Journal of Statistical Theory and Practice, 7:187–195, 2013
Copyright © Grace Scientific Publishing, LLC
ISSN: 1559-8608 print / 1559-8616 online
DOI: 10.1080/15598608.2013.771549

Revisiting the Likelihoodist Evidential Account [Comment on "A Likelihood Paradigm for Clinical Trials"]

ARIS SPANOS

Department of Economics, Virginia Tech, Blacksburg, Virginia, USA

1. Introduction

The primary objective of the article by Zhang and Zhang (2013) is to propose an extension of the *Law of Likelihood* (LL) with a view to address one of its key weaknesses: how to compare composite hypotheses using the likelihoodist ratio.

To place this weakness in a historical context, as noted by Zhang and Zhang; Hacking (1965) was the first to articulate the current form of the LL: "the data provide evidence supporting one parameter value θ_1 over another value θ_2 if $L(\theta_1) > L(\theta_2)$ and the strength of that evidence is measured by the LR $L(\theta_1)/L(\theta_2)$."

A major advantage of the LL, as seen at the time, was its "evidential" output based on the use of the likelihood ratio $LR = L(\theta_1)/L(\theta_2)$ as providing a measure of the "the strength of evidence" *for* H_1. Such an evidential interpretation was clearly missing from the frequentist approach based on the Neyman–Pearson (N-P) accept/reject rules or Fisher's p value. Attempts to frame one using the p value led to numerous confusions and misinterpretations; see Mayo and Spanos (2006).

What is often not reported in the statistics literature is that Hacking, when reviewing a book by Edwards (1972), *Likelihood*, had a definite change of mind about the pertinence of the LL:

> The only great thinker who tried it out was Fisher, and he was ambivalent. Allan Birnbaum and myself are very favourably reported in this book for things we have said about likelihood, but Birnbaum has given it up and I have become pretty dubious. (Hacking 1972, 137)

Birnbaum's change of mind seems to be primarily due to this particular key weakness, because in Birnbaum (1968, 1970), he reverted back to the LL comparing two simples hypotheses. This interpretation was also adopted by Royall (1997), emphasizing that: "it [LL] does not apply to composite hypotheses generally" (18). His explanation for this key weakness is that composite hypotheses raise the problem of a *disjunction* of different values of θ, rendering any pairwise comparison of simple hypotheses between different subsets

Address correspondence to: Aris Spanos, Department of Economics, Virginia Tech (0316), Blacksburg, VA 24061, USA. Email: aris@vt.edu

problematic: "The law of likelihood does not allow us to characterize the evidence in terms of the hypotheses: H_1: $\theta \leq .2$ and H_2: $\theta > .2$," in the context of a simple Bernoulli model.

2. A Generalized Law of Likelihood (GLL)

The preceding sketchy historical perspective indicates that addressing this key problem is very important for the likelihoodist approach to inference and evidence. The solution proposed by the authors in this article is straightforward and intuitively appealing:

Impose enough restrictions to render the likelihood function *well-behaved*, and then extend the ratio $LR = \frac{L(\theta_1;\mathbf{z}_0)}{L(\theta_2;\mathbf{z}_0)}$ for comparing two simple hypotheses, say; H_1: $\theta_1 = .2$ and H_2: $\theta_2 = .3$, to the composite hypotheses:

$$H_1\colon \theta{\in}\Theta_1 = [0, .2]\,\text{and}\,H_2\colon \theta{\in}\Theta_2 = (.2, 1],$$

using the *Generalized Law of Likelihood*:

$$GLR = \frac{sup_{\theta\in\Theta_1}L(\theta;\mathbf{z}_0)}{sup_{\theta\in\Theta_2}L(\theta;\mathbf{z}_0)} = \frac{L(\hat{\theta}_1;\mathbf{z}_0)}{L(\hat{\theta}_2;\mathbf{z}_0)} \tag{1}$$

That is, address the key problem by reducing a composite hypothesis into a simple one using the values θ with the highest likelihood within each subset (Θ_1, Θ_2) of the parameter space $(\Theta = \Theta_1 \cup \Theta_2)$. To sidestep the problem of a *disjunction* of hypotheses the authors add a minimax-type condition in the form of the following axiom:

> **Axiom** 1. If $\inf_{\theta\in\Theta_1} L(\theta;\mathbf{z}_0) > \sup_{\theta\in\Theta_2} L(\theta;\mathbf{z}_0)$, then there is evidence supporting H_1 over H_2

To make a case for the "sensibleness" of the GLL the authors relate their suggested likelihoodist procedure to well-known frequentist procedures like the likelihood ratio test, confidence intervals, significance testing based on the *p* value, and Wald-type testing procedures, using asymptotic approximation results. This discussion constitutes both the strength and the main weakness of the case for GLL made in the article.

The strength of the GLL stems from the fact that if one were to take the connections between the GLL and the already-mentioned frequentist procedures at face value, the proposed procedure does not seem unreasonable. The GLL appears to retain the desirable property of "good likelihoodist performance probabilities" (Royall 1997, 2–3), while shaking off the undesirable ones like the relevance of the sample space beyond $\mathbf{z}_0$ or controlling error probabilities. Indeed, the authors make a reasonably strong case in the context of the likelihoodist approach, and I found myself agreeing with many of their arguments, including their discussion of the connection between the GLL and the Bayesian approach.

A closer look at the GLL, however, reveals that it does not adequately address the various problems associated with the likelihoodist evidential interpretation. If anything, bringing out the connections between the GLL and the mentioned frequentist procedures highlights these weaknesses more clearly.

3. Problems and Challenges for the GLL

First, there is nothing in the proposed GLL likelihoodist procedure that addresses the *maximally likely value* problem:

(a) There is always a maximally likely value, $\theta^{\blacklozenge} = \widehat{\theta}_{MLE}(\mathbf{z}_0)$, irrespective of the null hypothesis or any substantive values of interest; see Mayo (1996, 200). It is well known that the particular value $\theta^{\blacklozenge}$ is usually accidental because it depends on the specific value $\mathbf{z}_0$ taken by the sample $\mathbf{Z}$, and the probability that $\theta^{\blacklozenge}$ will coincide with θ^*, the "true" value of θ, is zero! Learning from data about θ^*, however, is the primary aim of frequentist inference, and hence the need for *error probabilities* to calibrate the reliability of any inference based on $\widehat{\theta}_{MLE}(\mathbf{Z})$.

By ignoring the relevant error probabilities, the GLL will always provide evidence for the hypothesis $H_i : \theta \in \Theta_i$, such that $\theta^{\blacklozenge} \in \Theta_i$, $i = 1$ or 2. This means that whatever the value of θ^* or the value θ_0 of particular interest, there will always be an alternative hypothesis, depending on whether the value $\theta^{\blacklozenge}$ falls within the Θ_1 or Θ_2 subset of the parameter space.

Second, the proposed procedure is vulnerable to the simple null versus composite alternative hypothesis problem. The authors admit that using the GLL, (b) " it is impossible to obtain empirical evidence supporting a single parameter value in a smooth model over its complement" (XX).

This is a serious problem for the GLL because in scientific research there is often a particular value of substantive interest, say θ_0, and a key role of hypothesis testing is to answer the question: "When do data $\mathbf{z}_0$ provide evidence for or against the hypothesis $\theta = \theta_0$?" In the case of evidence against $\theta = \theta_0$ one would be interested in the discrepancy from it warranted by data $\mathbf{z}_0$.

Third, the GLL is highly vulnerable to the following two fallacies that have bedeviled frequentist testing since the 1930s:

(c) *The fallacy of acceptance*: (mis)interpreting accept H_1 [no evidence against H_1] as evidence for H_1.

(d) *The fallacy of rejection*: (mis)interpreting reject H_1 [evidence against H_1] as evidence *for* a particular H_2.

Due to the fact that the GLL reduces a composite hypothesis to a simple one using the value of $\theta \in \Theta_i$, $i = 1, 2$, corresponding to $\sup_{\theta \in \Theta_i} L(\theta; \mathbf{z}_0)$, it is highly likely that whichever hypothesis (H_i: $\theta \in \Theta_i$) GLL provides strong evidence for (against) will include values for which there is evidence against (for), rendering such inference highly vulnerable to the fallacies of acceptance and rejection. It should be noted that axiom 1 does nothing to sidestep these fallacies; see section 5.

In section 4 it is demonstrated that there is frequentist evidential interpretation, based on severity evaluation (Mayo 1996), that can be used to avoid these fallacies. Ironically, this evidential interpretation calls upon the two properties cast off by the GLL as undesirable: the relevance of the sample space beyond $\mathbf{z}_0$ and "controlling" the error probabilities. The latter can be traced back to the birth of N-P testing:

> But if we accept the criterion suggested by the method of likelihood it is still necessary to determine its sampling distribution in order to control the error involved in rejecting a true null hypothesis, because a knowledge of lambda alone is not adequate to insure control of this error. (Neyman and Pearson 1930, 80)

Fourth, the 'good likelihoodist performance' probabilities, including those of misleading evidence, first proposed by Royall (1997), and repeated by proponents of the likelihoodist evidential account (Blume 2011, 497–499, inter alia), are highly misleading when viewed as appraising the performance of a legitimate frequentist testing procedure as

Table 1
Physician's diagnostic test

	Test result (T)	
	Positive (+)	Negative (–)
Disease (D), yes (+)	0.95	0.05
Disease (D), no (–)	0.02	0.98

framed by the N-P approach. I recognize that this criticism is rather unfair on the authors of this article, but is important to mention it as a general criticism of the likelihoodist evidential account more generally.

A glance at the physician's diagnostic test, based on Royall's example in Table 1, reveals that it is only superficially related to a legitimate N-P testing procedure, because none of the key components that define a proper N-P test is present:

1. Appropriate data.
2. Legitimate hypotheses.
3. A test statistic and rejection region.
4. Sampling distributions.
5. Relevant error probabilities.

When viewed in the context of frequentist inference, the various probabilities used to extol the virtues of the LL, like $\Pr(T+|D+)/\Pr(T+|D-)$ and $\Pr(T-|D-)/\Pr(T-|D+)$ amount to nothing more than *deductive* calculations of conditional probabilities associated with particular *events* (not hypotheses) within the context of a *known* model (no unknown θ in the bivariate distribution in Table 1). Indeed, these probabilities have nothing to do with legitimate hypotheses, tests or error probabilities.

Legitimate frequentist N-P hypotheses are framed in terms of the unknown parameter(s) θ, and they pose questions relating to the true θ. Moreover, legitimate error probabilities (a) are never conditional; they are evaluated under different scenarios relating to particular values of θ; (b) are always assigned to inference procedures (never to hypotheses or events); and (c) invariably depend on the sample size $n > 1$: the large n problem. The likelihoodist false positive and false negative probabilities have only superficial resemblance to type I and II error probabilities since they ignore n, as well as component 1–5; see Spanos (2010) for further discussion.

The use of the preceding example reminds me of somebody using table-football (foosball) in order to shed light on the skills and strategies needed by a player and the team to play soccer effectively, just because there are 11 players associated with both games!

4. A Frequentist Evidential Interpretation

To render the discussion less abstract, let me focus on an example used in the article:

The trial enrolled 164 children with nephroblastoma, who were randomly assigned to either chemotherapy or radiation therapy. The primary objective of the trial was to demonstrate that chemotherapy is noninferior to radiation therapy with respect to the response rate. More precisely, noninferiority here means that the response rate for chemotherapy is not lower than that for radiation therapy by more than a margin of 10%, which is considered the smallest clinically meaningful difference between two groups. The observed response rates were 94.3% (83/88) for chemotherapy and 90.8% (69/76) for radiation therapy. (x)

For the random variables X_k, response to chemotherapy, and Y_k, response to radiation, the relevant statistical model comprises two independent *Bernoulli IID models:*

$$\text{Chemotherapy}: X_k \sim \text{BerIID}\left(\theta_1,\ \theta_1(1-\theta_1)\right), \quad k = 1, 2, \ldots, n_1, \ldots,$$

$$\text{Radiation}: Y_k \sim \text{BerIID}\left(\theta_2,\ \theta_2(1-\theta_2)\right), \quad k = 1, 2, \ldots, n_2, \ldots$$

and the substantive hypotheses of interest can be framed in the form of the statistical hypotheses:

$$H_0: \theta \leq 0 \text{ vs. } H_1: \theta > 0, \qquad \text{where } \theta: = \theta_1 - \theta_2. \tag{3}$$

Since the sampling distributions of the best estimators of (θ_1, θ_2) are

$$\widehat{\theta_1} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} X_k \sim \text{Bin}\left(\theta_1, \tfrac{\theta_1(1-\theta_1)}{n_1}; n_1\right), \quad \widehat{\theta_2} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} \quad Y_k \sim \text{Bin}\left(\theta_2, \tfrac{\theta_2(1-\theta_2)}{n_2}; n_2\right)$$

where 'Bin' denotes a scaled Binomial distribution, one can construct an α-significance level test for (3) based on

$$d(\mathbf{Z}) = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{Var(\hat{\theta}_1 - \hat{\theta}_2)}} = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} \quad C_1: = \{\mathbf{z}: d(\mathbf{z}) > c_\alpha\}, \tag{4}$$

where $\mathbf{Z}: = (X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2})$. For "large enough" n_1 and n_2, one can approximate the sampling distributions of this test statistic, under the null and alternatives, using the Normal distribution:

$$d(\mathbf{Z}) \overset{H_0}{\approx} \text{N}(0,1), \quad d(\mathbf{Z}) \overset{\theta=\gamma_1}{\approx} \text{N}(\delta_1, 1),\ \delta_1 = \frac{\gamma_1}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}}, \text{ for } \gamma_1 > 0.$$

It is important to note that in the context of this test the *noninferiority* hypothesis as described by the authors,

> the response rate for chemotherapy is not lower than that for radiation therapy by more than a margin of 10%, which is considered the smallest clinically meaningful difference between two groups. (X)

is viewed as rendering the discrepancy $\gamma_1^* = -.1$ of substantive interest. The framing in (3) should be contrasted with the treatment of this substantive information by the authors, where γ_1^* is inserted into the framing of the hypotheses of interest; see section 4.

Applying the test (4) to the preceding data yields:

$$d(\mathbf{Z}) = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}} = \frac{(83/88) - (69/76)}{\sqrt{\frac{(83/88)(1-(83/88))}{88} + \frac{(69/76)(1-(69/76))}{76}}} = \frac{.0353}{.04134}$$

$$= .854[.197],$$

where the number in square brackets denotes the p-value, which indicates acceptance of H_0 at any traditional significance level $.025 < \alpha < .1$. Does this mean that data $\mathbf{z}_0$ provide evidence for H_0? Or more accurately, does $\mathbf{z}_0$ provide evidence for *no substantive discrepancy* from H_0? Not necessarily! The leap from accept H_0 to evidence *for* H_0 is vulnerable to the fallacy of acceptance mentioned earlier.

To circumvent the fallacy of acceptance one needs to go beyond the p-value and the coarse N-P "accept/reject" results, and use the *post-data severe testing* evaluation aiming to delimit the discrepancy $\gamma_1 > 0$ warranted by data $\mathbf{z}_0$. Severity provides an evidential account by evaluating (post-data) the N-P accept/reject results with a view to establish the (smallest/largest) discrepancy γ from H_0 warranted by data $\mathbf{z}_0$. It is by definition *directional* since the sign ($\pm$) of $d(\mathbf{z}_0)$ indicates the relevant direction of departure from H_0, as indicated by $\mathbf{z}_0$.

A hypothesis or a claim H passes a *severe test* T with data $\mathbf{z}_0$ if

(S-1) $\mathbf{z}_0$ accords with H, (for a suitable notion of accordance) and
(S-2) with very high probability, test T would have produced a result that accords less well with H than $\mathbf{z}_0$ does, if H were false.

Severity, denoted by $SEV(T, \mathbf{z}_0, H)$, should be viewed as an feature of a test T as it relates to a particular data $\mathbf{z}_0$ *and* a specific claim H.

In the case of the preceding numerical example, the test $T_\alpha\colon = \{d(\mathbf{Z}), C_1(\alpha)\}$ in (4) accepts H_0, and the idea is to establish the *smallest* discrepancy $\gamma_1 > 0$ from H_0 (as indicated by $d(\mathbf{z}_0) = .854 > 0$) warranted by data $\mathbf{z}_0$, by evaluating (post-data) the relevant inferential claim $\theta \leq \gamma_1$. Note that in the case of "reject H_0," severity establishes the *largest* discrepancy warranted by $\mathbf{z}_0$, and the relevant inferential claim is $\theta > \gamma_1$; see Mayo and Spanos (2006).

Condition (S-1) is satisfied since $\mathbf{z}_0$ accords with H_0 because $d(\mathbf{z}_0) < c_\alpha$, but in addition, condition (S-2) calls for evaluating the probability of "all outcomes $\mathbf{z}$ for which test T_α accords less well with H_0 than $\mathbf{z}_0$ does," that is, ($\mathbf{z}$: $d(\mathbf{z}) > d(\mathbf{z}_0)$), under the hypothetical scenario that "$\theta \leq \gamma_1$ is false" or equivalently "$\theta > \gamma_1$ is true":

$$SEV(T_\alpha; \mathbf{x}_0; \theta \leq \gamma_1) = (\mathbf{z}\colon d(\mathbf{Z}) > d(\mathbf{z}_0);\ \theta > \gamma_1), \text{ for } \gamma_1 > 0. \tag{5}$$

The threshold $d(\mathbf{z}_0)$ render this a *post-data* error probability. Its evaluation relies on the sampling distribution (4) and different discrepancies ($\gamma \geq 0$), yielding the results in Table 2; see also Figure 1. Note that Sev$(T_\alpha; \mathbf{z}_0; \theta \leq \gamma_1)$ is evaluated at $\theta = \gamma_1$ because the probability increases with γ_1. As mentioned earlier; the aim in the case of accept H_0 is to find the "smallest" discrepancy γ warranted by data $\mathbf{z}_0$ with high enough severity.

Assuming a severity threshold of, say, .95 is high enough, the preceding results indicate that the "smallest" discrepancy warranted by $\mathbf{z}_0$, is:

$$\gamma_1 \geq .105 \text{ with } SEV(T_\alpha; \mathbf{z}_0; \theta \leq .105) = .95. \tag{6}$$

Table 2
Severity evaluation in the case of "accept H_0"

	Relevant inferential claim: $\theta \leq 0 + \gamma_1$												
γ_1	.01	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.15
Sev$(\theta \leq \gamma_1)$	.270	.449	.545	.639	.725	.799	.860	.907	.941	.965	.980	.989	.997

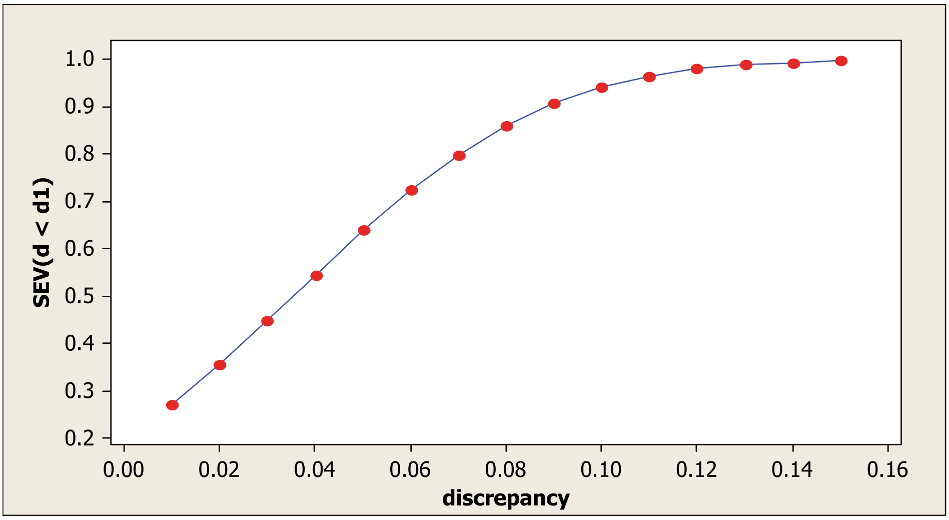

Figure 1. Severity evaluation for "accept H_0" (color figure available online).

That is, data $\mathbf{z}_0$ (via test T_α) provide evidence (with severity at least .95) for the presence of a discrepancy as large as $\gamma \geq .105$. Is this discrepancy substantively insignificant? Only substantive subject matter information can answer that question. In this sense, the severity evidential account outputs the discrepancy γ^* warranted by data $\mathbf{z}_0$ at a certain severity threshold, and calls upon:

(a) the subject matter information to determine whether the discrepancy in (6) is substantively significant or not, and
(b) basic logic to draw any further logical implications that follow from the relevant inferential claim ($\theta \leq 0 + \gamma_1$) in conjunction with the warranted discrepancy in (6).

How does the preceding evidential interpretation of the N-P test's accept H_0 differ from the one based on the GLL? The key difference is that the severity evaluation is firmly attached to the test T_α itself as it relates to the relevant inferential claim, $\theta \leq 0 + \gamma_1$, with a view to providing an evidential interpretation of the test's result based on the discrepancy from the null γ^* warranted by data $\mathbf{z}_0$. Whether data $\mathbf{z}_0$ provide evidence for or against a particular hypothesis H (or claim) depends crucially on the generic capacity (power) of the test in question to detect discrepancies from H_0. This stems from the intuition that a rejection of H_0 based on a test with low power (e.g. a small n) for detecting a particular discrepancy γ provides *stronger* evidence for the presence of γ than using a test with much higher power (e.g. a large n); see Mayo (1996). Indeed, the post-data severity evaluation of the accept/reject results provides a frequentist evidential account based on harnessing this intuition, by custom-tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{z}_0$. For the case reject H_0, see Mayo and Spanos (2006).

Viewed from the severity perspective, the problem with using a small p-value as a basis for inferring evidence *for* a particular alternative H_1 stems from the fact that the p-value only indicates the presence of "some" discrepancy from H_0, but provides no information about its magnitude; the latter requires summoning the generic capacity of the test; (Mayo and Spanos 2011). Given that $d(\mathbf{z}_0) = .854 > 0$ indicates a positive discrepancy from $\theta = 0$, the answer provided by the severity assessment in (6) makes perfectly good sense, intuitively.

5. The Likelihoodist Evidential Account Revisited

In contrast to the severity evidential account, the GLL attaches likelihoods to different values of $\theta \in \Theta$, and uses the GLR in (1) as a measure of the strength of evidence for and against H_1 and H_2. In the preceding example, the authors conclude:

Under the GLL, the non-inferiority hypothesis is strongly supported with a GLR of 138 based on the profile likelihood. In fact, with a higher observed response rate in the chemotherapy group, there is even evidence supporting the superiority of chemotherapy to radiation therapy. This latter piece of evidence is rather weak, though, with a GLR = 1.4.

Following Zhang (2006, 948), the "noninferiority" hypothesis (H_2):

$$H_1\colon \theta \in \Theta_1 = [-1, -.1] \text{ vs. } H_2\colon \theta \in \Theta_2 = (-.1, 1], \quad \text{for}\, \theta = \theta_1 - \theta_2,$$

in which case the authors' Figure. 1 suggests that the GLL yields:

$$GLR = \frac{\sup_{\theta \in \Theta_2} L(\theta; \mathbf{z}_0)}{\sup_{\theta \in \Theta_1} L(\theta; \mathbf{z}_0)} = \frac{L(\theta = .0353; \mathbf{z}_0)}{L(\theta = -.1; \mathbf{z}_0)} = \frac{1}{.00472} = 211.86,$$

indicating stronger evidence *for* H_2 than the GLR =138 reported by the authors. Similarly, framing the "superiority of chemotherapy to radiation therapy" as H_2 in:

$$H_1\colon \theta \in \Theta_1 = [-1, 0] \text{ vs.} H_2\colon \theta \in \Theta_2 = (0, 1], \quad \text{for}\, \theta = \theta_1 - \theta_2,$$

their Figure 1 confirms closely the authors' evaluation since:

$$GLR = \frac{\sup_{\theta \in (0,\, 1]} L(\theta; \mathbf{z}_0)}{\sup_{\theta \in (-.1, 0]} L(\theta; \mathbf{z}_0)} = \frac{L(\theta = .0353; \mathbf{z}_0)}{L(\theta = 0; \mathbf{z}_0)} = \frac{1}{.6945} = 1.44.$$

It is important to emphasize that in both GLR evaluations the problem with the maximally likely value, $\widehat{\theta}_{MLE}(\mathbf{z}_0) = .0353$, plays a crucial role. Depending on whether the partition line defining Θ_1 and Θ_2 is on the left or right of the line $\theta^{\blacklozenge} = .0353$ determines the evidence for or against the two hypotheses, and nothing else matters! For example, placing the partition line at $\theta = .1$ (indicated by severity) brings out the conflict between the two evidential accounts. For the hypotheses:

$$H_1\colon \theta \in \Theta_1 = [-1, .1] \text{ vs. } H_2\colon \theta \in \Theta_2 = (.1, 1], \ \text{for}\, \theta = \theta_1 - \theta_2,$$

$$GLR = \frac{\sup_{\theta \in [-1,\, 1]} L(\theta; \mathbf{z}_0)}{\sup_{\theta \in (.1, 1]} L(\theta; \mathbf{z}_0)} = \frac{L(\theta = .0353; \mathbf{z}_0)}{L(\theta = .1; \mathbf{z}_0)} = \frac{1}{.294} = 3.4,$$

indicates evidence *against* Θ_2, which includes the discrepancy $\gamma_1 = .1$ with .94 severity. Similarly, for the inferential claim, $\theta \leq -.1$, associated with the noninferiority threshold, the severity is tiny: $SEV(T_\alpha; \mathbf{x}_0; \theta \leq -.1) = .0005$. Moreover, the inferential claim, $\theta \leq .0353$, associated with the maximally likely value $\theta^{\blacklozenge} = .0353$ does not have high severity since $SEV(T_\alpha; \mathbf{x}_0; \theta \leq .0353) = .5$. In general, any composite hypothesis for which GLR provides strong evidence *for* (against) will always include values of θ for which the

LR indicates evidence *against* (for), rendering GLR highly susceptible to the fallacies of acceptance and rejection; See Spanos (2013).

There are two fundamental differences between the two perspectives. First, the severity evaluation is always attached to a relevant inferential claim and not to particular values of θ. Second, the GLL ignores the sampling distributions and the associated error probabilities, by invoking the *likelihood principle* (Berger and Wolpert, 1988), which asserts that no other value $\mathbf{z}$ of the sample $\mathbf{Z}$, apart from data $\mathbf{z}_0$, is relevant for inference purposes. As a result, the GLL ignores the generic capacity of a test (its power) when transforming the GLR result into evidence *for* H_1 or H_2. For instance, the evidential result of the GLR would have been the same whether the estimates $(\widehat{\theta_1} - \widehat{\theta_2}) = .0353$ and $\sqrt{Var(\widehat{\theta_1} - \widehat{\theta_2})} = .04134$ were based on $n = 10$ or $n = 10000$! The evidential result of the severity evaluation would have been very different in the two cases because the value of n affects the power of the test and thus the warranted discrepancy from the null. Intuitively, what goes wrong is that the GLR uses Euclidean geometry to evaluate evidence *for* H_1 or H_2, when in fact the statistical testing space is *curved*, with the curvature determined primarily by the generic capacity of the test to detect discrepancies from H_1.

References

Berger, J. O., and R. W. Wolpert. 1988. *The likelihood principle*, IMS Lecture Notes—Monograph series, 2nd ed., vol. 6, Hayward, CA, Institute of Mathematical Sciences.

Birnbaum A. 1968. Likelihood. In *International encyclopedia of the social sciences*, vol. 9, 299–301. New York, Macmillan and the Free Press.

Birbaum, A. 1970. Statistical methods in scientific inference (letter to the editor) *Nature*, 225, 1033.

Blume, J. D. 2011. Likelihood and its evidential framework. In *Handbook of philosophy of science, vol. 7: Philosophy of statistics*, ed. D. Gabbay, P. Thagard, and J. Woods, 493–511. Amsterdam, Elsevier.

Edwards, A. W. F. 1972. *Likelihood*. Baltimore, MD, Johns Hopkins University Press.

Hacking, I. 1965. *Logic of statistical inference*. Cambridge, UK, Cambridge University Press.

Hacking, I. 1972. Review of ‘Likelihood: An account of the statistical concept of likelihood and its application to scientific inference’ by A. F. Edwards. *Philos. Sci.*, 23, 132–137.

Mayo, D. G. 1996. *Error and the growth of experimental knowledge*, Chicago, University of Chicago Press.

Mayo, D. G., and A. Spanos. 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction, *Bri. J. Philos. Sci.*, 57, 323–357.

Mayo, D. G., and A. Spanos. 2011. Error statistics. In *Handbook of philosophy of science, vol. 7: Philosophy of statistics*, ed. D. Gabbay, P. Thagard, and J. Woods, 15–196. Amsterdam, Elsevier.

Neyman, J., and E. S. Pearson, 1930. On the problem of two samples. *Bull. Acad. Polish Sci. Lett.*, 73–96.

Royall, R. 1997. *Statistical evidence: A likelihood paradigm*. New York, Chapman & Hall.

Spanos, A. 2010. Is frequentist testing vulnerable to the base-rate fallacy? *Philos. Sci.*, 77, 565–583.

Spanos, A. 2013. Who should be afraid of the Jeffreys-Lindley paradox? *Philos. Sci.*, 80, 73–93.

Zhang, Z. 2006. Non-inferiority testing with a variable margin. *Biometrical J.*, 48, 948–965.

Zhang, Z., and B. Zhang. 2013. A likelihood paradigm for clinical trials. *J. Stat. Theory Pract.*, 7(2), 157–177.