

# Statistical Inference

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

Simon Shaw

University of Bath

APTS, 14-18 December 2020

# Introduction

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

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

- 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 something 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  $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x | \theta)\}$ . For the bijective  $g : \mathcal{X} \rightarrow \mathcal{Y}$ , let  $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_Y(y | \theta)\}$ , 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  $f_{\mathcal{X}_1}(x_1 | \theta) = c(x_1, x_2)f_{\mathcal{X}_2}(x_2 | \theta)$ , that is

$$L_{\mathcal{X}_1}(\theta; x_1) = c(x_1, x_2)L_{\mathcal{X}_2}(\theta; x_2)$$

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  $L_{\mathcal{X}}(\theta; x)$ .

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.

# The Sufficiency Principle

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

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

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

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

## Principle 6: Strong Sufficiency Principle, SSP

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

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

### Proof

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

- 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  $X_1, X_2, \dots$  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  $p_j(x_1, \dots, x_j)$ .
  - We assume, resources being finite, that the experiment **must stop** at specified time  $m$ , if it has not stopped already, hence  $p_m(x_1, \dots, x_m) = 0$ .
- The **stopping rule** may then be denoted as  $\tau = (p_1, \dots, p_m)$ . This gives an experiment  $\mathcal{E}^\tau$  with, for  $n = 1, 2, \dots$ ,  $f_n(x_1, \dots, x_n | \theta)$  where consistency requires that

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

# 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**  $x = (T, H, T, T, H, H, T, H, H, H)$ .
- We may feel that the evidence for  $\theta$ , 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

<sup>a</sup> In a sequential experiment  $\mathcal{E}^\tau$ ,  $\text{Ev}(\mathcal{E}^\tau, (x_1, \dots, x_n))$  does not depend on the stopping rule  $\tau$ .

<sup>a</sup>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  $\tau$ , are **irrelevant** for making inferences about  $\theta$ , once the observations  $(x_1, \dots, x_n)$  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  $p_1 = \dots = p_{n-1} = 1$  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  $\tau$  be an arbitrary stopping rule, and consider the outcome  $(x_1, \dots, x_n)$ , which we will denote as  $x_{1:n}$ .

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

Consequently, the probability of this outcome under  $\tau$  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

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

Now observe that this equation has the form

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

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.  $\square$

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  $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y | \theta)\}$  exactly when  $f_{X,Y}$  factorises as

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

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



- 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  $x = (x_1, \dots, x_n)$ ,  $n$  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  $\text{Ev}(\mathcal{E}, (x, y)) = \text{Ev}(\mathcal{E}^{X|Y}, x)$ .

- 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

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

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  $\theta$  given the data  $x$  are represented by  $\pi(\theta | x)$ .
- **Any** inferences made about  $\theta$  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

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

- 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**.

- 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  $T(X)$ ,  $MSE(T | \theta) = Var(T | \theta) + bias(T | \theta)^2$  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

$$f(x_1, x_2 | \theta) \propto \exp\{-\bar{x} - \theta)^2\}.$$

- Consider the alternate model for the **same** parameter  $\theta$

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

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

- Suppose that  $\text{Ev}(\mathcal{E}, x)$  depends on the value of  $f_X(x' | \theta)$  for some  $x' \neq x$ . Then, typically,  $\text{Ev}$  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

$$f_1(x' | \theta) = \alpha f_X(x' | \theta) + \beta f_X(\tilde{x} | \theta)$$

$$f_1(\tilde{x} | \theta) = (1 - \alpha) f_X(x' | \theta) + (1 - \beta) f_X(\tilde{x} | \theta)$$

- ▶ By suitable choice of  $\alpha, \beta$  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  $E_v$  but instead describes desirable properties of  $E_v$ .
- What is evaluated is the algorithm, the method by which  $(\mathcal{E}, x)$  is turned into an inference about the parameter  $\theta$ .
- 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.

# Introduction

- **Statistical Decision Theory** allows us to consider ways to construct the **Ev** function that reflects our needs, which will vary from application to application, and which assesses the consequences of making a **good or bad** inference.
- The set of possible inferences, or **decisions**, is termed the **decision space**, denoted  $\mathcal{D}$ .
- For each  $d \in \mathcal{D}$ , we want a way to assess the consequence of how good or bad the **choice** of decision  $d$  was under the **event**  $\theta$ .

## Definition (Loss function)

A loss function is any function  $L$  from  $\Theta \times \mathcal{D}$  to  $[0, \infty)$ .

- The loss function measures the **penalty** or error,  $L(\theta, d)$  of the **decision**  $d$  when the **parameter** takes the value  $\theta$ .
- Thus, larger values indicate worse consequences.

The three main types of inference about  $\theta$  are

- ① point estimation,
- ② set estimation,
- ③ hypothesis testing.

It is a great conceptual and practical simplification that Statistical Decision Theory **distinguishes** between these three types simply according to their **decision spaces**.

Type of inference	Decision space $\mathcal{D}$
Point estimation	The parameter space, $\Theta$ .
Set estimation	A set of subsets of $\Theta$ .
Hypothesis testing	A specified partition of $\Theta$ , denoted $\mathcal{H}$ .

# Bayesian statistical decision theory

In a Bayesian approach, a **statistical decision problem**  $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$  has the following ingredients.

- 1 The possible values of the parameter:  $\Theta$ , the **parameter space**.
- 2 The set of possible decisions:  $\mathcal{D}$ , the **decision space**.
- 3 The **probability distribution** on  $\Theta$ ,  $\pi(\theta)$ . For example,
  - 1 this could be a **prior** distribution,  $\pi(\theta) = f(\theta)$ .
  - 2 this could be a **posterior** distribution,  $\pi(\theta) = f(\theta | x)$  following the receipt of some **data**  $x$ .
  - 3 this could be a **posterior** distribution  $\pi(\theta) = f(\theta | x, y)$  following the receipt of some **data**  $x, y$ .
- 4 The **loss function**  $L(\theta, d)$ .

In this setting, **only**  $\theta$  is **random** and we can calculate the **expected loss**, or **risk**.

## Definition (Risk)

The **risk** of decision  $d \in \mathcal{D}$  under the distribution  $\pi(\theta)$  is

$$\rho(\pi(\theta), d) = \int_{\theta} L(\theta, d)\pi(\theta) d\theta.$$

We choose  $d$  to **minimise** this risk.

## Definition (Bayes rule and Bayes risk)

The **Bayes risk**  $\rho^*(\pi)$  minimises the expected loss,

$$\rho^*(\pi) = \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

with respect to  $\pi(\theta)$ . A decision  $d^* \in \mathcal{D}$  for which  $\rho(\pi, d^*) = \rho^*(\pi)$  is a **Bayes rule** against  $\pi(\theta)$ .

The Bayes rule may not be unique, and in weird cases it might not exist. We **solve**  $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$  by **finding**  $\rho^*(\pi)$  and (at least one)  $d^*$ .

## Example - quadratic loss

Suppose that  $\Theta \subset \mathbb{R}$  and we wish to find a **point estimate** for  $\theta$ . We consider the loss function  $L(\theta, d) = (\theta - d)^2$ .

- The **risk** of decision  $d$  is

$$\begin{aligned}\rho(\pi, d) &= \mathbb{E}\{L(\theta, d) \mid \theta \sim \pi(\theta)\} = \mathbb{E}_{(\pi)}\{(\theta - d)^2\} \\ &= \mathbb{E}_{(\pi)}(\theta^2) - 2d\mathbb{E}_{(\pi)}(\theta) + d^2,\end{aligned}$$

where  $\mathbb{E}_{(\pi)}(\cdot)$  denotes the expectation with respect to  $\pi(\theta)$ .

- Differentiating with respect to  $d$  we have

$$\frac{\partial}{\partial d}\rho(\pi, d) = -2\mathbb{E}_{(\pi)}(\theta) + 2d.$$

- So, the **Bayes rule** is  $d^* = \mathbb{E}_{(\pi)}(\theta)$ .

## Example - quadratic loss (continued)

- The corresponding Bayes risk is

$$\begin{aligned}
 \rho^*(\pi) &= \rho(\pi, d^*) = \mathbb{E}_{(\pi)}(\theta^2) - 2d^*\mathbb{E}_{(\pi)}(\theta) + (d^*)^2 \\
 &= \text{Var}_{(\pi)}(\theta) + (d^* - \mathbb{E}_{(\pi)}(\theta))^2 \\
 &= \text{Var}_{(\pi)}(\theta)
 \end{aligned}$$

where  $\text{Var}_{(\pi)}(\theta)$  is the variance of  $\theta$  computed with respect to  $\pi(\theta)$ .

- If  $\pi(\theta) = f(\theta)$ , a prior for  $\theta$ , then the Bayes rule of an immediate decision is  $d^* = \mathbb{E}(\theta)$  with corresponding Bayes risk  $\rho^* = \text{Var}(\theta)$ .
- If we observe sample data  $x$  then the Bayes rule given this sample information is  $d^* = \mathbb{E}(\theta | X)$  with corresponding Bayes risk  $\rho^* = \text{Var}(\theta | X)$  as  $\pi(\theta) = f(\theta | x)$ .



- Typically we solve:
  - $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$ , the **immediate decision** problem,
  - $[\Theta, \mathcal{D}, f(\theta | x), L(\theta, d)]$ , the decision problem **after sample information**.
- We may also want to consider the **risk of the sampling procedure**, before observing the sample, to decide whether or not to sample.
- We now consider both  $\theta$  and  $X$  as **random**.
- For each **possible sample**, we need to specify which decision to make.

### Definition (Decision rule)

A decision rule  $\delta(x)$  is a function from  $\mathcal{X}$  into  $\mathcal{D}$ ,

$$\delta : \mathcal{X} \rightarrow \mathcal{D}.$$

If  $X = x$  is the observed value of the sample information then  $\delta(x)$  is the decision that **will be taken**. The collection of all decision rules is denoted by  $\Delta$  so that  $\delta \in \Delta \Rightarrow \delta(x) \in \mathcal{D} \forall x \in \mathcal{X}$ .

- We wish to solve the problem  $[\Theta, \Delta, f(\theta, x), L(\theta, \delta(x))]$ .

### Definition (Bayes (decision) rule and risk of the sampling procedure)

The decision rule  $\delta^*$  is a **Bayes (decision) rule** exactly when

$$\mathbb{E}\{L(\theta, \delta^*(X))\} \leq \mathbb{E}\{L(\theta, \delta(X))\}$$

for all  $\delta(x) \in \mathcal{D}$ . The corresponding risk  $\rho^* = \mathbb{E}\{L(\theta, \delta^*(X))\}$  is termed the **risk of the sampling procedure**.

- If the sample information consists of  $X = (X_1, \dots, X_n)$  then  $\rho^*$  will be a function of  $n$  and so can be used to help determine **sample size choice**.

## Bayes rule theorem, BRT

Suppose that a Bayes rule exists for  $[\Theta, \mathcal{D}, f(\theta | x), L(\theta, d)]$ . Then

$$\delta^*(x) = \arg \min_{d \in \mathcal{D}} \mathbb{E}(L(\theta, d) | X = x).$$

## Proof

Let  $\delta$  be arbitrary. Then

$$\begin{aligned} \mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_{\theta} L(\theta, \delta(x)) f(\theta, x) d\theta dx \\ &= \int_x \int_{\theta} L(\theta, \delta(x)) f(\theta | x) f(x) d\theta dx \\ &= \int_x \left\{ \int_{\theta} L(\theta, \delta(x)) f(\theta | x) d\theta \right\} f(x) dx \\ &= \int_x \mathbb{E}\{L(\theta, \delta(x)) | X\} f(x) dx \end{aligned}$$

## Proof continued

Now, as  $f(x) > 0$ , the  $\delta^* \in \Delta$  which minimises  $\mathbb{E}\{L(\theta, \delta(X))\}$  may equivalently be found as the  $\delta^*$  which satisfies

$$\rho(f(\theta), \delta^*) = \inf_{\delta(x) \in \mathcal{D}} \mathbb{E}\{L(\theta, \delta(x)) | X\},$$

giving the result. □

- The minimisation of expected loss over the space of **all** functions from  $\mathcal{X}$  to  $\mathcal{D}$  can be achieved by the **pointwise minimisation** over  $\mathcal{D}$  of the expected loss **conditional** on  $X = x$ .
- The risk of the sampling procedure is  $\rho^* = \mathbb{E}[\mathbb{E}\{L(\theta, \delta^*(x)) | X\}]$ .

## Example - quadratic loss

We have  $\delta^* = \mathbb{E}(\theta | X)$  and  $\rho^* = \mathbb{E}\{\text{Var}(\theta | X)\}$ .

We could consider  $\Delta$ , the **set of decision rules**, to be our possible **set of inferences** about  $\theta$  when the sample is observed so that  $Ev(\mathcal{E}, x)$  is  $\delta^*(x)$ . We thus have the following result.

### Theorem

The Bayes rule for the posterior decision respects the strong likelihood principle.

### Proof

If we have two Bayesian models with the **same** prior distribution then if  $f_{X_1}(x_1 | \theta) = c(x_1, x_2)f_{X_2}(x_2 | \theta)$  the corresponding posterior distributions are the **same** and so the corresponding Bayes rule (and risk) is the same.  $\square$

# Admissible rules

- Bayes rules rely upon a **prior distribution** for  $\theta$ : the risk is a function of  $d$  only.
- In **classical statistics**, there is **no distribution** for  $\theta$  and so another approach is needed.

## Definition (The classical risk)

For a decision rule  $\delta(x)$ , the classical risk for the model  $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x | \theta)\}$  is

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f_X(x | \theta) dx.$$

- The classical risk is thus, for each  $\delta$ , a **function** of  $\theta$ .

## Example

Let  $X = (X_1, \dots, X_n)$  where  $X_i \sim N(\theta, \sigma^2)$  and  $\sigma^2$  is known. Suppose that  $L(\theta, d) = (\theta - d)^2$  and consider a conjugate prior  $\theta \sim N(\mu_0, \sigma_0^2)$ . Possible decision functions include:

- 1  $\delta_1(x) = \bar{x}$ , the **sample mean**.
- 2  $\delta_2(x) = \text{med}\{x_1, \dots, x_n\} = \tilde{x}$ , the **sample median**.
- 3  $\delta_3(x) = \mu_0$ , the **prior mean**.
- 4  $\delta_4(x) = \mu_n$ , the **posterior mean** where

$$\mu_n = \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \left( \frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2} \right),$$

the weighted average of the prior and sample mean accorded to their respective precisions.

## Example - continued

The respective classical risks are

- ①  $R(\theta, \delta_1) = \frac{\sigma^2}{n}$ , a **constant** for  $\theta$ , since  $\bar{X} \sim N(\theta, \sigma^2/n)$ .
- ②  $R(\theta, \delta_2) = \frac{\pi\sigma^2}{2n}$ , a **constant** for  $\theta$ , since  $\tilde{X} \sim N(\theta, \pi\sigma^2/2n)$  (approximately).
- ③  $R(\theta, \delta_3) = (\theta - \mu_0)^2 = \sigma_0^2 \left( \frac{\theta - \mu_0}{\sigma_0} \right)^2$ .
- ④  $R(\theta, \delta_4) = \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-2} \left\{ \frac{1}{\sigma_0^2} \left( \frac{\theta - \mu_0}{\sigma_0} \right)^2 + \frac{n}{\sigma^2} \right\}$ .

Which decision do we choose? We observe that  $R(\theta, \delta_1) < R(\theta, \delta_2)$  for **all**  $\theta \in \Theta$  but other comparisons depend upon  $\theta$ .

- The accepted approach for classical statisticians is to narrow the set of possible decision rules by **ruling out** those that are obviously **bad**.



## Definition (Admissible decision rule)

A decision rule  $\delta_0$  is **inadmissible** if there exists a decision rule  $\delta_1$  which **dominates** it, that is

$$R(\theta, \delta_1) \leq R(\theta, \delta_0)$$

for all  $\theta \in \Theta$  with  $R(\theta, \delta_1) < R(\theta, \delta_0)$  for **at least one** value  $\theta_0 \in \Theta$ . If no such  $\delta_1$  exists then  $\delta_0$  is **admissible**.

- If  $\delta_0$  is **dominated** by  $\delta_1$  then the classical risk of  $\delta_0$  is **never smaller** than that of  $\delta_1$  and  $\delta_1$  has a **smaller** risk for  $\theta_0$ .
- Thus, you would **never** want to use  $\delta_0$ .<sup>1</sup>
- The accepted approach is to **reduce** the set of possible decision rules under consideration by only **using admissible rules**.

---

<sup>1</sup>Here I am assuming that all other considerations are the same in the two cases: e.g. for all  $x \in \mathcal{X}$ ,  $\delta_1(x)$  and  $\delta_0(x)$  take about the same amount of resource to compute. ↻ 🔍

- We now show that **admissible rules** can be related to a **Bayes rule**  $\delta^*$  for a **prior distribution**  $\pi(\theta)$ .

## Theorem

If a prior distribution  $\pi(\theta)$  is strictly positive for all  $\Theta$  with finite Bayes risk and the classical risk,  $R(\theta, \delta)$ , is a continuous function of  $\theta$  for all  $\delta$ , then the **Bayes rule**  $\delta^*$  is **admissible**.

## Proof (Robert, 2007)

Letting  $f(\theta, x) = f_X(x | \theta)\pi(\theta)$  we have

$$\begin{aligned}\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_{\theta} L(\theta, \delta(x)) f(\theta, x) d\theta dx \\ &= \int_{\theta} \left\{ \int_x L(\theta, \delta(x)) f_X(x | \theta) dx \right\} \pi(\theta) d\theta \\ &= \int_{\theta} R(\theta, \delta) \pi(\theta) d\theta\end{aligned}$$

## Proof continued

- Suppose that the Bayes rule  $\delta^*$  is inadmissible and dominated by  $\delta_1$ .
- Thus, in an open set  $C$  of  $\theta$ ,  $R(\theta, \delta_1) < R(\theta, \delta^*)$  with  $R(\theta, \delta_1) \leq R(\theta, \delta^*)$  elsewhere.
- Consequently,  $\mathbb{E}\{L(\theta, \delta_1(X))\} < \mathbb{E}\{L(\theta, \delta^*(X))\}$  which is a contradiction to  $\delta^*$  being the Bayes rule. □

- The relationship between a Bayes rule with prior  $\pi(\theta)$  and an admissible decision rule is even stronger.
- The following result was derived by [Abraham Wald \(1902-1950\)](#)

## Wald's Complete Class Theorem, CCT

In the case where the parameter space  $\Theta$  and sample space  $\mathcal{X}$  are finite, a decision rule  $\delta$  is admissible if and only if it is a Bayes rule for some prior distribution  $\pi(\theta)$  with strictly positive values.

- An illuminating blackboard proof of this result can be found in [Cox and Hinkley \(1974, Section 11.6\)](#).
- There are [generalisations](#) of this theorem to non-finite decision sets, parameter spaces, and sample spaces but the results are [highly technical](#).
- We'll proceed [assuming](#) the more general result, which is that [a decision rule is admissible if and only if it is a Bayes rule for some prior distribution  \$\pi\(\theta\)\$](#) , which holds for practical purposes.

So what does the CCT say?

- 1 [Admissible decision rules respect the SLP](#). This follows from the fact that admissible rules are Bayes rules which respect the SLP. This provides support for using admissible decision rules.
- 2 If you select a [Bayes rule](#) according to some positive prior distribution  $\pi(\theta)$  then you [cannot](#) ever choose an [inadmissible](#) decision rule.

## Point estimation

- We now look at possible choices of loss functions for different types of inference.
- For **point estimation** the decision space is  $\mathcal{D} = \Theta$ , and the loss function  $L(\theta, d)$  represents the (negative) consequence of choosing  $d$  as a **point estimate** of  $\theta$ .
- It will not be often that an obvious loss function  $L : \Theta \times \Theta \rightarrow \mathbb{R}$  presents itself. There is a need for a **generic** loss function which is acceptable over a **wide range** of applications.

Suppose that  $\Theta$  is a **convex subset** of  $\mathbb{R}^P$ . A natural choice is a **convex loss function**,

$$L(\theta, d) = h(d - \theta)$$

where  $h : \mathbb{R}^P \rightarrow \mathbb{R}$  is a smooth non-negative convex function with  $h(0) = 0$ .

- This type of loss function asserts that small errors are much more tolerable than large ones.
- One possible further restriction is that  $h$  is an **even function**,  $h(d - \theta) = h(\theta - d)$ .
- In this case,  $L(\theta, \theta + \epsilon) = L(\theta, \theta - \epsilon)$  so that **under-estimation** incurs the **same** loss as **over-estimation**.
- We saw previously, that for **quadratic loss**  $\Theta \subset \mathbb{R}$ ,  $L(\theta, d) = (\theta - d)^2$ , the Bayes rule was the **expectation** of  $\pi(\theta)$ . As we will see, this attractive feature can be extended to more dimensions.
- There are many situations where this is **not** appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.

The **bilinear loss function** for  $\Theta \subset \mathbb{R}$  is, for  $\alpha, \beta > 0$ ,

$$L(\theta, d) = \begin{cases} \alpha(\theta - d) & \text{if } d \leq \theta, \\ \beta(d - \theta) & \text{if } d \geq \theta. \end{cases}$$

- The Bayes rule is a  $\frac{\alpha}{\alpha+\beta}$ -**fractile** of  $\pi(\theta)$ .
- If  $\alpha = \beta = 1$  then  $L(\theta, d) = |\theta - d|$ , the **absolute loss** which gives a Bayes rule of the **median** of  $\pi(\theta)$ .
- $|\theta - d|$  is smaller than  $(\theta - d)^2$  for  $|\theta - d| < 1$  and so absolute loss is smaller than quadratic loss for large deviations. Thus, it takes less account of the tails of  $\pi(\theta)$  leading to the choice of the median.
- If  $\alpha > \beta$ , so  $\frac{\alpha}{\alpha+\beta} > 0.5$ , then under-estimation is penalised more than over-estimation and so that Bayes rule is more likely to be an over-estimate.

## Example

If  $\Theta \in \mathbb{R}^p$ , the Bayes rule  $\delta^*$  associated with the distribution  $\pi(\theta)$  and the quadratic loss

$$L(\theta, d) = (d - \theta)^T Q (d - \theta)$$

is the expectation  $\mathbb{E}_{(\pi)}(\theta)$  for every positive-definite symmetric  $p \times p$  matrix  $Q$ .

## Example (Robert, 2007), $Q = \Sigma^{-1}$

Suppose  $X \sim N_p(\theta, \Sigma)$  where the known variance matrix  $\Sigma$  is diagonal with elements  $\sigma_i^2$  for each  $i$ . Then  $\mathcal{D} = \mathbb{R}^p$ . A possible loss function is

$$L(\theta, d) = \sum_{i=1}^p \left( \frac{d_i - \theta_i}{\sigma_i} \right)^2$$

so that the total loss is the sum of the squared component-wise errors.



- As the Bayes rule for  $L(\theta, d) = (d - \theta)^T Q (d - \theta)$  does not depend upon  $Q$ , it is the same for an uncountably large class of loss functions.
- If we apply the Complete Class Theorem to this result we see that for quadratic loss, a point estimator for  $\theta$  is admissible if and only if it is the conditional expectation with respect to some positive prior distribution  $\pi(\theta)$ .
- The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

# Set estimation

- For set estimation the **decision space** is a **set of subsets** of  $\Theta$  so that each  $d \subset \Theta$ .
- There are two contradictory requirements for set estimators of  $\Theta$ .
  - 1 We want the sets to be small.
  - 2 We also want them to contain  $\theta$ .
- A simple way to represent these two requirements is to consider the loss function

$$L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$$

for some  $\kappa > 0$  where  $|d|$  is the **volume** of  $d$ .

- The value of  $\kappa$  controls the **trade-off** between the two requirements.
  - ▶ If  $\kappa \downarrow 0$  then minimising the expected loss will always produce the **empty set**.
  - ▶ If  $\kappa \uparrow \infty$  then minimising the expected loss will always produce  $\Theta$ .

- For loss functions of the form  $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$  we'll show there is a simple necessary condition for a rule to be a Bayes rule.

### Definition (Level set)

A set  $d \subset \Theta$  is a **level set** of the posterior distribution exactly when  $d = \{\theta : \pi(\theta | x) \geq k\}$  for some  $k$ .

### Theorem (Level set property, LSP)

If  $\delta^*$  is a **Bayes rule** for  $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$  then it is a **level set** of the posterior distribution.

### Proof

Note that

$$\begin{aligned}\mathbb{E}\{L(\theta, d) | X\} &= |d| + \kappa(1 - \mathbb{E}(\mathbb{1}_{\theta \in d} | X)) \\ &= |d| + \kappa\mathbb{P}(\theta \notin d | X).\end{aligned}$$

## Proof continued

- For fixed  $x$ , we show that if  $d$  is **not** a level set of the posterior distribution then there is a  $d' \neq d$  which has a **smaller** expected loss so that  $\delta^*(x) \neq d$ .
- Suppose that  $d$  is **not a level set** of  $\pi(\theta | x)$ . Then there is a  $\theta \in d$  and  $\theta' \notin d$  for which  $\pi(\theta' | x) > \pi(\theta | x)$ .
- Let  $d' = d \cup d\theta' \setminus d\theta$  where  $d\theta$  is the tiny region of  $\Theta$  around  $\theta$  and  $d\theta'$  is the tiny region of  $\Theta$  around  $\theta'$  for which  $|d\theta| = |d\theta'|$ .
- Then  $|d'| = |d|$  but

$$\mathbb{P}(\theta \notin d' | X) < \mathbb{P}(\theta \notin d | X)$$

Thus,  $\mathbb{E}\{L(\theta, d') | X\} < \mathbb{E}\{L(\theta, d) | X\}$  showing that  $\delta^*(x) \neq d$ .  $\square$

- The **Level Set Property Theorem** states that  $\delta$  having the level set property is **necessary** for  $\delta$  to be a **Bayes rule** for loss functions of the form  $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ .
- The **Complete Class Theorem** states that being a **Bayes rule** is a **necessary** condition for  $\delta$  to be **admissible**.
- Being a **level set of a posterior** distribution for **some prior** distribution  $\pi(\theta)$  is a **necessary** condition for being **admissible** for loss functions of this form.
- **Bayesian HPD regions** satisfy the necessary condition for being a set estimator.
- **Classical set estimators** achieve a similar outcome if they are **level sets of the likelihood function**, because the posterior is proportional to the likelihood under a uniform prior distribution.<sup>2</sup>

---

<sup>2</sup>In the case where  $\Theta$  is unbounded, this prior distribution may have to be truncated to be proper.

# Hypothesis tests

- For hypothesis tests, the decision space is a **partition** of  $\Theta$ , denoted

$$\mathcal{H} := \{H_0, H_1, \dots, H_d\}.$$

- Each element of  $\mathcal{H}$  is termed a **hypothesis**.
- The loss function  $L(\theta, H_i)$  represents the (negative) consequences of choosing element  $H_i$ , when the true value of the parameter is  $\theta$ .
- It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(\theta, H_i) = \min_j L(\theta, H_j)$$

on the grounds that an **incorrect** choice of element **should never** incur a **smaller** loss than the **correct** choice.

- Consider the test of  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$  where  $\Theta_1 = \Theta \setminus \Theta_0$ . Let  $\mathcal{D} = \{d_0, d_1\}$  where  $d_i$  corresponds to accepting  $H_i$ . A generic loss function is the 0-1 ('zero-one') loss function

$$L(\theta, d_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ 1 & \text{if } \theta \notin \Theta_i. \end{cases}$$

- The classical risk is the probability of making a wrong decision,

$$R(\theta, \delta) = \begin{cases} \mathbb{P}(\delta(X) = d_1 \mid \theta) & \text{if } \theta \in \Theta_0, \\ \mathbb{P}(\delta(X) = d_0 \mid \theta) & \text{if } \theta \in \Theta_1, \end{cases}$$

which correspond to the familiar Type I and Type II errors.

- The Bayes rule is to choose  $H_0$  if  $\mathbb{P}_\pi(\theta \in \Theta_0) > \mathbb{P}_\pi(\theta \in \Theta_1)$  and  $H_1$  otherwise, where  $\mathbb{P}_\pi(\cdot)$  is the probability when  $\theta \sim \pi(\theta)$ .
- Hence, if  $\pi(\theta) = f(\theta \mid x)$ , the Bayes rule is to choose the hypothesis with the largest posterior probability.

- This approach can be naturally extended to multiple hypotheses  $\mathcal{H} = \{H_0, H_1, \dots, H_d\}$  which partition  $\Theta$  by taking

$$L(\theta, H_i) = 1 - \mathbb{1}_{\{\theta \in H_i\}}.$$

i.e., zero if  $\theta \in H_i$ , and one if it is not.

- For the posterior decision, the **Bayes rule** is to select the hypothesis with the **largest posterior probability**.
- However, this loss function is hard to defend as being realistic.
- If we choose  $H_i$  and it turns out that  $\theta \notin H_i$  then the inference is wrong and the loss is the same irrespective of where  $\theta$  lies.
- An alternative approach is to co-opt the theory of **set estimators**.
- The statistician can use her set estimator  $\delta$  to make at least some distinctions between the members of  $\mathcal{H}$ :
  - ▶ **Accept**  $H_i$  exactly when  $\delta(x) \subset H_i$ ,
  - ▶ **Reject**  $H_i$  exactly when  $\delta(x) \cap H_i = \emptyset$ ,
  - ▶ **Undecided** about  $H_i$  otherwise.



# Confidence procedures and confidence sets

- We consider **interval estimation**, or more generally **set estimation**.
- Under the model  $\mathcal{E} = \{\mathcal{X}, \Theta, f_{\mathcal{X}}(x | \theta)\}$ , for given data  $\mathbf{X} = \mathbf{x}$ , we wish to construct a set  $\mathbf{C} = \mathbf{C}(\mathbf{x}) \subset \Theta$  and the **inference** is the statement that  $\theta \in \mathbf{C}$ .
- If  $\theta \in \mathbb{R}$  then the set estimate is typically an **interval**.

## Definition (Confidence procedure)

A **random set**  $\mathbf{C}(\mathbf{X})$  is a level- $(1 - \alpha)$  **confidence procedure** exactly when

$$\mathbb{P}(\theta \in \mathbf{C}(\mathbf{X}) | \theta) \geq 1 - \alpha$$

for all  $\theta \in \Theta$ .  $\mathbf{C}$  is an **exact** level- $(1 - \alpha)$  confidence procedure if the probability **equals**  $(1 - \alpha)$  for all  $\theta$ .

- The value  $\mathbb{P}(\theta \in C(X) | \theta)$  is termed the **coverage** of  $C$  at  $\theta$ .
- Exact is a special case: typically  $\mathbb{P}(\theta \in C(X) | \theta)$  will depend upon  $\theta$ .
- The procedure is thus **conservative**: for a given  $\theta_0$  the **coverage** may be much **higher** than  $(1 - \alpha)$ .

### Uniform example

- Let  $X_1, \dots, X_n$  be independent and identically distributed  $\text{Unif}(0, \theta)$  random variables where  $\theta > 0$ . Let  $Y = \max\{X_1, \dots, X_n\}$ .
- We consider two possible sets:  $(aY, bY)$  where  $1 \leq a < b$  and  $(Y + c, Y + d)$  where  $0 \leq c < d$ .
  - 1  $\mathbb{P}(\theta \in (aY, bY) | \theta) = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n$ . Thus, the coverage probability of the interval **does not depend** upon  $\theta$ .
  - 2  $\mathbb{P}(\theta \in (Y + c, Y + d) | \theta) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n$ . In this case, the coverage probability of the interval **does depend** upon  $\theta$ .

- We distinguish between the confidence **procedure**  $C$ , which is a **random interval** and so a function for each possible  $x$ , and the result when  $C$  is **evaluated** at the **observation**  $x$ , which is a **set** in  $\Theta$ .

### Definition (Confidence set)

The observed  $C(x)$  is a level- $(1 - \alpha)$  confidence set exactly when the random  $C(X)$  is a level- $(1 - \alpha)$  confidence procedure.

- If  $\Theta \subset \mathbb{R}$  and  $C(x)$  is **convex**, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value.
- The **challenge** with confidence procedures is to construct one with a **specified level**: to do this we **start with the level** and then construct a  $C$  guaranteed to have this level.

## Definition (Family of confidence procedures)

- $C(X; \alpha)$  is a **family** of confidence procedures exactly when  $C(X; \alpha)$  is a level- $(1 - \alpha)$  confidence procedure for **every**  $\alpha \in [0, 1]$ .
- $C$  is a **nesting family** exactly when  $\alpha < \alpha'$  implies that  $C(x; \alpha') \subset C(x; \alpha)$ .
- If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

# Constructing confidence procedures

- The general approach to construct a confidence procedure is to **invert a test statistic**.
- In the Uniform example, the coverage of the procedure  $(aY, bY)$  does not depend upon  $\theta$  because the coverage probability could be expressed in terms of  $T = Y/\theta$  where the distribution of  $T$  did **not depend** upon  $\theta$ .
  - ▶  $T$  is an example of a **pivot** and confidence procedures are straightforward to compute from a pivot.
  - ▶ However, a drawback to this approach in general is that there is **no hard and fast method** for finding a pivot.
- An alternate method which does work generally is to exploit the property that **every confidence procedure** corresponds to a **hypothesis test** and vice versa.

Consider a hypothesis test where we have to decide either to **accept** that an hypothesis  $H_0$  is true or to **reject**  $H_0$  in favour of an alternative hypothesis  $H_1$  based on a sample  $x \in \mathcal{X}$ .

- The set of  $x$  for which  $H_0$  is rejected is called the **rejection region**.
- The complement, where  $H_0$  is accepted, is the **acceptance region**.
- A hypothesis test can be constructed from **any statistic**  $T = T(X)$ .

### Definition (Likelihood Ratio Test, LRT)

The likelihood ratio test (LRT) statistic for testing  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$ , where  $\Theta_0 \cup \Theta_0^c = \Theta$ , is

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L_X(\theta; x)}{\sup_{\theta \in \Theta} L_X(\theta; x)}.$$

A LRT at significance level  $\alpha$  has a **rejection region** of the form  $\{x : \lambda(x) \leq c\}$  where  $0 \leq c \leq 1$  is chosen so that  $\mathbb{P}(\text{Reject } H_0 \mid \theta) \leq \alpha$  for all  $\theta \in \Theta_0$ .

## Example

- Let  $X = (X_1, \dots, X_n)$  and suppose that the  $X_i$  are independent and identically distributed  $N(\theta, \sigma^2)$  random variables where  $\sigma^2$  is known.
- Consider the likelihood ratio test for  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ . Then, as the maximum likelihood estimate (mle) of  $\theta$  is  $\bar{x}$ ,

$$\begin{aligned} \lambda(\mathbf{x}) &= \frac{L_X(\theta_0; \mathbf{x})}{L_X(\bar{x}; \mathbf{x})} = \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n ((x_i - \theta_0)^2 - (x_i - \bar{x})^2) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} n(\bar{x} - \theta_0)^2 \right\}. \end{aligned}$$

Notice that, under  $H_0$ ,  $\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} \sim N(0, 1)$  so that

$$-2 \log \lambda(X) = \frac{n(\bar{X} - \theta_0)^2}{\sigma^2} \sim \chi_1^2,$$

the chi-squared distribution with one degree of freedom.

## Example continued

- The **rejection region** is  $\{x : \lambda(x) \leq c\} = \{x : -2 \log \lambda(x) \geq k\}$ .
- Setting  $k = \chi_{1,\alpha}^2$ , where  $\mathbb{P}(\chi_1^2 \geq \chi_{1,\alpha}^2) = \alpha$ , gives a test at the **exact** significance level  $\alpha$ .

The **acceptance region** of this test is  $\{x : -2 \log \lambda(x) < \chi_{1,\alpha}^2\}$  where

$$\mathbb{P}\left(\frac{n(\bar{X} - \theta_0)^2}{\sigma^2} < \chi_{1,\alpha}^2 \mid \theta = \theta_0\right) = 1 - \alpha.$$

This holds for all  $\theta_0$  and so, additionally rearranging,

$$\mathbb{P}\left(\bar{X} - \sqrt{\chi_{1,\alpha}^2} \frac{\sigma}{\sqrt{n}} < \theta < \bar{X} + \sqrt{\chi_{1,\alpha}^2} \frac{\sigma}{\sqrt{n}} \mid \theta\right) = 1 - \alpha.$$

Thus,  $C(X) = (\bar{X} - \sqrt{\chi_{1,\alpha}^2} \frac{\sigma}{\sqrt{n}}, \bar{X} + \sqrt{\chi_{1,\alpha}^2} \frac{\sigma}{\sqrt{n}})$  is an **exact** level- $(1 - \alpha)$  **confidence procedure** with  $C(x)$  the corresponding confidence set.



- Note that we obtained the level- $(1 - \alpha)$  **confidence procedure** by **inverting** the **acceptance region** of the level  $\alpha$  **significance test**.
- This correspondence, or **duality**, between acceptance regions of tests and confidence sets is a **general property**.

### Theorem (Duality of Acceptance Regions and Confidence Sets)

- 1 For each  $\theta_0 \in \Theta$ , let  $A(\theta_0)$  be the **acceptance region** of a test of  $H_0 : \theta = \theta_0$  at significance level  $\alpha$ . For each  $x \in \mathcal{X}$ , define  $C(x) = \{\theta_0 : x \in A(\theta_0)\}$ . Then  $C(X)$  is a level- $(1 - \alpha)$  **confidence procedure**.
- 2 Let  $C(X)$  be a level- $(1 - \alpha)$  **confidence procedure** and, for any  $\theta_0 \in \Theta$ , define  $A(\theta_0) = \{x : \theta_0 \in C(x)\}$ . Then  $A(\theta_0)$  is the **acceptance region** of a test of  $H_0 : \theta = \theta_0$  at **significance level**  $\alpha$ .

## Proof

- ① As we have a level  $\alpha$  test for each  $\theta_0 \in \Theta$  then  $\mathbb{P}(X \in A(\theta_0) | \theta = \theta_0) \geq 1 - \alpha$ . Since  $\theta_0$  is arbitrary we may write  $\theta$  instead of  $\theta_0$  and so, for all  $\theta \in \Theta$ ,

$$\mathbb{P}(\theta \in C(X) | \theta) = \mathbb{P}(X \in A(\theta) | \theta) \geq 1 - \alpha.$$

Hence,  $C(X)$  is a level- $(1 - \alpha)$  confidence procedure.

- ② For a test of  $H_0 : \theta = \theta_0$ , the probability of a Type I error (rejecting  $H_0$  when it is true) is

$$\mathbb{P}(X \notin A(\theta_0) | \theta = \theta_0) = \mathbb{P}(\theta_0 \notin C(X), | \theta = \theta_0) \leq \alpha$$

since  $C(X)$  is a level- $(1 - \alpha)$  confidence procedure. Hence, we have a test at significance level  $\alpha$ . □

A possibly easier way to understand the relationship between significance tests and confidence sets is by defining the set  $\{(x, \theta) : (x, \theta) \in \tilde{C}\}$  in the space  $\mathcal{X} \times \Theta$  where  $\tilde{C}$  is also a set in  $\mathcal{X} \times \Theta$ .

- For fixed  $x$ , define the confidence set as  $C(x) = \{\theta : (x, \theta) \in \tilde{C}\}$ .
- For fixed  $\theta$ , define the acceptance region as  $A(\theta) = \{x : (x, \theta) \in \tilde{C}\}$ .

### Example revisited

Letting  $x = (x_1, \dots, x_n)$ , with  $z_{\alpha/2}^2 = \chi_{1,\alpha}^2$ , define the set

$$\{(x, \theta) : (x, \theta) \in \tilde{C}\} = \{(x, \theta) : -z_{\alpha/2}\sigma/\sqrt{n} < \bar{x} - \theta < z_{\alpha/2}\sigma/\sqrt{n}\}.$$

The confidence set is then

$$C(x) = \{\theta : \bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \theta < \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\}$$

and acceptance region

$$A(\theta) = \{x : \theta - z_{\alpha/2}\sigma/\sqrt{n} < \bar{x} < \theta + z_{\alpha/2}\sigma/\sqrt{n}\}.$$

## Good choices of confidence procedures

- In the previous chapter, we showed that, under the generic loss  $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ , a necessary condition for admissibility was that  $d$  was a **level set** of the **posterior** distribution.
- We now proceed by consider confidence procedures that satisfy a **level set** property for the **likelihood**  $L_X(\theta; x) = f_X(x | \theta)$ .

### Definition (Level set property, LSP)

A confidence procedure  $C$  has the level set property exactly when

$$C(x) = \{\theta : f_X(x | \theta) > g(x)\}$$

for some  $g : \mathcal{X} \rightarrow \mathbb{R}$ .

We now show that we can construct a family of confidence procedures with the LSP. The result has pedagogic value, because it can be used to generate an uncountable number of families of confidence procedures, each with the level set property.

## Theorem

Let  $h$  be any probability density function for  $X$ . Then

$$C_h(x; \alpha) := \{\theta \in \Theta : f_X(x | \theta) > \alpha h(x)\}$$

is a family of confidence procedures, with the LSP.

## Proof

First notice that if we let  $\mathcal{X}(\theta) := \{x \in \mathcal{X} : f_X(x | \theta) > 0\}$  then

$$\begin{aligned} \mathbb{E}(h(X)/f_X(X | \theta) | \theta) &= \int_{x \in \mathcal{X}(\theta)} \frac{h(x)}{f_X(x | \theta)} f_X(x | \theta) dx \\ &= \int_{x \in \mathcal{X}(\theta)} h(x) \leq 1 \end{aligned}$$

because  $h$  is a probability density function.

## Proof continued

Now,

$$\mathbb{P}(f_X(X|\theta)/h(X) \leq u | \theta) = \mathbb{P}(h(X)/f_X(X|\theta) \geq 1/u | \theta) \quad (2)$$

$$\leq \frac{\mathbb{E}(h(X)/f_X(X|\theta) | \theta)}{1/u} \quad (3)$$

$$\leq \frac{1}{1/u} = u$$

where (3) follows from (2) by [Markov's inequality](#).<sup>a</sup> □

---

<sup>a</sup>If  $X$  is a nonnegative random variable and  $a > 0$  then  $\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$ .


- If we let  $g(x; \theta) = f_X(x|\theta)/h(x)$ , which may be infinite, then  $\mathbb{P}(g(X; \theta) \leq u | \theta) \leq u$ .
- We will see later that this implies that  $g(x; \theta)$  is [super-uniform](#).

- Among the interesting choices for  $h$ , one possibility is  $h(x) = f_X(x | \theta_0)$ , for some  $\theta_0 \in \Theta$ .
- As  $f_X(x | \theta_0) > \alpha f_X(x | \theta_0)$  we can **construct** a level- $(1 - \alpha)$  **confidence procedure** whose confidence sets will **always** contain  $\theta_0$ .
- This suggests an issue with confidence procedures: two statisticians may come to two **different** conclusions about  $H_0 : \theta = \theta_0$  depending on the intervals **they construct**.
- This illustrates why it is important to be able to **account** for the **choices** you make as a statistician.
- The theorem utilises Markov's Inequality which is a **very slack** result. It is likely that the **coverage** of the corresponding family of confidence procedures will be **much larger** than  $(1 - \alpha)$ .
- A more desirable strategy would be to use an **exact family** of confidence procedures which satisfy the **LSP**, if one existed.

# The linear model

- We'll briefly discuss the **linear model** and construct an **exact family** of confidence procedures which satisfy the **LSP**.
- Let  $Y = (Y_1, \dots, Y_n)$  be an  $n$ -vector of observables with  $Y = X\theta + \epsilon$ .
  - ▶  $X$  is an  $(n \times p)$  matrix<sup>3</sup> of **regressors**,
  - ▶  $\theta$  is a  $p$ -vector of **regression coefficients**,
  - ▶  $\epsilon$  is an  $n$ -vector of **residuals**.
- Assume that  $\epsilon \sim N_n(0, \sigma^2 I_n)$ , the  $n$ -dimensional **multivariate normal** distribution, where  $\sigma^2$  is **known** and  $I_n$  is the  $(n \times n)$  **identity matrix**.
- From properties of the multivariate normal distribution, it follows that  $Y \sim N_n(X\theta, \sigma^2 I_n)$ .

---

<sup>3</sup>We typically use  $X$  to denote a generic random variable and so it is not ideal to use it here for a specified matrix but this is the standard notation for `linear_models`. 



Now,

$$L_Y(\theta; y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\theta)^T (y - X\theta) \right\}.$$

Let  $\hat{\theta} = \hat{\theta}(y) = (X^T X)^{-1} X^T y$  then

$$\begin{aligned} (y - X\theta)^T (y - X\theta) &= (y - X\hat{\theta} + X\hat{\theta} - X\theta)^T (y - X\hat{\theta} + X\hat{\theta} - X\theta) \\ &= (y - X\hat{\theta})^T (y - X\hat{\theta}) + (X\hat{\theta} - X\theta)^T (X\hat{\theta} - X\theta) \\ &= (y - X\hat{\theta})^T (y - X\hat{\theta}) + (\hat{\theta} - \theta)^T X^T X (\hat{\theta} - \theta). \end{aligned}$$

Thus,  $(y - X\theta)^T (y - X\theta)$  is **minimised** when  $\theta = \hat{\theta}$  and so,  $\hat{\theta} = (X^T X)^{-1} X^T y$  is the **mle** of  $\theta$ . The likelihood ratio is

$$\begin{aligned} \lambda(y) &= \frac{L_Y(\theta; y)}{L_Y(\hat{\theta}; y)} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left[ (y - X\theta)^T (y - X\theta) - (y - X\hat{\theta})^T (y - X\hat{\theta}) \right] \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (\hat{\theta} - \theta)^T X^T X (\hat{\theta} - \theta) \right\} \end{aligned}$$

- Thus,  $-2 \log \lambda(y) = \frac{1}{\sigma^2} (\hat{\theta} - \theta)^T X^T X (\hat{\theta} - \theta)$ .
- As  $\hat{\theta}(Y) = (X^T X)^{-1} X^T Y$  then, as  $Y \sim N_n(X\theta, \sigma^2 I_n)$ ,

$$\hat{\theta}(Y) \sim N_p \left( \theta, \sigma^2 (X^T X)^{-1} \right)$$

- Consequently,  $-2 \log \lambda(Y) \sim \chi_p^2$ .

Hence, with  $\mathbb{P}(\chi_p^2 \geq \chi_{p,\alpha}^2) = \alpha$ ,

$$\begin{aligned} C(y; \alpha) &= \left\{ \theta \in \mathbb{R}^p : -2 \log \lambda(y) = -2 \log \frac{f_Y(y | \theta, \sigma^2)}{f_Y(y | \hat{\theta}, \sigma^2)} < \chi_{p,\alpha}^2 \right\} \\ &= \left\{ \theta \in \mathbb{R}^p : f_Y(y | \theta, \sigma^2) > \exp \left( -\frac{\chi_{p,\alpha}^2}{2} \right) f_Y(y | \hat{\theta}, \sigma^2) \right\} \end{aligned}$$

is a family of **exact confidence procedures** for  $\theta$  which has the **LSP**.

## Wilks confidence procedures

- This outcome, where we can find a family of exact confidence procedures with the LSP, is **more-or-less unique** to the regression parameters of the **linear model**.
- It is however found, **approximately**, in the **large  $n$**  behaviour of a much wider class of models.

### Wilks' Theorem

Let  $X = (X_1, \dots, X_n)$  where each  $X_i$  is independent and identically distributed,  $X_i \sim f(x_i | \theta)$ , where  $f$  is a **regular model** and the **parameter space**  $\Theta$  is an open convex subset of  $\mathbb{R}^p$  (and invariant to  $n$ ). The distribution of the statistic  $-2 \log \lambda(X)$  converges to a **chi-squared** distribution with  $p$  degrees of freedom as  $n \rightarrow \infty$ .

- A working guideline to regular model is that  $f$  must be smooth and differentiable in  $\theta$ ; in particular, the support must not depend on  $\theta$ .

- The result dates back to Wilks (1938) and, as such, the resultant confidence procedures are often termed **Wilks confidence procedures**.
- Thus, if the conditions of Wilks' Theorem are met,

$$C(x; \alpha) = \left\{ \theta \in \mathbb{R}^p : f_X(x | \theta) > \exp\left(-\frac{\chi_{p,\alpha}^2}{2}\right) f_X(x | \hat{\theta}) \right\}$$

is a family of **approximately exact** confidence procedures which satisfy the LSP.

- For a given model, the pertinent question is whether or not the approximation is a good one.
- We are thus interested in the **level error**, the difference between the **nominal level**, typically  $(1 - \alpha)$  everywhere, and the **actual level**, the actual minimum coverage everywhere,

$$\text{level error} = \text{nominal level} - \text{actual level}.$$

- Methods, such as **bootstrap calibration**, described in DiCiccio and Efron (1996), exist which attempt to **correct** for the level error.

## Significance procedures and duality

- A hypothesis test of  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$ , where  $\Theta_0 \cup \Theta_0^c = \Theta$ , at significance level of 5% (or any other specified value) returns one bit of information, either we accept  $H_0$  or reject  $H_0$ .
- We do not know whether the decision was borderline or nearly conclusive; i.e. whether, for rejection,  $H_0$  and  $C(x; 0.05)$  were close, or well-separated.
- Of more interest is to consider the smallest value of  $\alpha$  for which  $C(x; \alpha)$  does not intersect  $H_0$ . This value is termed the  $p$ -value.

### Definition ( $p$ -value)

A  $p$ -value  $p(X)$  is a statistic satisfying  $p(x) \in [0, 1]$  for every  $x \in \mathcal{X}$ . Small values of  $p(x)$  support the hypothesis that  $H_1$  is true. A  $p$ -value is valid if, for every  $\theta \in \Theta_0$  and every  $\alpha \in [0, 1]$ ,

$$\mathbb{P}(p(X) \leq \alpha \mid \theta) \leq \alpha.$$

- If  $p(X)$  is a valid  $p$ -value then a **significance test** that rejects  $H_0$  if and only if  $p(X) \leq \alpha$  is a test with **significance level**  $\alpha$ .
- In this part we introduce the idea of **significance procedure** at level  $\alpha$ , deriving a **duality** between it and a level  $1 - \alpha$  **confidence procedure**.
- Let  $X$  and  $Y$  be two **scalar** random variables. Then  $X$  **stochastically dominates**  $Y$  exactly when  $\mathbb{P}(X \leq v) \leq \mathbb{P}(Y \leq v)$  for all  $v \in \mathbb{R}$ .
- If  $U \sim \text{Unif}(0, 1)$  then  $\mathbb{P}(U \leq u) = u$  for  $u \in [0, 1]$ . With this in mind, we make the following definition.

### Definition (Super-uniform)

The random variable  $X$  is **super-uniform** exactly when it **stochastically dominates** a standard **uniform** random variable. That is

$$\mathbb{P}(X \leq u) \leq u$$

for all  $u \in [0, 1]$ .

- Thus, for  $\theta \in \Theta_0$ , the  $p$ -value  $p(X)$  is **super-uniform**.

- We now define a significance procedure. Note the similarities with the definitions of a confidence procedure which are not coincidental.

### Definition (Significance procedure)

- 1  $p : \mathcal{X} \rightarrow \mathbb{R}$  is a **significance procedure** for  $\theta_0 \in \Theta$  exactly when  $p(X)$  is **super-uniform** under  $\theta_0$ . If  $p(X)$  is **uniform** under  $\theta_0$ , then  $p$  is an **exact** significance procedure for  $\theta_0$ .
  - 2 For  $X = x$ ,  $p(x)$  is a **significance level** or (observed)  $p$ -value for  $\theta_0$  exactly when  $p$  is a **significance procedure** for  $\theta_0$ .
  - 3  $p : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$  is a **family of significance procedures** exactly when  $p(x; \theta_0)$  is a **significance procedure** for  $\theta_0$  for every  $\theta_0 \in \Theta$ .
- We now show that there is a duality between significance procedures and confidence procedures.

## Duality Theorem

- ① Let  $p$  be a family of **significance procedures**. Then

$$C(x; \alpha) := \{\theta \in \Theta : p(x; \theta) > \alpha\}$$

is a nesting family of **confidence procedures**.

- ② Conversely, let  $C$  be a nesting family of **confidence procedures**. Then

$$p(x; \theta_0) := \inf\{\alpha : \theta_0 \notin C(x; \alpha)\}$$

is a family of **significance procedures**.

If **either** is **exact**, then the **other** is **exact** as well.

## Proof

- If  $p$  is a family of significance procedures then for any  $\theta \in \Theta$ ,

$$\mathbb{P}(\theta \in C(X; \alpha) | \theta) = \mathbb{P}(p(X; \theta) > \alpha | \theta) = 1 - \mathbb{P}(p(X; \theta) \leq \alpha | \theta).$$



## Proof continued

- Now, as  $p$  is **super-uniform** for  $\theta$  then  $\mathbb{P}(p(X; \theta) \leq \alpha | \theta) \leq \alpha$ . Thus,  $\mathbb{P}(\theta \in C(X; \alpha) | \theta) \geq 1 - \alpha$ . Hence,  $C(X; \alpha)$  is a level- $(1 - \alpha)$  **confidence procedure**.
- If  $\alpha' > \alpha$  then if  $\theta \in C(x; \alpha')$  we have  $p(x; \theta) > \alpha' > \alpha$  and so  $\theta \in C(x; \alpha)$  and so  $C$  is **nesting**.
- If  $p$  is **exact** then the inequalities can be replaced by equalities and so  $C$  is also **exact**.

We thus have 1.

- Now, if  $C$  is a **nesting** family of confidence procedures then<sup>a</sup>

$$\inf\{\alpha : \theta_0 \notin C(x; \alpha)\} \leq u \iff \theta_0 \notin C(x; u).$$

<sup>a</sup>Here we're finessing the issue of the boundary of  $C$  by assuming that if  $\alpha^* := \inf\{\alpha : \theta_0 \notin C(x; \alpha)\}$  then  $\theta_0 \notin C(x; \alpha^*)$ .

## Proof continued

- Let  $\theta_0$  and  $u \in [0, 1]$  be arbitrary. Then,

$$\mathbb{P}(p(X; \theta_0) \leq u | \theta_0) = \mathbb{P}(\theta_0 \notin C(X; u) | \theta_0) \leq u$$

as  $C(X; u)$  is a level- $(1 - u)$  confidence procedure. Thus,  $p$  is super-uniform.

- If  $C$  is exact, then the inequality is replaced by an equality, and hence  $p$  is exact as well. □

# Families of significance procedures

- We now consider a very **general** way to construct a family of significance procedures.
- We will then show how to use **simulation** to compute the family.

## Theorem

Let  $t : \mathcal{X} \rightarrow \mathbb{R}$  be a statistic. For each  $x \in \mathcal{X}$  and  $\theta_0 \in \Theta$  define

$$p_t(x; \theta_0) := \mathbb{P}(t(X) \geq t(x) \mid \theta_0).$$

Then  $p_t$  is a family of **significance procedures**. If the distribution function of  $t(X)$  is **continuous**, then  $p_t$  is **exact**.

## Proof (Casella and Berger, 2002)

- Now

$$p_t(x; \theta_0) = \mathbb{P}(t(X) \geq t(x) | \theta_0) = \mathbb{P}(-t(X) \leq -t(x) | \theta_0).$$

- Let  $F$  denote the distribution function of  $Y(X) = -t(X)$  then  $p_t(x; \theta_0) = F(-t(x) | \theta_0)$ .
- Assume that  $t(X)$  is continuous so that  $Y(X) = -t(X)$  is continuous. Using the Probability Integral Transform,

$$\begin{aligned} \mathbb{P}(p_t(X; \theta_0) \leq \alpha | \theta_0) &= \mathbb{P}(F(Y) \leq \alpha | \theta_0) \\ &= \mathbb{P}(Y \leq F^{-1}(\alpha) | \theta_0) = F(F^{-1}(\alpha)) = \alpha. \end{aligned}$$

Hence,  $p_t$  is uniform under  $\theta_0$ .

- If  $t(X)$  is not continuous then, via the Probability Integral Transform,  $\mathbb{P}(F(Y) \leq \alpha | \theta_0) \leq \alpha$  and so  $p_t(X; \theta_0)$  is super-uniform under  $\theta_0$ .  $\square$

- So there is a family of significance procedures for **each** possible function  $t : \mathcal{X} \rightarrow \mathbb{R}$ .
- Clearly only a tiny fraction of these can be useful functions, and the rest must be useless.
- Some, like  $t(x) = c$  for some constant  $c$ , are always useless. Others, like  $t(x) = \sin(x)$  might sometimes be a little bit useful, while others, like  $t(x) = \sum_i x_i$  might be quite useful - but it all depends on the circumstances.
- Some **additional criteria** are required to separate out **good** from **poor** choices of the test statistic  $t$ , when using the construction in the theorem.

The most pertinent criterion is:

- Select a test statistic for which  $t(X)$  which will tend to be larger for decision-relevant departures from  $\theta_0$ .

### Example

For the likelihood ratio,  $\lambda(x)$ , small observed values of  $\lambda(x)$  support departures from  $\theta_0$ . Thus,  $t(X) = -2 \log \lambda(X)$ , is a test statistic for which large values support departures from  $\theta_0$ .

- Large values of  $t(X)$  will correspond to small values of the  $p$ -value, supporting the hypothesis that  $H_1$  is true.
- This criterion ensures that  $p_t(X; \theta_0)$  will tend to be smaller under decision-relevant departures from  $\theta_0$ ; small  $p$ -values are more interesting, precisely because significance procedures are super-uniform under  $\theta_0$ .

## Computing p-values

Only in very special cases will it be possible to find a **closed-form expression** for  $p_t$  from which we can compute the **p-value**  $p_t(x; \theta_0)$ .

### Theorem (Adapted from Besag and Clifford, 1989)

For any finite sequence of scalar random variables  $X_0, X_1, \dots, X_m$ , define the **rank** of  $X_0$  in the sequence as

$$R := \sum_{i=1}^m \mathbb{1}_{\{X_i \leq X_0\}}.$$

If  $X_0, X_1, \dots, X_m$  are **exchangeable**<sup>a</sup> then  $R$  has a **discrete uniform distribution** on the integers  $\{0, 1, \dots, m\}$ , and  $(R + 1)/(m + 1)$  has a **super-uniform** distribution.

---

<sup>a</sup>If  $X_0, X_1, \dots, X_m$  are exchangeable then their joint density function satisfies  $f(x_0, \dots, x_m) = f(x_{\pi(0)}, \dots, x_{\pi(m)})$  for all permutations  $\pi$  defined on the set  $\{0, \dots, m\}$ .

## Proof

By exchangeability,  $X_0$  has the **same probability** of having rank  $r$  as any of the other  $X_i$ s, for **any**  $r$ , and therefore

$$\mathbb{P}(R = r) = \frac{1}{m+1}$$

for  $r \in \{0, 1, \dots, m\}$  and zero otherwise, proving the first claim. For the second claim,

$$\mathbb{P}\left(\frac{R+1}{m+1} \leq u\right) = \mathbb{P}(R+1 \leq u(m+1)) = \mathbb{P}(R+1 \leq \lfloor u(m+1) \rfloor)$$

since  $R$  is an **integer** and  $\lfloor x \rfloor$  denotes the **largest integer no larger than**  $x$ .



## Proof continued

Hence,

$$\mathbb{P}\left(\frac{R+1}{m+1} \leq u\right) = \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \mathbb{P}(R=r) \quad (4)$$

$$= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} \quad (5)$$

$$= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u,$$

as required where equation (5) follows from (4) by [exchangeability](#).  $\square$

- We utilise this result to compute the **p-value**  $p_t(x; \theta_0)$  corresponding to the test statistic  $t(X)$  at  $\theta_0$ .
- Fix the test statistic  $t(x)$  and define  $T_i = t(X_i)$  where  $X_1, \dots, X_m$  are independent and identically distributed random variables with density  $f_X(\cdot | \theta_0)$ .
- Typically, we may have to use **simulation** to obtain the sample and we'll need to specify  $\theta_0$  for this.
- Notice that  $t(X), T_1, \dots, T_m$  are exchangeable and thus  $-t(X), -T_1, \dots, -T_m$  are **exchangeable**.
- Let

$$R_t(x; \theta_0) := \sum_{i=1}^m \mathbb{1}_{\{-T_i \leq -t(x)\}} = \sum_{i=1}^m \mathbb{1}_{\{T_i \geq t(x)\}},$$

then the previous theorem implies that

$$P_t(x; \theta_0) := \frac{R_t(x; \theta_0) + 1}{m + 1}$$

has a **super-uniform** distribution under  $X \sim f_X(\cdot | \theta_0)$ .

- Note that  $\mathbb{P}(T \geq t(x) | \theta_0) = \mathbb{E}(\mathbb{1}_{\{T \geq t(x)\}})$ .
- Hence, the **Weak Law of Large Numbers (WLLN)** implies that

$$\begin{aligned}
 \lim_{m \rightarrow \infty} P_t(x; \theta_0) &= \lim_{m \rightarrow \infty} \frac{R_t(x; \theta_0) + 1}{m + 1} \\
 &= \lim_{m \rightarrow \infty} \frac{R_t(x; \theta_0)}{m} \\
 &= \lim_{m \rightarrow \infty} \frac{\sum_{i=1}^m \mathbb{1}_{\{T_i \geq t(x)\}}}{m} \\
 &= \mathbb{P}(T \geq t(x) | \theta_0) = p_t(x; \theta_0).
 \end{aligned}$$

- Therefore, not only is  $P_t(x; \theta_0)$  **super-uniform** under  $\theta_0$ , so that  $P_t$  is a family of significance procedures for every  $m$ , but the **limiting value** of  $P_t(x; \theta_0)$  as  $m$  becomes large is  $p_t(x; \theta_0)$ .
- In summary, if you can **simulate** from your model under  $\theta_0$  then you can produce a  $p$ -value for **any test statistic**  $t$ , namely  $P_t(x; \theta_0)$ , and if you can simulate cheaply, so that the number of simulations  $m$  is large, then  $P_t(x; \theta_0) \approx p_t(x; \theta_0)$ .

- However, this simulation-based approach is not well-adapted to constructing **confidence sets**.
- Let  $C_t$  be the family of **confidence procedures** induced by  $p_t$  using **duality**.
- With **one set** of  $m$  simulations, we can answer "Is  $\theta_0 \in C_t(x; \alpha)$ ?"
  - ▶ These simulations give a value  $P_t(x; \theta_0)$  which is either larger or not larger than  $\alpha$ .
  - ▶ If  $P_t(x; \theta_0) > \alpha$  then  $\theta_0 \in C_t(x; \alpha)$ , and otherwise it is not.
- However, this is **not an effective way** to enumerate all of the points in  $C_t(x; \alpha)$  since we would need to do  $m$  **simulations** for **each point** in  $\Theta$ .

## Concluding remarks

- It is a very common observation, made repeatedly over the last 50 years see, for example, Rubin (1984), that clients think more like Bayesians than classicists.
- For example,  $\mathbb{P}(\theta \in C(X; \alpha) | \theta) \geq 1 - \alpha$  is often interpreted as a probability over  $\theta$  for the observed  $C(x; \alpha)$ .
- Classical statisticians thus have to wrestle with the issue that their clients will likely misinterpret their results.
- This can be potentially disastrous for  $p$ -values.
  - ▶ A  $p$ -value  $p(x; \theta_0)$  refers only to  $\theta_0$ , making no reference at all to other hypotheses about  $\theta$ .
  - ▶ A posterior probability  $\pi(\theta_0 | x)$  contrasts  $\theta_0$  with the other values in  $\Theta$  which  $\theta$  might have taken.
  - ▶ The two outcomes can be radically different, as first captured in Lindley's paradox (Lindley, 1957).

## References

- Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā* 37(1), 1–71.
- Besag, J. and P. Clifford (1989). Generalized Monte Carlo significance tests. *Biometrika* 76(4), 633–642.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association* 57, 269–306.
- Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association* 67, 858–861.
- Casella, G. and R.L. Berger (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA, USA: Duxbury.
- Cox, D.R. and D.V. Hinkley (1974). *Theoretical Statistics*. London, UK: Chapman and Hall.
- Dawid, A.P. (1977). Conformity of inference patterns. In J.R. Barra et al. (Eds.), *Recent Developments in Statistics*. pp. 245–256. Amsterdam, The Netherlands: North-Holland Publishing Company.

## References continued

- DiCiccio, T.J. and B. Efron (1996). Bootstrap confidence intervals *Statistical Science* 11(3), 189–228.
- Lindley, D.V. (1957). A statistical paradox. *Biometrika* 44, 187–192.
- Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, USA: Springer.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* 12(4), 1151–1172.
- Savage, L.J. et al. (1962). *The Foundations of Statistical Inference*. London, UK: Methuen.
- Smith, J.Q. (2010). *Bayesian Decision Analysis: Principle and Practice*. Cambridge, UK: Cambridge University Press.
- Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9(1), 60–62.