Why I Am Not a Likelihoodist

Greg Gandenberger

October 17, 2013

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 Please do not cite it without permission.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

1 Introduction

Likelihoodism is an alternative to Bayesian and frequentist approaches to statistics. Likelihoodists agree with Bayesians in accepting the Likelihood Principle and the Law of Likelihood, but agree with frequentists in denying that appeals to prior probabilities that cannot be interpreted objectively are appropriate for use in science. They thus combine what seem to many to be the best features of each of those methodologies.

I contend that despite its *prima facie* appeal, likelihoodism is not a viable alternative to Bayesianism and frequentism because it provides no alternative to those methodologies for using data to guide our beliefs and actions. Likelihoodists themselves admit that their methods merely characterize data as evidence, without providing immediate guidance for belief or action, but claim that characterizing data as evidence is valuable in itself (see e.g. Royall 1997, Ch. 1).

The claim that a statistical methodology that in itself provides no guidance for belief or action is useful is remarkably indifferent to practical considera-

tions. Beyond pointing out this fact, I do not know how to argue against that claim directly. My purpose in this paper is to argue against various possible likelihoodist attempts to make it seem plausible. In Section 2, I argue against the assumption that characterizing data as evidence must be valuable in itself because there is a close relationship between evidence and rational belief. In Section 3, I criticize likelihoodist appeal to analogies between evidential favoring and physical magnitudes such as heat or temperature. In Section 4, I argue that it does not help the likelihoodist cause to say that the Law of Likelihood provides a *ceteris paribus* norm of belief or action.

Others (e.g. Mayo 1996, Ch. 10) have argued that likelihoodism is false. I contend not that it is false, but that it is idle. I am far from the first to make this claim (see e.g. Savage 1962), but I am the first to present many of the arguments given below.

2 The Relationship Between Evidence and Rational Belief Fails to Vindicate Likelihoodism

It is quite plausible, especially from an empiricist perspective, to think that characterizing data as evidence is a first step toward using that data to guide one's beliefs and actions. As Hume wrote in his *Treatise*, "a wise man proportions his belief to the evidence" (1 3.13.7). (I presume that a wise woman does likewise.)

My argument is quite compatible with this view. The trouble with likelihoodism is not that likelihood functions are useless, but that likelihoodists provide no constructive alternative to Bayesian and frequentist ways of using them. As far as the relationship between evidence and rational belief is concerned, likelihoodists simply endorse a Bayesian account when prior probabil-

ities are available and provide no account of rational belief at all when prior probabilities are not available. Sober (2008, 32–5) characterizes this restraint as proper epistemic *modesty*, but it is also a form of epistemic *impotence*.

I clarify these claims in Subsection 2.1 by distinguishing between incremental and absolute notions of evidence and arguing that likelihoodist principles explicate an incremental notion where an absolute notion is needed. I articulate a further difficulty for the idea that the relationship between evidence and rational belief in Subsection 2.2, namely that the evidential import of what one learns upon receiving a datum is often very different from the evidential import of the datum itself.

2.1 Absolute vs. Incremental Notions of Evidence

Since Carnap (1962, new preface), it has been standard for philosophers of science to distinguish between two notions of confirmation that are today called *absolute* and *incremental*. The absolute notion concerns the "firmness" of a proposition in light of one's total evidence, whereas the incremental notion concerns the "increase in firmness" due to some particular datum. Bayesians generally maintain that E absolutely confirms H to degree k relative to background knowledge K if and only if $\Pr(H|E\&K) \geq k$ (Fitelson, 2001, 4). E incrementally confirms H relative to background knowledge K if and only if $\Pr(H|E\&K) > \Pr(H|K)$.

We can make the same distinction between notions of evidential favoring by saying that E favors H_1 over H_2 evidentially in an absolute sense to the extent that E makes H_1 more belief-worthy than H_2, and that E favors H_1 over H_2 evidentially in an incremental sense to the extent that E increases the belief-worthiness of H_1 relative to that of H_2 (all relative to some state of background knowledge K).

A likelihood ratio is a measure of evidential favoring in an incremental rather than an absolute sense. One standard justification for the Law of Likelihood is that a likelihood ratio for a pair of hypotheses on some datum is the ratio of the posterior odds for those hypotheses given that datum under Bayesian conditioning to the prior odds for those hypotheses without that datum. In that way, a likelihood ratio reflects the degree to which a datum *changes* the relative belief-worthiness of two hypotheses, rather than simply reflecting the relative belief-worthiness of those hypotheses in light of that datum and background knowledge. Likelihoodists maintain that likelihood ratios mean the same as measures of evidential favoring regardless of whether prior probabilities are available or not, which implies that they are measures of evidential favoring in an incremental sense whether prior probabilities are available or not.

Philosophers of science have long sought a "logic of induction" that would generalize the logic of deduction by allowing one to compute the *degree* to which an arbitrary set of premises entails an arbitrary conclusion. Some claim that Bayesianism is such a logic (e.g. Howson 1997), while others are dissatisfied with Bayesianism because assertions of prior probabilities are in general not truth-valued propositions[1]. The Law of Likelihood is appealing because it provides a measure of evidential favoring that does not involve prior probabilities, which seems *prima facie* to be what non-Bayesian philosophers of science who seek a logic of induction are looking for. But the Law of Likelihood is not a logic of induction for two reasons. First, it is a measure of evidential *favoring*, which is a three-place relation between a datum and a pair of hypotheses rather than a two-place relation between a datum and a single hypothesis. More importantly, it is a measure of evidential favoring in an incremental sense, so it allows one to

[1] Assertions of prior probabilities are truth-valued propositions in degenerate cases in which they take the value 0 or 1 and scientifically rare cases in which they can be interpreted as objective properties (such as propensities or long-run frequencies) of the system in question. They are not truth-valued propositions when they represent intermediate degrees of belief.

compute the degree to which an arbitrary set of premises entails an arbitrary conclusion only relative to a prior belief state.

The distinction between incremental and absolute notions of evidential support is no minor quibble. As the following example illustrates, an arbitrarily large degree of incremental evidential favoring may not suffice for a high degree of absolute evidential favoring even when neither of the hypotheses in question is silly or outlandish.

> **Example 1.** Suppose one were to examine a series of n radioactive isotopes and record a "1" in a sequence for each isotope that decayed within the half-life of its isotope species and a "0" for each isotope that did not decay in that time period. This procedure would yield a sequence of zeros and ones of length n, such as $x_0 = 01000100$ for $n = 8$. According to the Law of Likelihood, this sequence inevitably favors the hypothesis that the data had to turn out just as it did over the hypothesis that radioactive decay is a genuinely indeterministic hypothesis with probability $1/2$ of occurring within the half-life of the species of the isotope in question, to the degree 2^n. For instance, the Law of Likelihood says that x_0 favors the hypothesis that the second and sixth isotopes were bound to decay within the half-lives of their respective species while the other isotopes were bound not to decay over the hypothesis that each had equal chances of decaying and not decaying, to the degree $2^8 = 256$. Yet, according to both intuition and the probability calculus (see below), x_0 does not necessarily make the former deterministic hypothesis that predicted the data more belief-worthy than the latter indeterministic hypothesis.

Likelihoodists customarily interpret a likelihood ratio of 8 as "fairly strong" evidence and a likelihood ratio of 32 as "strong" evidence (Royall, 2000, 761). Thus, a datum with a likelihood ratio of 256 would generally be considered very

strong evidence indeed. Moreover, one could make this number as large as one likes in this example by increasing n, and the same points would apply.

Examples like Example 1 that involve comparing some hypothesis according to which the data were indeterministic to the hypothesis that the data were bound to come out exactly as they did have often been presented as counterexamples to the Likelihood Principle and the Law of Likelihood (e.g. Birnbaum 1964, 12–3). They do show that the Law of Likelihood is false for evidential favoring in an absolute sense: no degree of likelihoodist evidential favoring suffices in general to make the favored hypothesis more belief-worthy than the disfavored hypothesis. A Bayesian analysis clarifies this point. As Royall (1997, 13–5) explains in the context of a similar example, the probability for a Bayesian that radioactive decay is indeterministic is unchanged in this experiment regardless of the result if the only hypotheses on the table are the hypothesis that each isotope has probability one-half of decaying and the set of all hypotheses according to which some sequence of zeros and ones of length n was bound to occur and the hypotheses in that latter set all have equal prior probabilities. The datum refutes all but one of the deterministic hypotheses and has the effect under Bayesian updating of reassigning the probability masses of those hypotheses to the deterministic hypothesis that remains unrefuted.

Examples like Example 1 do not show that the Law of Likelihood is false for evidential favoring in an incremental sense. The arguments I give in Chapter 1 for the Likelihood Principle and in Chapter 2 for the Law of Likelihood still apply. Moreover, the likelihood ratio is still for a Bayesian the ratio of the posterior odds to the prior odds, which seems a sensible measure of increase in relative belief-worthiness. Moreover, Bayesian updating yields a reasonable result in Example 1 in that it conforms to the intuition that observations from this experiment have nothing to say one way or the other about whether radioactive

decay is deterministic or indeterministic.

Again, the point of this example in the present context is that an appropriate measure of evidential favoring in an incremental sense may differ enormously from an appropriate measure of evidential favoring in an absolute sense. Thus, the fact that the Law of Likelihood is true only of evidential favoring in an incremental sense is not a minor point given that the logic of induction that many non-Bayesian philosophers of science seek is related to a notion of evidential support in an absolute sense.

Because likelihoodism is true of evidential favoring only in an incremental sense, it does not characterize data as evidence in a sense to which Hume's dictum applies. Given a likelihoodist characterization of data as evidence, the most one could say would be that "a wise person proportions *shifts in* his or her beliefs to the evidence." Likelihoodist methods say nothing about what one should believe after receiving the evidence without reference to what one believed before receiving the evidence. They merely provide a measure of the degree to which the datum in question warrants a shift in one's beliefs vis-á-vis the pair of hypotheses in question.

I discuss an interesting response to this argument in Section 3. Before getting to that response, however, I would like to discuss a further difficulty for the idea that the relationship between evidence and rational belief vindicates likelihoodism.

2.2 The Evidential Meaning of What?

Likelihoodists claim that it is useful to report a measure of the degree to which some datum E favors one hypothesis of interest H_1 over a competitor of interest H_2 even when prior probabilities for those hypotheses are not available. One difficulty for this claim in addition to those discussed above is that the degree

to which E itself favors H_1 over H_2 can differ dramatically from the degree to which *learning* E favors H_1 over H_2 for a given individual.

Consider the following example

> **Example 2.** Jane the psychologist publishes results from an experiment that essentially amounts to asking a subject to guess one hundred times whether a fair coin will land heads or tails when flipped in order to investigate whether his or her probability of success p is greater than 0.5 or not. (Such an experiment could be used to investigate whether the subject has ESP or not, although of course one would want to rule out possible alternatives to ESP for explaining a subject's success.) Let X be a random variable the value of which is the number of times the subject guesses correctly. Jane reports that $X = 60$ and that this result yields a likelihood ratio of 8 for the hypothesis that $p = 0.6$ against the probability that $p = 0.5$. By standard likelihoodist conventions, this result counts as somewhat strong evidence for $p = 0.6$ against $p = 0.5$.

I am not concerned with the fact that likelihoodists classify this result as somewhat strong evidence despite the fact that most of us would not take it to make the hypothesis that the subject has ESP remotely belief-worthy. Of course, a likelihoodist is not committed to the claim that one should believe that the subject has ESP in light of this result. They acknowledge that prior probabilities as well as likelihoods are relevant for belief updating. They claim only to characterize data as evidence.

I am concerned instead with the fact that the likelihood ratio of the data the study reports can be quite different from the likelihood ratio of the data the reader of the study receives. Jane reports $X = 60$, but the reader learns not $X = 60$ itself, but rather [Jane reports $X = 60$]. The likelihood ratios of $p = 0.6$ against $p = 0.5$ for these respective data can be quite different.

They can be different not only because Jane might be dishonest, but also because Jane could have cherry-picked the datum she reported. Suppose Jane's procedure was to run the experiment described ten times and report the largest number of successes recorded. If the reader knows this information, then for him or her learning [Jane reports $X = 60$] amounts to learning $Y = 60$, where Y is a random variable the value of which is the maximum number of times that the subject guesses correctly in ten trials. The likelihood ratio of $p = 0.6$ against $p = 0.5$ for $Y = 60$ is not 8, but 1/30. Whereas, according to likelihoodist conventions, $X = 60$ is fairly strong evidence for $p = 0.6$ relative to $p = 0.5$, $Y = 60$ points in the opposite direction and nearly qualifies as "strong" evidence in that direction.

To make matters worse, suppose the Jane simply reported $X = 60$ and that neither Jane nor one's background knowledge provided any information about the procedure she used to generate this result. Then while the Law of Likelihood would allow one to characterize as evidence $X = 60$ itself, it would not allow one to characterize as evidence what one actually learns from Jane's report, namely that Jane reports $X = 60$. What is the use of using the Law of Likelihood to to characterize $X = 60$ as evidence when the evidential import of that datum could be quite different from the evidential import of what one actually learned?

The same points apply to a Bayesian approach, in which one conditions on what one learns. Conditioning on $X = 60$ may yield very different results from conditioning on a report that $X = 60$, and the latter is typically much harder to do properly. This fact is not an objection to Bayesianism in principle, but it is a great difficulty for Bayesianism in practice.

Likelihoodists are supposed to be gloriously free of the frequentist need to account for features of the data-generating process such as cherry-picking of results. Indeed, one can use the Law of Likelihood to characterize a reported

datum as evidence regardless of how that datum was generated. But that characterization is of dubious value not only because it does not tell one anything directly about what to believe or do, but also because it may not accurately reflect even the evidential meaning of what one actually learned.

3 Analogies between Evidential Favoring and Physical Measurement Fail to Vindicate Likelihoodism

Hacking coined the phrase "Law of Likelihood" in his (1965), but he used it to refer only to the claim that E favors H_1 over H_2 if $\Pr(E; H_1) > Pr(E; H_2)$. Edwards (1972) seems to have been the first to combine this qualitative claim with the quantitative claim that the likelihood ratio is a measure of that favoring.[2]

In a review of (Edwards 1972), Hacking expressed doubts about both the quantitative and the qualitative parts of the Law of Likelihood and argues specifically against the assumption of the quantitative claim that a likelihood ratio "means the same" in different contexts (1972, 136). Edwards had addressed this concern in his (1972) by arguing that the likelihood ratio as a measure of evidential favoring "will acquire a meaning as experience of its use accumulates" in the same way that the Fahrenheit scale had acquired a meaning as a measure of temperature (33). Hacking reported that he did not find this kind of argument "intrinsically disturbing," but argued that as a matter of fact likelihood ratios are not always commensurable across experiments. Many years later, Royall (1997) responded to Hacking's objection by comparing "units" of

[2]Edwards actually preferred to use the natural logarithm of the likelihood ratio so that the degree of evidential favoring of a conjunction of independent data points would be the sum of the measures of their individual degrees of evidential favoring. Because the natural logarithm is a monotonic function, this preference does not constitute a genuine disagreement with the Law of Likelihood.

evidential favoring to the BTU, a measure of heat.

Hacking acknowledged that there is a Bayesian argument for the commensurability of likelihood ratios, but likelihoodism is supposed to "[come] into its own" when Bayesianism does not apply because prior probabilities are not available (Sober, 2008, 38). Why think that likelihood ratios are commensurable across experiments in those cases? This question from Hacking is closely related to the concern I raised at the end of Subsection 2.1: what is evidential favoring in the sense in which likelihood ratios measure it, if not the degree to which the datum in question warrants a shift in one's beliefs with respect to the pair of hypotheses in question? An adequate answer to my question would also address Hacking's question by indicating something that likelihood ratios represent which would be constant across experiments with a common likelihood ratio. Conversely, an adequate response to Hacking's question would either answer my question or indicate that no answer is needed.

In Subsection 3.1, I maintain that Hacking's argument against the commensurability of likelihood ratios is weak, but the burden of proof is on the likelihoodist to show that they are commensurable rather than on Hacking to show that they are not. In Subsections 3.2 and 3.3, I argue that Edwards's and Royall's analogies between evidential favoring and heat or temperature fail adequately to address Hacking's concerns, and thus fail to vindicate likelihoodism as a viable alternative to Bayesian and frequentist methodologies.

3.1 Hacking's Argument Against the Commensurability of Likelihood Ratios

In his review of (Edwards 1972), Hacking writes that he "know[s] of no compelling argument" that a given likelihood ratio "means the same" in different

contexts. In this respect, a likelihood ratio is (at least prima facie)[3] different from a physical probability: if two independent, repeatable event types have the same physical probability, then they tend to occur equally often, roughly speaking. Of course, likelihood ratios are commensurable for a Bayesian, but what a likelihoodist needs is a kind of commensurability that does not depend on the availability of prior probabilities.

After saying that he sees no justification for assuming commensurability, Hacking attempts to use a pair of examples to argue against it. The first example he calls the "tank problem." Suppose we want to estimate the number of tanks in an enemy army. The tanks have serial numbers that start at 0001. We capture a tank at random and note that its serial number is 2176. The Law of Likelihood implies that this observation favors over all other possibilities the hypothesis that the total number of tanks in the enemy army is 2176. The second example Hacking uses we might call the "grating problem." Suppose we want to measure the width of a grating. We use a technique that produces normally distributed measurements with mean equal to the true width some and known variance. We make a measurement and note its value. The Law of Likelihood implies that this observation favors over all other possibilities the hypothesis that the true width of the grating is this observed value.

The claim that the observation favors the observed value over all other possibilities is intuitive in the grating problem but counterintuitive in the tank problem. But one can argue with some plausibility that our intuitions in the tank problem are misleading. While it would a striking coincidence if the tank we captured was the last one manufactured, it is no more improbable that we would capture the 2176th tank if there were 2176 tanks in all than that we would

[3]This claim about physical probabilities faces its share of challenges, of course (see e.g. Hájek 1996, 2009), but let it stand; the point of invoking physical probabilities here is simply to evoke some intuitions about the sort of property that Hacking takes likelihood ratios to lack.

capture the 2176th tank if there were 3000 in all. In fact, capturing the 2176th tank is more probable in the first scenario, as the likelihood ratio faithfully reports. We must not confuse "strikingness" with probability. Moreover, the claim that 2176 is the most evidentially favored estimate does not imply that it is the estimate one ought to report and use. Utilities are also relevant. One would expect that in a real version of the tank problem one would be particularly concerned not to underestimate the the enemy's forces, which would lead one to report a less evidentially favored estimate larger than 2176 in accordance with intuitions. One must not confused "most evidentially favored estimate" with "best estimate, all things considered."

These distinctions help dispel the intuition that 2176 is not in fact the most evidentially favored estimate in the tank problem. But Hacking advances a different objection: he reports that he has "no inclination to say" that the observation of a tank with serial number 2176 favors the hypothesis that there are 2176 tanks over the hypothesis that there 3000 tanks to the same degree that the observation of a particular value in the grating case favors the hypothesis that the true width has the value observed over an alternative hypothesis that generates the same likelihood ratio.

It helps in considering this objection to provide specific numbers in the grating example. Suppose that the observed measurement is 100 cm and the variance of the measuring technique is 1 cm. Then the Law of Likelihood says that the tank observation favors the hypothesis that there are 2176 tanks to the hypothesis that there are 3000 total tanks to the same degree that the width measurement favors the hypothesis that the width is 100 cm over the hypothesis that it is 99.2 cm. That degree is very small, namely 1.4. The claim that the degree of evidential favoring in these two cases is very small seems right. My intuition would not have told me on its own that the degree of favoring is exactly

the same in the two cases, but it does not rebel against that claim.

Perhaps Hacking's point is merely that it would typically be reasonable to use 100 cm as an estimate of the grating width but not to use 2176 as an estimate of the number of enemy tanks. But the same likelihood ratio is not supposed to "mean the same" across experiments in terms of its implications for our beliefs and actions. If this is Hacking's point, then he is subject to Royall's response that doubts about the commensurability of likelihood ratios across contexts "come from failure to distinguish between the strength of the evidence, which is constant, and its implications, which vary according to the context of each application" (1997, 12). Hacking could respond that likelihoodists have provided no clear non-Bayesian account of what it is that they take to be constant across applications with the same likelihood ratio. That response simply takes us back to the question I raised at the end of Subsection 2.1: what do likelihoood ratios represent, if not the degree to which the datum in question warrants shifting one's beliefs vis-á-vis the pair of hypotheses in question? Hacking's appeal to the tank and grating examples fails to advance the dialectic. But the burden of proof remains on the likelihoodist. Next, I consider in turn Edwards's and Royall's attempts to meet this burden.

3.2 Edwards's Analogy between Evidential Favoring and Temperature is Inadequate

In his (1972), Edwards argues for using (logarithms of) likelihood ratios as a measure of evidential favoring, or "support" as he calls it. He gives three responses to the objection that this measure "does not have any 'meaning' " (33). His first two responses are clearly inadequate. First, he claims that the fact that the measure of support lacks a probability interpretation is not damaging because it is meant to appeal to those who are suspicious of probability statements

that cannot be interpreted as objective frequencies. This response is inadequate because the objection is not simply that the measure of support lacks a probability interpretation, but that it seems to lack any interpretation whatsoever except the Bayesian interpretation as a measure of the change in belief that the data warrant. Second, Edwards claims that a likelihood ratios does have an "operational interpretation," as "the ratio of the frequencies with which, in the long run, the two hypotheses [in question] will deliver the observed data." This response is inadequate because the ratio of frequencies to which Edwards refers is a ratio of a frequency in one possible world to a frequency in a different possible world. Thus, his "operational interpretation" is not operational in any strong sense in which the existence of that interpretation should be comforting.

Edwards's third response is ultimately inadequate as well, but it cannot be dismissed as easily as the first two. He claims that the measure of support he proposes will, like our measures of temperature, "acquire a meaning" over time. He presents this response as follows (33):

> ...provided that support enables us to operate a logically-sound system of inference of undoubted relevance to the assessment of rival hypotheses, it will acquire a meaning as experience of its use accumulates. For many years temperature, as measured by Fahrenheit, had no "meaning" other than as an arbitrary scale conforming to an ordered sequence. Boiling water is not to be regarded as 6.6 times as hot as freezing water. But the measurement of temperature was nevertheless very important to the advancement of physics, and led ultimately, through the concepts of absolute zero and molecular movement, to a much deeper understanding of heat. The numerical assessment of rival hypotheses may be expected to be of equal benefit.

In this passage, Edwards points out that the notion of temperature and its measurement played a role in the development of a deeper understanding of heat. In the process, the notion of temperature acquired more "meaning." Edwards conjectures that the notion of support and its measurement will follow an analogous path.

This claim is dubious, and more than four decades of work since the publication of Edwards's book have not borne it out. Temperature is a property of matter on a macroscopic scale. Understanding it required developing theories of matter on a microscopic scale. Those theories involved the notion of molecular movement, which led naturally to the notion of absolute zero. The discovery of absolute zero made it possible to develop the Kelvin scale, in which zero corresponds to the coldest possible temperature and ratios between temperatures correspond to ratios of average kinetic energies. And theories of molecular motion have been enormously fruitful not only in thermodynamics, but also in many other areas of physical science.

Why should we think that the notion of support will be similarly fruitful? Temperature may not be entirely unique in its fecundity, but it is unusual. For that reason alone, the claim that support will be similarly fruitful is doubtful. Moreover, there is no plausible candidate for an analogue in the case of support to the theories of matter on a microscopic scale the development of which was necessary for understanding temperature. In addition, the log-likelihood ratio scale is already analogous to the Kelvin scale in that it has a meaningful zero analogous to absolute zero (the point of evidential neutrality) and meaningful ratios—a ratio of one log-likelihood ratio to another indicates the number of independent replications of the latter that would be needed to produce the same degree of support as the latter. For these reasons, the case of temperature fails to suggest a mechanism by which the use of the notion of support would

lead to advances in methodology.

Edwards also uses the analogy between support and temperature to argue that the fact that support is not the only factor relevant for assessing hypotheses does not mean that it is not useful for that purpose, any more than the fact that temperature is not the only factor relevant for our feeling of warmth does not mean that it is not useful (33–4):

> ...just as our feeling of warmth does not depend on the air temperature alone, so our assessment of hypotheses will not depend on the support alone. In the former case our impression will be affected by the wind, the humidity, the sun, our clothing, and a host of other factors; but the temperature will inform us about one particular factor. In the latter case, though by the likelihood axiom the support will inform us fully of the contribution to our judgement that the data can make, we shall also be influenced by the simplicity of the hypotheses, by their relevance to other situations, and by a multitude of subtle considerations that defy explicit statement.

There are at least two problems with this passage as a response to the worry we are considering. First, Edwards provides no alternative to Bayesianism for combining support with other factors such as simplicity in evaluating hypotheses. Second, the analogy between support and temperature here is mistaken and misleading. Edwards seems to be suggesting that one can "develop a feel for" what likelihoods mean for the assessment of hypotheses, in the same way that an American travelling in Europe can develop a feel for what Centigrade temperatures mean for, say, deciding what to wear on a given day. But as I argued in Subsection 2.1, support warrants *changes* in belief, in much the same way that *heat* produces changes in temperature. Thus, deciding what to believe based on support in Edwards's sense is more like deciding what to wear on the

basis of information about recent flows of heat into one's vicinity than it is like deciding what to wear on the basis of the temperature. To use information about heat flows in deciding what to wear, one would want to refer to the prior temperature. In the same way, to use information about support to decide what to believe, one would want to refer to prior probabilities, as likelihoodists themselves insist when trying to defend their theory against counterexamples like Example 1. One cannot simply "develop a feel" for using support to assess hypotheses without reference to prior probabilities, because the prior probabilities are crucial.

Log-likelihood ratios do have a meaning in the presence of prior probabilities as the difference between the posterior log-odds and the prior log-odds of the pair of hypotheses in question. Edwards's analogy between log-likelihood ratios and temperatures fails to vindicate the crucial likelihoodist claim that the former have some definite meaning that does not appeal to prior probabilities even implicitly.

3.3 Royall's Analogy between Evidential Favoring and Heat is Inadequate

Royall (1997) improves upon Edwards's treatment in that he compares evidential favoring to heat rather than temperature. The argument he makes using this analogy is different from Edwards's, but it too fails to vindicate likelihoodism as a genuine alternative to Bayesianism.

Royall responds to Hacking's objection that it is not clear that the same likelihood "represents the same strength of evidence in all contexts" first by pointing out that Bayes's theorem guarantees that a likelihood ratio of k corresponds to a k-fold increase in probability ratio, "whether the prior probabilities are known or not" (11). Of course, this response is not adequate for those who

think that prior probabilities are often not merely unknown, but either nonexistent or scientifically irrelevant. Royall responds to this concern by claiming that the likelihood ratio measure of evidential favoring "retains its meaning" in the absence of prior probabilities in the same way that a unit of heat defined in terms of its effect on a unit mass of water at a particular temperature retains its meaning in the absence of such water (11–2):

> The numerical value of the likelihood ratio, which is given a precise interpretation in [cases in which known prior probabilities have a frequency interpretation] (via Bayes's theorem), retains that meaning more generally.... The situation is analogous to that in physics where a unit of thermal energy, the BTU, is given concrete meaning in terms of water—one BTU is that amount of energy required to raise the temperature of one pound of water at 39.2°F by 1°F. But it is meaningful to measure thermal energy in BTUs in rating air conditioners and in other situations where there is no water at 39.2°F to be heated. Likewise the likelihood ratio, given a concrete meaning in terms of prior probabilities, retains that meaning in their absence.

This analogy is faulty. Even though the BTU is defined with reference to one pound of water at 39.2° F, the heat equation applies to any substance at any temperature in the absence of phase transitions: heat equals specific heat times mass times the change in temperature. The BTU "retains its meaning" in the absence of water at 39.2° F because the heat equation tells us what the significance of a BTU is in other settings. The reference to one pound of water at 39.2° F is used only as a convention for picking out one of the countably many permissible scales for measuring heat. By contrast, the best analogue to the heat equation for evidential favoring, the log-odds form of Bayes's theorem, simply

does not apply to hypotheses that lack prior probabilities: "bannage" equals the change in log-odds, where bannage is I.J. Good's term for the logarithm of the likelihood ratio of the evidence for the pair of hypotheses in question1. This equation tells us nothing about what the significance of a ban is in the absence of prior probabilities. Nor does any other true equation. As a result, a reference to prior probabilities in an account of the likelihoodist notion of evidential favoring is essential in a way that a reference to water at 39.2° F in an account of heat is not.

Royall claims that worries about the commensurability of likelihood ratios "come from failure to distinguish between the strength of the evidence, which is constant, and its implications, which vary according to the context of each application" (12). But such worries run deeper than the distinction between strength of evidence and its implications. The log-odds form of Bayes's theorem indicates a sense in which the same likelihood ratio represents the same strength of evidence in all contexts, provided that prior probabilities are available. But when prior probabilities are not available, likelihoodism is idle. To show that likelihoodism is a viable alternative to Bayesianism, one needs to show that likelihood ratios mean something when prior probabilities are absent. Royall's analogy between evidential favor and heat, like Edwards's analogy between support and temperature, fails to address this challenge.

4 *Ceteris Paribus* Cannot Save Likelihoodism

Likelihoodists admit—as they must in light of cases like Example 1—that a large likelihood ratio for H_1 against H_2 is not sufficient to make H_1 more believeworthy than H_2. But perhaps a large likelihood ratio is sufficient *ceteris paribus*, that is, "all else being equal."

This proposal is too vague to use or evaluate without an account of what "all

else being equal" means. The challenge for its proponents is to specify what "all else being equal" means in a way that justifies it without turning likelihoodism into nothing more than an approximation of a fragment of Bayesianism.

For instance, the most obvious specification is that "all else being equal" means that the prior probabilities of the two hypotheses in question are nearly equal.[4] But this proposal turns likelihoodism into nothing more than an approximation of a fragment of Bayesianism: it applies only to a small subset of the cases in which Bayesianism applies, and in those case it yields the same results that a Bayesianism analysis would—unless, of course, the departure from strict equality between the prior probabilities makes a difference to the outcome, in which case the likelihoodist conclusion is suspect because it conflicts with the Bayesian conclusion.

Even the staunchest subjective Bayesians admit that efforts to assign prior probabilities to propositions are typically subject to some kind of uncertainty or imprecision (e.g. De Finetti and Savage 1962, 95). One might think that likelihoodism would be useful in cases in which one's prior information is roughly symmetric with respect to a pair of hypotheses, but one is not willing to assign sharp prior probabilities to them. The problem with this idea is that there are broadly Bayesian methods for dealing with non-sharp prior probabilities, including robust Bayesian methods (see e.g. Berger 1994) for checking the sensitivity of a conclusion to the choice of prior and methods that use convex sets of probability distributions (see e.g. Walley 1991). Thus, this idea again turns likelihoodism into a mere approximation of a fragment of Bayesianism broadly construed—unless the more fine-grained information about one's epistemic state that these broadly Bayesian approaches can use makes a difference to the con-

[4]For ease of exposition, I assume that we are using a purely epistemic notion of belief-worthiness in which utilities play no part. I could accommodate the claim that there is no such notion simply by making appropriate stipulations about the relevant utilities as well as the prior probabilities.

clusion one reaches, in which case the likelihoodist approach is suspect.

Likelihoodism is not the most promising response to the concern that prior probabilities might be vague or only vaguely known. It is better motivated by a desire for a methodology that keeps the objective import of data separate from the influence of one's idiosyncratic opinions, whether they be precise or not. The problem is that examples like Example 1 show that the way in which likelihoodists characterize the objective import of data is not a good guide for belief and action without reference to prior probabilities.

There is a methodology that purports both to keep the objective import of data separate from the influence of one's idiosyncratic opinions and to be a good guide for belief and action: frequentism. I have argued at length that frequentist methods fail to respect evidential equivalence, but a frequentist in the Neyman-Pearson tradition could maintain that what matters is not to respect evidential equivalence, but—roughly—to use methods that have good long-run operating characteristics (see e.g. Neyman and Pearson 1933, 291).

One could try to defend *ceteris paribus* likelihoodism by arguing, as Royall (2000) does, that likelihoodism provides "a frequentist methodology that avoids the logical inconsistencies pervading current methods while maintaining the essential properties that have made those methods into important scientific tools" (31). In other words, likelihoodist methods are warranted by their long-run operating characteristics in the same way that frequentists take their methods to be, without being subject to the many objections that frequentist methods face (such as that they violate the Likelihood Principle). This fact seems to provide some reason to believe that likelihoodist methods can in fact reasonably be used to guide belief in cases in which frequentists take their methods to be reasonable—when "all else" is "equal" (or better, not unequal) in the sense that empirically warranted, objectively interpretable prior probabilities that could

distinguish among the hypotheses under consideration are not available.

This proposal has the virtue that it does not turn likelihoodism into a mere approximation of a fragment of Bayesianism. But it does not work because likelihoodist methods are not warranted by their operating characteristics in the same way that frequentists take their methods to be. The strongest kind of frequentist justification a method can have is that it is the best among some class of methods on some (typically worst-case) performance criterion. For instance, a uniformly most powerful level α hypothesis test is justified by the fact that it would reject the null hypothesis at least as often in repeated applications in the long run as any other test under any scenario within the model in which that hypothesis is false among all tests that would reject it no more than $100\alpha\%$ of the time if it were true. (In this case, the class of methods is those with Type I error rate no greater than α, and the performance criterion is that of being a most powerful test under each alternative hypothesis. This criterion is not merely worst-case, but it includes the worst case.) The key feature of this kind of justification for present purposes is that it is *comparative*. It answers questions of the form, "Why use method μ rather than some other method?" by pointing out some respect in which μ is better than every other method.

There are many objections to appeals to frequentist performance, including objections to the performance criteria used and worries about the relevance of those criteria to the evaluation of a single case. Those objections are not relevant to my present claim, which is that such justifications do not apply to likelihoodist methods even if they are legitimate.

The problem with likelihoodist performance justifications from a frequentist perspective is that they are not comparative. They merely show that some method performs well in some sense in repeated applications in the long run. They do not show that some method is better than its competitors in any way.

Thus, they are insufficient to answer questions of the form, "Why use method μ rather than some other method?"

The most often cited likelihoodist performance justification is the *universal bound*: $\Pr(X = x : \Pr(x|H_2)/\Pr(x|H_1) \geq k; H_1) \leq 1/k$. That is, for fixed H_1 and H_2, the probability that an experiment yields a likelihood ratio of at least k for H_2 against H_1 when H_1 is true is at most $1/k$. Thus, the Law of Likelihood has the nice property that for fixed H_1 and H_2, one of which is true, an arbitrary experiment is highly unlikely to produce a result that is according to the Law of Likelihood strong evidence for the one that is false.

This result is indeed nice, but it does not suffice to warrant using the Law of Likelihood rather than some other method because it does not compare the Law of Likelihood to any other method. There could be (and often are) many other methods in addition to the Law of Likelihood that also achieve the universal bound. Why should we use the Law of Likelihood rather than one of those other methods? Likelihoodists do not have a performance-based answer to that question.

Tighter bounds on the probability of misleading evidence than the universal bound are available in some cases (Royall 2000), but those bounds are also non-comparative. In order to appeal to performance characteristics to justify their methods, likelihoodists need to show that they achieve better performance than other possible methods in some respect.

One might object that methods with frequentist justifications are often based on likelihood ratios. For instance, the Karlin-Rubin theorem states that when one's hypothesis space has a monotone likelihood ratio, a test that rejects a point null against a one-sided composite alternative if and only if the likelihood ratio for the null against a pre-specified element of the alternative falls below a given cutoff value is a uniformly most powerful level α test, where α is its Type I error

rate. It might seem that this result vindicates likelihoodism. However, it applies only when the null hypothesis and test procedure are predesignated. Thus, it does not vindicate the likelihoodist practice of interpreting any likelihood ratio on likes as a measure of evidential favoring.

Royall regards the universal bound and other likelihoodist performance criteria as relevant for pre-trial planning rather than for post-trial evaluation and does not give them great emphasis in his efforts to promote a likelihoodist approach to statistics. Nonetheless, he sometimes suggests in his writings that such results justify likelihoodist methods in the same way that frequentist performance characteristics purport to justify frequentist methods. It is this claim that I am rejecting. Likelihoodism is not a frequentist methodology.

I can find no interpretation of the claim that likelihoodism is a good guide for belief or action *ceteris paribus* that is justified, yet makes likelihoodism more than an approximation to a fragment of Bayesianism (in a broad sense that includes robust Bayesianism and Bayesianism for imprecise probabilities).

Those who are familiar with the literature in this area might think that there is a stronger argument against the claim that likelihoodism is a good guide for belief or action *ceteris paribus* than I have given here that appeals to examples like Stone's "Flatland" (1976) in which likelihoods fail to distinguish among methods one of which dominates the others in coverage probability. In my view, such an argument would not be compelling because such examples require the unrealistic idealization of an infinite sample space. I discuss this issue in an appendix.

5 Conclusion

Likelihoodist methods by themselves provide no guidance for belief or action. I have considered and rejected several possible likelihoodist responses to the

highly plausible claim that this fact means that likelihoodism is not a viable alternative to Bayesian and frequentist methodologies.

References

Berger, James O. 1994. "An overview of robust Bayesian analysis." *Test* 3:5–59.

Birnbaum, Allan. 1964. "The Anomalous Concept of Statistical Evidence: Axioms, Interpretations, and Elementary Exposition." Technical Report IMM-NYU 332, New York University Courant Institute of Mathematical Sciences.

Carnap, R. 1962. *Logical Foundations of Probability.* University of Chicago Press.

De Finetti, Bruno and Savage, Leonard J. 1962. "Sul modo di scegliere le probabilità iniziali." *Biblioteca del Metron* 1:81–147.

Edwards, A.W.F. 1972. *Likelihood. An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference. [Mit Fig. U. Tab.].* Cambridge University Press.

Fitelson, Branden. 2001. *Studies in Bayesian confirmation theory.* Ph.D. thesis, University of Wisconsin.

Fraser, DAS, Monette, G, and Ng, KW. 1985. "Marginalization, likelihood and structured models." In *Multivariate analysis–VI: proceedings of the Sixth International Symposium on Multivariate Analysis*, volume 6, 209. North Holland.

Hacking, I. 1965. *Logic of Statistical Inference.* Cambridge University Press.

Hájek, Alan. 1996. ""Mises Redux" – Redux: Fifteen Arguments against Finite Frequentism." *Erkenntnis (1975-)* 45:209–227.

—. 2009. "Fifteen Arguments against Hypothetical Frequentism." *Erkenntnis (1975-)* 70:211–235.

Hill, BruceM. 1980. "On some statistical paradoxes and non-conglomerability." *Trabajos de Estadistica Y de Investigacion Operativa* 31:39–66.

Howson, Colin. 1997. "A Logic of Induction." *Philosophy of Science* 64:pp. 268–290.

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations. University of Chicago Press.

Neyman, J. and Pearson, E. S. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231:289–337.

Royall, Richard. 2000. "On the Probability of Observing Misleading Statistical Evidence." *Journal of the American Statistical Association* 95:pp. 760–768.

Royall, R.M. 1997. *Statistical Evidence: A Likelihood Paradigm*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Savage, L. J. 1962. "On the Foundations of Statistical Inference: Discussion." *Journal of the American Statistical Association* 57:307–8.

Sober, Elliott. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press.

Stone, Mervyn. 1976. "Strong Inconsistency from Uniform Priors." *Journal of the American Statistical Association* 71:114–116.

Walley, P. 1991. *Statistical reasoning with imprecise probabilities*. Monographs on statistics and applied probability. Chapman and Hall.

A Why "Flatland" and Similar Examples Are Not Fatal to *Ceteris Paribus* Likelihoodism

In the main text, I argued against the claim that "*ceteris paribus* likelihoodism" (the view that the Law of Likelihood provides a *ceteris paribus* norm for belief and action) is a viable alternative to Bayesianism and frequentism by explaining that I know of no interpretation of *ceteris paribus* likelihoodism that makes it justified without turning it into nothing more than an approximation of a fragment of Bayesianism. One might think that the following is a stronger argument against *ceteris paribus* likelihoodism: if having background knowledge and utilities that are symmetric with respect to H_1 and H_2 is sufficient to satisfy the *ceteris paribus* clause—as seems quite plausible—then *ceteris paribus* likelihoodism implies that one should be indifferent between a pair of estimators one of which strictly dominates the other in probability of success. For there is hypothetical scenario in which for any value x of some random variable X whose probability distribution is parameterized by θ, one's background knowledge and utilities are completely symmetric with respect to the hypotheses $\theta = \theta^*(x)$ and $\theta = \theta'(x)$, and $\Pr(x; \theta = \theta^*(x))/\Pr(x; \theta = \theta'(x)) = 1$, yet $\Pr(\theta^*(X) = \theta) = 3/4$ while $\Pr(\theta'(X) = \theta) = 1/4$ regardless of the true value of θ.

One could claim in response that being indifferent between the estimates θ^*(x) and θ'(x) of θ for all values x of X is not the same as being indifferent between the estimators θ^*(X) and θ'(X), but those two attitudes are indistinguishable behaviorally and have the same bad pragmatic consequences.

I will argue that the examples illustrating this type of phenomenon do not provide strong arguments against *ceteris paribus* likelihoodism because they are unrealistic: they require infinite sample spaces, which do not arise in real experiments because real measuring processes are discrete and bounded. It has

been claimed that responses of this kind "merely avoid the logical content of the problem" (Hill, 1980). This comment is accurate but not troubling. Hypothetical problems involving infinite sample spaces raise extremely interesting puzzles and may have some theoretical importance. There are simply outside the scope of my present concerns.

A hypothetical scenario in which this phenomenon arises is Stone's (1976) "Flatland" example. A quick gloss of this example can be given as follows. A sailor takes a number of steps along a two-dimensional grid, buries a treasure, takes one more step in a direction determined by the outcome of a roll of a fair four-sided die, and then dies. He carried with him a string that he kept taut. One's datum x is the path of that string. The parameter one wishes to estimate is the location of the treasure θ. $\Pr(x;\theta) = 1/4$ for θ one step north, south, east or west of the end of the string and 0 for all other values of θ. Thus, the Law of Likelihood implies that x is evidentially neutral between the hypothesis $\theta^*(x)$ that θ is one step back along the path x and the hypothesis $\theta'(x)$ that θ is one step forward along the path x in the direction of the final step. In addition, one's background knowledge is symmetric with respect to $\theta^*(x)$ and $\theta(x)$ because the scenario says nothing about how θ is generated, and we can simply stipulate that one's utilities are symmetric with respect to θ*(X) and θ'(X), say payoff 0 or 1 according to whether one's estimate is true or not. However, for any infinite sequence of θs (random or nonrandom), the estimator θ*(X) which says that θ is one step back along the path gets the right answer 3/4 of the time in the long run, while the estimator θ'(X) that θ is one step forward along the path in the direction of the final step gets the right answer 1/4 of the time in the long run: for any θ, $\Pr(\theta^*(X) = \theta;\theta) = 3/4$, while $\Pr(\theta'(X) = \theta;\theta) = 1/4$.

The argument just given does not work against orthodox (countably additive) Bayesianism because orthodox Bayesians must give higher posterior prob-

ability to $\theta^*(x)$ than to $\theta'(x)$ for some possible values x of X. Thus, it does not work against the Law of Likelihood understood simply as the claim that it is appropriate to use the phrase "the degree to which x favors H_1 over H_2" for the ratio of the posterior odds of H_1 to H_2 given $X = x$ to the prior odds of H_1 to H_2. Of course, the acceptability of the Law of Likelihood understood in that way does not vindicate *ceteris paribus* likelihoodism as a genuine alternative to Bayesianism.

The weakness of this argument is that it does not worlk if the two-dimensional grid along which the sailor walks is finite. Real sample spaces (unlike our idealized models of them) are finite, so a likelihoodist could reasonably claim that the Law of Likelihood is a *ceteris paribus* norm of inference for data from any experiment that we could actually perform. Other similar examples (e.g. Fraser et al. 1985, 213–4) have the same shortcoming, and not by accident: such examples require non-conglomerability over the sample space, which can only arise when the sample space is infinite. The fact that *ceteris paribus* likelihoodism encounters difficulties in impossible idealized thought experiments is irrelevant to practical methodology.