Statistical Inference

https://people.bath.ac.uk/masss/APTS/lecture7.pdf

Simon Shaw

University of Bath

Lecture Seven, 20 December 2019

Yesterday's lecture

- Family of confidence procedures: occurs when $C(X;\alpha)$ is a level-$(1-\alpha)$ confidence procedure, so $\mathbb{P}(\theta \in C(X;\alpha) \mid \theta) \geq 1-\alpha$, for every $\alpha \in [0,1]$.
- 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 $u \in [0,1]$.
- $p : \mathcal{X} \to \mathbb{R}$ 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.
- 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.
- $p : \mathcal{X} \times \Theta \to \mathbb{R}$ is a family of significance procedures exactly when $p(x;\theta_0)$ is a significance procedure for θ_0 for every $\theta_0 \in \Theta$.

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 $t : \mathcal{X} \rightarrow \mathbb{R}$.
- 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 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 $f_X(\cdot \,|\, \theta_0)$.
- 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

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

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

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

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

- 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 α.
  - 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 Θ.
- We'll omit the section looking at generalisations, including marginalisation.

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, $\mathbb{P}(\theta \in C(X;\alpha) \mid \theta) \geq 1 - \alpha$ is often interpreted as a probability over θ 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 θ_0, making no reference at all to other hypotheses about θ.
    - A posterior probability $\pi(\theta_0 \mid x)$ contrasts θ_0 with the other values in Θ which θ might have taken.
    - The two outcomes can be radically different, as first captured in Lindley's paradox (Lindley, 1957).