

# Statistical Inference

## Lecture Five

<https://people.bath.ac.uk/masss/APTS/2022-23/LectureFive.pdf>

Simon Shaw

University of Bath

APTS, 13-16 December 2022

# Overview of Lecture Five

- Confidence procedure: A random set  $C(X) \subset \Theta$  is a level- $(1 - \alpha)$  confidence procedure exactly when  $\mathbb{P}(\theta \in C(X) | \theta) \geq 1 - \alpha$ .
- Family of confidence procedures: occurs when  $C(X; \alpha)$  is a level- $(1 - \alpha)$  confidence procedure for every  $\alpha \in [0, 1]$ .
- $C$  is a nesting family if  $\alpha < \alpha'$  implies that  $C(x; \alpha') \subset C(x; \alpha)$ .
- The general approach to construct a confidence procedure is to invert a test statistic.
- Consider the likelihood ratio test (LRT) statistic

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

- Duality of acceptance regions and confidence sets.

- 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  $\delta^*(x) \neq d$ . □

- 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>1</sup>

---

<sup>1</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  $C(x; \alpha') \subset C(x; \alpha)$ .

- For  $X_1, \dots, X_n$  iid  $\text{Unif}(0, \theta)$ ,  $Y = \max\{X_1, \dots, X_n\}$  then

$$C(Y; \alpha) = \left( \left(1 - \frac{\alpha}{2}\right)^{-1/n} Y, \left(\frac{\alpha}{2}\right)^{-1/n} Y \right)$$

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

- For example, if  $n = 10$  then

$$C(y; 0.10) = (1.0051y, 1.3493y); \quad C(y; 0.05) = (1.0025y, 1.4461y).$$

- 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: pivotal quantities

- 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.

### Definition (Pivot)

A **pivot** for the model  $\mathcal{E} = \{\mathcal{X}, \Theta, f_{\mathcal{X}}(x | \theta)\}$  is a random variable  $Q(X_1, \dots, X_n, \theta)$  for which the **distribution** of  $Q$  **does not depend** upon  $\theta$ .

- For any set  $\mathcal{A}$ ,  $\mathbb{P}(Q(X_{(1:n)}, \theta) \in \mathcal{A} | \theta)$  does not depend on  $\theta$ .
- Hence,  $\{\theta : Q(x_{(1:n)}, \theta) \in \mathcal{A}\}$  is an **exact** confidence procedure.

## Example: location parameter

- for any pdf  $f(x)$ , the family of pdfs  $f(x - \theta)$ , indexed by  $\theta$ , is a **location family** with **standard** pdf  $f(x)$ .  $\theta$  is a **location parameter**.
- If  $X_1, \dots, X_n$  are iid from  $f(x - \theta)$  then the confidence set

$$C(x_1, \dots, x_n) = \{\theta : \bar{x} - a < \theta < \bar{x} + b\}$$

for constants  $a, b \geq 0$  has a **fixed coverage** for all  $\theta$  since

$$\begin{aligned} \mathbb{P}(\theta \in C(X_1, \dots, X_n) \mid \theta) &= \mathbb{P}(\bar{X} - a < \theta < \bar{X} + b \mid \theta) \\ &= \mathbb{P}\left(-b < \frac{1}{n} \sum_{i=1}^n Z_i < a \mid \theta\right) \end{aligned}$$

does not depend on  $\theta$  as each  $Z_i = X_i - \theta$  has pdf  $f(z)$  which does not depend on  $\theta$ .

- For location (and scale) parameters, we can easily find pivots but this is more difficult in general.
- An alternate method 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,

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

$$\leq \frac{\mathbb{E}(h(X)/f_X(X|\theta) | \theta)}{1/u} \quad (2)$$

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

where (2) follows from (1) 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>2</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>2</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**.