Model Verification and the Likelihood Principle

Samuel C. Fletcher*
Logic and Philosophy of Science
University of California, Irvine
scfletch@uci.edu

April 19, 2013

Abstract

The likelihood principle (LP) is typically understood as a constraint on any measure of evidence arising from a statistical experiment. It is not sufficiently often noted, however, that the LP assumes that the probability model giving rise to a particular concrete data set must be statistically adequate—it must "fit" the data sufficiently. In practice, though, scientists must make modeling *assumptions* whose adequacy can nevertheless then be verified using statistical tests. My present concern is to consider whether the LP applies to these techniques of model verification. If one does view model verification as part of the inferential procedures that the LP intends to constrain, then there are certain crucial tests of model verification that no known method satisfying the LP can perform. But if one does not, the degree to which these assumptions have been verified is bracketed from the evidential evaluation under the LP. Although I conclude from this that the LP cannot be a universal constraint on any measure of evidence, proponents of the LP may hold out for a restricted version thereof, either as a kind of "ideal" or as defining one among many different forms of evidence.

1 Introduction

One aspect of the longstanding debate over the nature of statistical evidence and inference centers on the likelihood principle (LP). Recall (Berger and Wolpert, 1988) that one can describe a statistical experiment as a triple $E = (X, \theta_0, \{P^\theta\}_{\theta\in\Theta})$, where X is a random variable with probability distribution $P^{\theta_0}(x)$ depending on a particular parameter $\theta_0 \in \Theta$. An inferential problem is then to determine *which* parameter in Θ generated a particular realization x of E—that is, a concrete data set x arising from the experiment E. Denote $e(E, x)$ as the evidence, inferential conclusions, report, etc., about θ_0 arising from x. The LP then constrains e as follows:[1]

*Thanks to Kent Johnson for patient guidance, and to audiences at UC Irvine and Virginia Tech—particularly Nicole Jinn at the latter—for their comments. Part of the research leading to this work was completed with the support of a National Science Foundation Graduate Research Fellowship.

[1]Not to be confused with the usage of Miller (1987) or Sober (2001), whose "likelihood principle" describes (versions of) a distinct principle sometimes called the "law of likelihood" (Hacking, 1965, Ch. 5).

Likelihood Principle Let $E_1 = (X_1, \theta_0, \{P_1^\theta\}_{\theta\in\Theta})$ and $E_2 = (X_2, \theta_0, \{P_2^\theta\}_{\theta\in\Theta})$ be two experiments with a common parameter space Θ. If, for realizations x_1 and x_2 of E_1 and E_2, respectively, there is a constant $C(x_1, x_2)$, not depending on θ, such that $P_1^\theta(x_1) = C(x_1, x_2)P_2^\theta(x_2)$ for all $\theta \in \Theta$, then any evidence function e must satisfy $e(E_1, x_1) = e(E_2, x_2)$.

Note that the parameters $\theta \in \Theta$ can denote different concrete models or hypotheses, so the LP codifies the intuition that (functions of) likelihood ratios measure the relative evidence a data set provides for one hypothesis over another. For instance, if $P_1^\theta(x_1)/P_1^{\theta'}(x_1) = k(\theta, \theta')$ and $P_1^\theta(x_1) = C(x_1, x_2)P_2^\theta(x_2)$ for all $\theta \in \Theta$, then $P_2^\theta(x_2)/P_2^{\theta'}(x_2) = k(\theta, \theta')$. As such it is intended as a minimal constraint on any "reasonable" notion of evidence.

The LP has received considerable attention since Allan Birnbaum's watershed essay, "On the Foundations of Statistical Inference" (1962), which showed (modulo some mathematical concerns later resolved) that the LP is equivalent to the conjunction of two intuitive principles of (weak) sufficiency and conditionality. The important and controversial connection between these principles (Berger and Wolpert, 1988) will not be my main focus here, although I will return to it briefly in §4. Rather, I wish primarily to examine the *domain of applicability* of the LP. In particular it is not often enough noted that the definition of an *experiment* in the LP plays a crucial role: up to a particular choice of θ, the probability model $P^\theta(x)$ must determine the probability (distribution) of X with complete accuracy. In practice, however, many (most?) problems of statistical inference involve substantive modeling assumptions whose *adequacy* is not known a priori. Here, a modeling assumption for a data set is statistically adequate when the assumption sufficiently fits the data, i.e., the data are consistent with what one would expect or predict from the assumption's truth, where this consistency can be measured in many ways according to taste and statistical framework (Mayo, 1996, p. 178–9).

Birnbaum was sensitive to this issue, explicitly limiting attention to experiments for which any such assumptions are not in question: "We deliberately delimit and *idealize* the present discussion by considering only models whose adequacy is postulated and is not in question" (Birnbaum, 1962, p. 274). Yet despite this restriction, he took accepting the LP to have "immediate radical consequences for the every-day practice as well as the theory of informative inference" (Birnbaum, 1962, p. 272), and commentators have emphasized the resulting far-reaching implications for the role of stopping rules, censoring mechanisms, and randomization in experimental design (Berger and Wolpert, 1988).[2] How can the LP have such implications for non-trivial statistical modeling if one must postulate that the substantive modeling assumptions are adequate? In such cases, it is indeed crucial for the modeler to verify, inasmuch as it is possible, the sufficient adequacy of those assumptions. Thus there seems to be a tension between the full demands of the LP's antecedent and the practical situations in which its proponents and detractors believe its consequent apply.

Some discussants of Birnbaum's essay recognized this tension explicitly. Because "Adequacy here refers to a correspondence between a theoretical [mathematical] model and

[2]This practical dimension was acknowledged immediately among some (though not all) of the discussants of Birnbaum's paper. For example, John W. Pratt wrote that "anyone who has taught a course in statistics of almost any kind would feel that full acceptance of the likelihood principle ought to change the course entirely" (Birnbaum, 1962, p. 316).

experimental fact," wrote Irwin Bross, "the adequacy of the model is always in question, and often the main point for a study is to test the adequacy of a model" (Birnbaum, 1962, p. 310). Consequently, Bross thought that LP never applies, a sentiment echoed a bit hyperbolically by Oscar Kempthorne: "It is a truism, I believe, that there is never an adequate mathematical statistical model for any actual situation" (Birnbaum, 1962, p. 319). While it is true that mathematical models of experiments in the world do not in general perfectly capture every feature of their targets, one only requires mere adequacy for the application of the LP. In other words, contra a strong reading of Bross's (and Kempthorne's) statement, it is legitimate to make idealizations when the adequacy of the idealized model is verified, as George Box emphasized in reply (Birnbaum, 1962, p. 311–312).[3]

But one can read in the second part of Bross's statement a suggestion regarding a different kind of worry that can arise when one seeks to apply the LP to the process of model verification itself.[4] The techniques used to verify the adequacy of statistical modeling assumptions are themselves statistical, and such verification is often as much a primary statistical goal as parametric inference. Does the LP constrain these techniques as well? If one takes seriously the purview of the LP to constrain measures of evidence that apply to *any* statistical inference, the answer must be affirmative. But I argue in §2 that this view implies that *no* techniques for model verification satisfying the LP are known to apply generally. Thus, insofar as securing the adequacy of a model is an important and ineliminable goal in statistical research, this *inferential* view of model verification is untenable with the LP.

In response, one may retreat to the weaker position holding that while the LP may not directly constrain techniques of model verification, it nevertheless applies to inferences regarding primary parameters of any statistical model whose adequacy has otherwise been verified. But this *non-inferential* view of model verification presents other problems for the general applicability of the LP. As I show in §3, given that establishing model adequacy is often a primary statistical goal, the difference between it and other inferential techniques based on the primary/secondary distinction is vague. But even if one can solve this problem of vagueness, the non-inferential conception of model verification brackets the relative security of the model's adequacy, which can be quantified statistically, from techniques bound by the LP. Thus there would still be aspects of any evidential evaluation that inferential techniques bound by the LP do not capture.

Consequently, whether one takes the inferential or non-inferential view of model verification, the LP cannot be a constraint on *any* measure of evidence, for either no such measure can satisfy it, or (among other problems) measures that do satisfy it cannot capture essential aspects of evidential bearing, respectively. I want to emphasize that I am not arguing for either the inferential or non-inferential view of model verification. (Indeed, I suspect that whether one seems more plausible will depend on the context of a particular statistical problem or inquiry.) Instead I want to point out that, whatever one's views about the role of model verification in inference, the LP cannot, as many commentators have assumed, constrain all the inferential procedures used in statistical problems with substantive modeling assumptions. Nevertheless, in §4 I contend that there may still be a role for evidence

[3]Understanding exactly how scientific models represent phenomena, approximately or otherwise, is subtle and important, but it will not be my focus here. For more, see Frigg and Hartmann (2012).

[4]Probing the adequacy of a statistical model goes under many names in the literature, including model critique, criticism, validation, verification, checking, and diagnostic and misspecification testing.

relations constrained by the LP if one takes it merely as a kind of "ideal" or if one gives up requiring that only these relations are viable. Then the LP becomes less of a constraint on measures of evidence as a way to define certain kinds of evidence with properties that are in many circumstances desirable.

2 The Inferential View of Model Verification

In order for the LP to apply to methods of model verification, those methods must take the form of a kind of parametric inference, since the LP explicitly constrains evidential relations between parameterized statistical models, their realizations in concrete data, and the particular parameter values of the former. Typically model verification is not articulated in this way, but there is no reason in principle why it cannot, just as the LP can in principle apply to nonparametric (but fully specified) statistical models (Berger and Wolpert, 1988, p. 43–44). Parameterize all possible models for some data set x by an index α and write the most general encompassing model as the sum

$$P^{\theta_\alpha}(x;\alpha) = \sum_{\beta} I_\alpha(\beta) P_\beta^{\theta_\beta}(x), \tag{1}$$

where $I_\alpha(\beta) = 1$ if $\beta = \alpha$ and 0 otherwise, β ranging over all index values. I have labeled the parameters θ by this index as well to allow for labeled parameterized family of models, but this distinction is inessential since one could just as well let the index specify a completely determined statistical model, one where θ_α is absorbed into α and the probability models P_α do not depend on the specification of any further parameters.

Procedures of model verification, then, can been seen as inferential in the sense that they select or determine the evidence for (or against) α from x in the same way as other procedures of statistical inference select or determine evidential relations for elements from the parameter space of θ_α from x. In doing so, model verification conceived inferentially winnows down the range of possible models to something more tractable. The inferential view also assimilates model verification with model selection, taking the relevant class of models from which to select (at least initially) to be *all* possible models.[5] It may be motivated by the view that the differences between techniques used to probe model assumptions, to compare the statistical fit of various models, and to make inferences about scientifically relevant parameters are at most differences in degree rather than differences in kind.

While well-defined, the sum (1) will in general be uncountable and unwieldy, and the details of virtually all of its components will be difficult to obtain. This presents a practical but nonetheless significant obstacle to applying the LP to techniques of model verification thus construed, since this application requires enough clarity on the form of the probability model for one to judge its proportionality to another. Further, it is difficult to conceive of a

[5]This may seem to conflict with Mayo and Spanos (2004), who wish to distinguish model verification (what they call misspecification testing) from model selection. But the conflict is illusory, since the basis for their distinction is that the latter "selects from an *assumed family of models*" (Mayo and Spanos, 2004, p. 1008)—if the assumed family consists in all possible models, then according to the inferential view the two dovetail. As discussed in more detail below, however, model verification thus conceived will not in general admit of severe tests.

situation where the LP would apply, since it would require another encompassing probability model that nonetheless shares the same huge parameter space and is proportional to but distinct from the first. But perhaps the most intractable problem for extending the purview of the LP to model verification is that in general model verification techniques cannot be comparative—they cannot assign merely relative measures of evidence to pairs of models. Essentially comparative methods face what Staley (2008) calls the problem of the unconsidered alternative: that a hypothesis is best supported among the available alternatives is generally insufficient for evidence *for* that hypothesis because there may be a much better supported hypothesis not included in the set under consideration.[6] Since comparative methods must work within a circumscribed class of statistical models, they cannot evaluate the statistical adequacy of that class itself.

Now, one could use comparative methods for model verification if one could compare *all* possible models, but as just remarked the details of nearly all of these, although well-defined, will be difficult to delineate. This delineation of a tractable proper subclass of models is fraught in part because it requires justifying why two otherwise statistically similar models, one lying within the boundaries of the subclass and one without, should fall where they do. Often the justification will be of pure convenience on the part of the modeler. In any case, once one has found as statistically adequate a class of models sufficiently narrow that their probability models are all tractable, one can well apply comparative techniques for further model selection. But, inasmuch as is possible, one must get to that point using non-comparative methods. The difficulty for the LP arises from the fact that these non-comparative methods seem to require the Fisherian logic of testing: one makes an assumption about the form or properties of the probability model, from which it follows that certain statistics follow certain sampling distributions which one constructs theoretically or estimates through simulation. To the degree that the statistics evaluated at the actual data are improbable, one then has reason to reject said assumption, the threshold for model respecification depending, of course, on the details of the scientific goals at hand.[7] Such methods are non-comparative because they do not require the specification of any explicitly defined alternatives to the assumptions being tested. Since it is well known that inferential procedures depending on sampling distributions do not in general satisfy the LP, the same follows for these techniques.

This perspective that model verification must (at least initially) depend on Fisherian tests with sampling distributions is new neither among error-statisticians (see Mayo (1996, p. 160–1), Spanos (1999, Ch. 15.2.1), and Mayo and Spanos (2011, §4)) nor, some will be surprised to note, among Bayesian statisticians (Box, 1980, Gelman and Shalizi, 2013). Besides using explicit test statistics, techniques of model verification also include informal graphical tests and simulation, but the underlying Fisherian logic is the same. Both of the latter are prominent in Bayesian model verification. To illustrate, one very general class of procedures is based on the posterior predictive distribution $P(x_{new}|x)$, the probability

[6] Although Staley does not explicitly draw a connection with the general problem of the *unconceived* alternative (Stanford, 2006), his problem in many ways seems to be its natural statistical counterpart.

[7] Contingent on the statistical adequacy of certain aspects of the model, the same methods can also be used to test for mistakes in the data, such as improperly recorded surveys in the social sciences or instrument malfunction in the physical and biological sciences. Throughout, however, I assume that the data has no such mistakes so that the model is always the object under scrutiny.

distribution for a new observation given the data already observed (Gelman et al., 2004, Ch. 6). Specifically, one uses the predictive distribution to generate *replications* of the original data—that is, new data sets whose covariates, if there are any, are identical to the ones that obtained in the actual data set. Graphically one can then visually compare histograms of replications—whether the raw replications themselves, or summary statistics or inferences—to the actual data. Large discrepancies, in proportion to their size, then give reason to reject the underlying model.

More quantitatively, one can define a measure of discrepancy between the observed data x and the inferred posterior model indexed by θ through a test statistic $T(x, \theta)$.[8] Then one can define the *posterior predictive p-value* (PPP), $P(T(x_{new}, \theta) \geq T(x, \theta)|x)$, which functions in much the same way as a classical p-value. (Here the probability is taken over the distribution of $x_{new}, \theta|x$.) If one has generated the replications through simulation, then this PPP can be estimated as just the proportion (i.e., relative frequency) of the replications that have a larger test statistic than that of the observed data.[9]

Note that techniques of model verification only test the fit between the data and the model, so passing such a test is not evidence for that *particular* model, since there may be many other, substantively different models whose assumptions would have been verified—in other words, they are not severe tests of modeling assumptions (Mayo and Spanos, 2011, §2.5). But methodologically they need not be. Model verification is not and cannot be a substitute for sophisticated statistical design and sensitivity to scientific context. Indeed, when a procedure of model verification finds that certain modeling assumptions are statistically inadequate, that procedure does not automatically provide alternative models to consider. Yet by understanding how different Fisherian tests are sensitive to different departures from the modeling assumptions they test, one has a better clue as to which classes of alternative models relaxing those assumptions might prove statistically adequate, regardless if one is using error statistical (e.g., Spanos (1999, Ch. 15.4), Mayo and Spanos (2004) and Staley (2012, §5)) or Bayesian (e.g., Box (1980, §4), Gelman et al. (2004, Ch. 6), and Gelman and Shalizi (2013, §4)) methodology.

On account of this fact—that the inadequacy of certain modeling assumptions found through the violation of a Fisherian test nearly always leads to model respecification and the attendant consideration of specific alternatives—some have suggested just skipping the non-comparative techniques for comparative ones. For example, while Berger and Wolpert (1988, §4.4.4) admit that Fisherian tests might be legitimate even for Bayesian model verification, they stress that it is almost always better to consider explicit alternatives, which allows for the sole use of techniques that satisfy the LP. Indeed, many Bayesian have objected that statistics for non-comparative model verification like the PPP cannot be assimilated into the Bayesian fold, thus must be discarded for comparative methods which probe the same discrepancies—see, for example, the discussions of O'Hagan and Lindley in Box (1980, p. 408, 423) and those of Bernardo, Lindley and O'Hagan in Bayarri and Berger (1999, 72–73, 75,

[8]Technically, $T(x, \theta)$ is not a statistic in the standard sense since it is a function of the probability distribution of the parameter θ as well as the data.

[9]There is considerable debate regarding the merits of the PPP over other candidates for Bayesian p-values, especially regarding the extent to which they make illicit "double use" of the data. See, e.g., Bayarri and Berger (1999) and references cited therein. What is important here is that all these candidates *do* make use of sampling distributions.

77).

While I think this response is coherent, it gives up completely on assessing the statistical adequacy of the class of alternatives, hence on answering the problem of the unconsidered alternative. Given that it seems uncontroversial that the assumptions of *deterministic* models can be verified non-comparatively, it is not clear why moving to statistical models should change this so drastically. After all, deterministic models are just statistical models lying on the edge of the appropriate probability simplex, so by a continuity argument one would expect that non-comparative methods would apply to statistical models as well. Further, without a way to compare the evidence for all possible models, one has no way of securing evidence from any model beyond intuitions that the delimited class of alternatives considered contains one that is sufficiently adequate. As Box stated colorfully in his response to O'Hagan,

> My difficulty with probability ratios (and likelihood ratios and predictive ratios) is of course that while I grant that it may be useful to know that it is a million times more probable that the first man I meet when I walk down the street will be called John Smith rather than Jeremiah Hezekiah Bramblebottom, this in itself tells me little about the chance of meeting a man called John Smith.
>
> (Box 1980, p. 427)

He goes on to remark that he finds it astonishing that anyone could consider all the possible models to be known a priori for any real scientific investigation, and that because Bayesian inference is conditional on the adequacy of a probability model, one must probe that assumption to secure inference.

Importantly, one can test an assumption in this way when it fixes a probability distribution for the relevant test statistic without doing so for the whole probability model for the data. Such nonparametric methods do not have clear comparative LP-satisfying counterparts,[10] and are crucial in combination with other omnibus and directed tests for strategic probing of modeling assumptions (Mayo and Spanos, 2004, p. 1023–1024). Thus, unless non-comparative inferential methods are developed that satisfy the LP, no general methods for the full demands of model verification can satisfy the LP if they are conceived of inferentially. Epistemologically, securing our evidence claims requires making them robust to uncertainty, and model verification secures evidence through winnowing the uncertainty regarding modeling assumptions (Staley, 2012). So, while some might be lead to the radical conclusion that the adequate fit of a class of models cannot be checked, it seems more plausible to accept and continue using non-comparative checks of fit while inferring that the LP is not a universal constraint on any measure of evidence.

3 The Non-Inferential View of Model Verification

In the face of the above problems with the inferential view of model verification, it seems that one must give up non-comparative assessments of a model's fit with a data set or admit that the LP cannot apply to those assessments, hence cannot apply universally. But an advocate of the LP as a constraint on evidence may respond that one can and should distinguish these

[10]So-called Bayesian non-parametric models allow flexibility with the number of latent variables, which are nevertheless constrained to follow parametric distributions.

assessment of fit—and perhaps techniques of model verification more generally—from the techniques of inference to which the LP really applies. This non-inferential view of model verification seeks to save the LP by restricting its scope—or equivalently, by denying that it should ever have applied to certain techniques in the first place.

There are several ways one might do this. One is to make a distinction between primary and secondary inferences, that is, between those of scientific interest in the problem at hand and those that are merely necessary in order to get at the former, like inferences on nuisance parameters and the testing of model assumptions. To a first approximation, this view structures statistical analysis as a two-step process, the first of which involves techniques of model verification to select a sufficiently statistically adequate model. The second step, in which the model is then subjected to inferential procedures for the parameters of scientific interest, would be the only one for which the LP applies.[11]

On this view, the antecedent of the LP commits one to hold off on its application until all of the modeling assumptions short of the primary scientific questions of inference are statistically adequate. It seems that Box took this view, for in response to a comment of Lindley's that his advocated model verification techniques violate the LP, he remarked that

> there is in my mind no question of abandoning the likelihood principle so far as estimation is concerned. ... But if we aim to judge whether samples resembling in some relevant respect the one we have observed are or are not rare, it seems to me essential to know (as part of the model) the rule by which the samples were generated (Box 1980, p. 427)

—that is, the sampling distribution of the statistics used in tests of modeling assumptions.

Despite the plausibility of this distinction, however, it is too vague to do the work needed to save the LP. Different parameters or features of a model may range in scientific interest hierarchically, in which case one may ask, how is one to decide what the appropriate cutoff between primary and secondary is supposed to be? Supposing the distinction to be conventional seems unsatisfactory, as the complexity of scientific modeling will demand that new conventions be invented endlessly. More importantly, it is not clear that a convention can really bear the epistemological weight that the distinction it stipulates imposes. Further complicating matters, there is no reason to believe that these parameters and features even form a linear order, that even if there is such an order, that this order is stable for a researcher over time or is invariant across researchers. Researchers can well disagree about which parameters are of importance, and the course of scientific investigation often changes their relative importance in marked ways.

Another option is to restrict the range of applicability of the LP to inferential questions where there is some known technique that satisfies those questions. Then the LP would only apply to circumstances where there are some competing measures of evidence, at least one of which satisfies its strictures. It would then demand that those techniques that satisfy it are always to be preferred, but would be silent when there are no such techniques. In particular, if there are no techniques to make a relevant comparative inference, as with checking certain

[11] In practice, one often performs primary inferences first and then circles back to test the assumptions of the resulting model. Moreover, these steps may be repeated to sequentially test different assumptions, but these complication are tangential to my argument here.

modeling assumptions, then the LP does not apply. This gives a relatively unambiguous criterion for distinguishing which inferential procedures are constrained by the LP and which are not, but it ossifies the distinction with respect to a particular epistemic situation. The state of one's statistical repertoire is dependent upon what one knows personally and what is known to the statistical community. Thus if a new technique satisfying the LP is developed for a particular problem for which one previously had only non-comparative techniques, all the analyses using the latter would become suspect.

There are firmer distinctions, however, on which the non-inferential view of model verification can stand. One is that the problematic techniques all use the Fisherian logic of testing, whereby one can disconfirm an assumption but only confirm the fit of an assumption with the data. That is, techniques of model verification by themselves are not severe tests of positive hypotheses and therefore do not have the full inferential capacities that should cohere with "true" measures of evidence. (Indeed, Bayarri and Berger (1999) (and others) call them measures of "surprise" since they merely indicate how strongly one should consider developing alternative models.) This is distinguished from the previous option by making particular techniques, not particular questions of inquiry, the distinct objects over which one asks if the LP applies. George Barnard, in his discussion of Birnbaums's essay, was one of the first to draw this distinction, pointing out that Fisherian tests of significance require less structure than a statistical experiment as defined in §1:

> In particular, [Fisherian] tests of significance arise, it seems to me, in situations where we do not have a parameter space of hypotheses; we have only a single hypothesis essentially, and the sample space then is the only space of variables present in the problem. The fact that the likelihood principle is inconsistent with significance test procedures in no way, to my mind, implies that significance tests should be thrown overboard; only that the domain of applicability of these two ideas should be carefully distinguished. (Birnbaum 1962, p. 308)

But because Bayesian inferential techniques are not typically framed in terms of severity,[12] it is no longer clear on this view how techniques of Bayesian model verification can be separated from traditional posterior inference. Perhaps the right distinction, then, is not between severe and non-severe tests as it is between comparative and non-comparative methods.

These problems might be overcome, or particular deficiencies accepted. *Any* distinction one wishes to draw to separate methods of model verification from those to which the LP should apply, however, faces the problem of accounting for how the different degrees to which assumptions may have been adequately verified must affect the evaluation of the evidence. For example, consider two data sets for a probability model that yield proportional likelihood functions but are compatible with (at least one of) the model's assumptions to a different degree. For primary inferences—or whichever category one intends to distinguish from model verification—the two data sets yield the same evidence by the LP, even if only one of them fits the model well.

One might protest that in this case, as in any, the value of the evidence is entirely conditional upon the statistical adequacy of the modeling assumptions. But any clear distinction between "adequate" and "inadequate" will virtually always be conventional, just

[12]See, however, Bandyopadhyay and Brittan (2006).

as the the 0.05 p-value threshold for rejecting a null hypothesis is conventional. Different scientific problems, even if they use formally the same statistical models, may demand different thresholds. Further, different scientists will have different senses for what the cutoff should be, even in the same problem. In other words, the degree of fit of the data with a model comes in degrees to which the evidential evaluation of the data should be sensitive. In particular, a strong lack of fit should preclude strong evidence. If, on the other hand, the degree of adequacy of the modeling assumptions is bracketed from the evidential evaluation, as the non-inferential view of model verification proposes, then there is no longer any way to take the former into account in the latter.

One can of course uphold the LP if one gives up letting the results of model verification influence assessments of the evidence, but then it is not clear why one would have bothered to check modeling assumptions in the first place. As with the radical response to the problems of the inferential view, however, this position admits no way to let the reliability of modeling assumptions affect assessments of the evidence. I take it to be uncontroversial that statisticians concerned with non-trivial statistical problems should care about the outcomes of their model checks because they should be concerned with the reliability of the statistical, hence scientific, assumptions undergirding any data analysis. In conflict with this, the non-inferential conception of model verification brackets information about reliability from LP-satisfying assessments of evidence.

4 Concluding Remarks on Evidence and Sufficiency

4.1 Idealistic and Pluralistic Revisions

An insistent advocate of the non-inferential view who also has affinity for the LP may retreat even further by conceding that whatever the evidence for various hypotheses might be, it is always contingent on the sufficient agreement of *all* parties that the model assumptions have been sufficiently verified. Then, to the *extent* to which this is case, one may be able to apply the LP. As stated, this view is a bit vague, but there is something attractive in it. Here is one possible explication: For all that I have argued above, an advocate of the LP still has the option to say that one can "recover" the LP as a constraint on any measure of evidence for a particular experimental outcome only in the limit as the relevant epistemic community judges the underlying modeling assumptions to be statistically adequate. In practice, though, the evidence will always be modulated by the uncertainty in those assumptions. Thus, one might affirm that the LP is a constraint on any "ideal" measure of evidence in the limit of "perfect adequacy" while admitting that actual evidential measures must depart from it somewhat.

But this is not the only direction one might explore. (Indeed, many opponents of the LP would not likely be satisfied with the above suggestion.) Disavowing the universal relevance of the above idealization, one might instead find a role for the LP not as a *constraint* on any measure of evidence, but merely as *defining* a class of evidential measures among possibly many others. On this pluralistic view, there are potentially several different types of evidence, the exploration of whose interrelationships may help resolve the sometimes contradictory intuitions feeding debates on the nature of scientific evidence. In particular, if the LP is no longer an all-encompassing restriction on every measure of evidence, there is no longer any

problem with the fact that in some circumstances of model verification it does not seem to apply.[13]

Relatedly, Royall (1997, p. 4) distinguishes between three kinds of questions that one can ask oneself subsequent to the observation of data:[14]

1. What should I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. How should I interpret this observation as evidence for various hypotheses?

Bayesian analysis concerns the first, Neyman-Pearson testing theory the second, and—Royall argues—pure likelihood methods the third. Importantly, the answer to the evidence question should bear on the belief and action questions. In light of the discussion so far, one might add a further question: what does this observation tell me about the reliability of my assumptions? Many methods might suggest themselves for answering this question, particularly error statistics and the kind of Fisherian testing that I have argued is crucial in many techniques of model verification because it assess the fit between the data and the model. The answer to the reliability question should in turn inform the answer to the evidence question, and in fact sufficient reliability should be a necessary condition for substantive evidence.[15] Following my discussion of the vagueness of the distinction between primary and secondary inferential goals in §3, there will, of course, be some vagueness regarding what part of inquiry falls under the evidence question and what part falls under the reliability question. But however one divides it in a particular circumstance, the answers to what one takes as questions of reliability will inform and ground the answers to what one takes as questions of evidence.

4.2 Return to Birnbaum and Sufficiency

As alluded in §1, before concluding I would like to return to Birnbaum's theorem: if considerations from model verification lead one to reject the general applicability of the LP, how do those considerations bear on the weak sufficiency and conditionality principles (Berger and Wolpert, 1988, p. 25–28)?

Weak Sufficiency Principle (WSP) For an experiment $E = (X, \theta_0, \{P^\theta\}_{\theta\in\Theta})$, consider any sufficient statistic $S(X)$ for $\theta \in \Theta$.[16] If $S(x_1) = S(x_2)$, then any measure of

[13]Cf. the discussion of Dempster in Birnbaum (1962, p. 318).

[14]I have adapted his questions somewhat to match the present discussion.

[15]Bandyopadhyay and Brittan (2006, p. 276), in the course of proposing a Bayesian account of severity, also recommend adding a fourth question to Royall's list: "when is a test *severe*?" A comparative analysis of their position would take me too far afield here, but one important difference is that they take severity to be a relationship between data and a hypothesis that obtains when the posterior probability of the hypothesis is high and the likelihood ratio of the hypothesis to its competitors is large. They accept the likelihood ratio as the right LP-satisfying measure of comparative evidence and take evidence to be prior to severity (Bandyopadhyay and Brittan, 2006, p. 290, fn. 58), which does not in general satisfy the LP because of its dependence on the posterior. Although I take techniques of model verification not to satisfy the LP in general, I am in contrast suggesting to take reliability to be a *precondition* of some kind for evidence.

[16]A statistic $S(X)$ is sufficient for a parameter θ, given a probability model $P^\theta(x)$, when $P^\theta(x|S(x))$ does not depend on θ.

evidence e must satisfy $e(E, x_1) = e(E, x_2)$.

Weak Conditionality Principle (WCP) Consider experiments, $E_1 = (X_1, \theta_0, \{P_1^\theta\}_{\theta\in\Theta})$ and $E_2 = (X_2, \theta_0, \{P_2^\theta\}_{\theta\in\Theta})$, where only the parameter space Θ need be common. Consider further the mixed experiment E_*, wherein $J = 1$ or 2 is observed, each having probability $\frac{1}{2}$ (independent of θ, X_1, X_2), and then experiment E_J is performed. Formally, $E_* = (X_*, \theta_0, \{P_*^\theta\}_{\theta\in\Theta})$, where $X_* = (J, X_J)$ and $P_*^\theta((j, x_j)) = \frac{1}{2}P_j^\theta(x_j)$. Then any measure of evidence e must satisfy $e(E_*, (j, x_j)) = e(E_j, x_j)$.

One often takes the WSP to state, intuitively, that all the information about θ present in the data x is contained in a sufficient statistic. The WCP, on the other hand, states that the information obtained about θ in a mixed experiment depends only on the subexperiment actually performed. Since Birnbaum's theorem states that the conjunction of the WSP and the WCP is equivalent to the LP, changing the scope of the latter must change the scope of at least one of the former.

While the foregoing arguments present no obvious threat to the WCP, I believe that they do bear upon the WSP in much the same way that they did the LP. Namely, a sufficient statistic evaluated on two distinct data sets may yield the same value (cf. proportional likelihoods) without those data sets being equally statistically adequate. One can read some precedent for this conclusion when Casella and Berger explain why they apparently take a non-inferential view of model verification:

> Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine *residuals* from a model ... Such a practice immediately violates the Sufficiency Principle, since the residuals are not based on sufficient statistics. (Of course, such a practice directly violates the Likelihood Principle also.) Thus, it must be realized that *before* considering the Sufficiency Principle (or the Likelihood Principle), we must be comfortable with the model. (Casella and Berger 2002, p. 295–6)

On the face of it, they seem to be arguing that model verification violates the SP because statistics used to test model assumptions, like residuals, are not sufficient statistics. This would be puzzling since the WSP applies but does not restrict inference to sufficient statistics. A more charitable reading is that the data sets x_1 and x_2 of two experiments E_1 and E_2 may satisfy $S(x_1) = S(x_2)$ for some sufficient statistic S while leading to radically different conclusions under tests of model assumptions. Thus the WSP is violated since substantive differences in statistical adequacy will lead to different evidential conclusions. For example, in a problem to estimate the mean of data assumed to be standard normal, $X_i \overset{iid}{\sim} N(\mu, 1)$, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient for μ. The WSP would require the same evidential conclusions from data sets with $\bar{x}_1 = \bar{x}_2$, even if they have sample variances of ≈ 1 and ≈ 100, respectively. The standard normal distribution is clearly inadequate to model the latter case, undermining the evidential claims about the parameter of interest (further respecification notwithstanding).

For these reasons, Casella and Berger seem to fall back to the non-inferential view of model verification, which takes the constraint of the LP to be contingent on the statistical adequacy of the model. But as discussed in §3, this view brackets information about

the statistical adequacy of the modeling assumptions from measures of evidence. Given their practical viewpoint,[17] however, I suspect that they would be attracted to many of the suggestions offered at the beginning of this section, whether idealistic or pluralistic.

References

Bandyopadhyay, Prasanta S. and Gordon G. Brittan, Jr. "Acceptability, Evidence, and Severity," *Synthese* 148: 259–293 (2006).

Bayarri, M. J. and James O. Berger. "Quantifying Surprise in the Data and Model Verification," *Bayesian Statistics 6*. Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford: Oxford University Press, 1999.

Berger, James O. and Robert L. Wolpert. *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics, 1988.

Birnbaum, Alan. "On the Foundations of Statistical Inference," *Journal of the American Statistical Association* 57: 269–326 (1962).

Box, George E. P. "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society A* 143.4: 383–430 (1980).

Casella, George and Roger L. Berger. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury, 2002.

Frigg, Roman and Stephan Hartmann. "Models in Science," *Stanford Encyclopedia of Philosophy*. June 25, 2012. `http://plato.stanford.edu/entries/models-science/`

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. *Bayesian Data Analysis*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC Press, 2004.

Gelman, Andrew and Cosma Rohilla Shalizi. "Philosophy and the practice of Bayesian statistics," *British Journal of Mathematical and Statistical Psychology* 66: 8–38 (2013).

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

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

Mayo, Deborah G. and Aris Spanos. "Methodology in Practice: Statistical Misspecification Testing," *Philosophy of Science* 71: 1007–1025 (2004).

Mayo, Deborah G. and Aris Spanos. "Error Statistics," *Handbook of the Philosophy of Science, Volume 7: Philosophy of Statistics*. Ed. Prasanta S. Bandyopadhyay and Malcolm R. Forster. Oxford: Elsevier, 2011.

[17] Concluding their discussion of the LP, they write, "At any rate, since many intuitively appealing inference procedures do violate the Likelihood Principle, it is not universally accepted by all statisticians. Yet it is mathematically appealing and does suggest a useful data reduction technique" (Casella and Berger, 2002, p. 296).

Miller, Richard W. *Fact and Method: Explanation, Confirmation and Reality in the Natural and Social Sciences*. Princeton: Princeton University Press, 1987.

Royall, Richard. *Statistical Evidence: A likelihood paradigm*. Boca Raton, FL: Chapman and Hall/CRC Press, 1997.

Sober, Elliott. "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause," *British Journal for the Philosophy of Science* 52: 331–346 (2001).

Spanos, Aris. *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press, 1999.

Staley, Kent. "Error-statistical elimination of alternative hypotheses," *Synthese* 163: 397–408 (2008).

Staley, Kent. "Strategies for securing evidence through model criticism," *European Journal for Philosophy of Science* 2: 21–43 (2012).

Stanford, P. Kyle. *Exceeding Our Grasp*. New York: Oxford University Press, 2006.