Statistical Inference
Lecture Three

https://people.bath.ac.uk/masss/APTS/apts.html

Simon Shaw

University of Bath

APTS, 14-18 December 2020

Overview of Lecture Three

So far we have explored a number of principles for statistical inference.

- Weak Conditionality Principle, WCP: if $\mathcal{E}^*$ is the mixture of the experiments $\mathcal{E}_1$, $\mathcal{E}_2$ according to mixture probabilities p_1, $p_2 = 1 - p_1$. then $\mathrm{Ev}(\mathcal{E}^*, (i, x_i)) = \mathrm{Ev}(\mathcal{E}_i, x_i)$.
- Strong Likelihood Principle, SLP: if $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$, for some function $c > 0$ for all $\theta \in \Theta$ then $\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2)$.

In this lecture we will introduce a final principle, and consider the likelihood principle in practice.

- Y is ancillary if $f_{X,Y}(x, y \mid \theta) = f_Y(y) f_{X|Y}(x \mid y, \theta)$.
- Strong Conditionality Principle, SCP: If Y is ancillary then $\mathrm{Ev}(\mathcal{E}, (x, y)) = \mathrm{Ev}(\mathcal{E}^{X|y}, x)$.
- Two Bayesian models with the same prior distribution, $\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \mid \theta), \pi(\theta)\}$ and $\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \mid \theta), \pi(\theta)\}$ have the same posterior distribution when $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$. Hence, the Bayesian approach satisfies the SLP.

A stronger form of the WCP

- We consider the concept of ancillarity.
- This has several different definitions in the Statistics literature; the one we use is close to that of Cox and Hinkley (1974, Section 2.2).

Definition (Ancillarity)

Y is ancillary in the experiment $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \mid \theta)\}$ exactly when $f_{X,Y}$ factorises as

$$f_{X,Y}(x, y \mid \theta) = f_Y(y) f_{X|Y}(x \mid y, \theta).$$

- The marginal distribution of Y is completely specified: it does not depend on θ.
- We could extend this to consider an extended parameter set, say (λ, θ) where λ is a nuisance parameter and θ is the parameter of interest.
- Ancillarity would be that f_Y doesn't depend on θ but may on λ whilst $f_{X|Y}$ depends on θ but doesn't depend on λ.

- Not all families of distributions will factorise in this way, but when they do, there are new possibilities for inference, based around stronger forms of the WCP.
- A familiar example is that of a random sample size: in a sample $x = (x_1, \ldots, x_n)$, n may be the outcome of a random variable N.
- We seldom concern ourselves with the distribution of N when we evaluate x; instead we treat N as known.
- Equivalently, we treat N as ancillary and condition on $N = n$.
- In this case, we might think that inferences drawn from observing (n, x) should be the same as those for x conditioned on $N = n$.

- When Y is ancillary, we can consider the conditional experiment

$$\mathcal{E}^{X\,|\,y} = \{\mathcal{X}, \Theta, f_{X\,|\,Y}(x\,|\,y, \theta)\}.$$

- That is, we treat Y as known, and treat X (conditional on $Y = y$) as the only random variable.

Principle 9: Strong Conditionality Principle, SCP

If Y is ancillary in $\mathcal{E}$, then $\mathrm{Ev}(\mathcal{E}, (x, y)) = \mathrm{Ev}(\mathcal{E}^{X|y}, x)$.

- The SCP is invoked (implicitly) when we perform a regression of Y on X: (X, Y) is random, but X is treated as ancillary for the parameters in $f_{Y|X}$. We model Y conditionally on X, treating X as known.
- Clearly the SCP implies the WCP, with the experiment indicator $I \in \{1, 2\}$ being ancillary, since p is known.

Theorem

SLP $\rightarrow$ SCP.

Proof

Suppose that Y is ancillary in $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta)\}$. Thus, for all $\theta \in \Theta$,

$$\begin{aligned} f_{X,Y}(x, y \,|\, \theta) &= f_Y(y) f_{X|Y}(x \,|\, y, \theta) \\ &= c(y) f_{X|Y}(x \,|\, y, \theta) \end{aligned}$$

Then the SLP implies that

$$\mathrm{Ev}\,(\mathcal{E}, (x, y)) = \mathrm{Ev}(\mathcal{E}^{X|y}, x),$$

as required. □

- From Birnbaum's Theorem, (WIP $\wedge$ WCP) $\leftrightarrow$ SLP so, as SLP $\rightarrow$ SCP, the WIP allows us to 'upgrade' the WCP to the SCP.

The Likelihood Principle in practice

- We consider whether there is any inferential approach which respects the SLP? Or do all inferential approaches respect it?

A Bayesian statistical model is the collection

$$\mathcal{E}_B = \{\mathcal{X}, \Theta, f_X(x \mid \theta), \pi(\theta)\}.$$

The posterior distribution is $\pi(\theta \mid x) = c(x) f_X(x \mid \theta)\pi(\theta)$ where $c(x)$ is the normalising constant,

$$c(x) = \left\{ \int_\Theta f_X(x \mid \theta)\pi(\theta)\, d\theta \right\}^{-1}.$$

- All knowledge about θ given the data x are represented by $\pi(\theta \mid x)$.
- Any inferences made about θ are derived from this distribution.

- Consider two Bayesian models with the same prior distribution, $\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \mid \theta), \pi(\theta)\}$ and $\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \mid \theta), \pi(\theta)\}$
- Suppose that $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$. Then

$$\begin{aligned} \pi_1(\theta \mid x_1) &= c(x_1) f_{X_1}(x_1 \mid \theta)\pi(\theta) &&= c(x_1) c(x_1, x_2) f_{X_2}(x_2 \mid \theta)\pi(\theta) \\ &&&= \pi_2(\theta \mid x_2) \end{aligned}$$

- Hence, the posterior distributions are the same. Consequently, the same inferences are drawn from either model and so the Bayesian approach satisfies the SLP.
- This assumes that $\pi(\theta)$ does not depend upon the form of the data.
- Some methods for making default choices for $\pi(\theta)$ depend on $f_X(x \mid \theta)$, notably Jeffreys priors and reference priors. These methods violate the SLP.

- Maximum likelihood estimation clearly satisfies the SLP and methods, such as penalised likelihood theory, have been generated to satisfy the SLP.
- However, inference tools used in the classical approach typically violate the SLP.
- Inference techniques depend upon the sampling distribution and so they depend on the whole sample space $\mathcal{X}$ and not just the observed $x \in \mathcal{X}$.
- Sampling distribution depends on values of f_X other than $L(\theta; x) = f_X(x \mid \theta)$.
- For a statistic $T(X)$, $MSE(T \mid \theta) = Var(T \mid \theta) + \text{bias}(T \mid \theta)^2$ depends upon the first and second moments of the distribution of $T \mid \theta$.

Example, Robert (2007)

- Suppose that X_1, X_2 are iid $N(\theta, 1)$ so that

$$f(x_1, x_2 \,|\, \theta) \;\propto\; \exp\left\{-(\overline{x} - \theta)^2\right\}.$$

- Consider the alternate model for the same parameter θ

$$g(x_1, x_2 \,|\, \theta) \;=\; \pi^{-\frac{3}{2}} \frac{\exp\left\{-(\overline{x} - \theta)^2\right\}}{1 + (x_1 - x_2)^2}$$

- Thus, $f(x_1, x_2 \,|\, \theta) \propto g(x_1, x_2 \,|\, \theta)$ as a function of θ. If the SLP is applied, then inference about θ should be the same in both models.
- The distribution of g is quite different from that of f and so estimators of θ will have different classical properties if they do not depend only on $\overline{x}$.
- For example, g has heavier tails than f and so respective confidence intervals may differ between the two.

- Suppose that $\mathrm{Ev}(\mathcal{E}, x)$ depends on the value of $f_X(x' \mid \theta)$ for some $x' \neq x$. Then, typically, Ev does not respect the SLP.
- We could create an alternate experiment $\mathcal{E}_1 = \{\mathcal{X}, \Theta, f_1(x \mid \theta)\}$ where:
  - $f_1(x \mid \theta) = f_X(x \mid \theta)$ for the observed x.
  - $f_1(x \mid \theta) \neq f_X(x \mid \theta)$ for all $x \in \mathcal{X}$.
- In particular, that $f_1(x' \mid \theta) \neq f_X(x' \mid \theta)$.
  - Let $\tilde{x} \neq x, x'$ and set

$$\begin{aligned} f_1(x' \mid \theta) &= \alpha f_X(x' \mid \theta) + \beta f_X(\tilde{x} \mid \theta) \\ f_1(\tilde{x} \mid \theta) &= (1-\alpha) f_X(x' \mid \theta) + (1-\beta) f_X(\tilde{x} \mid \theta) \end{aligned}$$

  - By suitable choice of α, β we can redistribute the mass to ensure $f_1(x' \mid \theta) \neq f_X(x' \mid \theta)$. We then let $f_1 = f_X$ elsewhere.
- Consequently, whilst $f_1(x \mid \theta) = f_X(x \mid \theta)$ we will not have that $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}_1, x)$ and so will violate the SLP.

The two main difficulties with violating the SLP are:

1. To reject the SLP is to reject at least one of the WIP and the WCP. Yet both of these principles seem self-evident. Therefore violating the SLP is either illogical or obtuse.
2. In their everyday practice, statisticians use the SRP (ignoring the intentions of the experimenter) which is not self-evident, but is implied by the SLP. If the SLP is violated, it needs an alternative justification which has not yet been forthcoming.

Reflections

- This chapter does not explain how to choose Ev but instead describes desirable properties of Ev.
- What is evaluated is the algorithm, the method by which $(\mathcal{E}, x)$ is turned into an inference about the parameter θ.
- It is quite possible that statisticians of quite different persuasions will produce effectively identical inferences from different algorithms.
- A Bayesian statistician might produce a 95% High Density Region, and a classical statistician a 95% confidence set, but they might be effectively the same set.
- Primary concern for the auditor is why the particular inference method was chosen and they might also ask if the statistician is worried about the SLP.
- Classical statistician might argue a long-run frequency property but the client might wonder about their interval.