Statistical Inference
Lecture Six

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

Simon Shaw

University of Bath

APTS, 13-17 December 2021

Overview of Lecture Six

In Lecture Five we introduced confidence procedures.

- Confidence procedure: A random set $C(X) \subset \Theta$ is a level-$(1-\alpha)$ confidence procedure exactly when $\mathbb{P}(\theta \in C(X) \mid \theta) \ \geq \ 1-\alpha$.
- Family of confidence procedures: occurs when $C(X; \alpha)$ is a level-$(1-\alpha)$ confidence procedure for every $\alpha \in [0, 1]$.
- Level set property, LSP: present for a confidence procedure C when $C(x) = \{\theta \, : \, f_X(x \mid \theta) > g(x)\}$ for some $g : \mathcal{X} \to \mathbb{R}$.

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

- 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.

Overview of Lecture Six continued

- A p-value $p(X)$ is a statistic satisfying, for every $\alpha \in [0, 1]$, $\mathbb{P}(p(X) \leq \alpha \,|\, \theta) \leq \alpha$. It is super-uniform.
- $p : \mathcal{X} \to \mathbb{R}$ is a significance procedure for $\theta_0 \in \Theta$ exactly when $p(X)$ is super-uniform under θ_0.
- We'll show there is a duality between significance procedures and confidence procedures.
- We'll show how to construct a family of significance procedures and how to use simulation to compute the family.

The linear model

- We'll briefly discuss the linear model and construct an exact family of confidence procedures which satisfy the LSP.
- Let $Y = (Y_1, \ldots, Y_n)$ be an n-vector of observables with $Y = X\theta + \epsilon$.
  - X is an $(n \times p)$ matrix[1] 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 $Y \sim N_n(X\theta, \sigma^2 I_n)$.

[1] 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 $\hat{\theta} = \hat{\theta}(y) = (X^TX)^{-1}X^Ty$ then

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

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

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

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

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

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

Hence, with $\mathbb{P}(\chi^2_p \geq \chi^2_{p,\alpha}) = \alpha$,

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

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 $X = (X_1, \ldots, X_n)$ 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 $n \to \infty$.

- 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,

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

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

$$\mathbb{P}(p(X) \leq \alpha \,|\, \theta) \quad \leq \quad \alpha.$$

- 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

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

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. $p : \mathcal{X} \to \mathbb{R}$ is a significance procedure for $\theta_0 \in \Theta$ exactly when $p(X)$ is super-uniform under θ_0. If $p(X)$ is uniform under θ_0, then p is an exact significance procedure for θ_0.
2. For $X = x$, $p(x)$ is a significance level or (observed) p-value for θ_0 exactly when p is a significance procedure for θ_0.
3. $p : \mathcal{X} \times \Theta \to \mathbb{R}$ is a family of significance procedures exactly when $p(x; \theta_0)$ is a significance procedure for θ_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 θ 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[a]

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

[a] 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 θ_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. $\square$

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

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

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 θ_0.
- It $t(X)$ 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 θ_0. $\square$

- So there is a family of significance procedures for each possible function $t : \mathcal{X} \to \mathbb{R}$.
- 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 θ_0.

Example

For the likelihood ratio, $\lambda(x)$, small observed values of $\lambda(x)$ support departures from θ_0. Thus, $t(X) = -2 \log \lambda(X)$, is a test statistic for which large values support departures from θ_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 θ_0; small p-values are more interesting, precisely because significance procedures are super-uniform under θ_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 $X_0, X_1, \ldots, X_m$, define the rank of X_0 in the sequence as

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

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

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

Proof

By exchangeability, X_0 has the same probability of having rank r as any of the other X_is, for any r, and therefore

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

for $r \in \{0, 1, \ldots, m\}$ 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,

$$\begin{aligned}
\mathbb{P}\left(\frac{R+1}{m+1} \leq u\right) &= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \mathbb{P}(R = r) && (1)\\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} && (2)\\
&= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u,
\end{aligned}$$

as required where equation (2) follows from (1) by exchangeability. □

- We utilise this result to compute the *p*-value $p_t(x; \theta_0)$ corresponding to the test statistic $t(X)$ at θ_0.
- Fix the test statistic $t(x)$ and define $T_i = t(X_i)$ where $X_1, \ldots, X_m$ are independent and identically distributed random variables with density $f_X(\cdot \mid \theta_0)$.
- Typically, we may have to use simulation to obtain the sample and we'll need to specify θ_0 for this.
- Notice that $t(X), T_1, \ldots, T_m$ are exchangeable and thus $-t(X), -T_1, \ldots, -T_m$ 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 $X \sim f_X(\cdot \mid \theta_0)$.

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

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

- Therefore, not only is $P_t(x;\theta_0)$ super-uniform under θ_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 θ_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 $P_t(x;\theta_0) \approx p_t(x;\theta_0)$.