Theory of Inference: Homework 5

Here are two exam-style revision questions, about p-values and confidence sets.

1. (a) Consider the general model in which $(X_1, \dots, X_m) \sim \mathrm{p}(\boldsymbol{x}; \theta)$ for $\theta \in \Omega$, with observables $Y_i := g_i(\boldsymbol{X})$ for $i = 1, \dots, n$, where the g_i are specified functions of $\boldsymbol{x}$. State the general formula for $\mathrm{p}(\boldsymbol{y}; \theta)$, and also the special case where $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{p}(x; \theta)$ and $Y_i = X_i$ for $i = 1, \dots, n$. [5 marks]

Answer. As usual in these answers, I write them independently of the notes, just for variety.

In general, applying the definition of probability and the FTP,

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{y}; \theta) &= \mathrm{E}\{\mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}}; \theta\} && \text{definition} \\
&= \mathrm{E}\{\mathbb{1}_{g_1(\boldsymbol{X}) \doteq y_1 \wedge \cdots \wedge g_n(\boldsymbol{X}) \doteq y_n}; \theta\} \\
&= \sum_{\boldsymbol{x}} \prod_{i=1}^{n} \mathbb{1}_{g_i(\boldsymbol{x}) \doteq y_i} \cdot \mathrm{p}(\boldsymbol{x}; \theta) && \text{FTP.}
\end{aligned}
$$

In the special case $\mathrm{p}(\boldsymbol{x}; \theta) = \prod_{j=1}^{m} \mathrm{p}(x_j; \theta)$ and $g_i(\boldsymbol{x}) = x_i$, and subsituting in gives

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{y}; \theta) &= \sum_{\boldsymbol{x}} \prod_{i=1}^{n} \mathbb{1}_{x_i \doteq y_i} \cdot \prod_{j=1}^{m} \mathrm{p}(x_j; \theta) \\
&= \sum_{x_1} \cdots \sum_{x_n} \prod_{i=1}^{n} \mathbb{1}_{x_i \doteq y_i} \, \mathrm{p}(x_i; \theta) \cdot \sum_{x_{n+1}} \cdots \sum_{x_m} \prod_{i=n+1}^{m} \mathrm{p}(x_i; \theta) \\
&= \sum_{x_1} \cdots \sum_{x_n} \prod_{i=1}^{n} \mathbb{1}_{x_i \doteq y_i} \; \mathrm{p}(x_i; \theta) \\
&= \prod_{i=1}^{n} \sum_{x_i} \mathbb{1}_{x_i \doteq y_i} \cdot \mathrm{p}(x_i; \theta) \\
&= \prod_{i=1}^{n} \mathrm{p}(y_i; \theta).
\end{aligned}
$$

[You don't need all the steps for a compelling answer, but you need most of them.]

(b) (i) Consider the model $\boldsymbol{Y} \sim \mathrm{p}(\boldsymbol{y};\theta)$ for $\theta \in \Omega$. Under what conditions is the statistic $p_0(\boldsymbol{y})$ a P-value for the simple hypothesis $H_0 : \theta = \theta_0$?

(ii) Let $t(\boldsymbol{y})$ be any statistic. Prove that

$$p_0(\boldsymbol{y}) := \Pr\left\{t(\boldsymbol{Y}) \geq t(\boldsymbol{y}); \theta_0\right\}$$

is a P-value for H_0. You may take as given the Probability Integral Transform (PIT), which states that if F_X is the distribution function of X, then $F_X(X)$ has a sub-uniform distribution.

(iii) Give an example of a P-value which is completely uninformative about H_0, and explain how this possibility affects our interpretation of P-values.

[10 marks]

Answer. p_0 is P-value exactly when $p_0(\boldsymbol{Y})$ has a sub-uniform distribution under H_0; i.e.

$$\Pr\{p_0(\boldsymbol{Y}) \leq u; \theta_0\} \leq u \qquad \text{for all } u \geq 0.$$

When the inequality is an equality for all u, then p_0 is an 'exact' P-value.

Let G_0 be the distribution function of $-t(\boldsymbol{Y})$ under H_0, so that

$$p_0(\boldsymbol{y}) = \Pr\left\{t(\boldsymbol{Y}) \geq t(\boldsymbol{y}); \theta_0\right\} = \Pr\left\{-t(\boldsymbol{Y}) \leq -t(\boldsymbol{y}); \theta_0\right\} = G_0(-t(\boldsymbol{y})).$$

Then

$$p_0(\boldsymbol{Y}) = G_0(-t(\boldsymbol{Y}))$$

and since $-t(\boldsymbol{Y})$ has distribution function G_0 under H_0, the result follows by the PIT.

One can construct a completely uninformative P-value by using a test statistic which does not depend on $\boldsymbol{y}$, such as $t(\boldsymbol{y}) = 1$. By extention, there must be many P-values which are nearly uninformative about H_0, and so on. So we see that the P-value needs to be carefully-chosen, in order to be informative about H_0.

(c) Let $p(\boldsymbol{y};\theta_0)$ be a P-value for $H_0 : \theta = \theta_0$, and suppose that this can be computed for each $\theta_0 \in \Omega_0 \subset \Omega$. Define what is meant by a P-value for

$H_0 : \theta \in \Omega_0$, and show that

$$p_{\Omega_0}(\boldsymbol{y}) := \sup_{\theta_0 \in \Omega_0} p(\boldsymbol{y}; \theta_0)$$

is such a P-value. [5 marks]

Answer. A P-value for the composite hypothesis $H_0 : \theta \in \Omega_0$ has a subuniform distribution under all possible values of $\theta_0 \in \Omega_0$. For any $\boldsymbol{y}$, we have, by construction,

$$p_{\Omega_0}(\boldsymbol{y}) \leq u \implies p(\boldsymbol{y}; \theta_0) \leq u \text{ for all } \theta \in \Omega_0.$$

Hence

$$\Pr\left\{p_{\Omega_0}(\boldsymbol{Y}) \leq u; \theta_0\right\} \leq \Pr\left\{p(\boldsymbol{Y}; \theta_0) \leq u; \theta_0\right\} \leq u \text{ for all } \theta_0 \in \Omega_0$$

as was to be shown. The first inequality follows from the monotonicity property of expectation, because if $A \implies B$ then $\mathbb{1}_A \leq \mathbb{1}_B$, and the second follows because $p(\boldsymbol{y}; \theta_0)$ is a P-value for each $\theta_0 \in \Omega_0$.

(d) You have computed $p_0(\boldsymbol{y}^{\text{obs}}) = 0.0135$ for some dataset $\boldsymbol{y}^{\text{obs}}$. Interpret this value for your non-statistical client in the case where p_0 is an exact P-value for H_0, and the case where p_0 is not an exact P-value. [5 marks]

Answer. Because p_0 is an exact P-value for H_0,

$$\Pr\left\{p_0(\boldsymbol{Y}) \leq 0.0135; H_0\right\} = 0.0135$$

so were H_0 to be 'true', then a rare event would have occurred, one that only happens once in $1/0.0135 \approx 1/0.014 = 1000/14 = 70$ repetitions, on average. This makes us suspect that H_0 may be false, given that there may well be a competing hypothesis under which this event is much less rare. The answer is the same when the P-value is not exact, but the probability under H_0 is possibly even smaller.

2. Consider a statistical model of the form $\boldsymbol{Y} \sim \mathrm{p}(\cdot; \theta)$ for $\theta \in \Omega \subset \mathbb{R}^p$.

(a) Let $\mathcal{C}$ be a function mapping from $\boldsymbol{\mathcal{Y}}$ to subsets of Ω. Define the *coverage* of $\mathcal{C}$ at θ. Define a level-β *confidence set* for θ. What special property does an *exact* confidence set have? [6 marks]

Answer. The coverage of $\mathcal{C}$ at θ is defined as

$$\operatorname{cov}(\theta) := \Pr\{\theta \in \mathcal{C}(\boldsymbol{Y}); \theta\}.$$

A level-β confidence set has coverage of at least β for all $\theta \in \Omega$. If $\mathcal{C}$ is an 'exact' level-β confidence set, then it coverage exactly β for every $\theta \in \Omega$.

(b) Propose an exact 95% confidence set for θ which is nonetheless entirely uninformative about θ. What do you conclude from the fact that this is possible? [6 marks]

Answer. If we allow ourselves an auxiliary uniform random variable U, then

$$\mathcal{C}(\boldsymbol{y}) := \begin{cases} \Omega & U \leq 0.95 \\ \emptyset & U > 0.95 \end{cases}$$

has a coverage of exactly 95%, and yet it is useless, because it does not depend on $\boldsymbol{Y}$ or $\mathrm{p}(\cdot; \theta)$ at all. If we do not allow ourselves a U, then we use $\boldsymbol{y}$ to seed a numerical (deterministic) random number generator, and set U equal to, say, the millionth value. Like P-values, this result indicates that confidence set $\mathcal{C}$ needs to be carefully chosen, in order to be informative about θ.

(c) State and prove the *marginalisation theorem* for confidence sets. [6 marks]

Answer. Let $g : \theta \mapsto \phi$ be a specified function, and $\mathcal{C}$ be a level-β confidence set for θ. Then $g\mathcal{C}$ is a level-β confidence for ϕ. By definition, $\theta \in \mathcal{C}(\boldsymbol{y})$ implies $\phi \in g\mathcal{C}(\boldsymbol{y})$ for all $\boldsymbol{y}$, and hence

$$\Pr\{\theta \in \mathcal{C}(\boldsymbol{Y}); \theta\} \leq \Pr\{\phi \in g\mathcal{C}(\boldsymbol{Y}); \theta\}$$

for each $\theta \in \Omega$. But since $\mathcal{C}$ is a level-β confidence set for θ, $\beta \leq \Pr\{\theta \in \mathcal{C}(\boldsymbol{Y}); \theta\}$ for all $\theta \in \Omega$, and hence $\beta \leq \Pr\{\phi \in g\mathcal{C}(\boldsymbol{Y}); \theta\}$ for all $\theta \in \Omega$, showing that $g\mathcal{C}$ is a level-β confidence set for ϕ.

[This is a good question for making sure that you really understand the difference between $\boldsymbol{y}$ and $\boldsymbol{Y}$. Also make sure you always put a '$;\theta$' into probability statements for $\boldsymbol{Y} \sim \mathrm{p}(\cdot; \theta)$.]

(d) Describe a general-purpose approach for computing a 95% confidence set for θ, based on level sets of the form

$$\mathcal{C}(\boldsymbol{y}) := \left\{\theta : \log \mathrm{p}(\boldsymbol{y}; \theta) \geq \log \mathrm{p}(\boldsymbol{y}; \hat{\theta}(\boldsymbol{y})) - k\right\},$$

where $\hat{\theta}(\boldsymbol{y})$ is the Maximum Likelihood (ML) estimator for θ. Include in your description a justification for the form given above, an explanation of *level error*, and a sampling-based approach for reducing level error. [7 marks]

Answer. Asymptotic theory suggests that, approximately,

$$2 \log \frac{\mathrm{p}(\boldsymbol{Y}; \hat{\theta}(\boldsymbol{Y}))}{\mathrm{p}(\boldsymbol{Y}; \theta)} \sim \chi^2_p \quad \text{under the model}$$

when size of $\boldsymbol{Y}$ is large, where p is the dimension of Ω. Hence

$$\Pr\left\{2 \log \frac{\mathrm{p}(\boldsymbol{Y}; \hat{\theta}(\boldsymbol{Y}))}{\mathrm{p}(\boldsymbol{Y}; \theta)} \leq \chi_p^{-2}(0.95); \theta\right\} \approx 0.95 \quad \text{for all } \theta \in \Omega$$

where $\chi_p^{-2}(0.95)$ is the 95th percentile of the χ^2_p distribution. Rearranging shows that

$$\Pr\left\{\log \mathrm{p}(\boldsymbol{Y}; \theta) \geq \log \mathrm{p}(\boldsymbol{Y}; \hat{\theta}(\boldsymbol{Y})) - \chi_p^{-2}(0.95)/2; \theta\right\} \approx 0.95 \quad \text{for all } \theta \in \Omega$$

i.e. $\mathcal{C}$ as defined above is approximately a 95% confidence set for θ, when $k \leftarrow \chi_p^{-2}(0.95)/2$.

Level error is the difference between the nominal coverage (95%) and the actual coverage, which may not be the same as the nominal coverage because the asymptotic conditions do not hold. Therefore using $k \leftarrow \chi_p^{-2}(0.95)/2$ might induce level error: a different value for k might be better.

For given observations $\boldsymbol{y}$, a sampling-based approach can be used to adjust k to get coverage of 95% at the value of the MLE, $\hat{\theta}(\boldsymbol{y})$. Many datasets $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(n)}$ are simulated independently from $\mathrm{p}(\cdot; \hat{\theta}(\boldsymbol{y}))$, and after that the value of k is adjusted until exactly 95% of the resulting confidence sets contain $\hat{\theta}(\boldsymbol{y})$, i.e. we adjust k until the coverage is 95% at $\hat{\theta}(\boldsymbol{y})$. By smoothness, we also expect the coverage to be roughly 95% in the region around $\hat{\theta}(\boldsymbol{y})$, and hence that our confidence is approximately an exact confidence set for θ, at least in that region.

If you would like to hand in this homework for marking, please do so by 5pm on Wed 6 May, in the box outside my offce.

Jonathan Rougier
Mar 2015