Statistical Inference

https://people.bath.ac.uk/masss/APTS/lecture2.pdf

Simon Shaw

University of Bath

Lecture Two, 17 December 2019

Yesterday's lecture

- We wish to consider inferences about a parameter θ given a parametric model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$

$$(\mathcal{E}, x) \xmapsto{\text{statistician, Ev}} \text{Inference about } \theta.$$

- Weak Indifference Principle, WIP: if $f_X(x \,|\, \theta) = f_X(x' \,|\, \theta)$ for all $\theta \in \Theta$ then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$.
- Distribution Principle, DP: if $\mathcal{E} = \mathcal{E}'$, then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}', x)$.
- Transformation Principle, TP: $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^g, g(x))$.
- $(\text{DP} \wedge \text{TP}) \to \text{WIP}$.
- Weak Conditionality Principle, WCP: let $\mathcal{E}^*$ be the mixture of $\mathcal{E}_1$, $\mathcal{E}_2$ according to probabilities p_1, p_2. Then $\text{Ev}\left(\mathcal{E}^*, (i, x_i)\right) = \text{Ev}(\mathcal{E}_i, x_i)$.
- Strong Likelihood Principle, SLP: if $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$, for some function $c > 0$ for all $\theta \in \Theta$ then $\text{Ev}(\mathcal{E}_1, x_1) = \text{Ev}(\mathcal{E}_2, x_2)$.
- Birnbaum's Theorem: $(\text{WIP} \wedge \text{WCP}) \leftrightarrow \text{SLP}$.

The Sufficiency Principle

- Recall the idea of sufficiency: if $S = s(X)$ is sufficient for θ then

$$f_X(x \mid \theta) \;=\; f_{X|S}(x \mid s, \theta) f_S(s \mid \theta)$$

where $f_{X|S}(x \mid s, \theta)$ does not depend upon θ.

- Consequently, consider the experiment $\mathcal{E}^S = \{s(\mathcal{X}), \Theta, f_S(s \mid \theta)\}$.

Principle 6: Strong Sufficiency Principle, SSP

If $S = s(X)$ is a sufficient statistic for $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^S, s(x))$.

Principle 7: Weak Sufficiency Principle, WSP

If $S = s(X)$ is a sufficient statistic for $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ and $s(x) = s(x')$ then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$.

Theorem

SLP $\rightarrow$ SSP $\rightarrow$ WSP $\rightarrow$ WIP.

Proof

As s is sufficient, $f_X(x \,|\, \theta) = cf_S(s \,|\, \theta)$ where $c = f_{X|S}(x \,|\, s, \theta)$ does not depend on θ. Applying the SLP, $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^S, s(x))$ which is the SSP. Note, that from the SSP,

$$\begin{aligned}
\text{Ev}(\mathcal{E}, x) &= \text{Ev}(\mathcal{E}^S, s(x)) && \text{(by the SSP)} \\
&= \text{Ev}(\mathcal{E}^S, s(x')) && \text{(as } s(x) = s(x')\text{)} \\
&= \text{Ev}(\mathcal{E}, x') && \text{(by the SSP)}
\end{aligned}$$

We thus have the WSP. Finally, if $f_X(x \,|\, \theta) = f_X(x' \,|\, \theta)$ as in the statement of WIP then $s(x) = x'$ is sufficient for x. Hence, from the WSP, $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$ giving the WIP. □

If we put together the last two theorems, we get the following corollary.

Corollary

$(\text{WIP} \wedge \text{WCP}) \rightarrow \text{SSP}$.

Proof

From Birnbaum's theorem, $(\text{WIP} \wedge \text{WCP}) \leftrightarrow \text{SLP}$ and from the previous theorem, $\text{SLP} \rightarrow \text{SSP}$. □

- Birnbaum's (1962) original result combined sufficiency and conditionality for the likelihood but he revised this to the WIP and WCP in later work.
- One advantage of this is that it reduces the dependency on sufficiency: Pitman-Koopman-Darmois Theorem states that sufficiency more-or-less characterises the exponential family.

Stopping rules

- Consider observing a sequence of random variables $X_1, X_2, \ldots$ where the number of observations is not fixed in advance but depends on the values seen so far.
    - At time j, the decision to observe X_{j+1} can be modelled by a probability $p_j(x_1, \ldots, x_j)$.
    - We assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already, hence $p_m(x_1, \ldots, x_m) = 0$.
- The stopping rule may then be denoted as $\tau = (p_1, \ldots, p_m)$. This gives an experiment $\mathcal{E}^{\tau}$ with, for $n = 1, 2, \ldots$, $f_n(x_1, \ldots, x_n \,|\, \theta)$ where consistency requires that

$$f_n(x_1, \ldots, x_n \,|\, \theta) \;=\; \sum_{x_{n+1}} \cdots \sum_{x_m} f_m(x_1, \ldots, x_n, x_{n+1}, \ldots x_m \,|\, \theta).$$

Motivation for the stopping rule principle (Basu, 1975)

- Consider four different coin-tossing experiments (with some finite limit on the number of tosses).
  - $\mathcal{E}_1$ Toss the coin exactly 10 times;
  - $\mathcal{E}_2$ Continue tossing until 6 heads appear;
  - $\mathcal{E}_3$ Continue tossing until 3 consecutive heads appear;
  - $\mathcal{E}_4$ Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.
- Suppose that all four experiments have the same outcome $x = (\text{T,H,T,T,H,H,T,H,H,H})$.
- We may feel that the evidence for θ, the probability of heads, is the same in every case.
  - Once the sequence of heads and tails is known, the intentions of the original experimenter (i.e. the experiment she was doing) are immaterial to inference about the probability of heads.
  - The simplest experiment $\mathcal{E}_1$ can be used for inference.

Principle 8: Stopping Rule Principle, SRP

[a] In a sequential experiment $\mathcal{E}^\tau$, $\text{Ev}\left(\mathcal{E}^\tau, (x_1, \ldots, x_n)\right)$ does not depend on the stopping rule τ.

[a] Basu (1975) claims the SRP is due to George Barnard (1915-2002)

- If it is accepted, the SRP is nothing short of revolutionary.
- It implies that the intentions of the experimenter, represented by τ, are irrelevant for making inferences about θ, once the observations $(x_1, \ldots, x_n)$ are known.
- Once the data is observed, we can ignore the sampling plan.
- The statistician could proceed as though the simplest possible stopping rule were in effect, which is $p_1 = \cdots = p_{n-1} = 1$ and $p_n = 0$, an experiment with n fixed in advance, $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n} \,|\, \theta)\}$.
- Can the SRP possibly be justified? Indeed it can.

Theorem

SLP $\rightarrow$ SRP.

Proof

Let τ be an arbitrary stopping rule, and consider the outcome $(x_1, \ldots, x_n)$, which we will denote as $x_{1:n}$.

- We take the first observation with probability one.
- For $j = 1, \ldots, n-1$, the $(j+1)$th observation is taken with probability $p_j(x_{1:j})$.
- We stop after the nth observation with probability $1 - p_n(x_{1:n})$.

Consequently, the probability of this outcome under τ is

$$f_\tau(x_{1:n} \,|\, \theta) \;=\; f_1(x_1 \,|\, \theta) \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j})\, f_{j+1}(x_{j+1} \,|\, x_{1:j}, \theta) \right\} (1 - p_n(x_{1:n}))$$

Proof continued

$$
\begin{aligned}
f_\tau(x_{1:n} \mid \theta) &= \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j}) \right\} (1 - p_n(x_{1:n}))\, f_1(x_1 \mid \theta) \prod_{j=2}^{n} f_j(x_j \mid x_{1:(j-1)}, \theta) \\
&= \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j}) \right\} (1 - p_n(x_{1:n})) f_n(x_{1:n} \mid \theta).
\end{aligned}
$$

Now observe that this equation has the form

$$
f_\tau(x_{1:n} \mid \theta) = c(x_{1:n}) f_n(x_{1:n} \mid \theta) \qquad (1)
$$

where $c(x_{1:n}) > 0$. Thus the SLP implies that $\text{Ev}(\mathcal{E}^\tau, x_{1:n}) = \text{Ev}(\mathcal{E}^n, x_{1:n})$ where $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n} \mid \theta)\}$. Since the choice of stopping rule was arbitrary, equation (1) holds for all stopping rules, showing that the choice of stopping rule is irrelevant. □

A comment from Leonard Jimmie Savage (1917-1971), one of the great statisticians of the Twentieth Century, captured the revolutionary and transformative nature of the SRP.

> *May I digress to say publicly that I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right. (Savage et al., 1962, p76)*

- We'll omit the section "A stronger form of the WCP" which looks at an extension of the WCP.

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, $\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \,|\, \theta), \pi(\theta)\}$ and $\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \,|\, \theta), \pi(\theta)\}$
- Suppose that $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$. Then

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

- 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 $f_X(x \,|\, \theta)$, notably Jeffreys priors and reference priors. These methods violate the SLP.