Assessing Procedures vs. Assessing Evidence
and the Principles of Sufficiency, Conditionality, and Likelihood

Michael Lavine

A fundamental idea in statistics and data science is that statistical procedures are judged by criteria such as misclassification rates, p-values, or convergence that measure how the procedure performs when applied to many possible data sets. But such measures gloss over quantifying the evidence in a particular data set. We show that assessing a procedure and assessing evidence are distinct. The main distinction is that procedures are assessed unconditionally, i.e., by averaging over many data sets, while evidence must be assessed conditionally by considering only the data at hand.

We present four examples to illustrate the difference between assessing a procedure and assessing evidence. Then we examine a fifth example in detail to introduce the Sufficiency and Conditionality Principles and show why evidence must be assessed conditionally on the data at hand,

Michael Lavine is program manager for probability and statistics in the US Army Research Office. His email address is `michael.l.lavine3.civ@mail.mil`.

DOI: https://doi.org/10.1090/noti2114

not averaged over many possible data sets. Finally we state the Likelihood Principle, its relationship to the Sufficiency and Conditionality Principles, and some of its implications.

Example 1 (Forensic evidence: fingerprints). After a fingerprint has been found at a crime scene and deemed to be of sufficiently high quality, it is common to take a fingerprint from a suspect to see whether it matches the print from the scene. Whether the task is accomplished by a forensic examiner or by an algorithm, it results in a classification of either match or mismatch. Using prints from a database where it is known whether two prints come from the same finger, we can estimate

$$\alpha = \Pr[\text{Type A error}] \quad \text{and} \quad \beta = \Pr[\text{Type B error}],$$

where the two types of error are classifying a true match as a mismatch and classifying a true mismatch as a match. Here α and β describe the procedure: how often, when averaged over the universe of fingerprint pairs, it makes an error of Type A or Type B. But those are averages over some pairs that are hard to classify and others that are easy. A pair that is hard to classify provides only weak evidence for its conclusion, while an easy pair provides strong evidence. Thus, α and β do not quantify the evidence in the particular fingerprint pair at hand. To quantify the evidence we need to know how hard this particular pair is to classify.

Example 1 illustrates a major theme of this paper, namely, that different data sets, even of the same type (here, pairs of fingerprints), contain evidence of different strengths. So assessing a procedure, which requires averaging over many possible data sets, differs from assessing the evidence in a single data set.

Throughout this paper we will be dealing with a random variable X that is an observation from a probability distribution F_{true} which is unknown but is assumed to belong to a given set of distributions $\{F\}$. Usually $\{F\}$ is indexed by a parameter θ in a parameter space Θ and we write $F_{\text{true}} \in \{F_\theta\}_{\theta\in\Theta}$. For densities we write f_θ. Thus F_{true} corresponds to a value θ_{true}, cdf $F_{\theta_{\text{true}}}$, and pdf $f_{\theta_{\text{true}}}$. The value θ_{true} is unknown and we use X to learn about it.

A key concept is the likelihood function

$$\ell(\theta) \equiv c f_\theta(x), \tag{1}$$

where x is the observed value of the random variable X and $c > 0$ is an arbitrary constant. $\ell(\theta)$ measures how well each value of θ describes the observation $X = x$. In (1), $\ell(\theta)$ is a function of θ, while x is understood to be given. $\ell(\theta)$ matters only up to an arbitrary multiplicative constant $c > 0$. That is, $\ell_1(\theta)$ and $\ell_2(\theta) = c\ell_1(\theta)$ are equivalent likelihood functions. It is often convenient to set c so that $\max_\theta \ell(\theta) = 1$.

Example 2 (Likelihood ratio as evidence: two classes). Suppose data sets of size n, $X = (X_1, \dots, X_n)$, are generated as independent observations from one of two distributions: either F_1, the uniform distribution on $[0, 1]$, or F_2, the uniform distribution on $[0, 1 + \epsilon]$. In this example, Θ is the set $\{1, 2\}$. For a given sample $x = (x_1, \dots, x_n)$, the evidence that it was generated from F_1 rather than F_2 is the likelihood ratio[1]

$$\text{LR} = \frac{\ell(1)}{\ell(2)} = \prod_{i=1}^{n} \frac{f_1(x_i)}{f_2(x_i)} = \begin{cases} (1+\epsilon)^n & \text{if} \max(x_1, \dots, x_n) \le 1, \\ 0 & \text{if} \max(x_1, \dots, x_n) > 1. \end{cases}$$

That is, the evidence is either weak ($\text{LR} \approx 1$) in favor of F_1 or conclusive ($\text{LR} = 0$) in favor of F_2, depending on the value of $\max x_i$. Thus, different data sets of size n have different strengths of evidence. Error probabilities and misclassification rates are averages over all possible data sets and do not quantify the evidence in any individual data set. Curiously, if $\epsilon \ll n^{-1}$, then $\Pr_2[\max x_i \le 1]$ is large and most data sets from F_2 favor F_1, albeit weakly.

Example 3 (Confidence intervals: location family). Example 3 examines samples of size $n = 3$ from the family

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{(x - \theta)^2 + 1},$$

the Cauchy density with unknown location parameter θ. In this example, $\Theta = \mathbb{R}$. Figure 1 shows four samples of size 3 from f_0. We pretend we don't know the true value of θ and are trying to learn about it. Each sample is plotted with its likelihood function and a 95% confidence interval[2] for θ. The confidence interval is $\text{median}(x_1, x_2, x_3) \pm 3.28$ because

$$\Pr_\theta\{[\text{median}(x_1, x_2, x_3) - 3.28, \\ \text{median}(x_1, x_2, x_3) + 3.28] \ni \theta\} \approx 0.95.$$

The confidence procedure yields four intervals that all have length 6.56 and confidence coefficient .95 even though some likelihood functions are sharply peaked and others are broader. The intervals' length, $2 \times 3.28 = 6.56$, is determined by taking a mean over all possible samples of size 3 w.r.t. the Cauchy distribution and is therefore the same for every sample. Yet the samples with sharp likelihood functions have strong evidence for θ, while the samples with broad likelihood functions have weak evidence.

Example 4 (Statistical consulting). A scientist is investigating a phenomenon that generates a random number X having a Poisson distribution with mean λ. The value of λ is unknown. In this example, $\Theta = \mathbb{R}^+$. The scientist

[1] *The Law of Likelihood asserts that evidence is measured by the likelihood ratio. See external sources such as* [Roy97] *for an explanation.*

[2] *A 95% confidence interval $CI(X)$ is an interval constructed according to a procedure having the property that* $\Pr_\theta[CI(X) \ni \theta \ge .95]$ *for all θ. Consult standard statistics texts for a further explanation of confidence intervals and their role in statistics.*

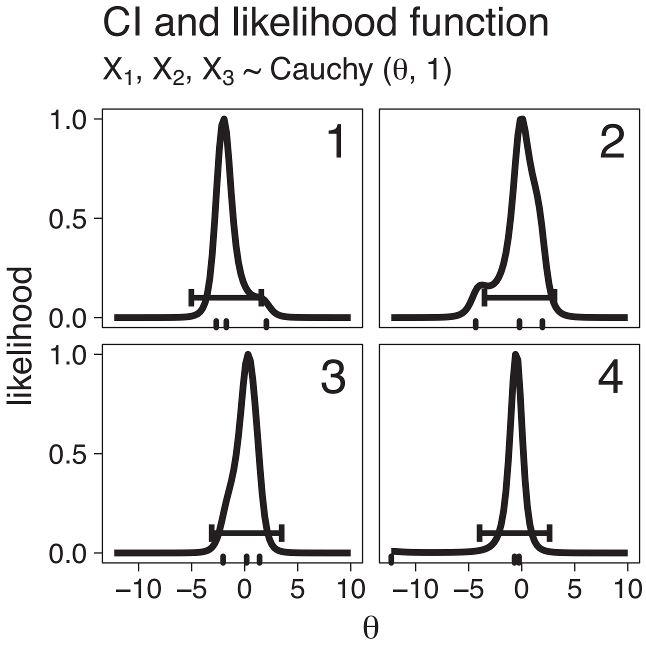

Figure 1. Four samples of size $n = 3$ from the Cauchy distribution with $\theta = 0$. Each panel shows a rug plot (tick marks on the horizontal axis) for the sample, a 95% confidence interval (horizontal bar at $y = 0.1$), and the likelihood function (solid curve) for θ. Sample 4 has a sharp likelihood function and hence contains strong information for θ. In contrast, the other samples have broader likelihood functions and weaker information for θ. The likelihood functions are scaled so each has a maximum value of 1.

brings a single observation $X = 1$ (experiments are expensive) to a statistician and asks what can be inferred. The statistician uses a standard procedure to give the scientist the 95% confidence interval shown in Figure 2. The scientist notices there are two values of λ, λ_1 and λ_2, such that

(a) λ_2 is in the confidence interval, but λ_1 is not and
(b) λ_1 describes the datum $x = 1$ better than λ_2 (because $\ell(\lambda_1) > \ell(\lambda_2)$)

and asks the statistician, *"Why do you call my attention to λ_2 but not λ_1?"*

Implicit in the scientist's question is the scientist's interest in knowing which values of λ describe the datum $x = 1$ well. The statistician used a standard procedure for constructing a confidence interval, but that doesn't address the scientist's interest, because the confidence interval and confidence coefficient of 0.95 come from averaging over all possible data sets while the scientist wants to draw an inference from this particular data set and because the confidence interval excludes some values of λ that describe the data well.

Examples 1–3 illustrate that different data sets even of the same size and type can have evidence of different strengths. Example 4 raises the question of what a scientist wants from the analysis of a data set: average performance or evidential strength.

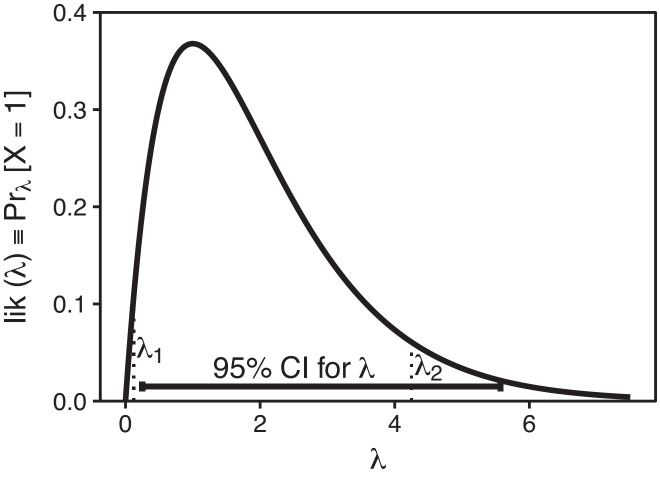

Figure 2. $X \sim \mathrm{Poi}_\lambda$. The datum is $x = 1$. The dark bar just above the horizontal axis is a 95% confidence interval for λ. The curve is the likelihood function for λ.

Now we take up one more example to introduce principles that measurements of evidence must follow. Suppose there is a repeatable experiment that results in a random X that is either success ($X = 1$) or failure ($X = 0$). Let X_i be the outcome of the ith repetition and assume the X_i's are mutually independent, all with the same probability of success θ_{true}. The X_i's are called Bernoulli trials. $\Theta = [0, 1]$. Two possible experiments to learn about θ_{true} are:

E1. Under **E1** we conduct two trials and record x_1 and x_2.

E2. Under **E2** we conduct as many trials as needed until the first success. If k is the eventual number of trials, then $x_1 = \cdots = x_{k-1} = 0$ and $x_k = 1$.

A third possible experiment is:

E3. Under **E3** we toss a coin. If the coin lands Heads we conduct **E1**; if the coin lands Tails we conduct **E2**.

Four possible scenarios are:

1. Conduct **E1** and observe the outcome $x_1 = 0$; $x_2 = 1$;
2. Conduct **E2** and observe the outcome $x_1 = 0$; $x_2 = 1$;
3. Conduct **E3** and observe the outcome Heads followed by $x_1 = 0$; $x_2 = 1$; and
4. Conduct **E3** and observe the outcome Tails followed by $x_1 = 0$; $x_2 = 1$.

Let Ev stand for "evidence" and, without yet defining what it means, consider whether the four scenarios have the same evidence regarding θ. That is, do we think the following equalities should hold:

$$\begin{aligned} \mathrm{Ev}(\mathbf{E1}, (0,1)) &\overset{?}{=} \mathrm{Ev}(\mathbf{E3}, (H, 0, 1)) \\ &\overset{?}{=} \mathrm{Ev}(\mathbf{E3}, (T, 0, 1)) \overset{?}{=} \mathrm{Ev}(\mathbf{E2}, (0,1))? \end{aligned} \quad (2)$$

The symbol $\overset{?}{=}$ in (2) means we are considering whether we should require equality to hold for any quantity we're willing to call "evidence." We now introduce two principles that argue for the equalities to hold.

Conditionality Principle (CP). CP describes the evidence in a situation where there are two possible experiments, E1 and E2, and we randomly choose which to perform. If we randomly choose E1 and observe x_1, the evidence is just the same as if we had always intended to perform E1 and had observed x_1. CP was first formally stated in [Bir62]. More recently Berger and Wolpert [BW88] state CP as:

> Suppose there are two experiments $E_1 = (X_1, \theta, \{f_\theta^1\})$ and $E_2 = (X_2, \theta, \{f_\theta^2\})$. Consider the mixed experiment E^*, whereby $J = 1$ or 2 is observed, …, and experiment E_j is then performed. …Then $\text{Ev}(E^*, (j, x_j)) = \text{Ev}(E_j, x_j)$.

Berger and Wolpert's E_1, E_2, and E^* are our **E1**, **E2**, and **E3**. Their X_1 and X_2 are our observations under **E1** and **E2**, respectively. Their J is our coin toss.

CP means we should accept as reasonable only those methods of measuring evidence that yield $\text{Ev}(E^*, (j, x_j)) = \text{Ev}(E_j, x_j)$. Many people agree that CP is a desirable property of any quantity we call "evidence." Readers should consider for themselves whether they agree.

According to CP, $\text{Ev}(\mathbf{E3}, (H, 0, 1)) = \text{Ev}(\mathbf{E1}, (0, 1))$ and $\text{Ev}(\mathbf{E3}, (T, 0, 1)) = \text{Ev}(\mathbf{E2}, (0, 1))$, so (2) becomes

$$\begin{aligned} \text{Ev}(\mathbf{E1}, (0,1)) &= \text{Ev}(\mathbf{E3}, (H, 0, 1)) \\ &\overset{?}{=} \text{Ev}(\mathbf{E3}, (T, 0, 1)) = \text{Ev}(\mathbf{E2}, (0, 1)). \end{aligned} \tag{3}$$

Sufficiency Principle (SP). A *statistic* is a function of a random variable, say $T(X)$. A *sufficient* statistic is one for which the conditional distribution $F_\theta(X \mid T(X))$ does not depend on θ. That is, $F_\theta(X \mid T(X))$ is the same for all $\theta \in \Theta$.

For an example of sufficiency, let $X = (X_1, \dots, X_n)$ be n i.i.d. observations from the $\text{N}(\theta, 1)$ distribution where θ is unknown. Then $\bar{X} \equiv \frac{1}{n} \sum X_i$ is a sufficient statistic for θ. For another example, let $X = (X_1, \dots, X_n)$ be n independent Bernoulli trials with the same unknown parameter θ. Then $\sum X_i$ is a sufficient statistic for θ.

When a sufficient statistic exists X can be decomposed as $X \equiv (T(X), A(X))$, where A represents all aspects of X other than $T(X)$. That is, X can be reconstructed from $T(X)$ and $A(X)$. In our example of a sufficient statistic for $\text{N}(\theta, 1)$, $A(X)$ is $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ or its equivalent. In our example of Bernoulli trials, $A(X)$ specifies which X_i's are 1's and which are 0's.[3]

Sufficiency is important in statistics because $T(X)$ carries all the information in X for θ. Because $F_\theta(X \mid T(X))$ does not depend on θ, $F_\theta(A(X) \mid T(X))$ also does not depend on θ, so $A(X)$ carries no additional information for θ. It is sufficient to base inference on just $T(X)$; other aspects of X may be ignored.

SP says that for an experiment E with observation X and sufficient statistic $T(X)$, if two outcomes x_1 and x_2 satisfy $T(x_1) = T(x_2)$, then $\text{Ev}(E, x_1) = \text{Ev}(E, x_2)$.

In **E3**, the sequence of Bernoulli trials is sufficient (not proven here), so $\text{Ev}(\mathbf{E3}, (H, 0, 1)) = \text{Ev}(\mathbf{E3}, (T, 0, 1))$ and (3) becomes

$$\begin{aligned} \text{Ev}(\mathbf{E1}, (0,1)) &= \text{Ev}(\mathbf{E3}, (H, 0, 1)) \\ &= \text{Ev}(\mathbf{E3}, (T, 0, 1)) = \text{Ev}(\mathbf{E2}, (0, 1)). \end{aligned} \tag{4}$$

That is, we accept as reasonable only those methods of quantifying evidence that imply (4).

A common statistical analysis is to partition Θ into $\Theta_1 \cup \Theta_2$ and test the hypothesis $H_1 : \theta_{\text{true}} \in \Theta_1$ versus $H_2 : \theta_{\text{true}} \in \Theta_2$. Once $X = x_{\text{obs}}$ has been observed we partition the possible outcomes of X into X_1, the set of x's that support Θ_1 more than the observed x_{obs}, and X_2, the set of x's that support Θ_2 at least as much as x_{obs}. Then the p-value[4] is

$$p = \sup_{\theta \in \Theta_1} \Pr{}_\theta[X_2].$$

Table 1 shows p-values under **E1**, **E2**, and **E3** for testing $H_1 : \theta_{\text{true}} \in \Theta_1 \equiv [.95, 1]$ vs. $H_2 : \theta_{\text{true}} \in \Theta_2 \equiv [0, .95)$. $p_1 \neq p_2 \neq p_3$, so (4) does not hold and p-values cannot be said to measure the evidence for H_1 vs. H_2.

Likelihood Principle (LP). An informal statement[5] of LP is

> All the information about θ …is contained in the likelihood function. [BW88]

LP is related to CP and SP by a theorem due to [Bir62],[6] $(SP, CP) \Leftrightarrow LP$, which says that CP and SP together imply and are implied by LP. The theorem is important because CP and SP each seem intuitively reasonable yet lead to LP, which not only seems unreasonable to many statisticians but says common statistical procedures like p-values and hypothesis tests do not measure evidence. LP says all the statistical evidence in a given data set is contained in the likelihood function $\ell(\theta) \propto f_\theta(x)$. $\ell(\theta)$ depends only on the observed x, which brings us back to a point made in the examples, viz., that evidence must be computed using the observed x and not by averaging over unobserved values of X. In contrast, performance measures such as p-values

[3] *See standard textbooks on mathematical statistics for further discussion of sufficiency. The factorization theorem and Basu's theorem are key concepts. Textbooks also contain examples of statistics that are sufficient for one parameter space Θ_1 but not for another Θ_2.*

[4] *Consult introductory statistics texts for more on p-values and their role in statistics.*

[5] *See* [BW88] *for formalities and the role of LP in statistics.*

[6] *Birnbaum's proof is essentially the argument leading from* (2) *to* (4) *for general experiments, not just for **E1**, **E2**, and **E3**.*

$$\begin{array}{c}\text{E1: } X \text{ is one of}\\ \left\{\begin{array}{ll} (\mathbf{0},\mathbf{0}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.0025}) \\ (\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.0475}) \\ (\mathbf{1},\mathbf{0}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.0475}) \\ (1,1) & \text{not relevant} \end{array}\right\}\\ p_1 = .0975\end{array}$$

$$\begin{array}{c}\text{E2: } X \text{ is one of}\\ \left\{\begin{array}{ll} (1) & \text{not relevant} \\ (\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.0475}) \\ (\mathbf{0},\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.002375}) \\ \vdots & (\mathbf{f}_{\mathbf{.95}} = \mathbf{...}) \end{array}\right\}\\ p_2 = .05\end{array}$$

$$\begin{array}{c}\text{E3: } X \text{ is one of}\\ \left\{\begin{array}{ll} (\mathbf{H},\mathbf{0},\mathbf{0}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.00125}) \\ (\mathbf{H},\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.02375}) \\ (\mathbf{H},\mathbf{1},\mathbf{0}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.02375}) \\ (H,1,1) & \text{not relevant} \\ (T,1) & \text{not relevant} \\ (\mathbf{T},\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.02375}) \\ (\mathbf{T},\mathbf{0},\mathbf{0},\mathbf{1}) & (\mathbf{f}_{\mathbf{.95}} = \mathbf{.0011875}) \\ \vdots & (\mathbf{f}_{\mathbf{.95}} = \mathbf{...}) \end{array}\right\}\\ p_3 = .07375\end{array}$$

Table 1. E1, **E2**, and **E3** p-values for testing $H_1 : \theta_{\text{true}} \geq .95$ vs. $H_2 : \theta_{\text{true}} < .95$ when the Bernoulli sequence $(0, 1)$ is observed. Each entry shows a possible outcome of the experiment and the supremum of its probability for $\theta \geq .95$. The p-value is the sum of the probabilities in bold. Note that $p_1 \neq p_2 \neq p_3$ even though the evidence is the same in each experiment.

and error rates average over unobserved values of X. For p-values specifically,

$$p = \sup_{\theta\in\Theta_1} \Pr{}_\theta[X_2] = \sup_{\theta\in\Theta_1} \mathbb{E}_\theta[\mathbf{1}_{X_2}(X)],$$

which is the supremum of an expectation over all possible values of X.

Some statistical ideas that follow LP are the maximum likelihood estimator, a likelihood interval, a likelihood ratio, and subjective Bayesian analysis. Some statistical ideas that don't follow LP are p-values, confidence intervals, misclassification rates, mean squared error, and bias.

Statisticians and data scientists often assert that we want procedures that work well and that we should quantify how well our procedures work. Granting that assertion, we should also ask, "procedures that work well in accomplishing what task?" This paper shows that popular procedures for testing hypotheses, finding confidence intervals, and making classifications are not accomplishing the task of assessing evidence; they are accomplishing something else. Quantifying how well those procedures work is different from quantifying evidence. It behooves us to understand what task we want to accomplish when analyzing any given data set.

Methods for quantifying non-LP procedures are common in statistics references. Methods for quantifying evidence are not, but can be found in the following references, among others, along with a deeper analysis of likelihood thought: [Blu02], [BSLM07], [GR88], [HB08], [Mel99], [Mel00], [MR97], [Paw01], [Roy86], [RT03], [Sev01], [Str18], and [TR95].

References

[Bir62] Allan Birnbaum, *On the foundations of statistical inference*, J. Amer. Statist. Assoc. **57** (1962), 269–326. MR138176

[Blu02] J. D. Blume, *Likelihood methods for measuring statistical evidence*, Statistics in Medicine **21** (2002), 2563–2569.

[BSLM07] Jeffrey D. Blume, Li Su, Remigio M. Olveda, and Stephen T. McGarvey, *Statistical evidence for GLM regression parameters: a robust likelihood approach*, Stat. Med. **26** (2007), no. 15, 2919–2936, DOI 10.1002/sim.2759. MR2370979

[BW88] James O. Berger and Robert L. Wolpert, *The likelihood principle*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, vol. 6, Institute of Mathematical Statistics, Hayward, CA, 1984. MR773665

[GR88] S. N. Goodman and R. Royall, *Evidence and scientific research*, American Journal of Public Health **78** (1988), 1568–1574.

[HB08] J. Hoch and J. D. Blume, *Measuring and illustrating statistical evidence in a cost-effectiveness analysis*, Journal of Health Economics **27** (2008), 476–495.

[Mel00] B. G. Mellen, *A likelihood approach to DNA evidence*, Statistical science in the courtroom. Statistics for social science and public policy, 2000.

[Mel99] Beverly G. Mellen, *Modeling epidemiologic typing data and likelihood inference for disease spread*, Journal of the American Statistical Association **94** (1999), 1015–1024.

[MR97] B. G. Mellen and R. M. Royall, *Measuring the strength of deoxyribonucleic acid evidence, and probabilities of strong implicating evidence*, Journal of the Royal Statistical Society (Series A) **160** (1997).

[Paw01] Yudi Pawitan, *In all likelihood: Statistical modelling and inference using likelihood*, Oxford University Press, 2001.

[Roy86] R. M. Royall, *The effect of sample size on the meaning of significance tests*, The American Statistician **40** (1986), 313–315.

[Roy97] Richard M. Royall, *Statistical evidence: A likelihood paradigm*, Monographs on Statistics and Applied Probability, vol. 71, Chapman & Hall, London, 1997. MR1629481

[RT03] Richard Royall and Tsung-Shan Tsou, *Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions*, J. R. Stat. Soc. Ser. B Stat. Methodol. **65** (2003), no. 2, 391–404, DOI 10.1111/1467-9868.00392. MR1983754

[Sev01] Thomas A. Severini, *Likelihood methods in statistics*, Oxford Statistical Science Series, vol. 22, Oxford University Press, Oxford, 2000. MR1854870

[Str18] L. J. Strug, *The evidence statistical paradigm in genetics*, Genetic Epidemiology **42** (2018), 590–607.

[TR95] Tsung-Shan Tsou and Richard M. Royall, *Robust likelihoods*, J. Amer. Statist. Assoc. **90** (1995), no. 429, 316–320. MR1325138

Michael Lavine

Credits

Opening graphic is courtesy of no_limit_pictures via Getty.
Figures 1–2 are courtesy of Michael Lavine.
Photo of Michael Lavine is courtesy of Louise Lavine.