

A note on Fisher information

<http://marekrychlik.com/node/20>

Fisher information is defined in terms of the likelihood function and the score function.

Given a family of *probability density functions* $f(x|\theta)$ parameterized by a parameter θ , the function $\theta \mapsto f(x|\theta)$ is called the **likelihood function** and denoted by $L(\theta|x)$ or $L(\theta; x)$.

The likelihood function is **not** a probability density function in the parameter space, but it sometimes can be normalized to such and interpreted in this manner.

The likelihood function is related to the Bayesian point of view on statistics. Its most fundamental application is to deriving **maximum likelihood estimators**. For instance, the sample mean is the maximum likelihood of the true mean of a population.

The **score function** is the following function:

$$\frac{\partial}{\partial \theta} \log L(\theta; X) = \frac{1}{L(\theta; X)} \cdot \frac{\partial L(\theta; X)}{\partial \theta}.$$

We note that this is a one-parameter family of random variables because of the dependence on the outcome X of the experiment.

The likelihood function is a function of the parameter and of the event $X = X(\omega)$. Thus, it is a family of random variables indexed with θ . Hence, it makes sense to talk about the expected value and variance of $L(\theta; X)$, where the probability measure is that corresponding to the parameter θ . Similarly, we may ask questions about the expected value and variance of the score function.

It is easy to see that the expected value of the score function is always zero.

The *Fisher information* is the variance of the score function. It is denoted by $\mathcal{I}(\theta)$.

Thus, explicitly written, **Fisher information** is defined as:

$$\mathcal{I}(\theta) = \mathbb{E} \left\{ \left[\frac{\partial}{\partial \theta} \log L(\theta; X) \right]^2 \middle| \theta \right\}.$$

For calculations, the following formula is important:

$$\mathcal{I}(\theta) = -\mathbb{E} \left\{ \frac{\partial^2}{\partial \theta^2} \log L(\theta; X) \middle| \theta \right\}.$$

which can be justified under some reasonable integrability and differentiability assumptions, so that one can integrate by parts.

Exercises:

- If $f(x|\sigma) = N(0, \sigma^2)$ then

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2}.$$

- Calculate $\mathcal{I}(\theta)$ for the Bernoulli process.
- Calculate $\uparrow(\theta)$ for other distributions.

The **Cramer-Rao Inequality** yields a uniform bound on the variance of an estimator $g(X)$ of a function $\psi(\theta)$ of a parameter of a probability density function $f(x|\theta)$:

$$\text{var}(g(X)) \geq \frac{|\psi'(\theta)|^2}{\mathcal{I}(\theta)}$$

In particular, if $g(X)$ is an *unbiased estimator* of θ , i.e.

$$\mathbb{E}(g(X)|\theta) = \theta$$

for all θ then

$$\text{var}(g(X)) \geq \frac{1}{\mathcal{I}(\theta)}$$

The **Hammersley–Chapman–Robbins Inequality** goes in the same direction, but it is often stronger and does not require differentiability of the distribution function:

$$\text{var}(g(X)) \geq \sup_{\Delta} \frac{|\psi(\theta + \Delta) - \psi(\theta)|^2}{\mathbb{E} \left\{ \left(\frac{f(x|\theta+\Delta)}{f(x|\theta)} - 1 \right)^2 \middle| \theta \right\}}.$$

The Cramer-Rao bound generalizes to random vectors, where the variance is replaced with covariance, and the Fisher information also becomes a symmetric matrix. The inequality is interpreted as positive definiteness of the difference of two symmetric matrices.