The Likelihood Principle

Surya Tokdar

The Likelihood Principle

- The Likelihood principle (LP) asserts that for inference on an unknown quantity θ, all of the evidence from any observation $X = x^*$ with distribution $X \sim p(x|\theta)$ lies in the likelihood function

$$L_{x^*}(\theta) \propto p(x^*|\theta), \quad \theta \in \Theta.$$

Understanding LP

- The interpretation of LP hinges on the rather subtle point of allowing any observable X to draw conclusions about θ.
- If there were two ways to gather information about θ, either with $X \sim p(x|\theta)$ or with $Y \sim \tilde{p}(y|\theta)$, and it happened that for the observations $X = x^*$ and $Y = y^*$ we had

$$L_{x^*}(\theta) = \text{const.} \times \tilde{L}_{y^*}(\theta), \quad \forall\theta \in \Theta$$

then our conclusions about θ should not depend on which observable we used.

An example

- Two researchers, Jerzy and Egon, each wants to determine whether more than half the students in a university support a recent government bill.
- Let $\theta \in (0, 1)$ be the proportion of students who support the bill.
- Jerzy decides to survey 18 students on this issue and finds 12 supporters.
- Egon decides to survey until he sees 12 supporters, and ends up surveying a total of 18 students.
- Jerzy's observable is an $X \sim Bin(18, \theta)$ and he observed $X = 12$.
- Egon's observable is an $Y \sim NBin(12, \theta)$, he observed $Y = 18$.

An example (contd.)

- Jerzy's likelihood function is:

$$L_{12}(\theta) = \binom{18}{12} \times \theta^{12}(1-\theta)^6$$

- Egon's likelihood function is:

$$\tilde{L}_{18}(\theta) = \binom{17}{11} \times \theta^{12}(1-\theta)^6$$

- So we indeed have $L_{12}(\theta) = \text{const} \times \tilde{L}_{18}(\theta)$, $\forall\theta$.

An example (contd.)

- Both the binomial and the negative binomial family are MLR, respectively, in X and Y.
- So for testing $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$ Jerzy's UMP test would reject if $X > c$ for some c. He'd report a p-value $P_{[\theta=0.5]}(X > 12) = 0.048$.
- Egon's UMP tests are given by reject if $Y < c$. So his reported p-value if $P_{[\theta=0.5]}(Y < 18) = 0.071$.

An example (contd.)

- LP is violated here is due to the fact that p-value is the probability under H_0 of observing evidence against H_0 that is more extreme than the one in the recorded data.
- Such calculations clearly care about other possible data than what has been currently observed.
- This is common to all classical methods and it is well documented the concern about data that have not been observed can lead to absurd inference based on data that has indeed been observed!

Example 2

- Suppose X_1 and X_2 are independent with $P(X_j = \theta \pm 1) = 1/2$ for some unknown $\theta \in \mathbb{R}$.
- The smallest 75% confidence interval for θ is

$$C(X_1, X_2) = \begin{cases} \text{the point } \frac{X_1+X_2}{2} & \text{if } X_1 \neq X_2 \\ \text{the point } X_1 - 1 & \text{if } X_1 \neq X_2 \end{cases},$$

so, $P_\theta(\theta \in C(X_1, X_2)) = 0.75$ for all θ.
- But once we observe X_1 and X_2, it is silly to report a 75% confidence. Instead we should report a confidence of
  1. 100% if $X_1 \neq X_2$.
  2. ≈50% if $X_1 = X_2$.
- The problem here lies in not conditioning the inference on the observed data – again a violation of LP.

Example 3 (Cox paradox)

- A laboratory has two instruments for performing the same task, one has accuracy ± 0.01 while the other has accuracy ± 0.05.
- What accuracy should a scientist who gets to use whichever instrument is available (w.p. 1/2)? The one that she used or the average accuracy?

Birnbaum's theorem

- Birnbaum (1962) proved that LP is equivalent to the following two principles
  (CP) Conditionality principle. Suppose there are two experiments E_1 and E_2 where the only unknown is the parameter θ, common to the two problems. Consider the mixed experiment E_* in which we select $i = 1$ or $i = 2$ with equal probabilities, then perform experiment E_i; then the resulting evidence about θ is that from experiment E_i, and we can ignore the existence of the other (unperformed) experiment.
  (SP) Sufficiency principle. Consider an experiment E and a sufficient statistic T. Then if $T(x_1) = T(x_2)$, the evidence about θ from observing x_1 is the same as the evidence about θ from observing x_2.
- Birnbaum showed LP $\iff$ CP + SP.

Birnbaum's formalization

- By an experiment E we'd mean a triplet $(\mathcal{X}, \Theta, f_\theta)$ of an outcome space $\mathcal{X}$, parameter space Θ and a sampling model given by pdfs/pmfs $f_\theta(x), x \in \mathcal{X}, \theta \in \Theta$.
- We use the notation $\mathrm{evd}(x, E)$ to denote evidence for θ from an observation x in experiment E.
- In CP with two basic experiments $E_1 = (\mathcal{X}_1, \Theta, f_\theta^1)$ and $E_2 = (\mathcal{X}_2, \Theta, f_\theta^2)$, the mixed experiment $E^* = (\mathcal{X}^*, \Theta, f_\theta^*)$ is given by:

$$\mathcal{X}^* = \{1, 2\} \times (\mathcal{X}_1 \cup \mathcal{X}_2)$$
$$f_\theta^*((i, x)) = \frac{1}{2} f_\theta^i(x)$$

Birnbaum's formalization (contd)

- Then CP is equivalent to : $\mathrm{evd}((i, x), E^*) = \mathrm{evd}(x, E_i)$.
- Also, SP says that for an experiment $E = (\mathcal{X}, \Theta, f_\theta)$ with a sufficient statistic T,

$$T(x_1) = T(x_2) \implies \mathrm{evd}(x_1, E) = \mathrm{evd}(x_2, E).$$

- LP states that for two experiments $E_1 = (\mathcal{X}_1, \Theta, f_\theta^1)$, $E_2 = (\mathcal{X}_2, \Theta, f_\theta^2)$, if $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy:

$$f_\theta^1(x_1) = c f_\theta^2(x_2), \quad \forall \theta \in \Theta$$

for some constant $c > 0$, then $\mathrm{evd}(x_1, E_1) = \mathrm{evd}(x_2, E_2)$.

Proof of CP + SP $\implies$ LP

- Suppose $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$ satisfy the LP condition for some $c > 0$.
- Define a statistic $T : \mathcal{X}^* \to \mathcal{X}^*$ as

$$T((i,x)) = \begin{cases} (1, x_1) & \text{if } i = 2, x = x_2 \\ (i, x) & \text{otherwise} \end{cases}$$

- Let $X^* \sim f_\theta^*$. We'll show that the distribution of X^* given $T(X^*)$ is free of θ. Indeed,
  1. if $T(X^*) \neq (1, x_1)$ then X^* must equal $T(X^*)$ w.p. 1.
  2. if $T(X^*) = (1, x_1)$ then X^* is either $(1, x_1)$ or $(2, x_2)$ with probabilities proportional to $\frac{1}{2} f_\theta^1(x_1)$ and $\frac{1}{2} f_\theta^2(x_2)$, i.e., with probabilities $\frac{c}{c+1}$ and $\frac{1}{c+1}$.
- So T is a sufficient statistic in E^*.

Proof of CP + SP $\implies$ LP (contd.)

- Therefore, because $T((1, x_1)) = T((2, x_2))$,

$$\begin{aligned} \mathrm{evd}(x_1, E_1) &= \mathrm{evd}((1, x_1), E^*) && \text{[by CP]} \\ &= \mathrm{evd}((2, x_2), E^*) && \text{[by SP]} \\ &= \mathrm{evd}(x_2, E_2) && \text{[by CP]} \end{aligned}$$

  as desired!

Stopping rules

- LP says that additional data which could have been collected, but have not been, do not impact the inference. This is most clearly visible and striking for sequential methods.
- Imagine that a client enters your statistical consulting office reporting that she has taken $n = 100$ observations from $X_j \overset{\text{IID}}{\sim} N(\theta, 1)$, and wants to test $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$ at level $\alpha = 0.05$.
- The classical procedure gives a p-value of $p = 2\Phi(-\sqrt{n}|\bar{x}_n|)$, and rejects H_0whenever $p \leq \alpha$ or, equivalently, when $\sqrt{n}\bar{x}_n \geq q_{N(1-\alpha/2)}$
- When you learn that her data show $\bar{x}_{100} = 0.20$, the problem seems easy – evidently the p-value is $p = 2\Phi(-2.00) = 0.0455 < \alpha$, leading to rejection.

Stopping rules (contd.)

- But when by chance you ask Why did you take $n = 100$ observations? and learn that the answer is Because that was enough to get significance, your answer has to change.
- If her intension was to reject if $\sqrt{100}|\bar{x}_{100}| \geq k = 1.96$ and otherwise to take another 100 observations and see if that leads to significance, i.e., to $\sqrt{200}|\bar{x}_{200}| \geq k$, then the true probability of a Type-I error is

$$p = P(|Z_1| > k \text{ or } |Z_1 + Z_2| > k\sqrt{2})$$

  or about 0.0768 for $k = 1.96$, so her test does not have its nominal size $\alpha = 0.05$.

Stopping rules (contd.)

- To achieve this size she would have to reject when either $\sqrt{100}|\bar{x}_{100}|$ of $\sqrt{200}|\bar{x}_{200}|$ exceeds $k = 2.12$.
- Since hers do not, we now must change our advice and say she cannot reject H_0!
- It is (or should be!) disturbing that the evidential import of her results should depend on her intentions, and not on the data and experiment. Even more alarming, most experiments are begun without a clear picture of when to stop taking data, so this silly example is in fact the usual situation.

Formalizing stopping rules

- Consider an infinite sequence of experiments $E_m = (\mathcal{X}_m, \Theta, f_\theta^m)$, $m = 1, 2, \cdots$.
- A stopping rule is a sequence of functions

$$\tau_m : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \to [0, 1]$$

  with the interpretation that we conduct the experiments sequentially, gathering data $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \cdots$ and deciding at every step m whether to stop with probability $\tau_m(x_1, \cdots, x_m)$ or otherwise to continue to the next step.
- A stopping rule is proper if it stops almost surely.

The Stopping Rule Principle

- If τ is proper, the the sequential experiments can be put together to define the stopping-rule experiment $E^{(\tau)} = (\mathcal{X}^{(\tau)}, \Theta, f_\theta^{(\tau)})$ where

$$\mathcal{X}^{(\tau)} = \{(m, x_1, x_2, \cdots, x_m) : m \in \mathbb{N}, x_i \in \mathcal{X}_i\}$$

$$f_\theta^{(\tau)}((m, x_1, \cdots, x_m)) = \tau_m(x_{1:m}) \left\{ \prod_{i=1}^{m-1} (1 - \tau_i(x_{1:i})) \right\} \prod_{i=1}^{m} f_\theta^i(x_i)$$

SRP (contd.)

- On the other hand, if we had decided beforehand to continue up to a fixed step m, then the corresponding m-step experiment is $E^{(m)} = (\mathcal{X}^{(m)}, \Theta, f_\theta^{(m)})$ where

$$\mathcal{X}^{(m)} = \{(x_1, x_2, \cdots, x_m) : x_i \in \mathcal{X}_i\}$$

$$f_\theta^{(m)}((x_1, \cdots, x_m)) = \prod_{i=1}^{m} f_\theta^i(x_i)$$

- The SRP states

$$\mathrm{evd}((m, x_1, \cdots, x_m), E^{(\tau)}) = \mathrm{evd}((x_1, \cdots, x_m), E^{(m)}).$$

- That is, once you stop at m, you can do inference pretending that you always wanted to do an m-step experiment.