M1 Data Science, University of Lille
Statistics 2 - Lecture notes

Emilie Kaufmann (CNRS, Univ. Lille)
emilie.kaufmann@univ-lille.fr

March 10, 2024

Pre-requisite In statistics 1, you have seen:

- classical distributions
- examples of estimators
- confidence intervals
- the statistical testing protocol (type I and type II error)
- example of classical tests

In statistics 2, we will revisit statistical estimation and testing with a focus on *optimality*.
We will notably discuss:

- different performance measure for estimators
- generic estimation strategies, notably the maximum likelihood principle
- asymptotic properties of estimators
- likelihood-ratio based testing procedures

Several examples will come from an important family of distributions called exponential families. Finally, if we have time we will also talk a bit about Bayesian statistics.

Chapter 1

Estimation

1.1 Statistical inference

In statistical inference, we observe a realization of some random variable (or random vector) X, called the observation, whose distribution over some space $\mathcal{X}$ is P_X. The goal is to discover ("infer") some properties of this underlying distribution, assuming that P_X belongs to some set of possible distributions, called the *statistical model*. Depending on the situation, we may make assumptions on the cumulative distribution function (cdf) of X, F_X or on its density f_X with respect to some reference measure and the statistical model may be a set of distribution, a set of cdfs or a set of pdfs parameterized by some parameter θ:

$$\mathcal{M} = \{P_\theta, \theta \in \Theta\}, \quad \mathcal{M} = \{F_\theta, \theta \in \Theta\} \quad \text{or} \quad \mathcal{M} = \{f_\theta, \theta \in \Theta\}.$$

When the parameter space $\Theta \subseteq \mathbb{R}^d$, the model is called parametric, otherwise it is non-parametric. Given the "true" parameter θ (i.e. θ such that $P_X = P_\theta$), the probability space on which X is defined is denoted by $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, and the corresponding expectation is denoted by $\mathbb{E}_\theta$.

The n-sample example Often the random variable X is of the form $X = (X_1, \dots, X_n)$ where the X_i are assumed to be iid realizations of the same distribution. These iid copies represent the repetition of some random experiment (for example the vote expressed by one individual in a population, or the effect of a treatment on one patient). These random variables X_i are defined on some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and will most of the time take values in $\mathbb{R}$ (we could consider some multi-dimensional outcomes in, e.g. two-sample testing problems).

In the n-sample setting, we denote by P the distribution of X_1 (which is the common distribution of all X_i's), by F the cdf of this distribution and by f its density (with respect to some reference measure ν), if it admits one. We will write indifferently

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P\,, \quad X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F \quad \text{or} \quad X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f\,.$$

In that case, the statistical model is typically expressed as possible candidates for P, F or f. Those also denoted by P_θ, F_θ and f_θ, respectively (by a slight abuse of notation), for some parameter θ belonging to the parameter space Θ.

Example 1. *Take a Gaussian n-sample with known variance 1 and unknown mean $\theta \in \mathbb{R}$: $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. Let f_θ be the density of a $\mathcal{N}(\theta, 1)$ variable with respect to the Lebesgue measure (in $\mathbb{R}$):*

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right).$$

If we look at the observation $X = (X_1, \dots, X_n)$, the statistical model $\mathcal{M}$ is a set of multivariate Gaussian distributions whose densities with respect to the Lebesgue measure in $\mathbb{R}^n$ is

$$f_\theta(x_1, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

for some parameter $\theta \in \mathbb{R}$.

In statistical inference, we are interested in making statements about the "true" parameter θ generating the data or about some *parameter of interest* which can be some function of θ, denoted by $g(\theta)$. This statement can be a guess for its value (estimation), an interval to which it belongs (confidence interval) or the answer to some question about this parameter (statistical test).

Example 2. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *The parameter of the model is $\theta = (\mu, \sigma)$ and the parameter space is $\Theta = \{(\mu, \sigma), \mu \in \mathbb{R}, \sigma > 0\}$. If we are solely interested in estimating the mean, the parameter of interest is μ and σ may be called a nuisance parameter.*

In some situations, we may be interested in estimating more complex functions of θ. For example, assume that X_i models the amount of antibodies produced 15 days after receiving a vaccine. For a given disease, the vaccine is considered efficient if this amount exceeds some threshold v. A possible parameter of interest is the probability of efficacy of the vaccine, p, which can be expressed as

$$p = \mathbb{P}(X_1 \geq v) = 1 - \mathbb{P}(X_1 < v) = 1 - \mathbb{P}\left(\frac{X_1 - \mu}{\sigma} < \frac{v - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{v - \mu}{\sigma}\right)$$

where Φ is the cdf of a $\mathcal{N}(0, 1)$ random variable.

Example 3 (regression model)**.** $Z_1, \dots, Z_n \overset{iid}{\sim} P$. $X_i = (Z_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ *such that*

$$Y_i = h(Z_i) + \varepsilon_i$$

where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $h : \mathcal{X} \to \mathcal{Y}$ is the regression function. The observation is $X = (X_1, \dots, X_n)$ and the parameters of the model are P (that could belong to some parametric class of probability distributions) and the regression function h (that could belong to a parametric families of functions, e.g. linear functions). In that case the "parameter" of interest is usually the regression function.

1.2 Performance of an estimator

An estimator of $g(\theta)$ is any function of the observation $\widehat{g} = h(X)$ that is supposed to be "close" to the parameter of interest $g(\theta)$. When $X = (X_1, \dots, X_n)$ has the n-sample structure, we will materialize the dependency in n of the estimator by writing $\widehat{g}_n = h(X_1, \dots, X_n)$.

From its definition, $\widehat{g}$ is a random variable (or a random vector, when we estimate a multi-dimensional parameter), hence its quality will be expressed in terms of some properties of its distribution, which should ideally be concentrated around $g(\theta)$. Two important characteristics of this distributions are its mean and its variance.

Definition 4. *The* bias *of estimator $\widehat{g}$ of $g(\theta)$ is defined as* $\mathrm{b}_\theta(\widehat{g}) = \mathbb{E}_\theta[\widehat{g}] - g(\theta)$.
When $\mathrm{b}_\theta(\widehat{g}) = 0$*, the estimator is called* unbiased.

Definition 5. *The* variance *of a real-valued estimator $\widehat{g}$ is* $\mathrm{Var}_\theta[\widehat{g}] := \mathbb{E}_\theta[(\widehat{g} - \mathbb{E}_\theta[\widehat{g}])^2]$.

A good (real-valued) estimator has ideally a small bias and a small variance, which indicates that on average, its value is close to $g(\theta)$ and that under different realizations of the experiments, its value would not change too much. The closeness from $\widehat{g}$ to $g(\theta)$ can also directly be measured using their average distance, a notion that can also be meaningful in the multi-dimensional setting.

Definition 6. *The* quadratic risk *of an estimator $\widehat{g}$ of $g(\theta) \in \mathbb{R}^p$ is*

$$\mathrm{R}_\theta(\widehat{g}) = \mathbb{E}_\theta\left[\|\widehat{g} - g(\theta)\|^2\right],$$

where $\|u\|$ is the Euclidian norm in $\mathbb{R}^p$, such that $\|u\|^2 = u^\top u$. In the one-dimensional case ($p = 1$), this quantity is sometimes called the mean-squared error*, and denoted by* $\mathrm{MSE}_\theta(\widehat{g})$.

Theorem 7 (bias-variance decomposition)**.** *Assume $g(\theta) \in \mathbb{R}$. We have*

$$\mathrm{R}_\theta(\widehat{g}) = (\mathrm{b}(\widehat{g}))^2 + \mathrm{Var}_\theta\left[\widehat{g}\right] .$$

Exercise 8. *Prove it.*

Comparing estimators with the quadratic risk The quadratic risk can be used to compare estimators, and we say that an estimator $\widehat{g}$ is better than an estimator $\widetilde{g}$ if for all $\theta \in \Theta$, $\mathrm{R}_\theta(\widehat{g}) \leq \mathrm{R}_\theta(\widetilde{g})$. However, this relationship is not a total order, as there may exists estimators for which $\mathrm{R}_{\theta_1}(\widehat{g}) \leq \mathrm{R}_{\theta_1}(\widetilde{g})$ for some parameter θ_1 but $\mathrm{R}_{\theta_2}(\widehat{g}) > \mathrm{R}_{\theta_2}(\widetilde{g})$ for a different parameter θ_2.

Definition 9. *An estimator $\widehat{g}$ of $g(\theta)$ is called* admissible *is there exists no estimator $\widetilde{g}$ which is strictly better than $\widehat{g}_n$ i.e. for which*

$$\forall \theta \in \Theta, \;\; \mathrm{R}_\theta(\widetilde{g}) \leq \mathrm{R}_\theta(\widehat{g})$$

and the inequality is strict for at least one value θ_0.

Influence of the sample size When X is a n-sample, the above properties for an estimator $\widehat{g}_n$ are all considering a fixed sample size n, and are not capturing another desirable property of an estimator: $\widehat{g}_n$ should get closer to $g(\theta)$ when the sample size n goes larger. We expect to $\widehat{g}_n$ to get closer to $g(\theta)$, meaning that its distribution concentrates for and more around $g(\theta)$. We will discuss these asymptotic properties in the next chapter.

Recap: Densities and Expectations All the criteria for evaluating estimators in this section are expressed with expectations. In general, if Z is a random variable taking values in $\mathcal{Z}$ whose distribution P has a density f with respect to some reference measure ν, we have, for all function ϕ,

$$\mathbb{E}[\phi(Z)] = \int_{\mathcal{Z}} \phi(x) f(x) d\nu(x).$$

We will mostly see examples of random variables defined on $\mathcal{Z} = \mathbb{R}^d$ whose distributions have a density with respect to the Lebesgue measure in $\mathbb{R}^d$, or of discrete random random variables (i.e. for which $\mathcal{Z}$

is discrete) that have a density with respect to the counting measure. In the discrete case, the density is simply defined, for all $z \in \mathcal{Z}$, by

$$f(z) = P(\{z\}) = \mathbb{P}_{Z\sim P}(Z = z) \ .$$

Back to our statistical model, in the most common n-sample case in which $X = (X_1, \dots, X_n) \overset{iid}{\sim} P_\theta$, we will often uncounter two cases.

Either $X_i \in \mathbb{R}$ and P_θ has a density with respect to the Lebesgue measure. Then

- for any $\phi : \mathbb{R}^n \to \mathbb{R}$, $\mathbb{E}_\theta[\phi(X)] = \int_{\mathbb{R}} \phi(x_1, \dots, x_n) f_\theta(x_1, \dots, x_n) dx_1 \dots dx_n$
- for any $\phi : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}_\theta[\phi(X_1)] = \int_{\mathbb{R}} \phi(u) f_\theta(u) du$

Or $X_i \in \mathcal{S}$ for some discrete set $\mathcal{S}$ (typically a subset of $\mathbb{N}$) and we have

- for any $\phi : \mathcal{S}^n \to \mathbb{R}$, $\mathbb{E}_\theta[\phi(X)] = \sum_{x\in\mathcal{S}^n} \phi(x_1, \dots, x_n) f_\theta(x_1, \dots, x_n)$
- for any $\phi : \mathcal{S} \to \mathbb{R}$, $\mathbb{E}_\theta[\phi(X_1)] = \sum_{u\in\mathcal{S}} \phi(u) f_\theta(u)$

1.3 Estimation procedures

1.3.1 The moment method

When $X = (X_1, \dots, X_n) \overset{iid}{\sim} P_\theta$, the moment method can be used when the parameter of interest $g(\theta)$ can be expressed as a function of the moments of X_1.

In the simple case, we have

$$g(\theta) = \mathbb{E}_\theta[\phi(X_1)]$$

for some function ϕ such that $\mathbb{E}[|\phi(X_1)|] < \infty$. Motivated by the law of large numbers, we define the moment estimator

$$\widehat{g}_n := \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$$

which satisfies $\widehat{g}_n \to g(\theta)$, $\mathbb{P}_\theta$ – a.s.. Hence, this estimator is naturally going to be close to $g(\theta)$ at least for a large sample size n.

More generally, suppose that we seek to estimate a multi-dimensional parameter $\theta = (\theta_1, \dots, \theta_k)$ and that for $1 \le j \le k$ the j-th moment can be expressed as some function of the parameter θ:

$$\mathbb{E}_\theta[X^j] = \alpha_j(\theta).$$

Letting $\widehat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$ the j-th sample moment, the moment estimator is defined as the solution $\widehat{\theta}_n$ of the system of equations

$$\alpha_1(\theta) = \widehat{\alpha}_1, \ \dots \ , \alpha_k(\theta) = \widehat{\alpha}_k.$$

Example 10. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *We can find the moment estimator for the parameter $\theta = (\mu, \sigma^2)$. There are two parameters so we can look at the first two moments.*

$$\begin{aligned} \mathbb{E}_\theta[X_1] &= \mu \\ \mathbb{E}_\theta[X_1^2] &= \mathrm{Var}_\theta[X_1] + (\mathbb{E}_\theta[X_1])^2 = \sigma^2 + \mu^2 \end{aligned}$$

The empirical first and second moments are $\frac{1}{n}\sum_{i=1}^n X_i$ *and* $\frac{1}{n}\sum_{i=1}^n X_i^2$ *so we get the system*

$$\begin{cases} \mu &= \frac{1}{n}\sum_{i=1}^n X_i \\ \mu^2+\sigma^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 \end{cases}$$

from which we get $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)$ *where*

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \textit{and} \quad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$$

We recognize the well-known empirical mean and empirical variance, which can also be rewritten

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{k=1}^n X_k\right)^2.$$

1.3.2 The plug-in method

The plug-in method is also suited for the n-sample setting, when the parameter of interest can be expressed as some functional of P, the distribution of X_1 (for example some moment of this distribution, or some quantile), we write

$$g(\theta) = H(P).$$

The idea is construct some empirical version of this distribution, denoted by $\widehat{P}_n$ and to "plug-in" this empirical distribution, that is to define

$$\widehat{g}_n = H(\widehat{P}_n).$$

We now describe this empirical distribution.

Definition 11. *Given a* n*-sample* $X = (X_1, \dots, X_n) \in \mathbb{R}^n$*, the empirical distribution* $\widehat{P}_n$ *is a probability measure on* $\mathbb{R}$ *defined as*

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

where δ_x *is the Dirac measure in* x*. The Dirac measure is defined, for all event* A *as* $\delta_x(A) = 1$ *if* $x \in A$*,* $\delta_x(A) = 0$ *otherwise. For any* $x \in \mathbb{R}$*, we have*

$$\widehat{P}_n(\{x\}) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(\{x\}) = \frac{\#\{i : X_i = x\}}{n}.$$

$\widehat{P}_n$ *is a discrete distribution whose (finite) support (= set of values that have non-zero probability in the discrete case) is the values in* $\{X_1, \dots, X_n\}$*.*

For any function ϕ, the expectation of $\phi(Z)$ when Z is distributed according to the empirical distribution $\widehat{P}_n$ is given by

$$\mathbb{E}_{Z\sim\widehat{P}_n}[\phi(Z)] = \sum_{x\in S}\phi(x)\widehat{P}_n(x) = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$$

where S is the support of $\widehat{P}_n$, i.e. the number of distinct values in the observation X. In particular, the cdf $\widehat{F}_n$ of the empirical distribution, which by definition is $\widehat{F}_n(x) = \mathbb{P}_{Z\sim\widehat{P}_n}(Z \le x) = \mathbb{E}_{Z\sim\widehat{P}_n}[\mathbb{1}(X_i \le x)]$ can be written

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \le x).$$

Remark 12. *When the functional $H(P)$ is defined as some expectation under P, the moment method and the plug-in method actually concide. Indeed, if*

$$g(\theta) = \mathbb{E}_{X\sim P}[\phi(X)]$$

the plug-in method yields

$$\widehat{g}_n = \mathbb{E}_{X\sim \widehat{P}_n}[\phi(X)] = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i).$$

We also call such estimators "empirical estimators".

But plug-in estimator can be more general when H is not defined as some expectation, for example we can define the empirical quantiles of a distribution to estimate its quantiles.

Exercise 13. *Using the plug-in approach, justify (again) the expression of the empirical mean and empirical variance of a distribution.*

1.3.3 Maximum Likelihood Estimation (MLE)

The maximum likelihood approach can be used to estimate $g(\theta) = \theta$ when the statistical model is of the form

$$\mathcal{M} = \{P_\theta : P_\theta \text{ has a density } f_\theta \text{ with respect to } \nu, \theta \in \Theta\}$$

where ν is a fixed reference measure (which is the same for all the distributions in the model). Such a model is called *dominated* (by the reference measure ν).

In most practical cases, this reference measure will be the Lebesgue measure in $\mathbb{R}^d$ (when the distributions are continuous) or the counting measure on discrete set (when the distributions are discrete). In that case, the density is given by $f_\theta(x) = \mathbb{P}_\theta(X = x)$.

Definition 14. *The likelihood of the observation X given a parameter θ is defined by*

$$L(X;\theta) = f_\theta(X).$$

In the n-sample case, due to independence, the log-likelihood can be written

$$L(X_1,\dots,X_n;\theta) = \prod_{i=1}^{n} f_\theta(X_i). \tag{1.1}$$

Example 15. *If $X_1,\dots,X_n \sim \mathcal{B}(\theta)$. The density of a Bernoulli distribution with parameter θ can be written*

$$f_\theta(x) = \theta\mathbb{1}(x=1) + (1-\theta)\mathbb{1}(x=0) = \theta^x(1-\theta)^{1-x}\mathbb{1}(x\in\{0,1\})$$

hence we have

$$L(X_1,\dots,X_n;\theta) = \prod_{i=1}^{n}\theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}$$

If $X_1,\dots,X_n \sim \mathcal{N}(\theta,\sigma^2)$, we get

$$L(X_1,\dots,X_n;\theta) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\theta)^2}{2\sigma^2}\right)$$

which can also be re-arranged.

The likelihood can be interpreted as the probability of making observation X if the underlying parameter is θ. Indeed, if $\mathcal{M}$ is a set of discrete distributions (i.e. when ν is the counting measure), we have $f_\theta(x_i) = \mathbb{P}_\theta(X_i = x_i)$. Due to independence, we have

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} f_\theta(x_i) = L(x;\theta) \quad \text{where } x = (x_1, \dots, x_n)$$

In the continuous case (i.e. when ν is the Lebesgue measure), the probability of a given $x = (x_1, \dots, x_n)$ is zero and we replace it by the value of the (joint) density in the point.

This observation motivates the maximum likelihood estimator as the estimator of $g(\theta) = \theta$ seeking the parameter θ for which the actual observation X is the most likely (i.e. which as it the largest "probability").

Definition 16. *A maximum likelihood estimator (MLE) of a parameter θ is an estimator satisfying*

$$\widehat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(X;\theta).$$

Remark 17. *As will be seen in some exercises, the maximum likelihood is not always unique.*

Computational considerations In simple case, the maximum likelihood can be computed explicitely, by finding the critical points (for which the derivative are zero) and proving that it is indeed a maximizer (e.g., by checking that the second derivative is negative in the critical point). In more complex cases, it can only be approximated using some optimization algorithm. In complex models (like the Gaussian mixture model), more fancy approximation schemes are needed, like the EM algorithm (Expectation Maximization) algorithm.

From a computational prespective (and due to the common product form of the likelihood, see (1.1)) it is often more convenient to maximize the logarithm of the likelihood (which then becomes a sum).

Definition 18. *The log-likelihood of the observation X given a parameter θ is denoted by*

$$\ell(X;\theta) = \log L(X;\theta).$$

Exercise 19. *Poisson distributions are often used to model count data (e.g. the number of monthly purchases of a customer on an e-commerce website may follow a Poisson distribution). A Poisson distribution with parameter $\lambda > 0$, denoted by $\mathcal{P}(\lambda)$, is a discrete distribution defined as*

$$\mathbb{P}(Z = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for all } k \in \mathbb{N}.$$

Compute the maximum likelihood estimator of λ given iid observations $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{P}(\lambda)$. What other method(s) could you use to obtain the same estimator?

Example 20. *In the logistic regression model, there are iid pairs of observations (X_i, Y_i) where X_i comes from some distribution on $\mathbb{R}^d$ that is assumed to have some density and $Y_i \in \{-1, 1\}$ is such that*

$$\mathbb{P}\left(Y_i = 1 | X_i = x\right) = \frac{1}{1 + e^{-x^\top \theta}}$$

where $\theta \in \mathbb{R}^d$ is a regression parameter.

To define the likelihood of the data, we admit that the density of $(X_1, Y_1) \in \mathbb{R}^d \times \{0,1\}$ *is*

$$f_\theta(x,y) = \mathbb{P}(Y_1 = y | X_1 = x) f(x).$$

You can verify that for all $x \in \mathbb{R}^d$ *and all* $y \in \{-1,1\}$, $\mathbb{P}(Y_1 = y|X_1 = x) = \frac{1}{1+e^{-yx^\top\theta}}$. *The likelihood can therefore be written*

$$L((X_1,Y_1),\dots,(X_n,Y_n)) = \prod_{i=1}^n f(X_i)\left(\frac{1}{1+e^{-Y_i(X_i^\top\theta)}}\right)$$

and a maximum likelihood estimator $\widehat{\theta}_n$ *satisfies*

$$\widehat{\theta}_n \in \underset{\theta\in\mathbb{R}^d}{\textit{argmin}} \sum_{i=1}^n \log\left(1+e^{-Y_i(X_i^\top\theta)}\right).$$

In this example, no closed-form expression exists for the MLE (unlike in a linear regression example), and we should resort to an optimization algorithm.

M-estimators The MLE estimator is actually an example of a more general family of estimators called M-estimators, that are obtained as the minimization of some cumulative loss function of the data. A M estimator is of the form

$$\widehat{\theta}_n \in \underset{\theta\in\mathbb{R}^d}{\text{argmin}}\, M_n(\theta) \quad \text{where} \quad M_n(\theta) = \sum_{i=1}^n m(X_i;\theta).$$

In the particular case of the MLE, we have $m(X;\theta) = -\log f_\theta(X)$.

1.4 Beyond the likelihood

Under some additional regularity conditions on some dominated model it is possible to define an important quantity called the Fisher information, which is useful to provide a lower bound on the quality of an (unbiased) estimator (see Section 1.5). The Fisher information will also be useful in the next chapter to characterize the asymptotic distribution of the maximum likelihood estimator.

To ease the presentation, we define everything in the single-parameter setting, that is when the parameter space Θ is a subset of $\mathbb{R}$. All this concepts can be extended to the multi-dimensional setting by replacing derivative with gradients, variances with covariances, and second derivative with Hessian. We will briefly discuss this extension afterwards.

Definition 21. *A (uni-dimensional) parameteric model* $\mathcal{M} = \{P_\theta, \theta \in \Theta \subseteq \mathbb{R}\}$ *is regular if*

1. *it is dominated by some reference measure* ν *and for all* θ, *the support of* f_θ, $S = \{x \in \mathcal{X} : f_\theta(x)\}$ *is independent of* θ
2. *for all* $x \in S$, $\theta \mapsto f_\theta(x)$ *is twice differentiable on* Θ *and its second derivative is continuous*
3. *for any event* $\mathcal{E}$, *we have*

$$\begin{aligned}\frac{\partial}{\partial\theta}\int_{\mathcal{E}} f_\theta(x)d\nu(x) &= \int_{\mathcal{E}}\frac{\partial}{\partial\theta}f_\theta(x)d\nu(x)\\ \frac{\partial^2}{\partial^2\theta}\int_{\mathcal{E}} f_\theta(x)d\nu(x) &= \int_{\mathcal{E}}\frac{\partial^2}{\partial^2\theta}f_\theta(x)d\nu(x)\end{aligned}$$

Example 22. *We can show that many classical parameteric model satisfy this assumption (e.g. Bernoulli models, Gaussian model, Poisson model). A counter-example that will be studied in an exercise is the family of uniform distributions on* $[0, \theta]$ *for* $\theta \in \mathbb{R}^+$*, which already violates assumption 1.*

1.4.1 The Fisher information

Definition 23. *The score function is defined as the derivative of the log-likelihood.*

$$s(X;\theta) = \frac{\partial \ell(X;\theta)}{\partial \theta} = \frac{1}{f_\theta(X)} \frac{\partial f_\theta(X)}{\partial \theta}$$

An important property of the score under a regular model is the following.

Lemma 24. *Under a regular model, for all* $\theta \in \Theta$, $\mathbb{E}_\theta[s(X;\theta)] = 0$.

Proof.

$$\begin{aligned}\mathbb{E}_\theta[s(X;\theta)] &= \int \frac{\partial \ell(x;\theta)}{\partial \theta} f_\theta(x) d\nu(x) = \int \frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) d\nu(x) = \int \frac{\partial}{\partial \theta} f_\theta(x) d\nu(x) \\ &\underset{(a)}{=} \int_S \frac{\partial}{\partial \theta} f_\theta(x) d\nu(x) \underset{(b)}{=} \frac{\partial}{\partial \theta} \left(\int_S f_\theta(x) d\nu(x) \right) \underset{(c)}{=} \frac{\partial}{\partial \theta}(1) = 0\end{aligned}$$

where (a) uses property 1. of a regular model, (b) uses property 3 and (c) uses that f_θ is a density.

□

The Fisher information matrix is defined as the variance of the score, which is equal to its second moment as the score is centered.

Definition 25. *In a regular model, the Fisher information of the observation* X *is defined as*

$$I^X(\theta) = \mathrm{Var}_\theta\left[s(X;\theta)\right] = \mathbb{E}_\theta\left[(s(X,\theta))^2\right] .$$

In the n*-sample case, we will write* $I_n(\theta)$ *to denote the Fisher information of the* n*-sample, and* $I(\theta)$ *the Fisher information of the observation made of a single realisation* $X_1 \sim P_\theta$*.*

1.4.2 Some properties of the Fisher information

Lemma 26. *Under a regular model, it holds that* $I^X(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(X;\theta)}{\partial^2 \theta}\right]$.

Exercise 27. *Prove it. Hint: start by computing the right-hand side, using property 3. of a regular model as in the proof of Lemma 24.*

The above lemma can be useful for the computation of the Fisher information. We now present another interesting property which is the additivity of the Fisher information. This property follows from the fact that the density of a couple of independent random variable is the product of their densities, and uses properties of the logarithm.

Lemma 28. *If X and Y are two independent random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, then*

$$I^{(X,Y)}(\theta) = I^X(\theta) + I^Y(\theta) .$$

It follows that for a n sample $X = (X_1, \dots, X_n) \overset{iid}{\sim} P_\theta$,

$$I_n(\theta) = I^X(\theta) = nI^{X_1}(\theta) = nI(\theta) .$$

Example 29. *Consider the Bernoulli model $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{B}(\theta)$. We have seen above that $I_n(\theta) = nI(\theta)$ where $I(\theta)$ is the Fisher information in a model with one Bernoulli observation X_1. In this model, we have*

$$\begin{aligned} L(X_1;\theta) &= \theta^{X_1}(1-\theta)^{1-X_1} \\ \ell(X_1;\theta) &= X_1 \log(\theta) + (1-X_1)\log(1-\theta) \\ \frac{\partial \ell(X_1;\theta)}{\partial \theta} &= \frac{X_1}{\theta} - \frac{1-X_1}{1-\theta} \\ \frac{\partial^2 \ell(X_1;\theta)}{\partial^2 \theta} &= -\frac{X_1}{\theta^2} + \frac{1-X_1}{(1-\theta)^2} \end{aligned}$$

hence

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(X_1;\theta)}{\partial^2 \theta}\right] = \mathbb{E}_\theta\left[\frac{X_1}{\theta^2} - \frac{1-X_1}{(1-\theta)^2}\right] = \frac{1}{\theta} - \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

Finally, using Lemma 28, we get $I_n(\theta) = \frac{n}{\theta(1-\theta)}$.

Extension to the multi-dimensional setting If $\theta = (\theta_1, \dots, \theta_d)$, the score is a vector in $\mathbb{R}^d$, defined as

$$s(X;\theta) = \nabla_\theta \ell(X;\theta) = \left(\frac{\partial \ell(X;\theta)}{\partial \theta_1}, \dots, \frac{\partial \ell(X;\theta)}{\partial \theta_d}\right)^\top .$$

In (an extension of the definition of a) regular model, the score satisfies $\mathbb{E}[s(X;\theta)] = 0$ and the Fisher information is defined as the (covariance) of the score, ie

$$I(\theta) = \mathbb{E}\left[(s(X,\theta))(s(X,\theta))^\top\right] .$$

The Fisher information is therefore a $d \times d$ matrix, and a counterpart of Lemma 26 can be proved:

$$I(\theta) = -\mathbb{E}\left[\left(\frac{\partial^2 \ell(X;\theta)}{\partial \theta_i \partial \theta_j}\right)_{\substack{1 \le i \le d \\ 1 \le j \le d}}\right] .$$

1.4.3 Interpretation of the Fisher information *(more advanced)*

The Fisher information will be shortely related to the minimal variance that a unbiased estimator can have. But we can still try to provide an explanation as to why it can be called "information".

First, due to its additivity property (Lemma 28), if we interpret $I(\theta)$ as an amount of "information" brought by one sample, we note that the Fisher information of a n-sample is the sum of all the information brought by individual samples. Moreover, another property is that given an observation X, any "summary" of this observation in the form of a statistic $S = s(X)$ has a smaller Fisher information.

Lemma 30. *For any statistic $S = s(X)$ of an observation X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, we have $I^S(\theta) \leq I^X(\theta)$.*

Proof. Let's write down the proof assuming that X takes values in a discrete space $\mathcal{X}$ (to avoid the concept of conditional density). X and $S = s(X)$ are clearly defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$. We can write

$$\mathbb{P}_\theta(X = x) \quad = \quad \mathbb{P}_\theta(X = x, S = s(x)) = \mathbb{P}_\theta(X = x | S = s(x)) \mathbb{P}_\theta(S = s(x))$$

Hence, for any $x \in \mathcal{X}$, writing $s = s(x)$, we have

$$f_\theta(x) = f_\theta(x|s) \widetilde{f}_\theta(s)$$

where we introduce f_θ the density of X, $\widetilde{f}_\theta$ the density of S and $f_\theta(x|s) := \mathbb{P}_\theta(X = x | S = s)$. Taking the logarithm and differentiating twice yields

$$\frac{\partial^2 \log f_\theta(x)}{\partial^2 \theta} = \frac{\partial^2 \log \widetilde{f}_\theta(s)}{\partial^2 \theta} + \frac{\partial^2 \log f_\theta(x|s)}{\partial^2 \theta}$$

and in particular

$$\frac{\partial^2 \log f_\theta(X)}{\partial^2 \theta} = \frac{\partial^2 \log \widetilde{f}_\theta(S)}{\partial^2 \theta} + \frac{\partial^2 \log f_\theta(X|S)}{\partial^2 \theta}$$

Taking the expectation and using Lemma 26 yields

$$I^X(\theta) = I^S(\theta) - \mathbb{E}_\theta \left[\frac{\partial^2 \log f_\theta(X|S)}{\partial^2 \theta} \right]$$

We conclude by noting that

$$-\mathbb{E}_\theta \left[\frac{\partial^2 \log f_\theta(X|S)}{\partial^2 \theta} \right] = \sum_s \mathbb{P}_\theta(S = s) \Bigg[\underbrace{-\mathbb{E}_\theta \left[\frac{\partial^2 \log \mathbb{P}_\theta(X|S = s)}{\partial^2 \theta} \right]}_{\geq 0} \Bigg]$$

and the term between brackets is positive as it is the Fisher information of the conditionnal distribution of X given $(S = s)$.

□

From this result a good statistic $S = s(X)$ is one that doesn't loose information, i.e. for which $I^S(\theta) = I^X(\theta)$. Sufficient statistic have this property, and are defined below.

Definition 31. *A statistic $S = s(X)$ is called sufficient for θ is the distribution of $X = (X_1, \ldots, X_n)$ contionally to S does not depend on θ.*

We admit the following characterization.

Theorem 32 (Neyman-Fisher)**.** *The statistic $S = s(X_1, \ldots, X_n)$ is sufficient for θ is there exists two positive functions g and h such that the density of X can be written*

$$f_\theta(x_1, \ldots, x_n) = g(x_1, \ldots, x_n) h(s(x_1, \ldots, x_n); \theta) \, .$$

1.4.4 The Kullback-Leibler divergence

We define another information theoretic quantity that is related to the likelihood (or actually rather to a likelihood ratio) and provides some notion of "distance" (although it is not a distance in the topological sense) between probability measures.

Definition 33. *For two probability measure P and Q that have a densities f and g with respect to the same probability measure ν and such that $g(x) = 0 \Rightarrow f(x) = 0$, we have*

$$\mathrm{KL}(P, Q) = \mathbb{E}_{X\sim P}\left[\log \frac{f(x)}{g(x)}\right].$$

In particular, if P_θ and $P_{\theta'}$ are two distributions in a regular model (actually assumption 1. in Definition 21 is sufficient), we can define

$$\mathrm{K}(\theta, \theta') := \mathrm{KL}(P_\theta, P_{\theta'}) = \mathbb{E}_\theta\left[\log \frac{f_\theta(X)}{f_{\theta'}(X)}\right].$$

Example 34. *The KL divergence between $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\mu', \sigma^2)$ is*

$$\mathrm{K}(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2}.$$

The KL divergence between two Bernoulli distributions of parameters θ and θ' is

$$\mathrm{K}(\theta, \theta') = \theta \log\left(\frac{\theta}{\theta'}\right) + (1-\theta)\log\left(\frac{1-\theta}{1-\theta'}\right).$$

1.5 The Cramer-Rao lower bound

The Fisher information defined in the previous section enables us (in the case of uni-dimensional estimation) to solve the following question: what is the minimal variance of an unbiased estimator? We consider this question for regular models.

Theorem 35. *Assume the statistical model is regular. Let $\widehat{g}$ be an estimator of $g(\theta) \in \mathbb{R}$ where g is differentiable. We assume that $\widehat{g} = h(X)$ is such that $\mathbb{E}_\theta[\widehat{g}_n] = g(\theta)$ (unbiased estimator) and*

$$\frac{\partial}{\partial\theta}\int h(x) f_\theta(x) d\nu(x) = \int h(x)\left(\frac{\partial}{\partial\theta} f_\theta(x)\right) d\nu(x)$$

Then, for all $\theta \in \Theta$,

$$\mathrm{Var}_\theta[\widehat{g}] \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

Proof. The idea of the proof is to differentiate $g(\theta) = \mathbb{E}_\theta[h(X)]$ and introduce the score. Using one of

the assumptions, we can write

$$\begin{aligned}
g'(\theta) &= \frac{\partial}{\partial\theta}\int_S h(x) f_\theta(x) d\nu(x) = \int_S h(x)\left(\frac{\partial}{\partial\theta} f_\theta(x)\right) d\nu(x) \\
&= \int_S h(x)\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) f_\theta(x) d\nu(x) \\
&\overset{(a)}{=} \int_S (h(x) - \mathbb{E}_\theta[h(X)])\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) f_\theta(x) d\nu(x) + \mathbb{E}_\theta[h(X)] \underbrace{\int_S \left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) f_\theta(x) d\nu(x)}_{=0} \\
&\overset{(b)}{=} \int_S (h(x) - \mathbb{E}_\theta[h(X)])\left(\frac{\partial}{\partial\theta}\log f_\theta(x) - \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right) f_\theta(x) d\nu(x)
\end{aligned}$$

where both (a) and (b) use that the expected score is zero by Lemma 24.

Now we assume that $\mathbb{E}_\theta[h^2(X)] < \infty$ (otherwise, the inequality in Theorem 35 is trivially true). Then we can use the Cauchy-Schwarz inequality to get

$$\begin{aligned}
|g'(\theta)| &\leq \sqrt{\mathbb{E}_\theta\left[(h(x) - \mathbb{E}_\theta[h(X)])^2\right]}\sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x) - \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right)^2\right]} \\
&\leq \sqrt{\mathrm{Var}_\theta[h(X)]}\sqrt{\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]} \\
&\leq \sqrt{\mathrm{Var}_\theta[h(X)]}\sqrt{I(\theta)}
\end{aligned}$$

where the last step uses the definition of the Fisher information.

□

An unbiased estimator that achieves the Cramer-Rao lower bound for all values of $\theta \in \Theta$ is called efficient (or uniformly efficient). The example below show that there exists efficient estimators.

Exercise 36. *Verify that in the Bernoulli model $X_1, \dots X_n \overset{iid}{\sim} \mathcal{B}(p)$ the MLE is an efficient estimator.*

1.6 Exponential families

Actually, the reason why in the Bernoulli model we can find an efficient estimator comes from the fact that the set of Bernoulli distributions is a particular example of exponential family. We define exponential families below, and we will see several of their properties in this class.

Definition 37. *An exponential family is a set of probability distibutions on some set $\mathcal{X}$ defined as*

$$\mathcal{P} = \{P_\theta, \theta \in \Theta : \quad P_\theta \text{ has a density } \quad f_\theta(x) = h(x)\exp\left(a(\theta)^\top T(x) - b(\theta)\right) \text{ wrt to } \nu\}$$

where ν is a reference measure (common to all distributions), $h : \mathcal{X} \to \mathbb{R}^+$ is a positive function, $a : \Theta \to \mathbb{R}^d$, $b : \Theta \to \mathbb{R}$ and $T : \mathcal{X} \to \mathbb{R}^d$ are some functions and $u^\top v = \sum_{i=1}^d u_i v_i$ is the scalar product in $\mathbb{R}^d$.

$T(x) \in \mathbb{R}^d$ is called the canonical statistic and d is the dimension of the exponential family. In a one-dimensional exponential family, the density can simply be expressed

$$f_\theta(x) = h(x)\exp\left(a(\theta)T(x) - b(\theta)\right).$$

Exercise 38. *Justify that family of Bernoulli distribution* $\mathcal{P} = \{\mathcal{B}(p), p \in (0,1)\}$ *form an exponential family (of dimension 1).*

Actually, we can prove that efficient estimator can only exist in some exponential families, and for a particular parameter to estimate. There are therefore not so much common. In the next chapter, we will define an asymptotic notion of efficiency, which can be easier to attain.

Chapter 2

Asymptotic properties of estimators

In this chapter, we focus on the n-sample case, in which $X = (X_1, \dots, X_n) \overset{iid}{\sim} P_\theta$. For each n, given an estimator $\widehat{g}_n = h(X_1, \dots, X_n)$ of a certain parameter of interest $g(\theta)$, we are interested in studying the sequence of estimators $(\widehat{g}_n)_n$ when the sample size n grows large. As the $\widehat{g}_n$ are random variables, we first recap the different notion of convergences, as well as some important results.

2.1 Refresher: Convergence of random variables

Definition 39. *Let $Z_1, Z_2, \dots$ be a sequence of random variable and let Z be another random variable. Let F_n denote the CDF of Z_n and let F denote the cdf of Z.*

1. ***Z_n converges to Z in probability*** *if, for every $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|Z_n - Z| > \varepsilon) = 0$.*
   We write $Z_n \xrightarrow{P} Z$.
2. ***Z_n converges to Z in distribution*** *if, $\lim_{n\to\infty} F_n(t) = F(t)$ for all t for which F is continuous.*
   We write $Z_n \rightsquigarrow Z$.
3. ***Z_n converges to Z almost surely*** *if $\mathbb{P}\left(\lim_{n\to\infty} Z_n = Z\right) = 1$. We write $Z_n \xrightarrow{a.s.} Z$.*
4. ***Z_n converges to Z in quadratic mean*** *if $\lim_{n\to\infty} \mathbb{E}\left[(Z_n - Z)^2\right] = 0$. We write $Z_n \xrightarrow{L^2} Z$.*

In statistics, the first two notions are the most common, and we will mostly discuss them in the following. The definitions above were all given for real-values random variables, but can be extended to the multi-dimensional setting. For the convergence in probability, the distance between Z_n and Z and $\mathbb{R}^d$ can no longer be measured with the absolute value, but given any distance d on $\mathbb{R}^d$ (for example the Euclidian distance), we define $Z_n \xrightarrow{P} Z$ is for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(d(Z_n, Z) > \varepsilon\right) = 0$.

The convergence in distribution in $\mathbb{R}^d$ can still be characterized by the cdf, but in this case, the cdf is a multi-variate function and we should have, for all $z = (z_1, \dots, z_d)$ in which F is continuous,

$$\lim_{n\to\infty} \mathbb{P}\left(Z_n^1 \le z_1, \dots, Z_n^d \le z_d\right) = \mathbb{P}(Z^1 \le z_1, \dots, Z^d \le z_d) = 0.$$

Example 40. *$Z_n \sim \mathcal{N}(0, \frac{1}{n})$. Justify that Z_n converges to 0 (the random variable that is constant and equal to zero) in distribution and in probability.*

2.1.1 Properties

The following relationship between the different convergence notions are useful.

Lemma 41. *1.* $X_n \xrightarrow{P} X$ *implies that* $X_n \rightsquigarrow X$

2. $X_n \xrightarrow{P} c$ *where* c *is a constant if and only if* $X_n \rightsquigarrow X$

3. $X_n \xrightarrow{L^2} X$ *implies that* $X_n \xrightarrow{P} X$

We note that (a) and (c) are not equivalences. In particular, beyond the case of convergence to constants, the convergence in distribution does not imply the convergence in probability. A (contrived) counter-example is the following: take any symmetric distribution Y, that is a distribution for which Y and $-Y$ have the same distribution (for example, a centered Gaussian distribution). Define $Z_n = Y$ for all n and $Z = -Y$. As the cdf and Z_n and that of Z are equal, we have in particular $Z_n \rightsquigarrow Z$. However, $\mathbb{P}(|Z_n - Z| > \varepsilon) = \mathbb{P}(|2Y| > \varepsilon)$ does not converge to zero for every ε (unless $Y = 0$ a.s.).

Lemma 42 (continuous mapping)**.** *Let* $g : \mathcal{X} \to \mathbb{R}$ *be a continuous function. Then*

- *If* $X_n \xrightarrow{P} X$ *then* $g(X_n) \xrightarrow{P} g(X)$
- *If* $X_n \rightsquigarrow X$ *then* $g(X_n) \rightsquigarrow g(X)$

Lemma 43 (Slutsky lemma)**.** *If* $X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$ *where* c *is a constant, then, for any continuous function* g*,*

$$g(X_n, Y_n) \rightsquigarrow g(X, c) .$$

In particular

- $X_n + Y_n \rightsquigarrow X + c$
- $X_n Y_n \rightsquigarrow cX$

Slutsky's lemma is a consequence of the fact that as a couple of random variables (X_n, Y_n) converges in distribution to (X, c) (and the fact that the continuous mapping lemma also apply to multi-variate random variables).

2.1.2 Two fundamental theorems

We recall here the two fundamental theorems in statistics: the law of large numbers and the central limit theorem. Given an iid sequence Z_i, they provide some convergence results for the empirical average

$$\widehat{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

Theorem 44 (Law of large numbers)**.** *If* $(Z_i)_{i \in \mathbb{N}}$ *is an iid sequence with* $\mathbb{E}[Z_1] < \infty$*, we have*

$$\frac{1}{n} \sum_{i=1}^{n} Z_i \xrightarrow{P} \mathbb{E}[Z_1]$$

Actually, a stronger version of this result (called the strong law of large numbers) holds under the same assumptions, in which the convergence in probability is replaced by an almost sure convergence.

Theorem 45 (Central limit theorem)**.** *If $(Z_i)_{i\in\mathbb{N}}$ is an iid sequence with $\mathbb{E}[Z_1^2] < \infty$, letting $\mu = \mathbb{E}[Z_1]$ and $\sigma^2 = \mathrm{Var}[Z_1]$, we have*

$$\sqrt{\frac{n}{\sigma^2}}\left(\widehat{Z}_n - \mu\right) \rightsquigarrow \mathcal{N}(0,1)$$

Under the hypotheses of the central limit theorem, $\widehat{Z}_n$ can be written

$$Z_n = \mu + \sqrt{\frac{\sigma^2}{n}} Y_n$$

where $Y_n \rightsquigarrow \mathcal{N}(0,1)$. Therefore, informally, the distribution of $\widehat{Z}_n$ is close to $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, a Gaussian distribution whose variance decays to zero and is therefore more and more concentrated around μ. We may write $\widehat{Z}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ and talk about the asymptotic distribution of $\widehat{Z}_n$.

2.2 Consistency and asymptotic normality

Definition 46. *An estimator $\widehat{g}_n$ of $g(\theta)$ is consistent if for every $\theta \in \Theta$, $\widehat{g}_n \xrightarrow{P} g(\theta)$.*

Consistency of estimators will often follow from the law of large numbers. When we further have an almost sure convergence, that is when $\widehat{g}_n \xrightarrow{a.s.} g(\theta)$, we shall say that $\widehat{g}_n$ is strongly consistent.

Lemma 41 and Lemma 42 also yield the following properties:

- If the quadratic risk $\mathrm{R}_\theta(\widehat{g}_n)$ goes to zero when n goes to infinity, $\widehat{g}_n$ is consistent.
- If $\widehat{\theta}_n$ is a consistent estimator of θ and g is a continuous mapping, then $\widehat{g}_n = g(\widehat{\theta}_n)$ is a consistent estimator of $g(\theta)$.

Example 47. *Justify that the empirical mean and empirical variance defined in the previous chapter are consistent estimators.*

Given a consistent estimator $\widehat{g}_n$, we may be interested in how fast $\widehat{g}_n - g(\theta)$ converges to zero. To do so, we will look at the limit distribution of (some re-normalization) of this random variable.

If we take the example of the empirical $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ estimator of the common mean μ of some n sample $(X_1, \ldots, X_n)$ that has variance σ^2, the Central Limit Theorem tells us that

$$\sqrt{n}\left(\widehat{\mu}_n - \mu\right) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

Here, the limit distribution is Gaussian, and the convergence speed is $\sqrt{n}$. Due to the generality of the Central Limit Theorem, we expect this Gaussian limit behavior to be a general pattern for estimators, hence the definition of asymptotic normality.

Definition 48. *An estimator is $\widehat{g}_n$ of $g(\theta)$ is asymptotically normal is it satisfies, for all $\theta \in \Theta$,*

$$\sqrt{n}\left(\widehat{g}_n - g(\theta)\right) \rightsquigarrow \mathcal{N}(0, \sigma_\theta^2)$$

where σ_θ^2 is called the asymptotic variance.

It is worth mentioning that there exists estimators that are have a limiting distribution, but are not asymptotically normal. It means that they satisfy something like

$$g(n)\left(\widehat{g}_n - g(\theta)\right) \rightsquigarrow Z$$

where $g(n)$ is some convergence speed (that can be different than $\sqrt{n}$) and Z is some fixed distribution (that is not necessarly Gaussian).

Example 49. *As studied in exercise, in the model $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{U}([0,\theta])$, the moment estimator is $\widehat{\theta}_n = \frac{2}{n}\sum_{i=1}^n X_i$ while the MLE is $\widetilde{\theta}_n = \max_{i=1..n} X_i$.*

Using the Central Limit Theorem (and the continuous mapping lemma), we can show that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \frac{\theta^2}{3}\right)$$

hence the moment estimator is asymptotically normal with asymptotic variance $\sigma_\theta = \frac{\theta^2}{3}$.

On the other hand, we computed the distribution of $\widetilde{\theta}_n$ in exercise, showing that

$$\mathbb{P}\left(\widetilde{\theta}_n \leq x\right) = \begin{cases} 1 & \text{if } x \geq \theta \\ \frac{x^n}{\theta^n}\mathbb{1}_{[0,\theta]}(x) & \text{else.} \end{cases}$$

Hence we have, for all $u > 0$,

$$\mathbb{P}\left(\theta - \widetilde{\theta}_n \geq u\right) = \left(1 - \frac{u}{\theta}\right)^n \mathbb{1}_{[0,\theta]}(u)$$

and finally, for all $t > 0$,

$$\mathbb{P}\left(n\left(\theta - \widetilde{\theta}_n\right) \geq t\right) = \left(1 - \frac{t}{n\theta}\right)^n \mathbb{1}_{[0,n\theta]}(t)$$

The limit of the right-hand side when n goes to infinity is equal to $e^{-\frac{t}{\theta}} = \mathbb{P}(Z > t)$ where Z is an exponential distribution with parameter $1/\theta$. Finally, one can write

$$n\left(\theta - \widetilde{\theta}_n\right) \rightsquigarrow \mathcal{E}\left(\theta^{-1}\right).$$

This provide another argument for using the MLE over the moment estimator in this particular case, as its asymptotic convergence is faster.

In Section 2.4, we will actually see that for regular models, the MLE is asymptotically Gaussian. The reason for this different behavior stems from the fact that the model considered here is not regular: one can indeed see that all the possible densities do not have the same support.

Comparing asymptotically normal estimators Between two asymptotically normal estimator, the one with smallest asymptotic variance σ_θ^2 is the one that converges "faster" to the parameter $g(\theta)$. This can be measured by the fact that, if we build asymptotic confidence intervals for $g(\theta)$ of level $1-\alpha$, using the estimator with smallest asymptotic variance will yield the smallest confidence region.

If two asymptotically normal estimators $\widehat{g}_n$ and $\widetilde{g}_n$ have respective asymptotic variances σ_θ^2 and $\widetilde{\sigma}_\theta^2$ and that $\sigma_\theta^2 \leq \widetilde{\sigma}_\theta^2$ for all $\theta \in \Theta$ (with at least one strict inequality), we say that $\widehat{g}_n$ is asymptotically more efficient than $\widetilde{g}_n$.

2.3 The $\triangle$-method

We now present a useful tool to compute asymptotic distributions of some transformation of an asymptotically normal estimator: the Δ-method. This result implies that under some mild conditions, if $\widehat{\theta}_n$ is an asymptotically normal estimator of θ, then $g(\widehat{\theta}_n)$ is an asymptotically normal estimator of $g(\widehat{\theta}_n)$.

Theorem 50. *Suppose that for some sequence of random variance* (Z_n)*,*

$$\sqrt{n}(Z_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

and that g *is a differentiable function such that* $g'(\mu) \neq 0$*. Then*

$$\sqrt{n}(Z_n - \mu) \rightsquigarrow \mathcal{N}\left(0, (g'(\mu))^2\sigma^2\right) .$$

In other words,

$$Z_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \textit{implies that} \quad g(Z_n) \approx \mathcal{N}\left(g(\mu), (g'(\mu))^2\frac{\sigma^2}{n}\right) .$$

2.4 Asymptotic properties of the Maximum Likelihood Estimator

2.5 Asymptotic efficiency

Chapter 3

Likelihood Ratio based Testing

Chapter 4

A glimpse of Bayesian statistics