Statistical Inference
Lecture Seven

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

Simon Shaw

University of Bath

APTS, 13-17 December 2021

Previously in Lecture Six

- 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.
- 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.

- 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 conclude by looking 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 $H_1 : \theta \neq \theta_1$.
- 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 $X \,|\, \theta \sim N(\theta, \sigma^2)$ 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,

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

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 $z = \sqrt{n}|\overline{x}|/\sigma$.
- 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

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

- 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 $1 - \alpha$.
- 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.
  - If I reject H_0 using a p-value then H_0 is a poor explanation for the observation.
- 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.

- This difference in approach can be seen if I consider what I need to compute the posterior probability when I have a p-value.
- Suppose I observe $p(x; H_0) \leq u$ then:

$$
\begin{aligned}
\mathbb{P}(H_0 \,|\, p(x; H_0) \leq u) &= \frac{\mathbb{P}(p(X; H_0) \leq u \,|\, H_0)\mathbb{P}(H_0)}{\mathbb{P}(p(X; H_0) \leq u)} \\
&\leq \frac{u\mathbb{P}(H_0)}{\mathbb{P}(p(X; H_0) \leq u)}
\end{aligned}
$$

since the p-value is super-uniform.

- Now,

$$
\mathbb{P}(p(X; H_0) \leq u) = \sum_{i=0}^{1} \mathbb{P}(p(X; H_0) \leq u \,|\, H_i)\mathbb{P}(H_i)
$$

and so I need to know the distribution of the p-value under H_1 to compute $\mathbb{P}(p(X; H_0) \leq u \,|\, H_1)$.

References

- Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā 37*(1), 1–71.
- Besag, J. and P. Clifford (1989). Generalized Monte Carlo significance tests. *Biometrika 76*(4), 633–642.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*, 269–306.
- Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association 67*, 858–861.
- Casella, G. and R.L. Berger (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA, USA: Duxbury.
- Cox, D.R. and D.V. Hinkley (1974). *Theoretical Statistics*. London, UK: Chapman and Hall.
- Dawid, A.P. (1977). Conformity of inference patterns. In J.R. Barra et al. (Eds.), *Recent Developments in Statistcs*. pp. 245–256. Amsterdam, The Netherlands: North-Holland Publishing Company.

References continued

- DiCiccio, T.J. and B. Efron (1996). Bootstrap confidence intervals *Statistical Science 11*(3), 189–228.
- Lindley, D.V. (1957). A statistical paradox. *Biometrika 44*, 187–192.
- Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, USA: Springer.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*(4), 1151–1172.
- Savage, L.J. et al. (1962). *The Foundations of Statistical Inference*. London, UK: Methuen.
- Smith, J.Q. (2010). *Bayesian Decision Analysis: Principle and Practice*. Cambridge, UK: Cambridge University Press.
- Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*(1), 60–62.