

Statistical Inference

Lecture Eight

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

Simon Shaw

University of Bath

APTS, 14-18 December 2020

Overview of Lecture Eight

In Lecture Seven we introduced confidence procedures.

- 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 $\alpha \in [0, 1]$.

In Lecture Eight we'll look at good choices of confidence procedures.

- Level set property, LSP: present for a confidence procedure C when $C(x) = \{\theta : f_X(x | \theta) > g(x)\}$ for some $g : \mathcal{X} \rightarrow \mathbb{R}$.
- For the linear model we can construct an exact family of confidence procedures which satisfy the LSP.
- Wilks Confidence procedures and the likelihood ratio test.
- Introduce the p -value.

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](#).^a □


^aIf 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 θ_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 $(n \times p)$ matrix¹ of **regressors**,
 - ▶ θ is a p -vector of **regression coefficients**,
 - ▶ ϵ is an n -vector of **residuals**.
- Assume that $\epsilon \sim N_n(0, \sigma^2 I_n)$, the n -dimensional **multivariate normal** distribution, where σ^2 is **known** and I_n is the $(n \times n)$ **identity matrix**.
- From properties of the multivariate normal distribution, it follows that $Y \sim N_n(X\theta, \sigma^2 I_n)$.

¹We 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 θ . 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 θ 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** Θ 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 θ ; in particular, the support must not depend on θ .

- 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 α 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** α .
- In this part we introduce the idea of **significance procedure** at level α , 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**.