Statistical Inference
Lecture One

https://people.bath.ac.uk/masss/APTS/2021-22/LectureOne.pdf

Simon Shaw

University of Bath

APTS, 13-17 December 2021

Overview of Lecture One

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

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

- We'll consider a series of statistical principles to guide the way to learn about θ.
- 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 $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^g, g(x))$.
- $(\text{DP} \wedge \text{TP}) \to \text{WIP}$.

Overview of Lecture One continued

- 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 $\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2)$.
- Birnbaum's Theorem: $(\mathrm{WIP} \wedge \mathrm{WCP}) \leftrightarrow \mathrm{SLP}$.
- Stopping Rule Principle, SRP: in a sequential experiment $\mathcal{E}^\tau$, $\mathrm{Ev}(\mathcal{E}^\tau, (x_1, \ldots, x_n))$ does not depend on the stopping rule τ.
- $\mathrm{SLP} \rightarrow \mathrm{SRP}$.

Introduction

- We wish to consider inferences about a parameter θ given a parametric model

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

- We assume that the model is true so that only $\theta \in \Theta$ is unknown. We wish to learn about θ from observations x (typically, vector valued) so that $\mathcal{E}$ represents a model for this experiment.

Smith (2010) considers that there are three players in an inference problem:

1. Client: person with the problem
2. Statistician: employed by the client to help solve the problem
3. Auditor: hired by the client to check the statistician's work

The statistician is thus responsible for explaining the rationale behind the choice of inference in a compelling way.

Reasoning about inferences

We consider a series of statistical principles to guide the way to learn about θ. The principles are meant to be either self-evident or logical implications of principles which are self-evident.
We shall assume that $\mathcal{X}$ is finite: Basu (1975) argues that "infinite and continuous models are to be looked upon as mere approximations to the finite realities."

- Inspiration of Allan Birnbaum (1923-1976) to see how to construct and reason about statistical principles given "evidence" from data.
- The model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ is accepted as a working hypothesis.
- How the statistician chooses her inference statements about the true value θ is entirely down to her and her client.
  - as a point or a set in Θ;
  - as a choice among alternative sets or actions;
  - or maybe as something more complicated, not ruling out visualisations.

- Following Dawid (1977), consider that the statistician defines, a priori, a set of possible inferences about θ
- Task is to choose an element of this set based on $\mathcal{E}$ and x.
- The statistician should see herself as a function Ev: a mapping from $(\mathcal{E}, x)$ into a predefined set of inferences about θ.

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

- For example, $\text{Ev}(\mathcal{E}, x)$ might be:
  - the maximum likelihood estimator of θ
  - a 95% confidence interval for θ
- Birnbaum called $\mathcal{E}$ the experiment, x the outcome, and Ev the evidence.

Note:

1. There can be different experiments with the same θ.
2. Under some outcomes, we would agree that it is self-evident that these different experiments provide the same evidence about θ.

Example

Consider two experiments with the same θ.

1. $X \sim Bin(n, \theta)$, so we observe x successes in n trials.
2. $Y \sim NBin(r, \theta)$, so we observe the rth success in the yth trial.

If we observe $x = r$ and $y = n$, do we make the same inference about θ in each case?

Consider two experiments $\mathcal{E}_1 = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \mid \theta)\}$ and $\mathcal{E}_2 = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \mid \theta)\}$.

Equivalence of evidence (Basu, 1975)

The equality or equivalence of $\text{Ev}(\mathcal{E}_1, x_1)$ and $\text{Ev}(\mathcal{E}_2, x_2)$ means that:

1. $\mathcal{E}_1$ and $\mathcal{E}_2$ are related to the same parameter θ.
2. Everything else being equal, the outcome x_1 from $\mathcal{E}_1$ warrants the same inference about θ as does the outcomes x_2 from $\mathcal{E}_2$.

- We now consider constructing statistical principles and demonstrate how these principles imply other principles.
- These principles all have the same form: under such and such conditions, the evidence about θ should be the same.
- Thus they serve only to rule out inferences that satisfy the conditions but have different evidences. They do not tell us how to do an inference, only what to avoid.

The principle of indifference

Principle 1: Weak Indifference Principle, WIP

Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. If $f_X(x \mid \theta) = f_X(x' \mid \theta)$ for all $\theta \in \Theta$ then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}, x')$.

- We are indifferent between two models of evidence if they differ only in the manner of the labelling of sample points.
- If $X = (X_1, \ldots, X_n)$ where the X_is are a series of independent Bernoulli trials with parameter θ then $f_X(x \mid \theta) = f_X(x' \mid \theta)$ if x and x' contain the same number of successes.

Principle 2: Distribution Principle, DP

If $\mathcal{E} = \mathcal{E}'$, then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}', x)$.

- Informally, (Dawid, 1977), only aspects of an experiment which are relevant to inference are the sample space and the family of distributions over it.

Principle 3: Transformation Principle, TP

Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. For the bijective $g : \mathcal{X} \to \mathcal{Y}$, let $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$, the same experiment as $\mathcal{E}$ but expressed in terms of $Y = g(X)$, rather than X. Then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}^g, g(x))$.

- Inferences should not depend on the way in which the sample space is labelled, for example, X or X^{-1}.

Theorem

(DP $\wedge$ TP) $\rightarrow$ WIP.

Proof

Fix $\mathcal{E}$, and suppose that $x, x' \in \mathcal{X}$ satisfy $f_X(x \mid \theta) = f_X(x' \mid \theta)$ for all $\theta \in \Theta$, as in the condition of the WIP.
Let $g : \mathcal{X} \rightarrow \mathcal{X}$ be the function which switches x for x', but leaves all of the other elements of $\mathcal{X}$ unchanged. Then $\mathcal{E} = \mathcal{E}^g$ and

$$\begin{aligned} \mathrm{Ev}(\mathcal{E}, x') &= \mathrm{Ev}(\mathcal{E}^g, x') \quad \text{[by the DP]} \\ &= \mathrm{Ev}(\mathcal{E}^g, g(x)) \\ &= \mathrm{Ev}(\mathcal{E}, x), \quad \text{[by the TP]} \end{aligned}$$

which gives the WIP. □

The Likelihood Principle

- Consider experiments $\mathcal{E}_i = \{\mathcal{X}_i, \Theta, f_{X_i}(x_i \mid \theta)\}$, $i = 1, 2, \ldots$, where the parameter space Θ is the same for each experiment.
- Let $p_1, p_2, \ldots$ be a set of known probabilities so that $p_i \geq 0$ and $\sum_i p_i = 1$.

Mixture experiment

The mixture $\mathcal{E}^*$ of the experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$ according to mixture probabilities $p_1, p_2, \ldots$ is the two-stage experiment

1. A random selection of one of the experiments: $\mathcal{E}_i$ is selected with probability p_i.
2. The experiment selected in stage 1. is performed.

Thus, each outcome of the experiment $\mathcal{E}^*$ is a pair (i, x_i), where $i = 1, 2, \ldots$ and $x_i \in \mathcal{X}_i$, and family of distributions

$$f^*((i, x_i) \mid \theta) \;=\; p_i f_{X_i}(x_i \mid \theta).$$

Principle 4: Weak Conditionality Principle, WCP

Let $\mathcal{E}^*$ be the mixture of the experiments $\mathcal{E}_1$, $\mathcal{E}_2$ according to mixture probabilities p_1, $p_2 = 1 - p_1$. Then $\text{Ev}(\mathcal{E}^*, (i, x_i)) = \text{Ev}(\mathcal{E}_i, x_i)$.

- The WCP says that inferences for θ depend only on the experiment performed and not which experiments could have been performed.
- Suppose that $\mathcal{E}_i$ is randomly chosen with probability p_i and x_i is observed.
- The WCP states that the same evidence about θ would have been obtained if it was decided non-randomly to perform $\mathcal{E}_i$ from the beginning and x_i is observed.

Principle 5: Strong Likelihood Principle, SLP

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter θ. If $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$, that is

$$L_{X_1}(\theta; x_1) \;=\; c(x_1, x_2) L_{X_2}(\theta; x_2)$$

for some function $c > 0$ for all $\theta \in \Theta$ then $\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2)$.

- The SLP states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same.
- A corollary of the SLP, obtained by setting $\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}$, is that $\mathrm{Ev}(\mathcal{E}, x)$ should depend on $\mathcal{E}$ and x only through $L_X(\theta; x)$.

Many classical statistical procedures violate the SLP and the following result was something of the bombshell, when it first emerged in the 1960s. The following form is due to Birnbaum (1972) and Basu (1975)

Birnbaum's Theorem

(WIP $\wedge$ WCP) $\leftrightarrow$ SLP.

Proof

Both SLP $\rightarrow$ WIP and SLP $\rightarrow$ WCP are straightforward. The trick is to prove (WIP $\wedge$ WCP) $\rightarrow$ SLP.

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter, and suppose that $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$ where the function $c > 0$. As the value c is known (as the data has been observed) then consider the mixture experiment with $p_1 = 1/(1+c)$ and $p_2 = c/(1+c)$.

Proof continued

$$f^*((1,x_1)\,|\,\theta) \;=\; \frac{1}{1+c} f_{X_1}(x_1\,|\,\theta) \;=\; \frac{c}{1+c} f_{X_2}(x_2\,|\,\theta) \;=\; f^*((2,x_2)\,|\,\theta)$$

Then the WIP implies that

$$\mathrm{Ev}\left(\mathcal{E}^*,(1,x_1)\right) \;=\; \mathrm{Ev}\left(\mathcal{E}^*,(2,x_2)\right).$$

Applying the WCP to each side we infer that

$$\mathrm{Ev}(\mathcal{E}_1,x_1) \;=\; \mathrm{Ev}(\mathcal{E}_2,x_2),$$

as required. □

Thus, either I accept the SLP, or I explain which of the two principles, WIP and WCP, I refute. Methods, which include many classical procedures, which violate the SLP face exactly this challenge.

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 $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}^S, s(x))$.

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 $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}, x')$.

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

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

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 $x = (\mathrm{T},\mathrm{H},\mathrm{T},\mathrm{T},\mathrm{H},\mathrm{H},\mathrm{T},\mathrm{H},\mathrm{H},\mathrm{H})$.
- 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 $\mathcal{E}^\tau$, $\mathrm{Ev}(\mathcal{E}^\tau, (x_1, \ldots, x_n))$ 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 $\rightarrow$ 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

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

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

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

where $c(x_{1:n}) > 0$. Thus the SLP implies that $\mathrm{Ev}(\mathcal{E}^\tau, x_{1:n}) = \mathrm{Ev}(\mathcal{E}^n, x_{1:n})$ 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)*