that $k_2 > k_1$ implies $\mathrm{E}\left|\bar{X}_{N_{k_2}}\right| < \mathrm{E}\left|\bar{X}_{N_{k_1}}\right|$.

Let $Y_n = \bar{X}_{-n}$, $S_n = \sum_{i=1}^{n} X_i$, $\mathcal{F}_n = \sigma(S_n, S_{n+1}, \ldots)$ (i.e., the σ-field generated by $S_n, S_{n+1}, \ldots$), and $\mathcal{G}_n = \mathcal{F}_{-n}$. The proof proceeds as follows.

1. $\{\ldots, Y_{-2}, Y_{-1}\}$ is a backward martingale with respect to $\mathcal{G}_n$.
2. $\{\ldots, |Y_{-2}|, |Y_{-1}|\}$ is a backward submartingale with respect to $\mathcal{G}_n$.
3. $\mathrm{E}|\bar{X}_{n+1}| < \mathrm{E}|\bar{X}_n|$ for all $n \geq 1$.
4. For any $k_2 > k_1$, $\mathrm{E}|\bar{X}_{N_{K_2}}| < \mathrm{E}|\bar{X}_{N_{K_1}}|$.

C.1 STEP 1

To be shown: $\{\ldots, Y_{-2}, Y_{-1}\}$ is a backward martingale with respect to $\mathcal{G}_n$.

That is:

(1) $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ for all $n < -1$;

(2) Y_n is measurable $\mathcal{G}_n$ for all $n \leq 1$;

(3) $\mathrm{E}[|Y_n|] < \infty$ for all $n \leq 1$;

(4) with probability 1, $\mathrm{E}[Y_{n+1}||\mathcal{G}_n] = Y_n$ for all $n < -1$.

Start with (1). $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ for all $n < -1$ if and only if $\mathcal{F}_n \subset \mathcal{F}_{n-1}$ for all $n > 1$. $\mathcal{F}_n = \sigma(S_n, S_{n+1}, \ldots) \subset \sigma(S_{n-1}, S_n, \ldots) = \mathcal{F}_{n-1}$, so this condition is satisfied.

Next consider (2). $S_n \in \mathcal{G}_n$, and $Y_n = S_n/n$, so Y_n is measurable $\mathcal{G}_n$ for all $n \leq 1$.

Now consider (3). Y_n is normally distributed with mean 0 and variance $1/|n|$. It follows that $|Y_n|$ follows a half-normal distribution with mean $\sqrt{2/|n|\pi}$, and thus that (3) is satisfied.

(4) can be shown as follows:

$$
\begin{aligned}
\mathrm{E}[Y_{n+1}||\mathcal{G}_n] &= \mathrm{E}[\bar{X}_{-(n+1)}||\mathcal{F}_{-n}] \\
&= \mathrm{E}[\bar{X}_{m-1}||\sigma(S_m, S_{m+1}, \ldots)] && \text{where } m = -n \\
&= \frac{1}{m-1}\mathrm{E}\left[S_m - X_m|S_m\right] \\
&= \frac{1}{m-1}\mathrm{E}[S_m|S_m] - \frac{1}{m-1}\mathrm{E}[X_m|S_m] \\
&= \frac{1}{m-1}S_m - \frac{1}{m-1}\left[\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}[X_m|S_m]\right] \\
&= \frac{m}{m-1}\bar{X}_m - \frac{1}{m-1}\left[\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}[X_i|S_m]\right] \\
&= \frac{m}{m-1}\bar{X}_m - \frac{1}{m-1}\left[\frac{1}{m}\mathrm{E}\left[\sum_{i=1}^{m}X_i\middle|S_m\right]\right] \\
&= \frac{m}{m-1}\bar{X}_m - \frac{1}{m-1}\left[\frac{1}{m}\mathrm{E}\left[S_m|S_m\right]\right] \\
&= \frac{m}{m-1}\bar{X}_m - \frac{1}{m-1}\bar{X}_m \\
&= \bar{X}_m \\
&= Y_n
\end{aligned}
$$

C.2 STEP 2

To be shown: $\{\ldots, |Y_{-2}|, |Y_{-1}|\}$ is a backward submartingale with respect to $\mathcal{G}_n$.

In fact, I will show a (well-known) more general fact: for any sequence $\{\ldots, Z_{-2}, Z_{-1}\}$ that is a backward martingale with respect to a filtration $\{\mathcal{H}_n\}$ and any convex function $\phi(\cdot)$, $\{\ldots, \phi(Z_{-2}), \phi(Z_{-1})\}$ is a backward submartingale with respect to $\{\mathcal{H}_n\}$.

The fact that $\{\ldots, Z_{-2}, Z_{-1}\}$ satisfies conditions (1)–(3) for being a martingale obviously implies that $\{\ldots, |Z_{-2}|, |Z_{-1}|\}$ does so as well: conditions (1) and (3) are unchanged, and condition (2) holds because the absolute value of a random variable that is measurable with respect to a given σ-field is also measurable with respect to that σ-field.

The fact that $\{\ldots, Z_{-2}, Z_{-1}\}$ satisfies the backward martingale condition $\mathrm{E}[Z_{n+1}||\mathcal{H}_n] = Z_n$ implies that $\{\ldots, |Z_{-2}|, |Z_{-1}|\}$ satisfies the backward submartingale condition $\mathrm{E}[|Z_{n+1}|||\mathcal{H}_n] \geq |Z_n|$ as a consequence of Jensen's inequality. Jensen's inequality says that for any random variable W and any convex function ϕ, $\mathrm{E}[\phi(W)] \geq \phi(\mathrm{E}[W])$. This result holds for conditional expectations as well. Thus, we have

$$\begin{aligned}\mathrm{E}[\phi(Z_{n+1})||\mathcal{H}_n] &\geq \phi(\mathrm{E}[Z_{n+1}]||\mathcal{H}_n]) \\ &= \phi(Z_n) \qquad\qquad \text{by the backward martingale property}\end{aligned}$$

Therefore, from the fact that $\{\ldots, Y_{-2}, Y_{-1}\}$ is a backward martingale with respect to $\mathcal{G}_n$, it follows that $\{\ldots, |Y_{-2}|, |Y_{-1}|\}$ is a backward submartingale with respect to $\mathcal{G}_n$.

C.3 STEP 3

To be shown: $\mathrm{E}|\bar{X}_{n+1}| < \mathrm{E}|\bar{X}_n|$ for all $n \geq 1$.

From the previous step, we have $\mathrm{E}[|Y_{n+1}|||\mathcal{G}_n] \geq |Y_n|$ for all $n < -1$. In fact, $\mathrm{E}[|Y_{n+1}|||\mathcal{G}_n] > |Y_n|$. After all, whether $|Y_{n+1}|$ is greater than, equal to, or less than $|Y_n|$ depends only on the value of $X_{-(n+1)}$. The only way for $|Y_{n+1}|$ to be less than $|Y_n|$ is for the sign of $X_{-(n+1)}$ to be opposite that of Y_n and for $|X_{-(n+1)}|$ to be less than $2|Y_n|$. The contribution to the expected value of $|Y_{n+1}|$ (given $\mathcal{G}_n$) made by these outcomes is exactly balanced by that of outcomes with the same sign as Y_n with $|X_{-(n+1)}| < 2|Y_n|$. The remaining outcomes with $|X_{-(n+1)}| \geq 2|Y_n|$ make a positive contribution to the expected value, producing the

strict inequality. Taking the expectation of both sides of this inequality and applying the double expectation formula to the left-hand side yields $\mathrm{E}|Y_{n+1}| \geq \mathrm{E}|Y_n|$, which is equivalent to $\mathrm{E}|\bar{X}_{-n-1}| \geq \mathrm{E}|\bar{X}_{-n}|$. Letting $m = -n - 1$ yields $\mathrm{E}|\bar{X}_m| \geq \mathrm{E}|\bar{X}_{m+1}|$ for all $m \geq 1$.

C.4 STEP 4

To be shown: For any $k_2 > k_1$, $\mathrm{E}|\bar{X}_{N_{K_2}}| \leq \mathrm{E}|\bar{X}_{N_{K_1}}|$.

A backward martingale is a closed martingale, and N_{k_1} and N_{k_2} are stopping times, so Doob's optional sampling theorem applies. Thus, from the fact that $\mathrm{E}|\bar{X}_{n+1}| < \mathrm{E}|\bar{X}_n|$ it follows that $\mathrm{E}|\bar{X}_{N_{K_2}}| < \mathrm{E}|\bar{X}_{N_{K_1}}|$

APPENDIX D

TECHNICAL DETAILS FROM STEIN'S EXAMPLE

In Stein's example (with slight refinements due to Berger and Wolpert [1988, 133–4] that do not affect the substance of the argument), Y is a random variable distributed according to the probability density function

$$cy^{-1}e^{-\frac{d^2}{2}(1-\frac{\theta}{y})^2}I_{(0,b\theta)}(y), \tag{D.1}$$

where b, c, d are known positive constants subject to the constraint

$$c = \frac{1}{\int_0^b y^{-1}e^{-\frac{d^2}{2}(1-y^{-1})^2}dy}$$

[Berger, 1980, 153].

Consider a pair of observations $X = x_0 = \sigma_0 d$ and $Y = y_0 = \sigma_0 d$. The respective likelihood functions $l_{x_0}(\theta)$ and $l_{y_0}(\theta)$ of those observations are proportional outside the interval $(0, \sigma_0 d/b]$:

$$l_{x_0}(\theta) \propto e^{-\frac{1}{2}(d-\frac{\theta}{\sigma_0})^2}I_{(0,\infty)}(\theta)$$
$$l_{y_0}(\theta) \propto e^{-\frac{1}{2}(d-\frac{\theta}{\sigma_0})^2}I_{(\sigma_0 d/b,\infty)}(\theta)$$

For fixed σ_0, one can make the interval $(0, \sigma_0 d/b]$ on which $l_{x_0}(\theta) \not\propto l_{y_0}(\theta)$ arbitrarily small by choice of d and b. Thus, it seems that an advocate of the Likelihood Principle would be hard pressed to deny that x_0 and y_0 are essentially evidentially equivalent.

Larger values of x_0 require larger values of b to keep the interval $(0, \sigma_0 d/b] = (0, \sigma_0 d/b]$ sufficiently small. Different value of b correspond to different random variables Y. Thus, there is no single random variable Y such that a reasonable Extended Likelihood Principle implies that $X = x_0$ is essentially evidentially equivalent to $Y = x_0$ for all values x_0 of X. As a result, even an advocate of the Likelihood Principle committed to using the 95% confidence interval $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ for all values of X is not committed to using the low-coverage interval $(Y - 1.96Y/d, Y + 1.96Y/d)$ for all values of some single random variable Y. For that reason, responding to Stein's example does not require arguing as Berger and Wolpert [1988, 133–5] and Grossman [2011a, 311–3] do that an advocate of the Likelihood Principle should not be committed to using the 95% confidence interval $(X - 1.96\sigma_0, X + 1.96\sigma_0)$ for all values of X. That being said, I have no objections to their arguments and would be happy to fall back on them if my response to Stein's example fails.

APPENDIX E

DERIVING THE ANOMALOUS LIKELIHOOD RATIO

Let $\Theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ be a random variable that gives the latitude of P, where $\Theta = 0$ at the equator and $\Theta = \frac{\pi}{2}$ at the north pole. Let $\Phi \in [0, 2\pi)$ be a random variable that gives the longitude of P, where $\Phi = 0$ at the prime meridian. A uniform probabiity distribution over the surface has in this coordinate system the probability density function $\cos\theta/4\pi$.

We can begin to calculate $\Pr(E|H_1)$ as follows:

$$\begin{aligned}\Pr(E|H_1) &= \Pr\left(-\frac{\pi}{180} \leq \Theta, \Phi \leq \frac{\pi}{180} \middle| \Theta = 0, \Phi \neq 0, \pi\right) \\ &= \Pr\left(-\frac{\pi}{180} \leq \Phi \leq \frac{\pi}{180} \middle| \Theta = 0, \Phi \neq 0, \pi\right)\end{aligned}$$

Now, the value of this second expression is not given by the simple ratio formula $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$: the right-hand side of that formula is undefined in this case because plugging in the relevant values yields $0/0$.

A standard way around this problem is to use an analogous ratio of probability *density* functions formula to produce a conditional probability density function and then to integrate that function over the relevant domain. In this case, the relevant conditional probability density function would be

$$
\begin{aligned}
f_1(\Phi|\Theta = 0, \Phi \neq 0, \pi) &= \frac{f_2(\Phi, \Theta = 0|\Phi \neq 0, \pi)}{f_3(\Theta = 0|\Phi \neq 0, \pi)} \\
&= \frac{f_2(\Phi, \Theta = 0|\Phi \neq 0, \pi)}{\int_0^{\pi} f_2(\phi, \Theta = 0|\phi \neq 0, \pi)\, d\phi + \int_{\pi}^{2\pi} f_2(\phi, \Theta = 0|\phi \neq 0, \pi)\, d\phi} \\
&= \frac{\cos(0)/4\pi}{\int_0^{\pi} \cos(0)/4\pi\, d\phi + \int_{\pi}^{2\pi} \cos(0)/4\pi\, d\phi} \\
&= \frac{1}{\int_0^{\pi}\, d\phi + \int_{\pi}^{2\pi}\, d\phi} \\
&= \frac{1}{2\pi}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\Pr(E|H_1) &= \int_{-\pi/180}^{\pi/180} \frac{1}{2\pi}\, d\phi \\
&= \frac{1}{180}
\end{aligned}
$$

Notice that we would have gotten the same result if we had not conditioned on $\Phi \neq 0, \pi$. The only difference this conditioning made in our calculations was that it led us to write $f_3(\Theta = 0|\Phi \neq 0, \pi)$ as a sum of two integrals from 0 to π and from π to 2π rather than as a single integral from 0 to 2π.

We can use the same technique to calculate $\Pr(E|H_2)$:

$$
\begin{aligned}
\Pr(E|H_2) &= \Pr\left(-\frac{\pi}{180} \leq \Theta, \Phi \leq \frac{\pi}{180} \middle| \Phi \in \{0 \cup \pi\}, \Theta \neq 0\right) \\
&= \frac{1}{2} \Pr\left(-\frac{\pi}{180} \leq \Theta \leq \frac{\pi}{180} \middle| \Phi \in \{0 \cup \pi\}, \Theta \neq 0\right) \qquad \text{(Symmetry)}
\end{aligned}
$$

$$
\begin{aligned}
f_4(\Theta|\Phi \in \{0 \cup \pi\}, \Theta \neq 0) &= \frac{f_5(\Theta, \Phi \in \{0 \cup \pi\}|\Theta \neq 0)}{f_6(\Phi \in \{0 \cup \pi\}|\Theta \neq 0)} \\
&= \frac{f_5(\Theta, \Phi \in \{0 \cup \pi\}|\Theta \neq 0)}{\int_{-\pi/2}^{0} f_5(\theta, \Phi \in \{0 \cup \pi\}|\theta \neq 0)\, d\theta + f_5(\theta, \Phi \in \{0 \cup \pi\}|\theta \neq 0)\, d\theta} \\
&= \frac{\cos\Theta/2\pi}{\int_{-\pi/2}^{0} \cos\theta/2\pi\, d\theta + \int_{0}^{\pi/2} \cos\theta/2\pi\, d\theta} \\
&= \frac{\cos\Theta}{\int_{-\pi/2}^{0} \cos\theta\, d\theta + \int_{0}^{\pi/2} \cos\theta\, d\theta} \\
&= \frac{\cos\Theta}{[\sin(0) - \sin(-\pi/2)] + [\sin(\pi/2) - \sin(0)]} \\
&= \frac{\cos\Theta}{2}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(E|H_2) &= \frac{1}{2}\int_{-\pi/180}^{\pi/180} \frac{\cos\theta}{2}\, d\theta \\
&= \frac{1}{4}[2\sin(\pi/180)] \\
&= \frac{\sin(\pi/180)}{2}
\end{aligned}
$$

Again, notice that conditioning on $\Theta \neq 0$ made no difference to the ultimate outcome of this calculation.

The resulting likelihood ratio is

$$
\begin{aligned}
\frac{\Pr(E|H_2)}{\Pr(E|H_1)} &= \frac{\sin(\pi/180)/2}{1/180} \\
&= 90\sin(\pi/180) \\
&= 1.57
\end{aligned}
$$

APPENDIX F

CONSTRUCTION TO COMPLETE PROOF THAT PRAFE-RULES CAN FORCE ONE TO VIOLATE (R1)

Here is the construction needed to complete the proof in Section 5.7 that any PRAFE-rule can force one to violate (R1). Having fixed a PRAFE-rule, let a be some value of x such that $f(x) > 1$ for all $x > a$ for that PRAFE-rule. Let $b = 2a/(2a + 1)$. Then assign probabilities to outcomes according to Table F1. Each of the four cells in the interior of that table give the probability that X_1 has the value for the given row and X_2 has the value for the given column. Thus, for instance, the bottom-right cell in the interior of the table says that the probability that $X_1 = X_2 = 1$ is $(1 - b)/2$. The values all the way to the right give the probability that X_1 has the value associated with that row. The values all the way at the bottom give the probability that X_2 has the value associated with that column. The 1 in the bottom-right cell indicates that the row and column totals both sum to 1, as they should.

Table F1: A probability assignment that yields the desired result

X_1 \ X_2	0	1	
0	$\frac{1-b}{4}$	b	$\frac{1+3b}{4}$
1	$\frac{1-b}{4}$	$\frac{1-b}{2}$	$\frac{3-3b}{4}$
	$\frac{1-b}{2}$	$\frac{1+b}{2}$	1

It is easy to show that this table does indeed specify a probability distribution. Because $a > 1$, $b = 2a/(2a+1)$ is strictly between 2/3 and 1. It follows that all of the probabilities in the table are between zero and one. In addition, the marginal probabilities (those shown in the rightmost column and the bottom row) are the sums of the relevant joint probabilities (those in the interior of the table), and the marginal probabilities sum to one. Thus, the distribution is additive, and the axioms of probability are satisfied.

I will now show that $\Pr(E|H_1)/\Pr(E|\neg H_1)$ and $\Pr(E|\neg H_2)/\Pr(E|H_2)$ are both greater than a, completing the proof.

$$
\begin{aligned}
\frac{\Pr(E|H_1)}{\Pr(E|\neg H_1)} &= \frac{\Pr(E \;\&\; H_1)}{\Pr(H_1)} \frac{\Pr(\neg H_1)}{\Pr(E \;\&\; \neg H_1)} \\
&= \frac{\Pr(X_1 = X_2 = 1)}{\Pr(X_1 = X_2 = 1)} \frac{\Pr(\neg(X_1 = X_2 = 1))}{\Pr(X_1 = 1 \;\&\; X_2 = 0)} \\
&= \frac{1 - \Pr(X_1 = X_2 = 1)}{\Pr(X_1 = 1 \;\&\; X_2 = 0)} \\
&= \frac{1 - (1-b)/2}{(1-b)/4} \\
&= \frac{4}{1-b} - 2 \\
&= \frac{4}{1 - (2a)/(2a+1)} - 2 \\
&= \frac{8a+4}{2a+1-2a} - 2 \\
&= 8a + 4 - 2 \\
&= 8a + 2 \\
&> a
\end{aligned}
$$

$$
\begin{aligned}
\frac{\Pr(E|\neg H_2)}{\Pr(E|H_2)} &= \frac{\Pr(E \;\&\; \neg H_2)}{\Pr(\neg H_2)} \frac{\Pr(H_2)}{\Pr(E \;\&\; H_2)} \\
&= \frac{\Pr(X_1 = 1 \;\&\; X_2 = 0)}{\Pr(X_2 = 0)} \frac{\Pr(X_2 = 1)}{\Pr(X_1 = X_2 = 1)} \\
&= \frac{(1-b)/4}{(1-b)/2} \frac{(1+b)/2}{(1-b)/2} \\
&= \frac{1}{2} \frac{1+b}{1-b} \\
&= \frac{1}{2} \frac{1 + 2a/(2a+1)}{1 - 2a/(2a+1)} \\
&= \frac{1}{2} \frac{2a + 1 + 2a}{2a + 1 - 2a} \\
&= \frac{1}{2}(4a + 1) \\
&= 2a + \frac{1}{2} \\
&> a
\end{aligned}
$$

APPENDIX G

DESCRIPTION OF THE EXPERIMENT SHOWING THAT ANY PRAFE-RULE CAN FORCE ONE TO VIOLATE EITHER (R2) OR (R3)

Completing the proof given in Section 5.8 that PRAFE-rules can force one to violate either (R2) or (R3) requires providing a description of an experiment that yields pieces of evidence that are individually non-neutral but collectively neutral with respect to H_1^r, H_2^r, and H_3^r; individually non-neutral between $H_1^{r'}$ and $H_3^{r'}$; and collectively neutral with respect to $H_1^{r'}$, $H_2^{r'}$, and $H_3^{r'}$. One experiment of this kind consists of rolling three times a three-sided die with weights that depend on r as shown in table 2.

Each possible outcome has a probability that varies with r, but the sequence of outcomes $\{1, 2, 3\}$, for instance, has probability $1/2 \times 1/3 \times 1/6 = 1/36$ regardless of r. Thus, that sequence contains three pieces of evidence that are individually non-neutral but collectively neutral with respect to H_1^r, H_2^r, and H_3^r.[1]

Because there is a one-to-one correspondence between values of r and values of r', we can also say that the die has weights that depend on r' as shown in Table G1.

If $H_1^{r'}$ is true (meaning that r' is in the interval $[1/2, 1)$), then r' is in either $[1/2, 2/3)$ or $[2/3, 1)$. Now, one cannot compute the likelihood $\Pr(E|H_1^{r'})$ on $E = 1$, $E = 2$, or $E = 3$ without assigning prior probabilities to $[1/2, 2/3)$ and $[2/3, 1)$ given $H_1^{r'}$. However, $\Pr(1|H_1^{r'})$ must be some weighted average of $\Pr(1|r' \in [1/2, 2/3)) = 1/3$ and $\Pr(1|r' \in [2/3, 1)) = 1/6$,

[1] If three pieces of relevant but collectively neutral evidence are not enough for sufficiently definite degrees of relative acceptance (see subsection 5.8.2), then we can imagine that instead of one roll of 1, one roll of 2, and one roll of 3, we have n rolls of each kind for n as large as one likes.

Table G1: Probabilities of possible die roll outcomes as a function of r'

If r' is in...	**...then** $\Pr(1) =$	**...then** $\Pr(2) =$	**...then** $\Pr(3) =$
$[1, 2)$	$1/2$	$1/3$	$1/6$
$[2/3, 1)$	$1/6$	$1/2$	$1/3$
$[1/2, 2/3)$	$1/3$	$1/6$	$1/2$

and thus must be less than $\Pr(1|H_3^{r'}) = \Pr(1|r' \in [3/2, 2)) = 1/2$. Thus, a likelihoodist can say that $E = 1$ favors $H_3^{r'}$ over $H_1^{r'}$ even if he or she is unwilling to assign prior probabilities and thus unable to assess the degree of favoring. $\Pr(\{1, 2, 3\}|H_1^{r'})$ must be some weighted average of $\Pr(\{1, 2, 3\}|r' \in [1/2, 2/3)) = 1/36$ and $\Pr(\{1, 2, 3\}|r' \in [2/3, 1)) = 1/36$, so it must be $1/36$. Thus, the evidence from the experiment is collectively but not individually neutral between $H_1^{r'}$ and $H_3^{r'}$. By a similar argument $\Pr(\{1, 2, 3\}|H_2^{r'}) = 1/36$, so the evidence is collectively neutral among $H_1^{r'}$, $H_2^{r'}$, and $H_3^{r'}$. This point completes the proof.

BIBLIOGRAPHY

Peter Armitage. Contribution to "consistency in statistical inference and decision". *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):30–1, 1961.

B. C. Arnold and C. A. Robertson. The conditional distribution of x given x=y can be almost anything! In *Advances on Theoretical and Methodological Aspects of Probability and Statistics*, pages 75–81. CRC Press, April 2003.

G. A. Barnard. The logic of statistical inference. *The British Journal for the Philosophy of Science*, 23(2):123–32, 1972.

Robert Bartlett, Dietrich Roloff, Richard Cornell, Alice Andrews, Peter Dillon, and Joseph Zwischenberger. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics*, 76(4):479–487, 1985.

D. Basu. Recovery of ancillary information. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 26(1):3–16, 1964.

D. Basu. Statistical information and likelihood [with discussion]. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 37(1):1–71, 1975.

M.J. Bayarri, M. H. DeGroot, and J. B. Kadane. In S.S. Gupta and J.O. Berger, editors, *Statistical Decision Theory and Related Topics IV*. Springer-Verlag, 1988.

Colin Begg. Investigating therapies of potentially great benefit: ECMO: Comment. *Statist. Sci.*, 4(4):320–322, 11 1989.

James Berger. *Statistical Decision Theory, Foundations, Concepts, and Methods*. Springer Series in Statistics: Probability and its Applications. Springer-Verlag, 1980.

James Berger. The Case for Objective Bayesian Analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

James Berger and Robert Wolpert. *The Likelihood Principle*, volume 6 of *Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH, 2nd edition, 1988.

Donald Berry. Investigating Therapies of Potentially Great Benefit: ECMO: Comment: Ethics and ECMO. *Statist. Sci.*, 4(4):306–310, 11 1989.

Joseph Bertrand. *Calcul des Probabilitiés*. Gauthier-Villars, Paris, 1889.

Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

Allan Birnbaum. The anomalous concept of statistical evidence: Axioms, interpretations, and elementary exposition. Technical Report IMM-NYU 332, New York University Courant Institute of Mathematical Sciences, October 1964.

Allan Birnbaum. On Durbin's modified principle of conditionality. *Journal of the American Statistical Association*, 65(329):402–3, 1970a.

Allan Birnbaum. Statistical methods in scientific inference. *Nature*, 225(5237):1033, Mar 1970b.

Allan Birnbaum. More on concepts of statistical evidence. *Journal of the American Statistical Association*, 67(340):858–61, 1972.

Allan Birnbaum. Comments on paper by J. D. Kalbfleisch. *Biometrika*, 62(2):262–4, 1975.

Allan Birnbaum. The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese*, 36(1): 19–49, 1977.

Émile Borel. *Eléments de la Théorie des Probabilités*. Paris, 1909.

Peter Brössel and Franz Huber. Bayesian confirmation: A means with no end. *The British Journal for the Philosophy of Science*, 2014.

George Casella and Roger Berger. *Statistical Inference*. Duxbury Advanced Series in Statistics and Decision Sciences. Thomson Learning, Pacific Grove, CA, 2nd edition, 2002.

Jake Chandler. Contrastive confirmation: Some competing accounts. *Synthese*, 190(1): 129–38, 2013.

David Cox. Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–72, 1958.

David Cox and David Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

David Cox and Deborah Mayo. Statistical scientist meets a philosopher of science: A conversation. *Rationality, Markets and Morals*, 2:103–14, 2011.

A. P. Dawid. Conformity of inference patterns. In J.R. Barra and the European Meeting of Statisticians, editors, *Recent Developments in Statistics: Proceedings of the European Meeting of Statisticians, Grenoble, 6-11 Sept., 1976*, pages 245–67. North-Holland, Amsterdam, 1977.

Bruno de Finetti. *Theory of Probability: a Critical Introductory Treatment*. Wiley series in probability and mathematical statistics. Wiley, 1975.

Lester Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3(1):89–99, 1975.

James Durbin. On Birnbaum's theorem on the relation between Sufficiency, Conditionality and Likelihood. *Journal of the American Statistical Association*, 65(329):395–8, 1970.

Kenny Easwaran. *The Foundations of Conditional Probability*. PhD thesis, University of California, Berkeley, 2008.

A.W.F. Edwards. *Likelihood. An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference.* Cambridge University Press, 1972.

Michael Evans, Donald Fraser, and Georges Monette. On principles and arguments to Likelihood. *Canadian Journal of Statistics*, 14(3):181–94, 1986.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.

R. A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A*, 144(852):285–307, 1934.

R. A. Fisher. Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):69–78, 1955.

Branden Fitelson. *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin, 2001.

Branden Fitelson. Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, 156: 473–489, 2007.

Branden Fitelson. Contrastive Bayesianism. In Martijn Blaauw, editor, *Contrastivism in Philosophy*, Routledge Studies in Contemporary Philosophy, chapter 3, pages 64–87. Routledge, 2013.

Malcolm Forster. Predictive accuracy as an achievable goal of science. *Philosophy of Science*, 69(S3):S124–34, 2002.

Malcolm Forster. Counterexamples to a likelihood theory of evidence. *Minds and Machines*, 16(3):319–338, August 2006.

Malcolm Forster and Elliott Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994.

DAS Fraser, G Monette, and KW Ng. Marginalization, likelihood and structured models. In *Multivariate analysis–VI: proceedings of the Sixth International Symposium on Multivariate Analysis*, volume 6, page 209. North Holland, 1985.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis, Second Edition*. Taylor & Francis, 2nd edition, 2003.

Andrew Gelman. It is not necessary that Bayesian methods conform to the Likelihood Principle. *Statistical Modeling, Causal Inference, and Social Science* (Blog), 2012. <andrewgelman.com/2012/10/it-not-necessary-that-Bayesian-methods-conform-to-the-likelihood-principle/>. November 1 comment.

Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

Gerd Gigerenzer. Mindless statistics. *Journal of Socio-Economics*, 33(5):587–606, 2004.

Joshua Greene, R. Brian Sommerville, Leigh Nystrom, John Darley, and Jonathan Cohen. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537): 2105–2108, 2001.

Steven Gross, Ellen Bifano, Diane D'Euqenio, David Hakanson, and Robert Hinqre. Prospective randomized controlled trial of conventional treatment or transport for ECMO in infants with severe persistent pulmonary hypertension (PPHN): Two year follow up. *Pediatric Research*, 36(1):17A, 1994.

Jason Grossman. Statistical inference: From data to simple hypotheses. Unpublished manuscript. Available at <bunny.xeny.net/linked/Grossman-statistical-inference.pdf>, 2011a.

Jason Grossman. The Likelihood Principle. In D.M. Gabbay, P.S. Bandyopadhyay, P. Thagard, and M.R. Forster, editors, *Philosophy of Statistics*, Handbook of the Philosophy of Science, pages 553–80. Elsevier, Amsterdam, 2011b.

Ian Hacking. *Logic of Statistical Inference*. Cambridge University Press, 1965.

Alan Hájek. What conditional probability could not be. *Synthese*, 137(3):273–323, 2003.

Bruce Hill. *The Likelihood Principle*, volume 6 of *Lecture Notes—Monograph Series*, chapter Discussion by Bruce M. Hill, pages 161–74.4. Institute of Mathematical Statistics, Beachwood, OH, 2nd edition, 1988.

David Hume. *Essays and Treatises on Several Subjects*, volume 2. Bell and Bradfute, Edinburgh, 1825.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

V. M. Joshi. Fallacy in the proof of Birnbaum's theorem. *Journal of Statistical Planning and Inference*, 26(1):111–2, 1990.

John D. Kalbfleisch. Sufficiency and Conditionality. *Biometrika*, 62(2):251–9, 1975.

Kevin Kelly. A new solution to the puzzle of simplicity. *Philosophy of Science*, 74(5):561–573, 2007.

J. Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72(360):789–808, 1977.

A. N. Kolmogorov. *Foundations of the Theory of Probability*. AMS Chelsea Publication. Chelsea Pub. Co., 1956.

Thomas Kuhn. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press, 1977.

Stephen Leeds. Other minds, support, and likelihoods, 2004. Unpublished manuscript. Available at <http://philsci-archive.pitt.edu/5472/>.

Yiyuan Li, Hong Li, Jean Decety, and Kang Lee. Experiencing a natural disaster alters childrens altruistic giving. *Psychological Science*, 24(9):1686–1695, 2013.

Edouard Machery. In defense of reverse inference. *The British Journal for the Philosophy of Science*, 65(2):251–267, 2014.

Deborah Mayo. Behavioristic, evidentialist, and learning models of statistical testing. *Philosophy of Science*, 52(4):493–516, 1985.

Deborah Mayo. Did Pearson reject the Neyman-Pearson philosophy of statistics? *Synthese*, 90(2):233–62, 1992.

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

Deborah Mayo. An error in the argument from Conditionality and Sufficiency to the Likelihood Principle. In Deborah Mayo and Aris Spanos, editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, chapter 7(III), pages 305–14. Cambridge University Press, 2009a.

Deborah Mayo. Learning from error, severe testing, and the growth of theoretical knowledge. In Deborah Mayo and Aris Spanos, editors, *Error and Inference: Recent Exchanges*

on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, chapter 1, pages 28–57. Cambridge University Press, 2009b.

Deborah Mayo. On the Birnbaum argument for the Strong Likelihood Principle. Unpublished. Available at <www.phil.vt.edu/dmayo/conference_2010/9-18-12MayoBirnbaum.pdf>, September 2012.

Deborah Mayo. On the Birnbaum argument for the Strong Likelihood Principle. *Statist. Sci.*, 29(2):227–239, 05 2014.

Deborah Mayo and Aris Spanos. *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press, 2009.

Deborah Mayo and Aris Spanos. Error statistics. In D.M. Gabbay, P.S. Bandyopadhyay, P. Thagard, and M.R. Forster, editors, *Philosophy of Statistics*, Handbook of the Philosophy of Science, pages 153–98. Elsevier, Amsterdam, 2011.

Peter Milne. log[p(h/eb)/p(h/b)] is the one true measure of confirmation. *Philosophy of Science*, 63(1):21–26, 1996.

Jerzy Neyman. "Inductive behavior" as a basic concept of philosophy of science. *Review of the International Statistical Institute*, 25(1/3):7–22, 1957.

Jerzy Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A(1/2):175–240, 1928.

Jerzy Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231: 289–337, 1933.

John Norton. Ignorance and indifference. *Philosophy of Science*, 75:45–68, 2008.

E. S. Pearson. Some thoughts on statistical inference. *The Annals of Mathematical Statistics*, 33(2):394–403, 1962.

K.R. Popper. *The Logic of Scientific Discovery*. Harper & Row, 1968.

Malempati Madhusudana Rao. *Conditional Measures and Applications*, volume 271. CRC Press, 2005.

Alfréd Rényi. On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 6 (3):285–335, 1955.

Michael Rescorla. Some epistemological ramifications of the Borel-Kolmogorov paradox. *Synthese*, pages 1–33, 2014.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

Richard Royall. Investigating therapies of potentially great benefit: ECMO: Comment. *Statist. Sci.*, 4(4):318–319, 1989.

Richard Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London, 1997.

Richard Royall. On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451):760–8, 2000.

L. J. Savage. *The Foundations of Statistical Inference: A Discussion*. Methuen's Monographs on Applied Probability and Statistics. Methuen, London, 1962.

L. J. Savage. Comments on a weakened principle of conditionality. *Journal of the American Statistical Association*, 65(329):399–401, 1970.

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

Teddy Seidenfeld. Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In Vincent F. Hendricks, Stig Andur Pederson, and Klaus Frovin Jørgensen, editors, *Probability Theory: Philosophy, Recent History and Relations to Science*. Synthese Library, Kluwer, 2001.

Elliott Sober. Parsimony in systematics: Philosophical issues. *Annual Review of Ecology and Systematics*, 14(1):335–357, 1983.

Elliott Sober. Is drift a serious alternative to natural selection as an explanation of complex adaptive traits? *Royal Institute of Philosophy Supplements*, 56:10–11, 2005.

Elliott Sober. *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press, 2008.

Aris Spanos. Statistical adequacy and the trustworthiness of empirical evidence: Statistical vs. substantive information. *Economic Modelling*, 27(6):1436–52, 2010.

Daniel Steel. Bayesian confirmation theory and the Likelihood Principle. *Synthese*, 156(1): 53–77, 2007.

Jacob Stegenga. Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4):497–507, 2011.

Charles Stein. A remark on the Likelihood Principle. *Journal of the Royal Statistical Society. Series A (General)*, 125(4):565–568, 1962.

Mervyn Stone. Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71(353):114–116, 1976.

Patrick Suppes. Models of data. *Logic, Methodology and Philosophy of Science*, 898:252–261, 1962.

UK Collaborative ECMO Trial Group. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. *International Journal of Gynecology & Obstetrics*, 55 (3):314–15, 1996.

Bas van Fraassen. *The Scientific Image*. Clarendon Press, Oxford, UK, 1980.

Richard von Mises. *Probability, Statistics, and Truth*. Courier Corporation, 1957.

James Ware. Investigating therapies of potentially great benefit: ECMO. *Statistical Science*, 4(4):298–306, 1989.

Larry Wasserman. Statistical principles? *Normal Deviate* (Blog), 2012. <normaldeviate.wordpress.com/2012/07/28/statistical-principles>. July 28 blog post.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in the Philosophy of Science. Oxford University Press, USA, 2003.