New Responses to Three Counterexamples to the Likelihood Principle

By Greg Gandenberger

March 14, 2014

Note: This document is a work in progress. Please do not cite it without permission. Comments are welcome at greg@gandenberger.org.

Contents

1 Introduction **2**

2 Why a counterexample to the Law of Likelihood is a *prima facie* counterexample to the Likelihood Principle **4**

3 Response to Fitelson's counterexample **6**

3.1 Objection 1: Response conflicts with constraints on evidential favoring 8

3.2 Objection 2: Response fails to address tacking paradox 10

3.2.1 Response 1: Bite the bullet 11

3.2.2 Response 2: Regard Law of Likelihood as explicating "r-favoring" 12

3.2.3 Response 3: Restrict Law of Likelihood to structurally identical alternatives 14

3.3 Objection 3: Response excludes cases of scientific interest 15
3.3.1 Cases involving competing causal claims 15
3.3.2 Cases involving nested models 18

4 Response to Armitage's counterexample 21
4.1 Two problems for attempts to use the Weak Repeated Sampling Principle as an objection to the Law of Likelihood 22
4.1.1 Problem 1: The Weak Repeated Sampling Principle is unreasonably strong 22
4.1.2 Problem 2: The Law of Likelihood itself cannot conflict with the Weak Repeated Sampling Principle 25
4.2 Why Armitage's example is no threat to the Law of Likelihood . 26

5 Response to Stein's counterexample 28

6 Conclusion 34

A Proof that (CE) and (†) are jointly incompatible with six leading measures of confirmation 38

B Technical details from Stein's example 42

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 Likeli-

hood 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.

I present an argument from the Likelihood Principle to the Law of Likelihood in Section 2. I then respond to one purported counterexample in each of Sections 3–5. Those counterexamples come from Fitelson (2007), Armitage (first presented in Smith 1961), and Stein (1962), respectively.

There are many purported counterexamples to the Likelihood Principle that I do not consider in this chapter. Many are variants on two examples I discuss in the next chapter of this dissertation, namely the example due to Birnbaum (1964, 12–3) that I label Example 1 in that chapter and the example due to Stone (1976) that I discuss in an appendix. The rest of the counterexamples of which I am aware are relatively simple and have been satisfactorily addressed elsewhere (see e.g. 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.

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 and 4, but not the one presented in Section 5) simply by denying the Law of Likelihood. I argue in this section that this response requires denying that there is a correct measure of evidential favoring. Some, particularly among Bayesians, will be willing to accept this conclusion. For others, a counterexample to the Law of Likelihood is a counterexample to the Likelihood Principle.

The Likelihood Principle says that the evidential meaning of datum E with respect to a set of hypothesis $\mathbf{H}$ each of which ascribes a determinate probability to E depends only on the probabilities which those hypotheses ascribe to E, up to a constant of proportionality. The Law of Likelihood is typically formulated[1] as the claim that E evidentially favors some H_1 of this kind over another H_2 of this kind if and only if the former gives higher probability to that datum than the latter, with the ratio of those probabilities measuring the degree of that favoring.[2] More compactly, E favors H_1 over H_2 if and only if the likelihood ratio $\mathcal{L} = \frac{\Pr(E;H_1)}{\Pr(E;H_2)} > 1$, and $\mathcal{L}$ measures the degree of that favoring. I take it that evidential favoring is defined only up to strictly monotone increasing

[1] I write "The Law of Likelihood is typically formulated as..." rather than "The Law of Likelihood states..." because I argue in Section 3 that the Law of Likelihood is appropriate only for mutually exclusive hypotheses, contrary to this formulation.

[2] Throughout this chapter, I assume that all hypotheses are simple, discrete statistical hypotheses in the sense that they ascribe a definite probability to the datum in question and that this probability is strictly between zero and one. Composite hypotheses are outside the scope of the Law of Likelihood. Continuous hypotheses raise technical issues that are not of primary interest here. Ignoring them is permissible because they are merely useful idealizations: all real data is discrete because of the finite precision of our measuring instruments.

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 ascribe a determinate probability to datum E, either E is neutral between H_1 to H_2 or it favors H_1 over H_2 to a determinate, real-valued degree or vice versa. The Likelihood Principle says that the evidential meaning of E with respect to a 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 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

[3] 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.

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 $\mathcal{L}$.

This argument shows that for anyone who believes that a correct measure of evidential favoring exists, a counterexample to the Law of Likelihood is a counterexample to the Likelihood Principle.

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 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.

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.

[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 $l_1 > l_2$, 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.

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 hypotheses that are genuine competitors in the sense that only one of them can be true. 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 Likelihoood to mutually exclusive hypotheses is indeed necessary to avoid counterexamples.

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 3.1, 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 3.2, I respond to the objection that it fails to solve the "tacking paradox." In Subsection 3.3, I

[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 nonmonotonic. 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.

respond to the objection that it excludes cases of genuine scientific interest.

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]

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.

[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.

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 A 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 seems to me 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 does not seem to me undesirable.

Restricting (†) in this way is not "too easy," but rather is well motivated by considerations of linguistic naturalness 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′).

3.2 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 1 that give rise to a version of the tacking paradox[9] One such example is as follows (Fitelson, 2013, 77):

> **Example 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.

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

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 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.

3.2.1 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 my intuitions about the tacking paradox do not make me unique, they do seem to place me in the minority. For those who differ from me in this respect, I offer two additional responses to the tacking paradox.

3.2.2 Response 2: Regard Law of Likelihood 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.

I have no objection to this response to the tacking paradox for the Law of Likelihood except that my own intuitions about evidential favoring do not require it. A reader whose intuitions differ from mine in this respect is welcome to adopt it.

A true likelihoodist would balk at regarding the Law of Likelihood as explicating r-favoring because r-favoring is defined in terms of inductive probabilities that likelihoodists regard as illegitimate in typical scientific applications. The goal of likelihoodism is to provide an account of scientific evidence that does not appeal to such probabilities. I will argue in the next chapter that likelihoodism fails to satisfy this goal, but for those who disagree with me on this issue there is a third possible response to the tacking paradox.

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 (suggested but not embraced by Steel 2007, 68) 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 rejects his own proposal to restrict the Law of Likelihood to structurally identical alternatives 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 hypoeses 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.

Developing this proposal fully 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. Regardless of the details, the worst case for the Law of Likelihood is that the tacking paradox shows that it does not apply directly to the evaluation of high-level theories. 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 everday 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.

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.

3.3.1 Cases involving competing causal claims

Machery (2013) 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.

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 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. Yet it does seem possible to test H_1 and H_2 against one another.

These statements can be generalized to any case involving claims of the form "X causes Y" and "Z causes Y." 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 such an argument is that it requires more than that one can both interpret claims of the form "X causes Y" and "Z causes Y" as com-

patible and regard them as being tested against one another. It requires that one can do both of these things *simultaneously*. It does not seem plausible to me that scientists can truly be said to test claims of the form "X is a cause of Y" and "Z is a cause Y" 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 Y" and "Z causes Y" 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 Y and Z does not" and "Z causes Y 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 Y" and "Z is the most important cause of Y" (or perhaps "X is a more important cause of Y than Z" and vice versa), where the notion of importance would generally be operationalized as the percent of variance in Y 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 Y" and "Z causes Y" against one another directly, but rather a more specific claim "in the spirit of 'X causes Y' " against a more specific claim "in the spirit of 'Z causes Y.' " 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 interpretation seems quite plausible in the case of Example 3. It is perfectly compatible with the possibility 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 must account.

A full defense of any particular interpretation of Example 3 lies outside the scope of this dissertation. Suffice it to say that there are plausible interpretations of this case according to which the claims Greene et al. actually test against one another in particular experiments are mutually exclusive. Thus, Greene et al.'s work does not seem to be a counterexample to my conjecture that some suitable interpretation of the claims "X causes Y" and "Z causes Y" can be found that make them mutually exclusive in every case in which scientists can truly be said to test such claims against one another.

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 mod-

els, there are many 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 favours 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 (in unusual cases only) 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.

Scientists may 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 (2011) 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 no 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]

[10] Formally, three or more data points will favor (QUAD) over (LIN) with probability one if (LIN) is true. It is *possible* for three or more data points to fall on a single straight line, but this event has probability zero if (LIN) is true because its error term has a continuous distribution.

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 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.

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 meaningfulness or appropriateness of distinctively Bayesian statements that involve subjective probabilities.

The upshot of Fitelson's counterexample to the Law of Likelihood 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.

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]

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

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

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, Pro-

[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 Jerzy 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). It is a broader claim than the Weak Repeated Sampling Principle in some loose sense, but it is not logically stronger because it does not saying anything about *how* behavior in hypothetical repetitions is to be used in choosing statistical procedures.

cedure 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 one to 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. 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 what frequentists call 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

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

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 at least as likely as not to deliver the wrong answer. Using a different kind of procedure will not help: the procedure that uses a simple cutoff c has the the lowest Type II error rate for each element of the alternative hypotheses among all procedures with no more than its Type I error rate.[14]

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 "more often than 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 does quality 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 $\mu_0 + \epsilon$ for any positive ϵ does not go to one and can be made as small as one likes in principle by increasing the sample size and decreasing c. 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 4.2 that it also defuses Armitage's purported

[14]The procedure that uses a simple cutoff c is in this sense the *uniformly most powerful* test of H_0 against its negation. This fact is a consequence of the Karlin-Rubin theorem, which is a generalization of the Neyman-Pearson lemma.

counterexample to the Law of Likelihood.

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 use maximum likelihood estimation except in special circumstances or for rhetorical or pragmatic reasons.

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 a conclusion 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.

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.[15]

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,[16] even when in fact $\mu = 0$. 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

[15] 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).

[16] 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 $\mu \neq 0$.

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 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 distribution 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 favors 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 $|\bar{x}_n|$). 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, which is not problematic at all.

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 (2011, 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 and is the shortest among all random intervals 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 ($\theta = 0$). 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 without justifying the choice of that random interval for that purpose.

Stein begins[17] 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[18]) 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.

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.

[17] 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.

[18] 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.

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.[19] 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

[19] 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 only "evidentialist" forms of frequentism in the Fisherian tradition because only those forms of frequentism can conflict with the Likelihood Principle.

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 $(X-1.96\sigma_0, X+1.96\sigma_0)$ 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 in the 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.[20]

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 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 Prin-

[20] 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 obviously is in this case.

ciple 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 (2011, 311–3) make strong arguments for this claim,[21] 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 single random variable Y. (See Appendix B 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.

6 Conclusion

I have provided new responses to three purported counterexamples to the Likelihood Principle. Together with the proof of the Likelihood Principle presented

[21] 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.

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 two chapters, I argue that likelihoodism is not a viable genuine alternative to Bayesianism and consider various additional objections to likelihoodism. In the subseqeunt chapter, I examine the prospects for reliabilist forms of frequentism that are consistent with the Likelihood Principle and the Law of Likelihood because they deny the evidentialist assumption that acceptable methods of statistical inference must conform to explications that succeed in rendering precise and systematic our informal evidential notions.

A Proof that (CE) and (†) are jointly incompatible with six leading measures of confirmation

Here I demonstrate the statement made on page 9 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 3.1

1. (d) $c(H, E) = \Pr(H|E) - \Pr(H)$.

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

2. (s) $c(H, E) = \Pr(H|E) - \Pr(H|\bar{E})$.

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

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

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

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

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

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

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

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

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

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

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

3. (c) $c(H, E) = \Pr(H\&E) - \Pr(H)\Pr(E)$.

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

(Shown in item 1 above.)

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

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

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

4. (n) $c(H,E) = \Pr(E|H) - \Pr(E|\bar{H})$

$$
\begin{aligned}
0.9801 &> 0.01 \\
(0.99)^2 &> 0.01 \\
\frac{.99}{.01} &> \frac{1}{.99} \\
\frac{\Pr(H_2|E)}{\Pr(H_2)} &> \frac{\Pr(H_1|E)}{\Pr(H_1)} \\
\frac{\Pr(H_2|E)\Pr(E)}{\Pr(H_2)} &> \frac{\Pr(H_1|E)\Pr(E)}{\Pr(H_1)} \\
\Pr(E|H_2) &> \Pr(E|H_1) \\
2\Pr(E|H_2) &> 2\Pr(E|H_1) \\
\Pr(E|H_2) - \Pr(E|H_1) &> \Pr(E|H_1) - \Pr(E|H_2) \\
\Pr(E|H_2) - \Pr(E|\bar{H_2}) &> \Pr(E|H_1) - \Pr(E|\bar{H_1}) \quad (H_1 \text{ and } H_2 \text{ are exhaustive}) \\
c(H_2,E) &> c(H_1,E)
\end{aligned}
$$

5. (l) $c(H,E) = \log\left[\frac{\Pr(E|H)}{\Pr E|\bar{H})}\right]$

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

6. (r) $c(H,E) = \log\left[\frac{\Pr(H|E)}{\Pr(H)}\right]$.

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

B 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

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

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 (2011, 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.