3 Sufficiency and Information

3.1 Sufficiency

Context:

- Parametric statistical model $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$
- $X_1, \ldots, X_n$ iid random sample of size n extracted from X
- Goal: use the information contained in $X_1, \ldots, X_n$ to produce inferential statements about the unknown parameter θ
- When we compute statistics, i.e., functions of the random sample, we are summarizing the information contained in the random sample. In what circumstances can we be assured that in the process we are not losing any relevant information about the parameter?

Example 3.1 *Let* $X \mid \lambda \sim Po(\lambda)$, *where* $\lambda > 0$. *Imagine, for example, that* $X =$ *"number of calls that a call center receives in a day". We observe a random sample of size* $n = 2$, *but we only know that the observed value of the statistic* $T = \sum_{i=1}^{n} X_i$ *is* $t = 31$. *In these circumstances, what can we say about the random sample* X_1, X_2*?*

$$P(X_1 = x_1, X_2 = x_2 \mid T = 31, \lambda) = \frac{P(X_1 = x_1, X_2 = x_2, T = 31 \mid \lambda)}{P(T = 31 \mid \lambda)}$$

$$= \begin{cases} 0 & if\ x_1 + x_2 \neq 31 \\ \frac{P(X_1 = x_1, X_2 = x_2 \mid \lambda)}{P(T = 31 \mid \lambda)} & if\ x_1 + x_2 = 31 \end{cases}$$

Example 3.1 *(cont.) Since* $T \mid \lambda \sim Po(2\lambda)$*, it follows that, if* $x_1 + x_2 = 31$*,*

$$\frac{P(X_1 = x_1, X_2 = x_2 \mid \lambda)}{P(T = 31 \mid \lambda)} = \frac{\prod_{i=1}^{2} e^{-\lambda}\ \lambda^{x_i}/x_i!}{e^{-2\lambda}(2\lambda)^{31}/31!} = \frac{31!}{(31 - x_1)!x_1!}\left(\frac{1}{2}\right)^{x_1}\left(1 - \frac{1}{2}\right)^{31-x_1}$$

that is,

$$X_1 \mid T = 31, \lambda \sim B(31, 1/2)$$

which means that the number of calls, amongst the 31, that correspond to the first component of the random sample have the same distribution as the number of tails obtained in 31 flips of a balanced coin!

Definition 3.1 [Sufficient statistic] *We say that a statistic T is sufficient for $\mathcal{F}$ (or, in a simplified manner, for θ), if the conditional distribution of the random sample given the observed value of T does not depend on the unknown parameter θ, for all θ.* ■

Example 3.2 *Suppose that $X \mid \theta \sim B(1, \theta)$ and that $X_1, \ldots, X_n$ corresponds to a random sample from X. Intuitively, it's clear that $T = \sum_{i=1}^{n} X_i$ should be sufficient for θ, since the order by which the successes are obtained should be irrelevant for any inferences about θ. In fact, and if $T(x_1, \ldots, x_n) = t$*

$$f(x_1, \ldots, x_n \mid t, \theta) = \frac{f(x_1, \ldots, x_n \mid \theta)}{f(t \mid \theta)} = \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \binom{n}{t}^{-1}$$

which does not depend on θ, which shows that our intuition is correct.

We should retain that, in general,

$$f(x_1, \ldots, x_n \mid t, \theta) = \frac{f(x_1, \ldots, x_n, t \mid \theta)}{f_T(t \mid \theta)}$$

$$= \begin{cases} 0 & \text{if } T(x_1, \ldots, x_n) \neq t \\ \frac{f(x_1, \ldots, x_n \mid \theta)}{f_T(t \mid \theta)} & \text{if } T(x_1, \ldots, x_n) = t \end{cases}$$

The definition of sufficient statistic is not very useful to discover sufficient statistics. However,

Theorem 3.1 [Halmos-Savage Factorization Criterion] *A statistic T is sufficient for θ if and only if there are non-negative functions g and h such that*

- *g depends on θ and on the random sample exclusively through the observed value of T*
- *h depends exclusively on the random sample*
- $f(x_1, \ldots, x_n \mid \theta) = g(T(x_1, \ldots, x_n); \theta) \ \times \ h(x_1, \ldots, x_n)$.

[Idea of the proof.]

Example 3.3 *Let $X \mid \lambda \sim Po(\lambda)$. Then, with $t = \sum_{i=1}^{n} x_i$,*

$$f(x_1, \ldots, x_n \mid \lambda) = \prod_{i=1}^{n} e^{-\lambda} \lambda_i^x / x_i! = \underbrace{e^{-n\lambda} \lambda^t}_{g(t;\lambda)} \times \underbrace{\prod x_i!^{-1}}_{h(x_1,\ldots,x_n)}$$

and the factorization criterion allows us to conclude that $T = \sum_{i=1}^{n} X_i$ is sufficient for λ.

Example 3.4 *Let $X_1, \ldots, X_n$ be a random sample of size n extracted from a population $X \mid \theta \sim U(0, \theta)$, $\theta > 0$. In this case, the support of the distribution of X depends on the unknown parameter, so we must be careful:*

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\theta} \; I_{[0,\theta]}(x_i) \\ &= \underbrace{\frac{1}{\theta^n} \; I_{(-\infty,\theta]}(x_{(n)})}_{g(x_{(n)};\theta)} \; \underbrace{I_{[0,+\infty)}(x_{(1)})}_{h(\mathbf{x})} \end{aligned}$$

and the factorization criterion allows us to conclude that the statistic $T = X_{(n)}$ is sufficient for θ.

Example 3.5 *Another example, the shifted exponential model:*
$f(x \mid \lambda, \delta) = \lambda \; e^{-\lambda(x-\delta)} \; I_{[\delta,+\infty)}(x)$

Partition induced in the sample space by a statistic:

- Any statistic $T : \mathcal{X} \to \mathbb{R}^q$ induces a partition in the sample space $\mathcal{X}$
- The finer the partition induced by T in $\mathcal{X}$, the less information is lost; the smaller is the data reduction operated by T
- A sufficient statistic operates a data reduction that does not involve loss of relevant information about the parameter; the partition it induces is also said to be sufficient
- The notion of partition is more general than that of statistic; different statistics can induce the same partition, in which case they are said to be equivalent (they are one-to-one)
- If the partition induced by T is finer than the partition induced by S, then S is a function of T (HW problem). In that case, if S is sufficient, so is T, that is

$$\left.\begin{array}{l} S = h(T) \\ S \text{ sufficient} \end{array}\right\} \Rightarrow T \text{ sufficient}$$

- However, if T is sufficient and $S = h(T)$, it's not a given that S is also sufficient, unless h is injective, in which case S and T are equivalent.
- We are interested in finding statistics which are sufficient but operate the greatest data reduction. In other words, statistics that induce the coarsest partition that is still sufficient.

Definition 3.2 (Minimal sufficient statistic) *A statistic T is said to be minimal sufficient for $\mathcal{F}$ if it is sufficient and, if S is any other sufficient statistic, then $T = h(S)$ for some h.* ■

How do we find minimal sufficient statistics?

- Consider the binary relation in $\mathcal{X}$ defined by

$$\mathbf{y} \; R \; \mathbf{x} \Leftrightarrow \forall \; \theta \in \Theta, f(\mathbf{y} \mid \theta) = c(\mathbf{y}, \mathbf{x}) f(\mathbf{x} \mid \theta)$$

  where $c(\mathbf{y}, \mathbf{x}) > 0$ does not depend on θ

- This binary relation is an equivalence relation, that is, it is symmetric, reflexive and transitive. Hence, it induces a partition in $\mathcal{X}$ with parts

$$\Pi_{\mathbf{x}} = \{\mathbf{y} \in \mathcal{X} : \mathbf{y} \; R \; \mathbf{x}\} \text{ for } \mathbf{x} : f(\mathbf{x} \mid \theta) > 0 \text{ for some } \theta \in \Theta$$

$$\Pi_0 = \{\mathbf{y} \in \mathcal{X} : f(\mathbf{y} \mid \theta) = 0 \; \forall \theta \in \Theta\}$$

Theorem 3.2 [Lehmann-Scheffè] *The partition with parts* Π_0 *and* $\{\Pi_{\mathbf{x}}\}$ *described above is minimal sufficient, and any statistic which induces it is minimal sufficient.*

Sketch of the proof: Consider the statistic G that to each $\mathbf{x} \in \mathcal{X}$ associates a representative of the element of the partition to which it belongs:

$$\mathbf{x} \in \mathcal{X} \mapsto \mathbf{x}_\Pi \in \Pi_{\mathbf{x}} = G(\mathbf{x})$$

Let $\mathbf{x}$ be an element of $\Pi_{\mathbf{x}}$. Then,

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= c(\mathbf{x}, \mathbf{x}_\Pi)\ f(\mathbf{x}_\Pi \mid \theta) \\ &= c(\mathbf{x}, G(\mathbf{x}))\ f(G(\mathbf{x}) \mid \theta) \end{aligned}$$

and the factorization criterion guarantees that G is sufficient.

Let U be another sufficient statistic. Denote by $\Pi^\star = \{\Pi^\star_{\mathbf{x}}\}$ the partition induced by U. Let $\mathbf{y} \in \Pi^\star_{\mathbf{x}}$. Since U is sufficient

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= g(U(\mathbf{x}); \theta)\ h(\mathbf{x}) \\ f(\mathbf{y} \mid \theta) &= g(U(\mathbf{y}); \theta)\ h(\mathbf{y}) \\ &= g(U(\mathbf{x}); \theta)\ h(\mathbf{y}) \end{aligned}$$

because $\mathbf{x}$ and $\mathbf{y}$ belong to the same element of the partition induced by U. Hence,

$$f(\mathbf{y} \mid \theta) = \frac{h(\mathbf{y})}{h(\mathbf{x})} f(\mathbf{x} \mid \theta)$$

and as a consequence $\mathbf{y} \in \Pi_{\mathbf{x}}$. Conclusion: $\Pi^\star_{\mathbf{x}} \subset \Pi_{\mathbf{x}}$, and hence U induces a finer partition than G, i.e., G is a function of U.

Example 3.6 *Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. Let $\theta = (\mu, \sigma^2)$. It's easy to see that*

$$\frac{f(\mathbf{y} \mid \theta)}{f(\mathbf{x} \mid \theta)} = \exp\left\{-\frac{1}{2\sigma^2}\left[\sum y_i^2 - \sum x_i^2 - 2n\mu(\bar{y} - \bar{x})\right]\right\}$$

does not depend on θ **if and only if**

$$\begin{cases} \sum y_i^2 = \sum x_i^2 \\ \bar{y} = \bar{x} \end{cases}$$

So, two samples $\mathbf{x}$ and $\mathbf{y}$ belong to the same element of the partition iff the above holds. According to the theorem we have just proved, this partition is minimal sufficient. Any statistics which induces it is minimal sufficient, for example, $(\sum X_i^2, \bar{X})$, but also $(S^2, \bar{X})$.

3.2 Ancilarity and Completeness

Definition 3.3 (Ancillary statistic) *A statistic T is said to be ancillary if its sampling distribution does not depend on the unknown parameter θ.* ■

Definition 3.4 (Location-scale family of distributions) *The location-scale family of distributions is composed by all the probability distributions such that the associated cumulative distribution function is of the form*

$$F(x \mid \delta, \lambda) = G\left(\frac{x-\lambda}{\delta}\right)$$

where G is a function that does not involve unknown parameters, $\lambda \in \mathbb{R}$ is the location parameter, and $\delta > 0$ is the scale parameter.

This family includes the location family (δ is known) and the scale family (λ is known.)

Example 3.7 *Suppose that* $X \sim N(\mu, \sigma^2)$. *Since*

$$F(x \mid \mu, \sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

it follows that the normal family of distributions is a member of the location-scale family with location parameter μ *and scale parameter* σ.

Example 3.8 *Let* $X \sim U(0, \theta)$, $\theta > 0$. *Then,*

$$F(x \mid \theta) = \begin{cases} 0 & x \leq 0 \\ x/\theta & 0 < x \leq \theta \\ 1 & x > \theta \end{cases}$$

hence, $F(x \mid \theta) = G(x/\theta)$ *with* G *known. As a consequence, the* $U(0, \theta)$, $\theta > 0$ *family of distributions is a member of the scale family, with scale parameter* θ

Remarks

- The distribution of X is part of the location-scale family with location parameter λ and scale parameter δ if and only if the distribution of $(X - \lambda)/\delta$ does not depend on unknown parameters — HW problem.
- If the distribution of X is a member of the location-scale family with location parameter λ and scale parameter δ then any statistic which is a function of $X_1, \ldots, X_n$ only through the vector

$$\left(\frac{X_i - \lambda}{\delta}, \ i = 1, \ldots, n \right)$$

is ancillary. ■

Example 3.9 *Let* $X \mid \theta \sim U(\theta - 1/2, \theta + 1/2)$, $\theta \in \mathbb{R}$. *It is easy to check (do it!) that* $X - \theta \mid \theta \sim U(-1/2, 1/2)$. *Hence, this distribution is part of the location family with location parameter* θ. *The statistic* $R = X_{(n)} - X_{(1)} = (X_{(n)} - \theta) - (X_{(1)} - \theta)$ *depends on the random sample only through* $(X_i - \theta,\ i = 1, \ldots, n)$. *Hence,* R *is ancillary, that is, its probability distribution does not depend on* θ. ■

It would be natural to expect a minimal sufficient statistic and an ancillary statistic to be independent. However, that's not the case in general:

Example 3.10 *Let* $X \mid \theta \sim U(\theta - 1/2, \theta + 1/2)$. *It is easy to verify that* $(X_{(1)}, X_{(n)})$ *is a minimal sufficient statistic for* θ. *Hence, so is* $(X_{(n)} - X_{(1)}, X_{(1)})$. *However, we have seen that* $R = X_{(n)} - X_{(1)}$ *is ancillary!*

Definition 3.5 (Complete statistic) *A statistic T is said to be complete if and only if*

$$E[h(T) \mid \theta] = 0 \ \forall \theta \in \Theta \Rightarrow h(T) \equiv 0 \ .$$

Notice: When T is complete and $h_1(T)$ and $h_2(T)$ are two functions of T that have the same expected value, then it must be the case that they are the same, $h_1(T) = h_2(T)$.

Theorem 3.3 *Any statistic that is sufficient and complete is minimal sufficient.*

Proof: Let T be sufficient and complete and suppose that there exists T_1 sufficient such that $T_1 = g(T)$; that is, T_1 is sufficient and induces a partition that is coarser than that of T.

Consider $h(T) = T - E[T \mid T_1]$ and notice that

- h is a function of T only, since $E[T \mid T_1]$ is a function of T_1 but T_1 is a function of T
- $h(T)$ is a statistic, because since T_1 is sufficient, $E[T \mid T_1]$ does not depend on θ
- $E_\theta[h(T)] = 0$ for all θ
- As a consequence, since T is complete, $T = E[T \mid T_1]$ and T is a function of T_1. Hence, T and T_1 are equivalent, which finishes the proof.

Example 3.11 *(The converse is not valid.) Let* $X \mid \theta \sim U(\theta - 1/2, \theta + 1/2)$*. Since* $R = X_{(n)} - X_{(1)}$ *is ancillary, we have that* $E[R] = c$*, that is, it does not depend on* θ*. The statistic* $T = (R, X_{(1)})$ *is minimal sufficient but* $h(T) = R - c$ *has zero expected value and it is not zero. Hence,* T *is not complete.* ■

It turns out that sufficient and complete statistics operate a data reduction that is more effective than that operated by minimal sufficient statistics that are not complete. That is the subject of the Theorem of Basu:

Theorem 3.4 (Basu's Theorem) *Let T be a sufficient and complete statistic. Then T is independent of any ancillary statistic.*

Proof: Let U be an ancillary statistic and A an arbitrary event. Consider $h_A(T) = P(U \in A \mid T)$. Notice that

- since T is sufficient, $h_A(T)$ is a statistic
- $h_A(T) = E[I_A(U) \mid T]$
- $E[h_A(T)] = P_\theta(U \in A)$ which does not depend on θ since U is ancillary
- $h_A(T) - P(U \in A)$ is hence a statistic with zero expectation for all θ, which because of the completeness of T implies that

$$P(U \in A \mid T = t) = P(U \in A)$$

and this, since A is arbitrary, means that U and T are independent.

Remarks:

- A complete and sufficient statistic does not contain any ancillary information, that is, the data reduction operated by this type of statistics is more effective than that operated by minimal sufficient statistics which are not complete;
- Either all the minimal sufficient statistics are complete, or there are no sufficient and complete statistics (homework problem);
- Basu's theorem is a very useful tool to prove the independence of two statistics in a particularly elegant way.

3.3 Exponential family

Definition 3.6 (Exponential family of distributions) *We say that a random vector X is distributed according to a member of the k-parametric exponential family if its pdf or pmf can be expressed in the form*

$$f(x \mid \theta) = c(\theta)\ h(x)\ \exp\left[\sum_{j=1}^{k} \omega_j(\theta)\ R_j(x)\right]$$

with support $\{x :\ f(x \mid \theta) > 0\}$ independent of $\theta = (\theta_1, \ldots, \theta_k)$. Also, $c(\theta) \geq 0$, $h(x) \geq 0$, and $R_j(x)$ are scalar functions of x.

The canonical form *of a distribution which belongs to the k-parameter exponential family is obtained through the so-called natural parametrization, $\alpha_j = \omega_j(\theta)$, $j = 1, \ldots, k$:*

$$f(x \mid \alpha) = d(\alpha)\ h(x)\ \ \exp\left[\sum_{j=1}^{k} \alpha_j\ R_j(x)\right]$$

where $\alpha = (\alpha_1, \ldots, \alpha_k) \in A$ is called the natural parameter and A being called the natural parametric space. ■

Example 3.12 *The binomial model $B(n;\theta)$ belongs to the one-parameter exponential family with natural parameter $\alpha = \ln[\theta/(1-\theta)]$; the natural parametric space is $A = \mathbb{R}$.*

Theorem 3.5 *The k-parameter exponential family structure is preserved under iid random sampling and there is a k-dimensional sufficient statistic regardless of the sample size.*

Let $X_1, \ldots, X_n$ a random sample of size n from a population which belongs to the k-parameter exponential family. Then,

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} c(\theta)\ h(x_i)\ \exp\left[\sum_{j=1}^{k} \omega_j(\theta)\ R_j(x_i)\right]$$

$$= [c(\theta)]^n\ \prod_{i=1}^{n} h(x_i)\ \exp\left[\sum_{j=1}^{k} \omega_j(\theta)\ T_j(x)\right]$$

where $T_j(x) = \sum_{i=1}^{n} R_j(x_i)$. So the joint distribution of the random sample still belongs to the k-parameter exponential family, and the factorization criterion guarantees that $T = (T_1(X), \ldots, T_k(X))$ is sufficient for θ.

Theorem 3.6 *The sufficient statistic $T = (T_1(X), \ldots, T_k(X))$ just described is complete if the natural parametric space A contains an open subset of $\mathbb{R}^k$.* ■

Example 3.13 *Consider the model $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. In this case,*

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-\mu^2/(2\sigma^2)}\ \exp\left[-\frac{1}{2\sigma^2}\ x^2 + \frac{\mu}{\sigma^2}\ x\right]$$

hence, the normal family of distributions belongs to the 2-parameter exponential family with natural parameter $(\alpha_1, \alpha_2) = (-1/(2\sigma^2), \mu/(2\sigma^2))$ and natural parametric space $A = \mathbb{R}_- \times \mathbb{R}$. The sufficient statistic is $(\sum X_i^2, \sum X_i)$, that is equivalent to $T = (S^2, \bar{X})$. Since the natural parameter space contains an open subset of $\mathbb{R}^2$, it follows that T is also complete.

Example 3.14 (In a normal population, $\bar{X}$ and S^2 are independent.) *We start by assuming that σ^2 is known: $\sigma^2 = \sigma_0^2$. Then, it is easy to see this model belongs to the one-parameter exponential family, and that $\bar{X}$ is sufficient and complete for μ. This normal model (with σ^2 known) is part of the location family with location parameter μ. Hence, since*

$$S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n = \sum_{i=1}^{n}[(X_i - \mu) - \sum_{i=1}^{n}(X_i - \mu)/n]^2/n$$

it is clear that S^2 is ancillary. Basu's theorem guarantees that S^2 and $\bar{X}$ are independent, that is,

$$P(\bar{X} \leq t, S^2 \leq s \mid \mu, \sigma_0^2) = P(\bar{X} \leq t \mid \mu, \sigma_0^2) \times P(S^2 \leq s \mid \mu, \sigma_0^2)$$

for all t, s, μ and σ_0^2. But σ_0^2 is arbitrary, and this shows the independence in the biparametric model.

Example 3.15 *(When we restrict a model, sufficiency is maintained; completeness not always.)*

Suppose that $X \sim N(\mu, \sigma^2)$ *but consider the sub-family* $\mu = \sigma^2$*. Clearly,* $T = (S^2, \bar{X})$ *is still sufficient for this sub-family. However, the natural parametric space no longer contains an open subset of* $\mathbb{R}^2$ *so that the previous theorem does not guarantee the completeness of* T*. Nevertheless, it is easy to see that* T *is in this setting not complete:* $\bar{X} - n\ S^2/(n-1)$ *is a function of* T *which has zero expected value but is obviously not null.*

3.4 Fisher information

Definition 3.7 Likelihood function: *When we observe the sample* $(x_1, \ldots, x_n)$*, the observed value of the random sample* $X_1, \ldots, X_n$ *with joint pdf or pmf function* $f(x_1, \ldots, x_n)$*, the corresponding likelihood function is a function of* θ *given by*

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) \ .$$

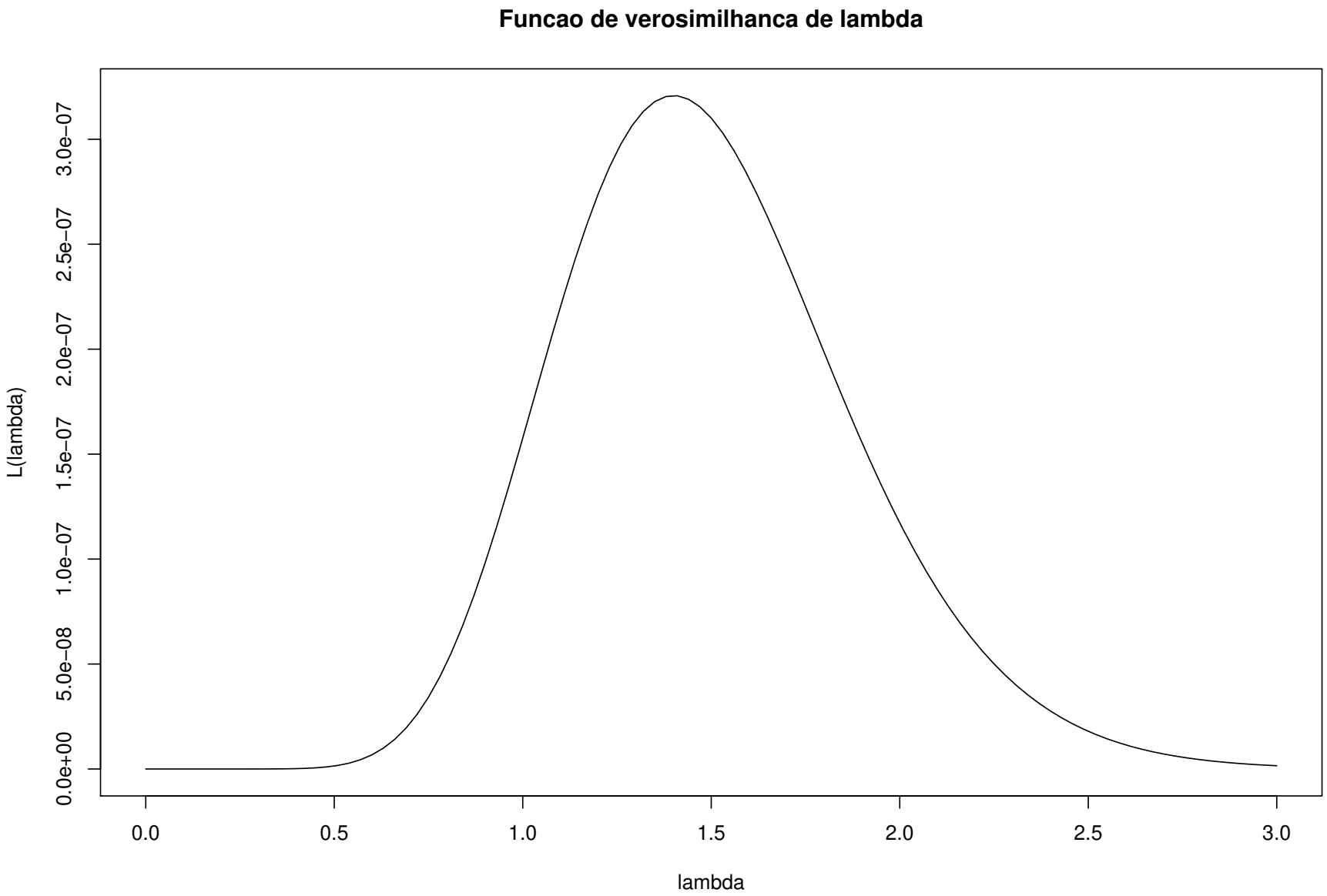
■

Remarks

- Since $x_1, \ldots, x_n$ is fixed, it's not unusual to write $L(\theta)$ instead of $L(\theta \mid x_1, \ldots, x_n)$. Note that L is a function of θ and not of $x_1, \ldots, x_n$.
- If $X_1, \ldots, X_n$ is an iid random sample, we have

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Example 3.16 *Suppose that* $X \mid \lambda \sim Po(\lambda)$. *We observe a random sample of size* $n = 10$ *and we conclude that* $\sum_{i=1}^{n} x_i = 14$ *and* $\prod x_i! = 288$. *We then have* $L(\lambda) = \exp(-10\lambda)\lambda^{14}/288$.


Funcao de verosimilhanca de lambda
L(lambda)
0.0e+00
5.0e−08
1.0e−07
1.5e−07
2.0e−07
2.5e−07
3.0e−07
0.0
0.5
1.0
1.5
2.0
2.5
3.0
lambda

Important observation: The likelihood function (as a function of θ) is not a probability (mass or density) function. Hence, it has no "natural" scale associated. In fact, and it's more appropriately defined as

$$L(\theta \mid x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n \mid \theta)$$

that is, up to a multiplying constant. This multiplying constant can depend on $x_1, \ldots, x_n$ but certainly not on θ. What bears meaning are ratios like

$$\frac{L(\theta^{'} \mid x_1, \ldots, x_n)}{L(\theta^{\star} \mid x_1, \ldots, x_n)}$$

which measure how likely is $\theta^{'}$ compared to $\theta^{\star}$ having observed the data $x_1, \ldots, x_n$.

With this definition, the likelihood function depends on the data only through the observed value of a sufficient statistic.

Let us consider the case $\Theta \subset \mathbb{R}$.

Regularity conditions

C1—Θ is an open interval of $\mathbb{R}$.

C2—The set $\{x : f(x \mid \theta) > 0\}$, i.e., the support of $f(\cdot \mid \theta)$, does not depend on θ.

C3—The function $f(x \mid \theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$, is differentiable in θ for all x

C4—We have $0 < E_\theta \left[\partial \ln f(X \mid \theta)/\partial\theta\right]^2 < +\infty$ for all θ

C5—It is OK to permute the symbols $\frac{\partial}{\partial\theta}$ and $\int dx$.

Remarks:

- C2 excludes models like $\mathrm{U}(0, \theta)$: $\theta > 0$
- C4 guarantees that the r.v. $S = \partial \ln f(X \mid \theta)/\partial\theta$ has finite second moment.

Definition 3.8 Score function: *Having observed the sample $x_1 \ldots, x_n$, the score function measures the relative variation of the likelihood function (as a function of θ)*

$$S(\theta \mid x_1, \ldots, x_n) = \frac{L'(\theta \mid x_1, \ldots, x_n)}{L(\theta \mid x_1, \ldots, x_n)} = \frac{\partial \ln L(\theta \mid x_1, \ldots, x_n)}{\partial \theta}.$$

■

Note: For an iid random sample, we have $S(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} S(\theta \mid x_i)$ where $S(\theta \mid x_i)$ is the score function associated with the ith observation.

Theorem 3.7 *Under the regularity conditions, we have that for all $\theta \in \Theta$*

$$E_\theta[S(\theta \mid X_1, \ldots, X_n)] = 0 \ .$$

■

Definition 3.9 Fisher information: *The Fisher information about θ contained in $X_1, \ldots, X_n$ is defined by*

$$I_{(X_1,\ldots,X_n)}(\theta) = E_\theta \left\{ [S(\theta \mid X_1, \ldots, X_n)]^2 \right\}$$

■

Properties under the regularity conditions

- $\mathrm{I}_{(X_1,\ldots,X_n)}(\theta) = \mathrm{Var}[S(\theta \mid X_1, \ldots, X_n)]$
- If $\boldsymbol{X} = (X_1, X_2)$ with X_1 and X_2 independent $\forall\theta$, then

$$\mathrm{I}_{\boldsymbol{X}}(\theta) = \sum_{i=1}^{2} \mathrm{I}_{X_i}(\theta) \ .$$

- Consequently, in the case of an iid random sample, $\mathrm{I}_{(X_1,\ldots,X_n)}(\theta) = n\mathrm{I}_{X_i}(\theta)$
- Useful formula to compute I:

$$\mathrm{I}_{(X_1,\ldots,X_n)}(\theta) = -E_\theta \left[\frac{\partial^2 \ln f(X_1, \ldots, X_n \mid \theta)}{\partial \theta^2} \right]$$

Theorem 3.8 *If $T(X_1, \ldots, X_n)$ is a statistic, and under regularity conditions,*

$$I_{(X_1,\ldots,X_n)}(\theta) \geq I_T(\theta)$$

and the equality holds if and only if T is sufficient for θ. ■

Theorem 3.9 *Consider the alternative parametrization $\theta = g(\phi)$ with g differentiable. Then, the Fisher information about ϕ in X, $I_X^\star(\phi)$, satisfies*

$$I_X^\star(\phi) = I_X(g(\phi)) \, [g'(\phi)]^2$$

with $I_X(\theta)$ representing the Fisher about θ in X. ■

Example 3.17 *Let $X_1, \ldots, X_n$ be an iid random sample from a $Po(\lambda)$, $\lambda > 0$, population. Then,*

$$S(\lambda \mid x_1, \ldots, x_n) = \frac{n}{\lambda}(\bar{x} - \lambda)$$

and hence

$$I_{(X_1,\ldots,X_n)}(\lambda) = \left(\frac{n}{\lambda}\right)^2 Var(\bar{X}) = \frac{n}{\lambda}.$$

It is easy to see that $I_{(X_1,\ldots,X_n)}(\lambda) = nI_{X_i}(\lambda)$.

Additionally,

$$\frac{\partial}{\partial\lambda} S(\lambda \mid x_1, \ldots, x_n) = -\frac{\sum x_i}{\lambda^2}$$

and hence

$$E_\theta\left[-\frac{\partial^2 \ln f(X_1, \ldots, X_n \mid \lambda)}{\partial\lambda^2}\right] = \frac{n}{\lambda} = I_{(X_1,\ldots,X_n)}(\lambda)$$

If we consider $\phi = 1/\lambda$, it is easy to see that $I^*_{(X_1,\ldots,X_n)}(\phi) = n/\phi^3$ both using the definition and using the theorem above. ■

The ideas and properties that we have described for the case $\Theta \subset \mathbb{R}$ can be naturally extended to the multiparametric case $\Theta \subset \mathbb{R}^k$:

- If the parameter is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, the score function is the $k \times 1$ vector

$$\begin{aligned} \boldsymbol{S}(x \mid \boldsymbol{\theta}) &= \frac{\partial \ln L(\boldsymbol{\theta} \mid x)}{\partial \boldsymbol{\theta}} \\ &= \left(\frac{\partial \ln L(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \theta_i}, i = 1, \ldots, k \right)' \end{aligned}$$

- The Fisher information is then the $k \times k$ matrix

$$\begin{aligned} \boldsymbol{I}_X(\boldsymbol{\theta}) =& E_{\boldsymbol{\theta}} \left[\boldsymbol{S}(X \mid \boldsymbol{\theta}) \boldsymbol{S}(X \mid \boldsymbol{\theta})' \right] \\ &= \left[E_{\boldsymbol{\theta}} \left[\frac{\partial \ln L(\boldsymbol{\theta} \mid X)}{\partial \theta_i} \frac{\partial \ln L(\boldsymbol{\theta} \mid X)}{\partial \theta_j} \right] \right]_{i,j=1,\ldots,k} \end{aligned}$$

Under multiparametric regularity conditions similar to C1–C4,

- $E_{\boldsymbol{\theta}}\left[\boldsymbol{S}(X \mid \boldsymbol{\theta})\right] = \mathbf{0}$.
- An alternative formula for the information matrix:

$$\begin{aligned}\boldsymbol{I}_X(\boldsymbol{\theta}) =& E_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ln L(\boldsymbol{\theta} \mid X)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \\ =& E_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ln L(\boldsymbol{\theta} \mid X)}{\partial \theta_i \partial \theta_j}\right]_{i,j}\end{aligned}$$

- if X is a random sample of size n,

$$\boldsymbol{I}_X(\boldsymbol{\theta}) = n\boldsymbol{I}_{X_1}(\boldsymbol{\theta}).$$

- In terms of the alternative parameterization $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$, with $\boldsymbol{g}(\boldsymbol{\phi}) = (g_1(\boldsymbol{\phi}), \ldots, g_k(\boldsymbol{\phi}))' = \boldsymbol{\theta}$, and g_i differentiable $\forall i$, we have that

$$\boldsymbol{I}^{\star}_X(\boldsymbol{\phi}) = \boldsymbol{J}'\boldsymbol{I}_X(\boldsymbol{g}(\boldsymbol{\phi}))\boldsymbol{J}$$

  where $\boldsymbol{J}$ is a matrix of order k whose ith line is

$$\left(\frac{\partial g_i(\boldsymbol{\phi})}{\partial \phi_1}, \ldots, \frac{\partial g_i(\boldsymbol{\phi})}{\partial \phi_k}\right)$$

Example 3.18 *If $X = (X_1, \ldots, X_n)'$ is a random sample of size n from the model* $\{\mathsf{Ga}(\alpha, \delta) : \alpha, \delta > 0\}$,

$$f(x \mid \alpha, \delta) = \frac{\delta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\delta x) I_{(0,\infty)(x)}$$

compute $I_X(\alpha, \delta)$.

First, $I_X(\alpha, \delta) = n\ I_{X_1}(\alpha, \delta)$. Then, it is easy to see that

$$\ln L(\alpha, \delta \mid x_1) = c + \alpha \ln \delta - \ln \Gamma(\alpha) + (\alpha - 1) \ln x_1 - \delta x_1$$

where c is a constant.

Hence,

$$\frac{\partial^2 \ln L}{\partial \alpha^2} = -\frac{d^2 \ln \Gamma(\alpha)}{d\alpha^2} = -\psi'(\alpha)$$
$$\frac{\partial^2 \ln L}{\partial \delta \partial \alpha} = \frac{1}{\delta}$$
$$\frac{\partial^2 \ln L}{\partial \delta^2} = -\frac{\alpha}{\delta^2},$$

where $\psi'(\alpha)$ is the so-called trigamma function, the second derivative of $\ln \Gamma(\alpha)$.

So,

$$I_{\boldsymbol{X}}(\alpha, \delta) = n \begin{pmatrix} \psi'(\alpha) & -\frac{1}{\delta} \\ -\frac{1}{\delta} & \frac{\alpha}{\delta^2} \end{pmatrix}.$$

What if we are interested in the alternative parametrization $(\alpha, 1/\delta)$?

Let $(\alpha, 1/\delta) = (\phi_1, \phi_2) = \boldsymbol{\phi}'$; then,

$$\begin{cases} \alpha = \phi_1 = g_1(\boldsymbol{\phi}) \\ \delta = 1/\phi_2 = g_2(\boldsymbol{\phi}) \end{cases}$$

and as such

$$\boldsymbol{J} = \begin{pmatrix} \frac{\partial g_1}{\partial \phi_1} & \frac{\partial g_1}{\partial \phi_2} \\ \frac{\partial g_2}{\partial \phi_1} & \frac{\partial g_2}{\partial \phi_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1/\phi_2^2 \end{pmatrix}$$

and after some algebra, we get that

$$I^{\star}_{\boldsymbol{X}}(\boldsymbol{\phi}) = n \begin{pmatrix} \psi'(\phi_1) & \frac{1}{\phi_2} \\ \frac{1}{\phi_2} & \frac{\phi_1}{\phi_2^2} \end{pmatrix}.$$