STAT 200B 2019 Week09

Soyeon Ahn

1 Fisher information (revisited.)

1.1 score

The score function
$$s(X;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$$

The Fisher information (based on n observations) is

$$\begin{aligned}
I_n(\theta) &= V_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right)\\
&= V_\theta\left(\sum_{i=1}^n s(X_i;\theta)\right) \text{(definition of the Fisher information)}\\
&= \sum_{i=1}^n V_\theta(s(X_i;\theta)) \quad \text{(if } X_1,\dots,X_n \text{ are independent)}\\
&= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1,\dots,X_n \text{ are identically distributed)}\\
&= nI_1(\theta) \quad \text{(when } n = 1\text{)}\\
&\equiv nI(\theta)
\end{aligned}$$

1.2 Fisher information

(Lemma 5.3. in Theory of Point Estimation by Lehmann and Casella)

The expectation of the score is zero by differentiating

$$\int f(x;\theta)dx = 1. \tag{1}$$

$$\begin{aligned}
\int \frac{\partial}{\partial\theta} f(x;\theta)dx &= \int \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} f(x;\theta)dx\\
&= \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right) f(x;\theta)dx\\
&= E\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]
\end{aligned}$$

If, in addition, the second derivative with respect to θ of $\log f(x;\theta)$ exists for all x and θ and the second derivative with respect to θ of the left side (1) can be obtained by differentiating twice under the integral sign, then

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

We have

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} - \left[\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\right]^2 \tag{2}$$

and

$$\begin{aligned} E_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)}\right] &= \int \frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx \\ &= \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = 0 \end{aligned}$$

If we take the expectation of both sides of (2),

$$\begin{aligned} I(\theta) = V_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right) &= E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2 \\ = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] \end{aligned}$$

The “observed” Fisher information

$$I_n^{obs}(\theta) = \frac{-\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log f(X_i;\theta)$$

measures the curvature of the log-likelihood function. In particular $I_n^{obs}(\hat{\theta})$ measures the curvature at the MLE. The more peaked $\ell_n(\theta)$ is around $\hat{\theta}$, the more ”information” the likelihood gives us. $I(\theta)$ measures the average value of this quantity.

1.3 Cramér-Rao bound, information equality

(Theorem 5.10 in Theory of Point Estimation by Lehmann and Casella and Theorem 3.3 in Mathematical Statistics by Jun Shao)

Suppose that $T(X)$ is an estimator with $E[T(X)] = g(\theta)$ being a differentiable function of θ. The inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ.

2 Prior

2.1 Different prior choices

- Subjective Bayesianism: The prior should reflect in as much detail as possible the researcher's prior knowledge of and uncertainties about the problem. These should be determined through *prior elicitiation.*
- Objective Bayesianism: The prior should incorporate as little information as possible. Priors with this property are known as *non-informative.*
- Robust Bayesianism: Reasonable people may hold different priors, and it is difficult to precisely express even one person's prior; we should therefore consider the *sensitivity* of our inferences to changes in the prior.

2.2 Noninformative prior

(Section 2.9 in Bayesian Data Analysis, by Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin)

The rationale for using noninformative (vague, flat, diffuse) prior distributions is often said to be 'to let the data speak for themselves', so that inferences are unaffected by information external to the current data.

2.3 Improper prior can lead to proper posterior

- proper prior: if it does not depend on data and integrates to 1.
- improper prior: if it does not integrated to a positive finite value.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. We could take $f(\theta) \propto 1$. This prior is called "improper", since $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$.

However, the posterior density $\propto \exp\left\{-\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n]\right\} \sim N(\bar{X}_n, 1/n)$ and the posterior is proper.

2.4 Flat priors are not invariant

$X \sim Bernoulli(p)$ and suppose we use the flat prior $f(p) = 1$. Now let $\psi = \log \frac{p}{1-p}$. This a transformation of p. The pdf of ψ is not flat:

$$\frac{\partial}{\partial \psi} \frac{\exp^{\psi}}{1 + \exp^{\psi}} = \frac{\exp^{\psi}}{(1 + \exp^{\psi})^2}$$

since

$$P(\Psi \leq \psi) = P\left(\frac{p}{1-p} \leq \psi\right) = P\left(p \leq \frac{\exp^{\psi}}{1 + \exp^{\psi}}\right) = \frac{\exp^{\psi}}{1 + \exp^{\psi}}$$

Flat priors are not transformation invariant.

3 Jeffreys' prior

Consider one-to-one transformations of the parameter $\psi = g(\theta)$. By transformation of variables, the prior density $f(\theta)$ is equivalent, in terms of expression the same beliefs, to the following prior density on ϕ:

$$f_\phi(\phi) = f_\theta(\theta) \left| \frac{\theta}{\phi} \right| = f_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|$$

Jeffreys' principle leads to defining the noninformative prior density as

$$f(\theta) \propto I(\theta)^{1/2}$$

To see transformation invariant, evaluate $I(\phi)$ at $\theta = g^{-1}$:

$$\begin{aligned} I(\phi) &= -E\left[\frac{\partial^2 \log f(y;\phi)}{\partial \phi^2}\right] \\ &= E\left[\frac{\partial^2 \log f(y;\theta = g^{-1}(\phi))}{\partial \theta^2} \left|\frac{\partial \theta}{\partial \phi}\right|^2\right] \\ &= I(\theta)\left|\frac{\partial \theta}{\partial \phi}\right|^2 \end{aligned}$$

Thus,

$$I(\phi)^{1/2} = I(\theta)^{1/2}\left|\frac{\partial \theta}{\partial \phi}\right|$$

3.1 Example: Bernoulli

Consider the Bernoulli $X \sim Bernoulli(p)$ and $I(p) = \frac{1}{p(1-p)}$

Jeffreys' prior is

$$f(p) \propto \sqrt{p(1-p)} = p^{-1/2}(1-p)^{-1/2} \sim Beta(1/2, 1/2)$$

A flat prior (proper prior) is

$$Uniform(0,1) \sim Beta(1,1)$$

A conjugate prior (proper prior) is

$$Beta(\alpha, \beta)$$

where $s = \sum_i x_i$, the number of success. and its posterior is

$$Beta(\alpha + s, \beta + n - s)$$

3.2 Example: Poisson

$X \sim Poisson(\lambda)$, $f(x;\lambda) \sim \lambda^x \exp^{-\lambda}$ and $I(\lambda) = \frac{1}{\lambda}$.

A conjugate prior is $Gamma(\alpha, \beta)$ (Please note that the parameter β is a rate.)

$$\propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp^{-\beta\lambda}$$

and its posterior is $Gamma(\alpha + \sum_i x_i, \beta + n)$

$$\propto \lambda^{\alpha+\sum_i x_i - 1} \exp^{-(\beta+n)\lambda}$$

If we use a flat prior (improper prior!), the posterior is $Gamma(1 + \sum_i x_i, n)$

If we use Jeffreys' prior (improper prior!),

$$f(p) \propto \frac{1}{\sqrt{\lambda}} \sim \lambda^{\frac{1}{2}-1}$$

the posterior is $Gamma(1/2 + \sum_i x_i, n)$

$$\propto \lambda^{\frac{1}{2}+\sum_i x_i - 1} \exp^{-(n)\lambda} \sim Gamma(\frac{1}{2} + \sum_i x_i, n)$$

Note that flat prior and Jeffreys prior can be regarded as $Gamma(1, 0)$, and $Gamma(\frac{1}{2}, 0)$ although they are not valid densities.