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

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

March 4, 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.

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 "converge" to $g(\theta)$ and to be able to measure the "convergence speed".

We will discuss these asymptotic properties in the next chapter.

1.3 Estimation procedures

1.3.1 The moment method

When $X = (X_1, \ldots, X_n) \overset{iid}{\sim} P_\theta$, 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

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

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

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

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 $X = (X_1, \dots, X_n)$, the empirical distribution P_n is defined as*

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

where δ_x is the Dirac measure in x. That is, $\widehat{P}_n$ is a distribution supported in $\{X_1, \dots, X_n\}$, the observed values in X and for every potential observation $x \in \mathbb{R}$, we have $\widehat{P}_n(x) = \frac{\#\{\text{nb of occurences of } x \text{ in } X\}}{n}$.

The cdf of the empirical distribution is called the empirical cdf and satisfies

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

Remark 12. *When the function $H(P)$ 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 measures 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

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

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*

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

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 (which requires to look at the sign of the second derivative). In more complex cases, it can only be approximated using some optimization algorithm. In complex models (see the Gaussian mixture example in exercice), 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 17. *The log-likelihood of the observation X given a parameter θ is denoted by*

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

Exercise 18. *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 $\lambda > 0$ is a discrete distribution supported on $\mathbb{N}$ defined as*

$$\mathbb{P}(Z = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

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 19. *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*

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

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

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

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-estimator, that are obtained as the minimization of some cumulative loss function of the data. A M estimator is of the form

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

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 (see Section 1.3.3) 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.

1.4.1 The Fisher information

Definition 20. *The score function is defined as the gradient of the log-likelihood.*

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

If $\theta = (\theta_1, \dots, \theta_d)$, we have

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

The expectation of the score can be computed as follows

$$\mathbb{E}_\theta[s(X;\theta)] = \int \nabla_\theta \ell(x;\theta) f_\theta(x) d\nu(x) = \int \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x) d\nu(x) = \int \nabla_\theta f_\theta(x) d\nu(x)$$

If all the densities f_θ have a support that is independent on θ, denoted by S and if we can invert the integral and the gradient (which will be the case in a "regular model" to be defined shortly), we can further write

$$\mathbb{E}_\theta[s(X;\theta)] = \int_S \nabla_\theta f_\theta(x) d\nu(x) = \nabla_\theta \left(\int_S f_\theta(x) d\nu(x) \right) = \nabla_\theta(1) = 0$$

and the score is centered. The Fisher information matrix can be defined for any model in which the score is centered, as the covariance matrix of the score.

Definition 21. *In a dominated model in which for all $\theta \in \Theta$, $\mathbb{E}_\theta[s(X;\theta)] = 0$ and $\mathbb{E}_\theta[\|s(X;\theta)\|^2] < \infty$, the Fisher information matrix is defined as*

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

When we seek to estimate a multivariate parameter $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, the score is a vector in $\mathbb{R}^d$ and the Fisher information is a (semi-definite positive) matrix. But when the parameter $\theta \in \mathbb{R}$, both the score and the Fisher information are just numbers, and we have

$$\begin{aligned} s(X;\theta) &= \frac{\partial \ell(X;\theta)}{\partial \theta} \\ I(\theta) &= \mathbb{E}[(s(X,\theta))^2] \end{aligned}$$

In the rest of this chapter, to ease the presentation, we will present all results in this uni-dimensional setting, but all of them can be extended to the multi-dimensional setting.

1.4.2 Some properties in regular models

The assumption of a regular model that we present below will be sufficient to have a centered score and therefore to be able to define the Fisher information as the variance of the score, but it will also be useful to establish other properties of the MLE (see the next chapter).

Definition 22. *A parameteric model $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is regular if*

1. *it is dominated by some reference measure ν and $S = \{x \in \mathcal{X} : f_\theta(x)\}$, the support of f_θ 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 23. *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.*

In the above section, we sketched the proof of the following result.

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

Under a regular model, we can also propose an alternative expression for the Fisher information that can be convenient for its computation.

Lemma 25. *Under a regular model, it holds that $I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2\ell(X;\theta)}{\partial^2\theta}\right]$.*

Exercise 26. *Prove it.*

The n-sample case When $X = (X_1,\dots,X_n) \sim P_\theta$, we will index by n the score and the Fisher information, ie we will write $s_n(X;\theta)$ and $I_n(\theta)$. The Fisher information can be easily related to the Fisher information in a model with a single observation $X_1 \sim P_\theta$. Indeed, due to the properties of the logarithm, the score is clearly additive:

$$s(X;\theta) = \sum_{i=1}^n s(X_i;\theta)$$

hence we have, due to independence

$$I_n(\theta) = \mathrm{Var}_\theta\left[\sum_{i=1}^n s(X_i;\theta)\right] = n\mathrm{Var}_\theta\left[(s(X_1;\theta))^2\right] = nI(\theta)\ .$$

Exercise 27. *Compute the Fisher information in a Bernoulli model, a Gaussian model and a Poisson model.*

1.4.3 The Kullback-Leibler divergence

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

Definition 28. *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 $g(x) = 0 \to f(x) = 0$, 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 22 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 29. *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 will enable 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 30. *Let $\widehat{g}$ be an estimator of $g(\theta) \in \mathbb{R}$ where g is differentiable. We make the following assumptions:*

- *For all $\theta \in \Theta$, the density functions have the same support $S = \{x \in \mathcal{X} : f_\theta(x) > 0\}$.*
- *For all $x \in \mathcal{X}$, $\theta \mapsto f_\theta(x)$ is differentiable on Θ*
- *$\widehat{g} = h(X)$ 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:

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int_S \left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right) f_\theta(x) d\nu(x) = 0\,.$$

Now we assume that $\mathbb{E}_\theta[h^2(X)] < \infty$ (otherwise, the inequality in Theorem 30 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.

□

Chapter 2

Asymptotic properties of estimators

Chapter 3

Likelihood Ratio based Testing

Chapter 4

A glimpse of Bayesian statistics