IV - Hypothesis testing and credible sets

For a given statistical model $\{f(\cdot|\theta), \theta \in \Theta\}$ the problem of hypothesis testing consists in answering the following question:

> Is the hypothesis that θ belongs to the subset $\Theta_0 \subset \Theta$ of the parameter space acceptable?

In short, we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta_1$, with $\Theta_1 = \Theta \setminus \Theta_0$.

Example: *Consider the following logistic regression model:*

$$\mathbb{P}(Y = 1|Z = z) = \exp(\beta_0 + \beta_1 z)/\{1 + \exp(\beta_0 + \beta_1 z)\}.$$

- *How gender affects a given behaviour?*

$$H_0 : \beta_1 > 0, \quad H_1 : \beta_1 \leq 0.$$

- *Does the presence of a nuclear plant increases the risk of leukemia in its vicinity?*

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 > 0.$$

- *Does weather affects the sales of a given product?*

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0.$$

Hypothesis testing as a decision theoretic problem

Hypothesis testing is clearly a decision theoretic problem (see Chapter 2) where the set of possible decisions contains only two elements

$$\mathcal{D} = \{0, 1\}.$$

By convention, 1 stands for the acceptance of H_0 and 0 for its rejection.

The most natural loss function in that particular context is the a_0–a_1 loss defined, for $(\theta, d) \in \Theta \times \mathcal{D}$ and $a_0, a_1 \geq 0$, by

$$L(\theta, d) = \begin{cases} 0 & \text{if } d = \mathbf{1}_{\Theta_0}(\theta), \\ a_0 & \text{if } \theta \in \Theta_0 \text{ and } d = 0, \\ a_1 & \text{if } \theta \in \Theta_1 \text{ and } d = 1. \end{cases}$$

Example: *Spam filtering consists in detecting and discarding automatically any unsolicited e-mail that gets into your mailbox. The most elaborate spam filters (e.g. Mozilla/Netscape) rely on a Bayesian modelling of the occurrence of particular terms. This is a testing problem (spam/not spam) in which the loss function is clearly not symmetrical (the 'cost' of discarding a non-spam e-mail being more important than of keeping a spam mail).*

In some contexts it is not necessarily clear what are the respective costs of each decision, so one may set arbitrarily $a_0 = a_1 = 1$ (and recover the 0–1 loss function introduced in Chapter 2).

Remark: Hypothesis testing can also be considered as the problem of estimating the function of θ defined by $g(\theta) := \mathbf{1}_{\Theta_0}(\theta)$.

Bayesian estimator and a_0–a_1 loss function

Proposition 4.1 *The decision rule $\delta^\pi : \mathcal{X} \to \{0, 1\}$ associated with the a_0–a_1 loss function is defined, for $x \in \mathcal{X}$, by*

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \pi(\Theta_0|x) > a_1/(a_0 + a_1), \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Done in class.

Remarks:

- If $a_0 = a_1$ this procedure simply consists in selecting the hypothesis with the highest posterior probability.
- This procedure only depends on a_0/a_1. The larger this ratio is, the more important a wrong answer under H_0 is relative to H_1.
- The quantity $a_1/(a_0 + a_1)$ is referred to as the acceptance level.
- Since the a_0–a_1 loss is bounded, as soon as $\pi(\theta)$ is a proper prior density the Bayes risk is finite (i.e. $r(\pi) < +\infty$) and thus δ^π is the Bayes estimator of $g(\theta) := \mathbf{1}_{\Theta_0}(\theta)$.

Notation: We recall that for a measurable set $A \subseteq \Theta$ we use the notation

$$\pi(A|x) = \int_A \pi(\theta|x)\mathrm{d}\theta, \quad \pi(A) = \int_A \pi(\theta)\mathrm{d}\theta.$$

The Bayes factor

Bayesian hypothesis testing (and Bayesian model choice, see Chapter 5) are often performed using the Bayes factor.

Definition 4.1 *The Bayes factor* $B_{01}^{\pi} : \mathcal{X} \to [0, +\infty)$ *for the test* $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_1$ *is defined by*

$$B_{01}^{\pi}(x) = \frac{\pi(\Theta_0|x)}{\pi(\Theta_1|x)} \frac{\pi(\Theta_1)}{\pi(\Theta_0)} \quad x \in \mathcal{X}.$$

The Bayes factor therefore measures the modification of the odds of H_0 against H_1 due to the observation of x. It allows for (partly) reducing the impact of the prior distribution on the decision and as such is considered as an 'objective' quantity.

To understand this last point let $\pi_i(\theta)$ be a prior density under H_i $(i = 0, 1)$, $\rho_0 \in (0, 1)$ and

$$\pi(\theta) = \rho_0 \pi_0(\theta) + (1 - \rho_0)\pi_1(\theta) \implies \pi(\Theta_0) = \rho_0.$$

Then,

$$B_{01}^{\pi}(x) = \frac{\int_\Theta f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\int_\Theta f(x|\theta)\pi_1(\theta)\mathrm{d}\theta} = \frac{m_0(x)}{m_1(x)} \tag{1}$$

where

$$m_i(x) = \int_\Theta f(x|\theta)\pi_i(\theta)\mathrm{d}\theta, \quad i = 0, 1. \tag{2}$$

Consequently, $B_{01}^{\pi}(x)$ does not depend on $\pi(\Theta_0)$, the prior probability that H_0 is true.

Remark: If $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, then

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

so that the Bayes factor reduces to the likelihood ratio.

Duality between loss and prior distribution

Proposition 4.2 *The decision rule $\delta^\pi : \mathcal{X} \to \{0, 1\}$ associated with the a_0–a_1 loss function can be alternatively defined, for $x \in \mathcal{X}$, by*

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } B^\pi_{01}(x) > \frac{a_1\pi(\Theta_1)}{a_0\pi(\Theta_0)}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Done in class.

Remark: In the expression of δ^π given in Proposition 4.1 the impact of $\pi(\Theta_0)$ on the testing procedure is hidden in the quantity $\pi(\Theta_0|x)$ while in the one given in Proposition 4.2 it is explicit.

Remark: $a_i\pi(\Theta_i)$ is the prior expected loss of not choosing H_i.

This alternative condition for accepting H_0 illustrates the fundamental duality between the prior distribution and the loss function.

Indeed, it is equivalent to:

- To modify the penalties as follows

  $$a_0 \to a_0\,\pi(\Theta_0), \quad a_1 \to a_1\,\pi(\Theta_1)$$

  and then set $\pi(\Theta_0) = \pi(\Theta_1) = 1/2$;

- To modify the prior probabilities as follows:

  $$\pi(\Theta_0) \to \frac{a_0\,\pi(\Theta_0)}{a_0\pi(\Theta_0) + a_1\pi(\Theta_1)}, \quad \pi(\Theta_1) \to \frac{a_1\,\pi(\Theta_1)}{a_0\pi(\Theta_0) + a_1\pi(\Theta_1)}$$

  and then set $a_0 = a_1$.

Jeffreys' scale

The strength of the evidence in favour of H_0 is sometimes measured through a scale proposed by Jeffreys (1961) which goes a follows:

- The evidence is poor if $\log_{10} B_{01}^{\pi}(x)$ is between 0 and 0.5;
- The evidence is substantial if $\log_{10} B_{01}^{\pi}(x)$ is between 0.5 and 1;
- The evidence is strong if $\log_{10} B_{01}^{\pi}(x)$ is between 1 and 2;
- The evidence is decisive if $\log_{10} B_{01}^{\pi}(x)$ is larger than 2;

Remark: The bounds separating one strength from another are mostly a matter of convention (and not derived using decision theory).

Point-null hypotheses

We now consider tests where $\Theta_0 = \{\theta^*\}$ for some $\theta^* \in \Theta$.

Some statisticians argue that point-null hypotheses are intrinsically absurd when Θ is a continuous space. Indeed, how can we determine that a parameter attains an *exact* value (in other words, that it is exactly known) from a finite amount of data?

In some cases however, point-null hypotheses make more sense because they are related to a *qualitative* aspect of the model.

Example: *In the Logistic regression model*

$$\mathbb{P}(Y = 1|Z = z) = \exp(\beta_0 + \beta_1 z)/\{1 + \exp(\beta_0 + \beta_1 z)\}$$

the hypothesis $H_0 : \beta_1 = 0$ *is equivalent to assuming that the factor* z *has no impact on the event* y.

In any cases point-null hypotheses are very popular in practice and thus there is a need for the Bayesian point-null hypothesis testing procedure that we introduce below.

Bayes factors for point-null hypotheses

If a given null hypothesis is reasonable for the problem at hand, then its prior probability must be positive (and thus $\pi(\theta)$ cannot be a continuous prior distribution).

Let $\rho_0 \in (0,1)$ be the prior probability that $\theta = \theta^*$ and $g_1(\theta)$ be a prior density on $\Theta_1 = \Theta \setminus \{\theta^*\}$.

Then, we consider the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{\theta^*\}}(\theta) + (1-\rho_0) g_1(\theta) \mathbf{1}_{\{\theta \neq \theta^*\}}(\theta), \quad \theta \in \Theta. \tag{3}$$

Remark: When we write $\pi(\theta)$ as in (3) we implicitly assume that $\mathrm{d}\theta$ is such that $\int_{\{\theta^*\}} \pi(\theta)\mathrm{d}\theta = \rho_0$.

In this case,

$$\pi(\{\theta^*\}|x) \propto \rho_0 f(x|\theta^*)$$

while

$$\pi(\Theta_1|x) \propto (1-\rho_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, \mathrm{d}\theta.$$

Using the notation introduced above,

$$\pi_0(\theta) = \mathbf{1}_{\{\theta^*\}}(\theta), \quad \pi_1(\theta) = g_1(\theta)\mathbf{1}_{\{\theta \neq \theta^*\}}(\theta)$$

and thus, by (1),

$$B_{01}^{\pi}(x) = \frac{f(x|\theta^*)}{m_1(x)}$$

where $m_1(x)$ is defined in (2); that is,

$$m_1(x) = \int_{\Theta} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta = \int_{\Theta_1} f(x|\theta) g_1(\theta)\mathrm{d}\theta.$$

Remark: As per above the Bayes factor does not depend on ρ_0.

Hypothesis testing with noninformative prior distributions

Remark first that the testing setting is not coherent with an absolute lack of information since it implies to partition the parameter space into two sets.

In addition, for point-null hypotheses, the set Θ_0 has in general measure zero under the Laplace and the Jeffreys prior densities.

Moreover, Laplace's prior (and more generally improper prior densities) poses problems for hypothesis testing as illustrated in the following example.

Example 4.1 *Let $f(\cdot|\,\theta)$ be the p.d.f. of the $\mathcal{N}_1(\theta, 1)$ distribution and consider the test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Then, if we use the prior $\pi(\theta)$ defined by*

$$\pi(\theta) = \frac{1}{2}\mathbf{1}_{\{0\}}(\theta) + \frac{1}{2}\mathbf{1}_{\{\theta\neq 0\}}(\theta), \quad \theta \in \Theta := \mathbb{R}$$

we have

$$\pi(\{0\}|x) = \frac{1}{1+\sqrt{2\pi}e^{x^2/2}} \leq \frac{1}{1+\sqrt{2\pi}} \approx 0.285, \quad \forall x \in \mathbb{R}^d.$$

The posterior probability $\pi(\{0\}|x)$ is therefore bounded above so that the posterior distribution is biased toward H_1. Consequently, if the loss function does not take this bias into account the null hypothesis will be often rejected.

More one testing with improper prior distributions

Example 4.1 (continued) *Assume now that $\pi(\theta)$ is defined by*

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{0\}}(\theta) + (1-\rho_0)\, c\, \mathbf{1}_{\{\theta \neq 0\}}(\theta), \quad \theta \in \Theta$$

for some constant $c > 0$ and where $\rho_0 \in (0,1)$.

Then, we have

$$B_{01}^{\pi}(x) = \frac{1}{c\sqrt{2\pi}\, e^{x^2/2}}$$

and therefore, by increasing (resp. decreasing) the parameter c, we can make the Bayes factor arbitrarily small (resp. large).

Conclusion: Improper prior densities should be avoided in the context of hypothesis testing.

Remark: Another reason why improper prior densities should be avoided in the context of hypothesis testing is that $\pi(\Theta_i) = +\infty$ for at least on $i \in \{0,1\}$, so that the testing procedure given in Proposition 4.2 is meaningless.

Testing with a vague prior distribution

Example 4.1 (end) *Assume now that* $\pi(\theta)$ *is defined by*

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{0\}}(\theta) + (1-\rho_0) g_1(\theta) \mathbf{1}_{\{\theta \neq 0\}}(\theta), \quad \theta \in \Theta := \mathbb{R}$$

where $g_1(\theta)$ *is the density of the* $\mathcal{N}_1(0, \sigma_0^2)$ *distribution. Then, we have*

$$\pi(\{0\}|x) = \left[1 + \frac{1-\rho_0}{\rho_0} \sqrt{\frac{1}{1+\sigma_0^2}} \exp\left(\frac{\sigma_0^2 x^2}{2(1+\sigma_0^2)}\right)\right]^{-1}, \quad x \in \mathbb{R}^d$$

which converges to 1 as $\sigma_0^2 \to +\infty$ *for every observation* x*!*

If instead we use for $g_1(\theta)$ *the Laplace's prior* $g_1(\theta) \equiv 1$ *we obtain, as shown above,*

$$\pi(\{0\}|x) = \left[1 + \sqrt{2\pi} e^{x^2/2}\right]^{-1}, \quad x \in \mathbb{R}^d.$$

Consequently, and contrary to what we saw for point estimation (see Chapter 3), limiting arguments (i.e. letting the prior variance going to infinity) do not allow to derive uninformative answers in the context of hypothesis testing.

Credible intervals and confidence intervals

A topic closely related to hypothesis testing is the derivation of regions of the parameter space that contain the 'most likely values' for the parameter, called credible sets in the Bayesian approach.

The frequentist counterpart of credible sets are the *confidence interval* which are usually derived from the asymptotic distribution of the estimator.

For instance, for an univariate parameter θ, an $(1-\alpha)$-confidence interval is typically

$$\left[\widehat{\theta}_n - q_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n + q_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right]$$

where n is the sample size, $\widehat{\theta}_n$ is the frequentist estimator of θ (e.g. the MLE), $\widehat{\sigma}_n^2$ is an estimator of the asymptotic variance of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, with θ_0 the "true" parameter value, and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the $\mathcal{N}_1(0,1)$ distribution, e.g. $q_{1-\alpha/2} = 1.96$ for $\alpha = 0.05$.

Reminder: In the frequentist perspective, the parameter is fixed and the confidence interval is random, having a probability of $(1-\alpha)$ to actually contain θ_0 (when we repeat the same experiment a great number of times). It is therefore not possible to interpret $(1-\alpha)$ as the probability that the parameter lies in the confidence interval *for the considered experiment.*

Credible sets

Definition 4.2 *Let $\alpha \in (0,1)$. A subset $C \subseteq \Theta$ is an $(1-\alpha)$-credible set for θ if*

$$\pi(C|x) \geq 1 - \alpha.$$

The notion of credible sets is not very useful by itself as there are obviously an infinite number of credible sets for a given α.

In an univariate setting, one may restrict its attention to the credible interval centred at a given Bayesian estimator. This is arbitrary however, as it depends on a particular choice of Bayesian estimator (posterior mean, median, etc.) and such intervals may not exist when Θ is bounded, as illustrated in the next example.

Example: Assume that the posterior distribution is the Beta(1,30) distribution. Then, the posterior mean is 1/31, and the posterior probability of being above 1/31 is approximatively 0.37. Therefore credible intervals centred at 1/31 exists only if $1-\alpha$ is smaller than 0.74.

Highest posterior density regions

A more satisfactory approach is to restrict our attention to the credible set that contains the 'most likely values'.

Definition 4.3 *The subset $C_\alpha(x)$ of the parameter space is a highest posterior density (HPD) region at level $(1-\alpha)$ if it is of the form*

$$\{\theta \in \Theta : \pi(\theta|x) > \gamma_\alpha\} \subset C_\alpha(x) \subset \{\theta \in \Theta : \pi(\theta|x) \geq \gamma_\alpha\}$$

where γ_α is the largest bound such that

$$\pi(C_\alpha(x)|x) \geq 1-\alpha.$$

Remarks:

- HPD regions minimize the volume among $(1-\alpha)$-credible regions.
- If $\pi(\theta|x)$ is a continuous density the HPD region is simply

$$C_\alpha(x) = \{\theta \in \Theta : \pi(\theta|x) \geq \gamma_\alpha\}.$$

- If the choice of an HPD region among $(1-\alpha)$-credible sets is natural, it is also justified (to some extend) from a decision theoretic perspective.