Statistical Science
2022, Vol. 37, No. 4, 625–627
https://doi.org/10.1214/22-STS862
Main article: https://doi.org/10.1214/20-STS811

Comments on *Confidence as Likelihood* by Pawitan and Lee in *Statistical Science*, November 2021

Michael Lavine and Jan F. Bjørnstad

Abstract. Pawitan and Lee (*Statist. Sci.* **36** (2021) 509–517) attempt to show a correspondence between confidence and likelihood, specifically, that "confidence is in fact an extended likelihood" (*Statist. Sci.* **36** (2021) 509–517, abstract). The word "extended" means that the likelihood function can accommodate unobserved random variables such as random effects and future values; see (*J. Amer. Statist. Assoc.* **91** (1996) 791–806) for details. Here, we argue that the extended likelihood presented by (*Statist. Sci.* **36** (2021) 509–517) is not the correct extended likelihood and does not justify interpreting confidence as likelihood.

Key words and phrases: Likelihood, confidence.

1. EXPERIMENTS, THE LIKELIHOOD FUNCTION AND THE LIKELIHOOD PRINCIPLE

An experiment is a triple $E = \{Y, (\Psi, \theta), P\}$ where Y denotes an observed random variable, Ψ denotes an unobserved random variable, θ represents unknown fixed parameters indexing the distribution of (Y, Ψ) and $P = \{p_\theta(y, \psi), \theta \in \Theta\}$. Ψ could be, for example, a random parameter, a random effect or a future observable. The likelihood function resulting from an experiment E is

$$L(\theta, \psi; y) \propto p_\theta(y, \psi), \tag{1}$$

a function of possible values (θ, ψ) after $Y = y$ has been observed; see Bjørnstad (1996) (hereafter B) for details. Likelihood functions are defined only up to constants of proportionality. Pawitan and Lee (2021) (hereafter PL) call (1) the *extended* likelihood function where "extended" refers to (1)'s inclusion of ψ. In keeping with B, we call it simply the likelihood function. When ψ is absent, then (1) reduces to the familiar

$$L(\theta; y) \propto p_\theta(y).$$

PL note that B proves what they call the extended likelihood principle (LP), namely that "if two experiments with identical unknown parameters [including Ψ] produce the same (or proportional) extended likelihoods, then they have the same evidence about the parameters." B adds "that by evidence we usually mean the inference made." An implication is that if two experiments yield proportional likelihood functions then they should also yield the same inference. Any inferential procedure that leads to different inferences does not follow LP.

Michael Lavine is Professor Emeritus, Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, Massachusetts 01003, USA (e-mail: lavine@math.umass.edu). Jan F. Bjørnstad is Professor Emeritus, University of Oslo and Statistics Norway (e-mail: Jan.Bjornstad@ssb.no).

2. PL'S ARGUMENT

PL begin with (1) and note that if there is no unknown fixed parameter and no data then (1) reduces to

$$L(\psi) = P(\Psi = \psi)\big(\text{PL 's (1)}\big). \tag{2}$$

PL then apply (2) to two cases.

Case 1 In PL's Section 2.1, U is the indicator of whether a confidence interval covers its target:

$$U = U(Y) \equiv \mathbf{1}_{\mathrm{CI}(Y)}(\theta_{\text{true}}) \tag{3}$$

where $\mathrm{CI}(Y)$ is the $1 - \alpha$ confidence interval constructed from Y and θ_{true} is the true but unknown value of θ. Equation (3) is the same as PL's first unnumbered equation in their Section 2.1 but makes the dependence on Y explicit. When $Y = y$ is observed, U is realized but still unobserved because it is a function of the unknown θ_{true}. U plays the role of Ψ.

Case 2 In PL's Section 2.3, $T = t$ is an estimate of θ. The right-hand side P-value function is $C(t, \theta) \equiv P_\theta(T^* \geq t)$ where T^* is a random variable with the same distribution as T. PL explain how $C(t, \theta)$ can be used to derive the confidence distribution. PL also say "When t

is random, the quantity $V = C(T, \theta)$ is a random variable." When $Y = y$ is observed V is realized but unobserved and plays the role of Ψ.

In continuous problems, U and V are pivotals because $P_\theta(U) = \text{Bern}(1-\alpha)$ and $P_\theta(V) = \text{Unif}(0,1)$ do not depend on θ. Therefore, according to PL, (2) applies to U and V:

$$L(u) = p(u) = \begin{cases} 1-\alpha & \text{for } u = 1, \\ \alpha & \text{for } u = 0 \end{cases} \quad \text{and} \tag{4}$$
$$L(v) = p(v) = 1,$$

and hence both coverage and confidence can be interpreted as likelihood. In our opinion, (4) is a misapplication of likelihood, as we shall now explain.

3. ON THE DEFINITION OF THE LIKELIHOOD FUNCTION

When PL invoke (4), they imply that U, V or any other random variable has a unique unambiguous likelihood function. However, it is experiments, not random variables, that induce likelihood functions. As B explains,

> "The essential feature of this definition [of likelihood] ... is that it depends on the following two factors:
>
> a. specification of the *complete* model for observable variables, unobserved variables of interest (both in a modeling sense and for inferential interest) and unknown parameters
> b. inferential aim of the statistical investigation."

Thus no random quantity Ψ, and no fixed parameter θ, has a likelihood function except in relation to an experiment, its model, and its inferential aim. Equation (4) is an invalid likelihood function because it ignores those requirements.

PL's U and V are deterministic functions of θ and Y, and hence redundant, so instead of (4) we could use the familiar likelihood function $L(\theta; y) \propto p_\theta(y)$, ignoring U and V without any loss of information. But if we take seriously the idea that θ along with either $\Psi = U$ or $\Psi = V$ are of interest then the likelihood function must incorporate the full model for both observed and unobserved random variables and we are led to

$$L(\theta, \psi; y) = p_\theta(y, \psi) = p_\theta(y)\, p_\theta(\psi \mid y), \tag{5}$$

where ψ is either u or v. Although $p_\theta(\psi)$ does not depend on θ, $p_\theta(\psi \mid y)$ *does* depend on θ and the reasoning that led from (1) to (4) does not apply.

For an example using $\Psi = U$,

$$\begin{aligned} L(\theta, u = 1; y) &= P_\theta(y, U = 1) \\ &= P_\theta(y) P_\theta(U = 1 \mid y) \\ &= \begin{cases} P_\theta(y) = L(\theta; y) & \theta \in \text{CI}(y), \\ 0 & \theta \notin \text{CI}(y), \end{cases} \end{aligned} \tag{6}$$

and

$$\begin{aligned} L(\theta, u = 0; y) &= P_\theta(y, U = 0) \\ &= P_\theta(y) P_\theta(U = 0 \mid y) \\ &= \begin{cases} 0 & \theta \in \text{CI}(y), \\ P_\theta(y) = L(\theta; y) & \theta \notin \text{CI}(y). \end{cases} \end{aligned}$$

Equation (6) has the usual interpretation of a likelihood function, namely that it quantifies how well a given combination of (θ, u) describes the data y relative to other combinations of (θ, u). For example, if $Y \sim \text{N}(\theta, 1)$ and $\text{CI}(y) = (y - 1.96, y + 1.96)$, then when $y = 2$,

$$\begin{aligned} &L(\theta = 0, u = 1; y = 2) = 0, \\ &L(\theta = 1, u = 1; y = 2) \approx 0.24, \\ &L(\theta = 0, u = 0; y = 2) \approx 0.05, \quad \text{and} \\ &L(\theta = 1, u = 0; y = 2) = 0, \end{aligned}$$

which has the interpretation that $(\theta = 1, u = 1)$ describes the data about five times better than $(\theta = 0, u = 0)$ and infinitely better than either $(\theta = 0, u = 1)$ or $(\theta = 1, u = 0)$, both of which are impossible.

This example shows that the correct likelihood approach to interval estimation is to define $\text{CI}(y)$ as a high likelihood interval *per* Fisher (1956, pp. 72–73).

For an example using $\Psi = V$, let E_1 be the experiment in which Y_1 is the number of successes in two Bernoulli trials and $\theta \in \Theta \equiv [0, 1]$ is the parameter of the Bernoulli distribution. $Y_1 \sim \text{Bin}(2, \theta)$. Suppose the observation is $y_1 = 1$. Let E_2 be the experiment in which Y_2 is the total number of Bernoulli trials needed in order to get the first failure. $Y_2 \sim \text{NegBin}(1, \theta)$. Suppose the observation is $y_2 = 2$.

The likelihood functions from the two experiments are identical: $L_1(\theta) \propto L_2(\theta) \propto \theta(1-\theta)$. Also, PL's functions are identical: $L_1(v) = L_2(v) \propto 1$.

To get the confidence distributions from E_1 and E_2, we follow PL's Section 2.3. Let $T = t$ be an estimate of θ, say the m.l.e. The right-hand side P-value function is $C(t, \theta) \equiv P_\theta(T^* \geq t)$ where T^* is a random variable with the same distribution as T. In both E_1 and E_2, the m.l.e. is $t = 0.5$ so

$$C_1(0.5, \theta) = 1 - P_\theta(Y_1 = 0) = 1 - (1-\theta)^2 = 2\theta - \theta^2$$

and

$$\begin{aligned} C_2(0.5, \theta) &= 1 - P_\theta(\text{first trial is a failure}) \\ &= 1 - (1-\theta) = \theta. \end{aligned}$$

Then the confidence densities are

$$c_1(0.5, \theta) = dC_1(0.5, \theta)/d\theta = 2 - 2\theta$$

and

$$c_2(0.5, \theta) = dC_2(0.5, \theta)/d\theta = 1.$$

$C_1 \neq C_2$ and $c_1 \neq c_2$. Therefore, inferences based on confidence distributions are different in E_1 and E_2 even though the likelihood function $\theta(1-\theta)$ and PL's function $L(v) \propto 1$ are the same in E_1 and E_2. Confidence does not follow the likelihood principle and cannot be interpreted as likelihood.

4. SUMMARY AND DISCUSSION

PL (Section 2.1) "start by asking if there is a probabilistic way to state our sense of uncertainty in an observed CI" in the usual parametric framework in which $Y \sim P_\theta$ for some $\theta \in \Theta$. It is well known that confidence does not correspond to the usual likelihood function $L(\theta; y) \propto p_\theta(y)$ but PL introduce a new element in the form of the random Ψ and investigate whether confidence can correspond to the likelihood function for Ψ. They use the function in (4) which, they note, depends on neither θ nor y. But $L(\psi) = P(\Psi)$ is not a *likelihood* function.

B's Section 4.1 addresses the choice of likelihood function and explains why it must be (1) and not (4). In applying (1) to PL's problems, we find that the full model includes a pair of random variables, either (Y, U) or (Y, V), so the likelihood function is either

$$L(\theta, u; y) = P_\theta(y, u) = P_\theta(u) P_\theta(y \mid u)$$

or

$$L(\theta, v; y) = P_\theta(y, v) = P_\theta(v) P_\theta(y \mid v)$$

and, while $P_\theta(u)$ and $P_\theta(v)$ do not depend on θ, $P_\theta(y \mid u)$ and $P_\theta(y \mid v)$ do depend on θ and must be included.

A key feature of (1) is that it depends on the observed y but not on other values that Y might have taken. In particular, and in contrast to frequentist theory, a likelihood function cannot depend on T^*, a "random variable with the same distribution as T[1]" because that dependence entails averaging over other, nonobserved, possible values of Y.

We claim that (4) does not account for the full model, does not address the inferential aims of the statistical investigation and is not the (extended) likelihood function.

[1] See the discussion of the P-value function above and in PL, page 511.

REFERENCES

BJØRNSTAD, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791–806. MR1395746 https://doi.org/10.2307/2291674

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

PAWITAN, Y. and LEE, Y. (2021). Confidence as likelihood. *Statist. Sci.* **36** 509–517. MR4323049 https://doi.org/10.1214/20-sts811