# The Measurement of Statistical Evidence
# Lecture 3 - part 1

Michael Evans
University of Toronto
http://www.utstat.utoronto.ca/mikevans/sta4522/STA4522.html

2021

**Example** - *"random" sequence in* $\{0,1\}^n$

- $(\{0,1\}^n)^*$ comprises $1/2$ of the elements of $\{0,1\}^n$ and, from last week's class, we know that the vast majority of elements of $(\{0,1\}^n)^*$ can be considered as random according to Kolmogorov's definition, at least for large $n$

- by symmetry the same argument applies to those elements of $\{0,1\}^n$ that end in 0 and so we conclude that, for large n, most elements of $\{0,1\}^n$ are indeed random in the sense of Kolmogorov and let's denote the subset of such sequences by $R_n \subset \{0,1\}^n$

- suppose we produce such a sequence $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ (say via coin tossing) that is supposed to be $i.i.d.$ Bernoulli$(1/2)$

- we can test this, but the usual statistical tests are just asking whether or not the sequence is $i.i.d.$ Bernoulli$(1/2)$ and, as discussed last week, this is not a test that the sequence is a random sequence (there is no such test and consider the Champernowne sequence which can be turned into a sequence of 0's and 1's in the obvious way and this sequence will pass the test for $i.i.d.$ Bernoulli$(1/2)$ for $n$ large enough)

- so why in the end do we consider that $x$ is a random sequence?

- note that, if we accept the $i.i.d.$ Bernoulli$(1/2)$ model, then each sequence in $\{0,1\}^n$ has probability (as belief) $1/2^n$ (even the sequences of all 1's, all 0's, alternating 0's and 1's etc.) and therefore, for large $n, P(R_n) \approx 1$

- in other words, in such a context, it is our **belief** (because that is what probability measures) that $x$ will be a random sequence

- again probability enters the picture, not as being intrinsically associated with randomness, but rather in expressing our belief that the sequence is random

- how large does $n$ have to be for this argument to make sense? I have no idea and my guess is that it can't be known ($K$ is not computable)

- so in the end the "belief" about randomness is an unquantifiable, untestable assertion about reality and the objectivity of the observed data

# Characterizing Statistical Evidence

- Chapter 3 discusses the various approaches that have been taken to characterizing the concept of *statistical evidence*

- recall the number of times you read "the evidence in the data says ...", "based on the evidence in the data ...", etc.

- but how is "the evidence in the data" defined or characterized?

> *The current state of affairs is ambiguous and there are different approaches to answering this question.*

- although we recognize that models and priors may not be "correct", all of our discussions in this part of the course will proceed as if they are correct (checking the model and prior comes later)

## 1. Pure Likelihood

- based only on the model $\{f_\theta : \theta \in \Theta\}$ and the data $x$, sometimes put together as a 2-tuple $I = (\{f_\theta : \theta \in \Theta\}, x)$ and called an *inference base*

- for inference base $I$ the *likelihood function* is defined as $L(\cdot \,|\, x) : \Theta \to [0, \infty)$ by $L(\theta \,|\, x) = cf_\theta(x)$ for **any** constant $c > 0$

- so if an inference depends in some way on $c$, it is **not** a likelihood inference

- the "likelihood function" is really an equivalence class of functions such that the ratio of any two versions is a constant

- a nice book (for a number of things) Royall, R. (1997) Statistical Evidence: A likelihood paradigm and another good book Edwards A.W.F. (1972) Likelihood

- why the constant $c$?

- to start, let's suppose interest is in answering **E** or **H** for $\theta$

- consider the discrete case so $f_\theta(x) =$ probability of observing $x$ when $\theta$ is true

- when $f_{\theta_1}(x) > f_{\theta_2}(x)$ (iff $L(\theta_1 \,|\, x) > L(\theta_2 \,|\, x)$) there is evidence in favor of $\theta_1$ over $\theta_2$ and the strength of this evidence is given by (*Law of Likelihood*) the *likelihood ratio* $f_{\theta_1}(x)/f_{\theta_2}(x) = L(\theta_1 \,|\, x)/L(\theta_2 \,|\, x)$

- so there is a preference ordering on $\Theta$ given by: $\theta_1$ *is preferred to* $\theta_2$ whenever $L(\theta_1 \,|\, x) > L(\theta_2 \,|\, x)$ and there is a measure of the strength of the evidence for one value over another given by $L(\theta_1 \,|\, x)/L(\theta_2 \,|\, x)$ and all of this is independent of $c$

- so there is no need, whether in the discrete or continuous case, to give an interpretation to the value $L(\theta \,|\, x)$ (e.g., it is not the evidence that $\theta$ is the true value)

- to answer **E** we *must* (according to the preference ordering) take $\theta(x) = \sup_\theta L(\theta \,|\, x) =$ the *maximum likelihood estimate* (MLE) of $\theta$

- the accuracy of $\theta(x)$ is then assessed by recording a set

$$C(x) = \{\theta : L(\theta \,|\, x) \geq k(x)\}$$

so $\theta(x) \in C(x)$ and looking at the "size" of $C(x)$

- but how do we choose $k(x)$ $(\leq L(\theta(x)\,|\,x))$?

- the *relative likelihood function* is given by

$$L_{rel}(\theta\,|\,x) = \frac{L(\theta\,|\,x)}{L(\theta(x)\,|\,x)}$$

(a likelihood ratio) and note that

$$C(x) = \{\theta : L(\theta\,|\,x) \geq k(x)\} = \left\{\theta : L_{rel}(\theta\,|\,x)) \geq \frac{k(x)}{L(\theta(x)\,|\,x)}\right\}$$

- the L of L implies that the interpretation of likelihood ratios is independent of any particular inference base

- so $k(x)/L(\theta(x)\,|\,x)$ is taken to be a constant say $1-\gamma$ for some $\gamma \in [0,1]$ and a $(1-\gamma)$-likelihood region is

$$C_\gamma(x) = \{\theta : L_{rel}(\theta\,|\,x)) \geq 1-\gamma\}$$

which is comprised of those values which have at least $100(1-\gamma)\%$ of the maximum support

- how do we choose $\gamma$?

- this is not at all clear as $\gamma$ is **not** a probability

- Royall argues, based on an urn model (a specific model and so violating L of L), that $\gamma = 7/8$ gives all those values for which there is "fairly strong evidence" that they correspond to the true value

- for **H**, where $H_0 : \theta = \theta_0$, record $L_{rel}(\theta_0 \,|\, x)$ and say there is strong evidence in favor of $H_0$ whenever $L_{rel}(\theta_0 \,|\, x) \geq 1/8$

**Example** *Is the Law of Likelihood reasonable?*

- Evans (1989) constructs an set of examples, indexed by $(k, \delta) \in [0, \infty)^2$, where $L(\theta \,|\, x)$ is positive on only two values of $\Theta$, say $\theta(x)$ and $\theta'(x)$ (the value not equal to the MLE having positive likelihood), where $C_\gamma(x) = \{\theta(x)\}$ for every $\gamma < 1$ (so MLE is very "accurate"), $L_{rel}(\theta'(x) \,|\, x) \to 0$ and $P_\theta(\theta'(x) = \theta) \to 1 - \delta$ for every $\theta$ as $k \to \infty$

- choose $(k, \delta)$ so we are virtually certain that $\theta'(x)$ is the true value and $\theta(x)$ is supported over $\theta'(x)$ say with likelihood ratio $10^{10}$

- this (and other such examples) suggest a problem with the L of L

- suppose interest is in $\psi = \Psi(\theta) \in \Psi$

- how do we get rid of the nuisance parameters to make likelihood inferences about $\Psi$?

- there doesn't seem to be a general way to construct a function $L^{\Psi}(\cdot \,|\, x) : \Psi \rightarrow [0, \infty)$ so that $L^{\Psi}(\psi \,|\, x)$ is a likelihood, namely, proportional to the probability of observing something

- the general approach is to use the *profile likelihood* defined as

$$L^{\Psi}(\psi \,|\, x) = \sup_{\theta \in \Psi^{-1}\{\psi\}} L(\theta \,|\, x)$$

and use $L^{\Psi}(\cdot \,|\, x)$ as a likelihood function to derive inferences for $\psi$

- under weak conditions, $C_{\Psi,\gamma}(x) = \Psi C_{\gamma}(x)$ ($L(\cdot \,|\, x)$ is continuous and $\Psi^{-1}\{\psi\}$ compact for each $\psi$)

**Lemma** If $\theta_\psi(x) = \arg\sup_{\theta \in \Psi^{-1}\{\psi\}} L(\theta \,|\, x)$ exists for each $\psi \in \Psi$, then $C_{\Psi,\gamma}(x) = \Psi C_\gamma(x)$.
Proof: Suppose $\theta \in C_\gamma(x)$ so $L(\theta \,|\, x) \geq (1-\gamma)L(\theta_{MLE}(x))$ which implies

$$L^\Psi(\Psi(\theta) \,|\, x) = \sup_{\theta' \in \Psi^{-1}\{\Psi(\theta)\}} L(\theta' \,|\, x) \geq (1-\gamma)L^\Psi(\psi_{MLE}(x) \,|\, x).$$

Therefore, $\Psi(\theta) \in C_{\Psi,\gamma}(x)$ and so $\Psi C_\gamma(x) \subset C_{\Psi,\gamma}(x)$. Now suppose $\psi \in C_{\Psi,\gamma}(x)$ so $L^\Psi(\psi \,|\, x) \geq (1-\gamma)L^\Psi(\psi_{MLE}(x) \,|\, x)$. This implies

$$L(\theta_\psi(x) \,|\, x) \geq (1-\gamma)L(\theta_{MLE}(x) \,|\, x)$$

and so $\theta_\psi(x) \in C_\gamma(x)$. This in turn implies that $\psi = \Psi(\theta_\psi(x)) \in \Psi C_\gamma(x)$.

**Corollary** The **profile** likelihood MLE $\psi(x)$ satisfies $\psi(x) = \Psi(\theta(x))$.
Proof: Take $\gamma = 0$.

**Corollary** Likelihood inferences are invariant, namely, inferences for reparameterizations ($\Psi$ is a bijection) transform appropriately.

**Example** *A profile likelihood is not generally a likelihood.*

- suppose model is given by

| $\theta$ | $f_\theta(1)$ | $f_\theta(2)$ |
|---|---|---|
| 0 | 1/2 | 1/2 |
| 1 | 1/3 | 2/3 |
| 2 | 1/5 | 4/5 |

$\Psi = I_A$ where $A = \{0, 1\}$ so $\Psi(\theta) = 0$ if $\theta = 2$ and $\Psi(\theta) = 1$ otherwise

$$\begin{aligned} L^{\Psi}(0\,|\,1) &= 1/5 \qquad L^{\Psi}(1\,|\,1) = 1/2 \\ L^{\Psi}(0\,|\,2) &= 4/5 \qquad L^{\Psi}(1\,|\,2) = 2/3 \end{aligned}$$

- for $L^{\Psi}(\cdot\,|\,x)$ to be a likelihood there has to be $T : \{1, 2\} \to \{1, 2\}$ such that the likelihoods based on the model induced by $T$ are proportional to $L^{\Psi}(\cdot\,|\,1)$ and $L^{\Psi}(\cdot\,|\,2)$

- since $L^{\Psi}(\cdot\,|\,1)$ and $L^{\Psi}(\cdot\,|\,2)$ are not proportional, $T$ must be 1-1 to produce two distinct likelihood functions

- when $T$ is 1-1 its model is given by same table and $L(\cdot\,|\,x) \neq L^{\Psi}(\cdot\,|\,x)$

- so the idea of profiling is not consistent with the basic idea of likelihood

**Example** *Prediction problems.*

- want to predict a value $y \in \mathcal{Y}$ having observed $x$ ($y$ could be a future value)

- suppose $y$ has model $\{g_\delta(\cdot \,|\, x) : \delta \in \Delta\}$ with $\delta = \Delta(\theta)$ and $\delta_{true} = \Delta(\theta_{true})$, then joint likelihood for $(\theta, y)$ is

$$L(\theta, y \,|\, x) = c g_{\Delta(\theta)}(y \,|\, x) f_\theta(x)$$

- interest is in $y$, so profile out $\theta$ to form a *predictive likelihood* for $y$

$$L^{\mathcal{Y}}(y \,|\, x) = \sup_\theta L(\theta, y \,|\, x)$$

- then $y(x)$ is the predictor of $y$ and profile likelihood regions for this quantity can be formed to assess its accuracy

- given $L(\theta, y \,|\, x)$ the profile likelihood for $\theta$ is

$$L^{\Theta}(\theta \,|\, x) = \sup_y L(\theta, y \,|\, x) \neq L(\theta \,|\, x)$$

and different inferences will be obtained

## Summary of the pure likelihood approach

*Positives*

(1) as we will see, the likelihood for the full model parameter does properly order the values $\theta \in \Theta$ with respect to the evidence

(2) inferences are invariant under reparameterizations

(3) it provides a step in the "right" direction for a full theory by concentrating on statistical evidence

(4) in many examples it seems to give reasonable inferences

*Negatives*

(1) it does not provide a characterization of statistical evidence from which a complete theory can be built, e.g., when does the data $x$ contain evidence that $\theta_0$ is the true value?

(2) in general it does not handle inferences about marginal parameters appropriately (profiling)

(3) there are problems with the L of L (there is no universal scale) as a measure of the strength of the evidence and there doesn't seem to be a way to use the "size" of likelihood regions to measure the accuracy of $\theta(x)$ that isn't parameterization dependent