Statistical Inference
Lecture Two

https://people.bath.ac.uk/masss/APTS/2022-23/LectureTwo.pdf

Simon Shaw

University of Bath

APTS, 13-16 December 2022

Overview of Lecture Two

In Lecture One we considered a number of statistical principles.

- Weak Indifference Principle, WIP: if $f_X(x \mid \theta) = f_X(x' \mid \theta)$ 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: for the bijective $g : \mathcal{X} \to \mathcal{Y}$, construct $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$. Then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^g, g(x))$.
- $(\text{DP} \wedge \text{TP}) \to \text{WIP}$.
- 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 $\text{Ev}\left(\mathcal{E}^*, (i, x_i)\right) = \text{Ev}(\mathcal{E}_i, x_i)$.
- Strong Likelihood Principle, SLP: if $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$, for some function $c > 0$ for all $\theta \in \Theta$ then $\text{Ev}(\mathcal{E}_1, x_1) = \text{Ev}(\mathcal{E}_2, x_2)$.
- Birnbaum's Theorem: $(\text{WIP} \wedge \text{WCP}) \leftrightarrow \text{SLP}$.

Overview of Lecture Two continued

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

- Strong Sufficiency Principle, SSP: if $S = s(X)$ is a sufficient statistic for $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^S, s(x))$.
- 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 $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$.
- SLP $\rightarrow$ SSP $\rightarrow$ WSP $\rightarrow$ WIP.
- Stopping Rule Principle, SRP: in a sequential experiment $\mathcal{E}^\tau$, $\text{Ev}\,(\mathcal{E}^\tau, (x_1, \ldots, x_n))$ does not depend on the stopping rule τ.
- SLP $\rightarrow$ SRP.
- 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 $\text{Ev}\,(\mathcal{E}, (x, y)) = \text{Ev}(\mathcal{E}^{X|y}, x)$.

The Sufficiency Principle

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

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

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, $f_X(x \mid \theta) = cf_S(s \mid \theta)$ where $c = f_{X|S}(x \mid s, \theta)$ 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 $f_X(x \mid \theta) = f_X(x' \mid \theta)$ 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. $\square$

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

Corollary

(WIP $\wedge$ WCP) $\rightarrow$ SSP.

Proof

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

- 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, $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n} \mid \theta)\}$.
- Can the SRP possibly be justified? Indeed it can.

Theorem

SLP $\to$ 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

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

Proof continued

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

Now observe that this equation has the form

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

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 $\mathcal{E}^n = \{\mathcal{X}_{1:n}, \Theta, f_n(x_{1:n} \,|\, \theta)\}$. 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)*

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

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

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

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

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

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

- All knowledge about θ given the data x are represented by $\pi(\theta \,|\, x)$.
- 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 $L(\theta; x) = f_X(x \,|\, \theta)$.
- For a statistic $T(X)$, $MSE(T \,|\, \theta) = Var(T \,|\, \theta) + \text{bias}(T \,|\, \theta)^2$ depends upon the first and second moments of the distribution of $T \,|\, \theta$.

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.

Binomial and Negative Binomial example

- Let $\mathcal{E}_1 = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$, where $X \,|\, \theta \sim Bin(n, \theta)$ so that

$$f_X(x \,|\, \theta) \;=\; \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

- Let $\mathcal{E}_2 = \{\mathcal{Y}, \Theta, f_Y(y \,|\, \theta)\}$, where $Y \,|\, \theta \sim Nbin(r, \theta)$, so that

$$f_Y(y \,|\, \theta) \;=\; \binom{y-1}{r-1}\theta^r(1-\theta)^{y-r}, \quad y = r, r+1, \ldots .$$

- Suppose we observe $x = r = 3$ and $y = n = 12$ then

$$f_X(3 \,|\, \theta) \;=\; \binom{12}{3}\theta^3(1-\theta)^9, \;\; f_Y(12 \,|\, \theta) \;=\; \binom{11}{2}\theta^3(1-\theta)^9$$

- Thus, $f_X(3 \,|\, \theta) \propto f_Y(12 \,|\, \theta)$.

- Consider the hypothesis test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta < \frac{1}{2}$ at significance level 5%.
- Let $\mathrm{Ev}(\mathcal{E}_1, 3)$ be the result of the hypothesis test for the Binomial model where small values of X support H_1

$$\mathbb{P}(X \leq 3 \,|\, \theta = 1/2) \;=\; \sum_{x=0}^{3} f_X(x \,|\, \theta = 1/2) \;=\; 0.0730.$$

- Thus, $\mathrm{Ev}(\mathcal{E}_1, 3)$ is to not reject H_0.
- Let $\mathrm{Ev}(\mathcal{E}_2, 12)$ be the result of the hypothesis test for the Negative Binomial model where large values of Y support H_1

$$\mathbb{P}(Y \geq 12 \,|\, \theta = 1/2) \;=\; \sum_{y=12}^{\infty} f_Y(y \,|\, \theta = 1/2) \;=\; 0.0327.$$

- Thus, $\mathrm{Ev}(\mathcal{E}_2, 12)$ is to reject H_0.
- This inference method does not respect the SLP: the choice of the model is relevant to the inference.

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

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

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