Statistical Inference
Lecture Seven

https://people.bath.ac.uk/masss/APTS/2022-23/LectureSeven.pdf

Simon Shaw

University of Bath

APTS, 13-16 December 2022

- 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.

Coming up in Lecture Seven

- 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.
- For any finite sequence of scalar exchangeable random variables $X_0, X_1, \ldots, X_m$, then if R is the rank of X_0 in the sequence then R has a discrete uniform distribution on the integers $\{0, 1, \ldots, m\}$, and $(R + 1)/(m + 1)$ has a super-uniform distribution.
- We utilise this result to compute the p-value $p_t(x; \theta_0)$ corresponding to the test statistic $t(X)$ at θ_0.
- We'll briefly look at Bayesian hypothesis testing and Lindley's Paradox.

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. □

- 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,

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

$$= \sum_{r=0}^{\lfloor u(m+1)\rfloor - 1} \frac{1}{m+1} \quad (2)$$

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

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

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

- 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.

Interpretations

- 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.
- We will now briefly look at Bayesian approaches to hypothesis testing.
- In this approach, we can calculate the posterior probability of each hypothesis.

- Consider a point-null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- A possible prior is a mixture of a point mass on θ_0 and a distribution, $\pi_1(\theta)$, under H_1:

$$\pi(\theta) \;=\; p_0 \mathbb{I}_{\{\theta=\theta_0\}} + (1-p_0)\pi_1(\theta)$$

  where $p_0 = \mathbb{P}(\theta = \theta_0)$.
- If $f_X(x \,|\, \theta)$ is the data generating model then the posterior probability of $\theta = \theta_0$ is

$$\mathbb{P}(\theta = \theta_0 \,|\, X) \;=\; \frac{p_0 f_X(x \,|\, \theta_0)}{\int f_X(x \,|\, \theta)\pi(\theta)d\theta} \;=\; \frac{p_0 f_X(x \,|\, \theta_0)}{p_0 f_X(x \,|\, \theta_0) + (1-p_0)f_1(x)}$$

  where $f_1(x)$ is the marginal distribution under H_1,

$$f_1(x) \;=\; \int_{\Theta_1} f_X(x \,|\, \theta)\pi_1(\theta)d\theta$$

- Thus, $\mathbb{P}(\theta = \theta_0 \,|\, X) = (1+y)^{-1}$ where

$$y \;=\; \frac{1-p_0}{p_0}\frac{f_1(x)}{f_X(x \,|\, \theta_0)}.$$

Example: normal model for $H_0 : \theta = 0$

- Let $\theta_0 = 0$ and suppose that $X \mid \theta \sim N(\theta, \sigma^2)$ for σ^2 known.
- For the prior under $H_1 : \theta \neq 0$ we assert $\theta \sim N(0, \sigma_0^2)$ where σ_0^2 is known.
- Thus,

$$
\begin{aligned}
f_X(x \mid \theta = 0) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}x^2\right\}, \\
f_1(x) &= \int_{-\infty}^{\infty} f_X(x \mid \theta)\pi_1(\theta)d\theta \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma\sigma_0} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2 - \frac{1}{2\sigma_0^2}\theta^2\right\} d\theta \\
&= \frac{(\sigma^2+\sigma_0^2)^{-\frac{1}{2}}}{\sqrt{2\pi}} \left\{-\frac{x^2}{2(\sigma^2+\sigma_0^2)}\right\}
\end{aligned}
$$

so that $f_1(x)$ is the pdf of $N(0, \sigma^2 + \sigma_0^2)$.

Example: normal model for $H_0 : \theta = 0$

- Hence, $\mathbb{P}(\theta = 0 \,|\, X = x) = (1 + y)^{-1}$ where

$$
\begin{aligned}
y &= \frac{1 - p_0}{p_0} \frac{f_1(x)}{f_X(x \,|\, \theta = 0)} \\
&= \left(\frac{1 - p_0}{p_0}\right) \left(\frac{\sigma^2}{\sigma^2 + \sigma_0^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2 + \sigma_0^2} - \frac{1}{\sigma^2}\right) x^2\right\} \\
&= \left(\frac{1 - p_0}{p_0}\right) \left(\frac{\sigma^2}{\sigma^2 + \sigma_0^2}\right)^{\frac{1}{2}} \exp\left\{\frac{\sigma_0^2 x^2}{2(\sigma^2 + \sigma_0^2)\sigma^2}\right\}
\end{aligned}
$$

- A disperse prior for $H_1 : \theta \neq 0$ is sometimes proposed and this can be achieved by increasing the prior variance σ_0^2.
- If $\sigma_0^2 \to \infty$ then $y \to 0$ and $\mathbb{P}(\theta = 0 \,|\, X = x) \to 1$ for all x. This may be an issue with using improper priors: a proper prior has σ_0^2 finite.
- Note that y increases in $|x|$ and so $\mathbb{P}(\theta = 0 \,|\, X = x)$ decreases.
- With a proper prior, as $|x| \to \infty$, $y \to \infty$ and $\mathbb{P}(\theta = 0 \,|\, X = x) \to 0$. The Bayesian analysis behaves reasonably.

- Now consider taking n iid observations and consider the posterior probability given $\overline{x}$.
- Notice that, as $\overline{X} \,|\, \theta \sim N(\theta, \sigma^2/n)$, our calculations will take the same form as previously but with x replaced by $\overline{x}$ and σ^2 by σ^2/n.
- Thus, $\mathbb{P}(\theta = 0 \,|\, \overline{X} = \overline{x}) = (1 + y_n)^{-1}$ where

$$
\begin{aligned}
y_n &= \left(\frac{1-p_0}{p_0}\right)\left(\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\right)^{\frac{1}{2}} \exp\left\{\frac{n^2\sigma_0^2\overline{x}^2}{2(\sigma^2 + n\sigma_0^2)\sigma^2}\right\} \\
&= \left(\frac{1-p_0}{p_0}\right)\left(\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\right)^{\frac{1}{2}} \exp\left\{\frac{n\sigma_0^2}{2(\sigma^2 + n\sigma_0^2)} z^2\right\}
\end{aligned}
$$

  and $z = \sqrt{n}|\overline{x}|/\sigma$.
- Note that if H_0 is true then $\sqrt{n}\overline{X}/\sigma \sim N(0,1)$ so $Z^2 \sim \chi^2_1$.
- Suppose that $z = \sqrt{n}|\overline{x}|/\sigma$ is fixed as we increase n. Then $y_n \to 0$ and hence $\mathbb{P}(\theta = 0 \,|\, \overline{X} = \overline{x}) \to 1$.
- The Bayesian model favours H_0 over H_1.

- Now let's consider the classical approach to this problem using a p-value.
- Consider the test statistic $|\overline{X}|$ which will be large for departures from $H_0 : \theta_0 = 0$. We have

$$
\begin{aligned}
p(|\overline{x}|; 0) &= \mathbb{P}(|\overline{X}| \geq |\overline{x}| \,|\, \theta = 0) \\
&= \mathbb{P}(\sqrt{n}|\overline{X}|/\sigma \geq z \,|\, \theta = 0).
\end{aligned}
$$

- Now, under H_0, $\sqrt{n}\overline{X}/\sigma \sim N(0, 1)$. If $z = \sqrt{n}|\overline{x}|/\sigma$ is fixed for all n then the p-value is fixed for all n.
- Thus, if $\alpha \geq p(|\overline{x}|; 0)$ we reject H_0 for all values of n at significance level α.
- This is an illustration of what is termed Lindley's paradox (Lindley, 1957).

Lindley's paradox

The main idea of this seeming paradox can be expressed as follows.

- For a normal model $N(\theta, \sigma^2)$ with known variance σ^2, consider the hypothesis test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- Assume $\mathbb{P}(\theta \in H_0) > 0$ and any regular prior on $\{\theta \neq \theta_0\}$. Then for any $\alpha \in [0, 1]$ we can find a sample size $n(\alpha)$ and iid data $x_1, \ldots, x_n$ such that:
  1. The sample mean $\overline{x}$ is significantly different from H_0 at level α.
  2. The posterior probability that $\theta = \theta_0$ is greater that $1 - \alpha$.
- In our example, if we set $\sigma^2 = \sigma_0^2 = 1$, $n = 16818$, and $\overline{x} = 1.96(16818)^{-\frac{1}{2}} = 0.015$ then $z = 1.96$ and $\mathbb{P}(\theta = 0 \,|\, \overline{X} = \overline{x}) = 0.95$
- The reasoning for this seeming paradox is that the classical and Bayesian approaches are asking different questions.

Concluding remarks: understanding the problem

- A p-value $p(x; \theta_0)$ refers only to θ_0, making no reference at all to other hypotheses about θ.
  - A p-value can be viewed as measuring the fit of a model, that under H_0, to the observed data.
  - If I reject H_0 using a p-value then H_0 is a poor explanation for the observation.
  - However, a large p-value indicates only that the data is not unusual under the model but it does not imply that the model is correct.
  - For example, there may be many other models defined by other hypotheses which may be exhibit greater consistency with the observed data.
- A posterior probability $\pi(\theta_0 \,|\, x)$ contrasts θ_0 with the other values in Θ which θ might have taken.
  - If I favour H_0 then H_0 is a better explanation for the data x than H_1.

- Wasserstein and Lazar (2016) is a statement from the American Statistical Association (ASA) on statistical significance and p-values.
- The statement gives six principles for the correct use and interpretation of p-values.
- These principles, in particular Principles 3 and 4, reflect values that should be at the heart of any work that we do.

Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

- Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.
- Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.

Proper inference requires full reporting and transparency.

- Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.
- Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.