

# Statistical Inference

<https://people.bath.ac.uk/mass/APTS/lecture1.pdf>

Simon Shaw

University of Bath

Lecture One, 16 December 2019

# Introduction

- We wish to consider inferences about a parameter  $\theta$  given a parametric model

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

- We assume that the model is **true** so that only  $\theta \in \Theta$  is unknown. We wish to learn about  $\theta$  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  $\theta$ . 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  $\mathcal{E} = \{\mathcal{X}, \Theta, f_{\mathcal{X}}(x | \theta)\}$  is accepted as a working hypothesis.
- How the statistician chooses her inference statements about the true value  $\theta$  is entirely down to her and her client.
  - ▶ as a point or a set in  $\Theta$ ;
  - ▶ as a choice among alternative sets or actions;
  - ▶ or maybe as some more complicated, not ruling out visualisations.

- Following Dawid (1977), consider that the statistician defines, *a priori*, a set of possible **inferences about  $\theta$**
- 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  $\theta$** .

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

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

Note:

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

### Example

Consider two experiments with the same  $\theta$ .

- 1  $X \sim \text{Bin}(n, \theta)$ , so we observe  $x$  successes in  $n$  trials.
- 2  $Y \sim \text{NBin}(r, \theta)$ , so we observe the  $r$ th success in the  $y$ th trial.

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

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

### 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  $\theta$ .
  - 2 Everything else being equal, the outcome  $x_1$  from  $\mathcal{E}_1$  warrants the same inference about  $\theta$  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  $\theta$  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  $\mathcal{E} = \{\mathcal{X}, \Theta, f_{\mathcal{X}}(x | \theta)\}$ . If  $f_{\mathcal{X}}(x | \theta) = f_{\mathcal{X}}(x' | \theta)$  for all  $\theta \in \Theta$  then  $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$ .

- We are indifferent between two models of evidence if they differ only in the manner of the labelling of sample points.
- If  $X = (X_1, \dots, X_n)$  where the  $X_i$ s are a series of independent Bernoulli trials with parameter  $\theta$  then  $f_{\mathcal{X}}(x | \theta) = f_{\mathcal{X}}(x' | \theta)$  if  $x$  and  $x'$  contain the same number of successes.

## Principle 2: Distribution Principle, DP

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

- 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  $\mathcal{E} = \{\mathcal{X}, \Theta, f_{\mathcal{X}}(x | \theta)\}$ . For the bijective  $g : \mathcal{X} \rightarrow \mathcal{Y}$ , let  $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_{\mathcal{Y}}(y | \theta)\}$ , the **same** experiment as  $\mathcal{E}$  but expressed in terms of  $Y = g(X)$ , rather than  $X$ . Then  $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}^g, g(x))$ .

- 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  $f_{\mathcal{X}}(x | \theta) = f_{\mathcal{X}}(x' | \theta)$  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

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

which gives the WIP. □

## The Likelihood Principle

- Consider experiments  $\mathcal{E}_i = \{\mathcal{X}_i, \Theta, f_{\mathcal{X}_i}(x_i | \theta)\}$ ,  $i = 1, 2, \dots$ , where the parameter space  $\Theta$  is the same for each experiment.
- Let  $p_1, p_2, \dots$  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  $\mathcal{E}_1, \mathcal{E}_2, \dots$  according to mixture probabilities  $p_1, p_2, \dots$  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  $i = 1, 2, \dots$  and  $x_i \in \mathcal{X}_i$ , and family of distributions

$$f^*((i, x_i) | \theta) = p_i f_{\mathcal{X}_i}(x_i | \theta).$$

## 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  $\theta$  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  $\theta$  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  $\theta$ . If  $x_1 \in \mathcal{X}_1$  and  $x_2 \in \mathcal{X}_2$  satisfy  $f_{X_1}(x_1 | \theta) = c(x_1, x_2)f_{X_2}(x_2 | \theta)$ , 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  $\text{Ev}(\mathcal{E}_1, x_1) = \text{Ev}(\mathcal{E}_2, x_2)$ .

- 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  $\text{Ev}(\mathcal{E}, x)$  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  $f_{X_1}(x_1 | \theta) = c(x_1, x_2)f_{X_2}(x_2 | \theta)$  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

$$\text{Ev}(\mathcal{E}^*, (1, x_1)) = \text{Ev}(\mathcal{E}^*, (2, x_2)).$$

Applying the **WCP** to each side we infer that

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

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.