Notes 9. The Likelihood Principle

(Adapted from Robert Wolpert's notes)

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.
- θ = the unknown proportion of students who support the bill.
- Jerzy
  - Survey 18 students and find $X = \#\text{supporters}$
  - Model $X \sim Bin(18, \theta)$. Obs $X = 12$.
- Egon
  - Survey until 6 non-supporters, $Y = \#\text{supporters}$
  - Model $Y \sim NBin(12, \theta)$, Obs $Y = 12$.

An example (contd.)

- Jerzy's likelihood function is:

$$L^{\text{J}}_{12}(\theta) = \binom{18}{12} \times \theta^{12}(1-\theta)^6$$

- Egon's likelihood function is:

$$L^{\text{E}}_{12}(\theta) = \binom{17}{5} \times \theta^{12}(1-\theta)^6$$

- So we indeed have $L^{\text{J}}_{12}(\theta) \propto L^{\text{E}}_{12}(\theta)$, $\forall \theta$. And LP demands that both Jerzy and Egon should draw same conclusions about p (if their prior beliefs were the same)

An example (contd.)

- Both the binomial and the negative binomial family are MLR, respectively, in X and Y, and their UMP tests will reject for large values of X and Y respectively.
- Jerzy's p-value:

$$\max_{p \le 0.5} P(X \ge 12|p) = P(X \ge 12|p = 0.5)$$
$$= 1 \text{ - pbinom}(11, 18, 0.5) = 0.12.$$

- Egon's p-value

$$\max_{p \le 0.5} P(Y \ge 12|p) = P(Y \ge 12|p = 0.5)$$
$$= 1 \text{ - pnbinom}(11, 6, 0.5) = 0.07$$

An example (contd.)

- LP is violated
- p-value = probability under H_0 of observing more extreme evidence against H_0 than what is observed
- Care about data that has not been observed
- Jeffreys said:

  A hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred

Example 2

- X_1, X_2 are IID: $P(X_j = \theta \pm 1) = 1/2$. $\theta \in \mathbb{R}$ unknown.
- Shortest 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. $\approx 50\%$ 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 lab with 2 instruments
- Accuracies = ± 0.01 and ± 0.05.
- A scientist gets to use whichever is available (w.p. $1/2$)
- What accuracy to report?
- Accuracy of 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 $E_1 = (\mathcal{X}_1, \Theta, f^1_\theta)$ and $E_2 = (\mathcal{X}_2, \Theta, f^2_\theta)$, the mixed experiment $E^* = (\mathcal{X}^*, \Theta, f^*_\theta)$ is given by:

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

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 $E_1 = (\mathcal{X}_1, \Theta, f^1_\theta)$, $E_2 = (\mathcal{X}_2, \Theta, f^2_\theta)$, if $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy:

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

  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 $\frac{1}{2}f^1_\theta(x_1)$ and $\frac{1}{2}f^2_\theta(x_2)$, 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!

LP & uncollected additional data

- 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.
- But first a story!

The voltmeter story (due to JW Pratt)

An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate volt-meter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean.

Voltmeter story (contd)

Later he visits the engineer's laboratory, and notices that the volt-meter reads only as far as 100, so the population appears to be "censored". This necessitates a new analysis, if the statistician is orthodox. However, the engineer says he has another meter, equally accurate over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all.

Voltmeter story (contd)

But the next day the engineer telephones and says, "I just discovered my high-range volt-meter was not working the day I did the experiment you analyzed for me.". The statistician ascertains that the engineer would not have held up experiment until the meter was fixed, and informs him that the a new analysis will be required.

Voltmeter story (contd)

The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope!"

Stopping rules

- 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_0 whenever $p \leq \alpha$ or, equivalently, when $\sqrt{n}|\bar{x}_n| \geq z(\alpha)$
- 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.