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

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

March 1, 2026

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 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.

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 The observation X will often be a random vector of the form $X = (X_1, \ldots, X_n)$ where the X_i are assumed to be *i.i.d.* (independent and identically distributed) realizations of some common 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 $\mathbb{R}$ or $\mathbb{R}^d$.

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

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

In that case, the statistical model is typically expressed as possible candidates for P, F or f directly, instead of possible candidates for $P^{\otimes n}$, $F^{\otimes n}$ and $f^{\otimes n}$. By a slight abuse of notation, we will also denote by P_θ, F_θ and f_θ the possible candidate for the distribution of X_1, for θ in some parameter space Θ.

Example 1.1. *We consider a Gaussian n-sample with unit variance and unknown mean. That is, we have $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ for $\theta \in \Theta = \mathbb{R}$. 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 $\mathcal{M}$ for X 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 1.2. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *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, $p = p(\mu, \sigma)$, 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 1.3 (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, both expressed with expectations.

1.2.1 Recap: Densities and 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,

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

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 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 encounter 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 $\phi : \mathcal{S} \to \mathbb{R}$, $\mathbb{E}_\theta[\phi(X_1)] = \sum_{u \in \mathcal{S}} \phi(u) f_\theta(u)$

1.2.2 Bias, Variance and Quadratic Risk

Definition 1.4. *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 1.5. *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 1.6. *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* $\mathrm{MSE}_\theta(\widehat{g})$*:*

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

Theorem 1.7 (bias-variance decomposition)**.** *Assume $g(\theta) \in \mathbb{R}$. We have*

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

Proof. When $g(\theta) \in \mathbb{R}$, we write

$$\begin{aligned}
\mathrm{R}_\theta(\widehat{g}) &= \mathbb{E}_\theta\left[(\widehat{g} - g(\theta))^2\right] = \mathbb{E}_\theta\left[((\widehat{g} - \mathbb{E}_\theta[\widehat{g}]) + (\mathbb{E}_\theta[\widehat{g}] - g(\theta)))^2\right] \\
&= \mathbb{E}_\theta\left[(\widehat{g} - \mathbb{E}_\theta[\widehat{g}])^2\right] + \mathbb{E}_\theta\left[(\mathbb{E}_\theta[\widehat{g}] - g(\theta))^2\right] + 2\mathbb{E}_\theta\left[(\widehat{g} - \mathbb{E}_\theta[\widehat{g}])(\mathbb{E}_\theta[\widehat{g}] - g(\theta))\right] \\
&= \mathbb{E}_\theta\left[(\widehat{g} - \mathbb{E}_\theta[\widehat{g}])^2\right] + (\mathbb{E}_\theta[\widehat{g}] - g(\theta))^2 + 2(\mathbb{E}_\theta[\widehat{g}] - g(\theta))\underbrace{\mathbb{E}_\theta\left[\widehat{g} - \mathbb{E}_\theta[\widehat{g}]\right]}_{=0} \\
&= \mathrm{Var}_\theta[\widehat{g}] + (\mathrm{b}(\widehat{g}))^2 .
\end{aligned}$$

□

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 $\theta_1 \in \Theta$ but $\mathrm{R}_{\theta_2}(\widehat{g}) > \mathrm{R}_{\theta_2}(\widetilde{g})$ for a different parameter $\theta_2 \in \Theta$.

Definition 1.8. *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.

Illustration Figure 1.1 below displays the distribution of three estimators. The first one seems to be unbiased, but has a large variance, the second one has a small negative bias and a smaller variance, and the third one has a small positive bias and an even smaller variance. Despite the bias, on this example the third estimator is probably the one with the smallest mean-squared error.

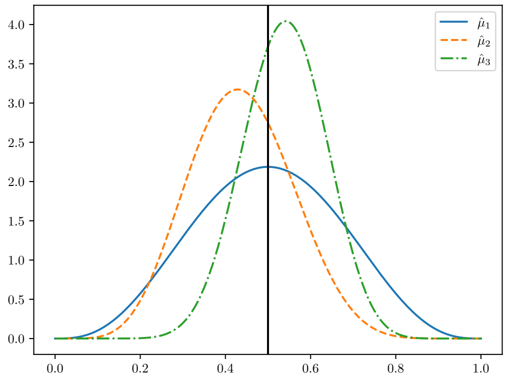

Figure 1.1: Comparing the distribution of three estimators of the same parameter

Influence of the sample size When X is a n-sample, the performance measures for an estimator $\widehat{g}_n$ are all defined for 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.

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 (some function of) X_1.

In the simplest case, the parameter of interest can directly be written as an expectation:

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

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.

Example 1.9 (the empirical cdf)**.** *Given iid samples $X_1 \dots, X_n$ from some distribution P in $\mathbb{R}$ whose cdf is F, we want to estimate the function F, that is, for each value $x \in \mathbb{R}$ we want to estimate the quantity $F(x) = \mathbb{P}(X \le x) = \mathbb{E}[\mathbb{1}(X \le x)]$. As this quantity can be written as an expectation, its moment estimator is simply*

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

The function $x \mapsto \widehat{F}_n(x)$ is called empirical cdf of $X = (X_1, \dots, X_n)$.

We can further compute the bias and the bias of this estimator

$$b(\widehat{F}_n(x)) = \mathbb{E}\left[\widehat{F}_n(x)\right] - F(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{1}(X_i \le x)\right] - F(x) = F(x) - F(x) = 0$$

and its variance:

$$\begin{aligned}\mathrm{Var}[\widehat{F}_n(x)] &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}\mathbb{1}(X_i \le x)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\mathbb{1}(X_i \le x)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}F(x)(1-F(x)) = \frac{F(x)(1-F(x))}{n}\,.\end{aligned}$$

For the variance computation we have used that the variance of the sum of independent random variables is the sum of their variances, and that $\mathbb{1}(X_i \le x)$ which takes value in $\{0,1\}$ is a Bernoulli distribution with mean $p = \mathbb{P}(X_i \le x) = F(x)$, whose variance if $p(1-p)$.

More generally, suppose that we seek to estimate a multi-dimensional parameter $\theta = (\theta_1, \dots, \theta_k)^\top \in \mathbb{R}^k$ 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 1.10. $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. *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.*

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

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 of equations*

$$\left\{ \begin{array}{rcl} \mu & = & \frac{1}{n}\sum_{i=1}^n X_i \\ \mu^2 + \sigma^2 & = & \frac{1}{n}\sum_{i=1}^n X_i^2 \end{array} \right.$$

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 (unadjusted) empirical variance, which can also be rewritten

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

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 the distribution of X_1 (for example some moment of this distribution, or some quantile). This means that we can write

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

where P is the cdf of X_1.

A plug-in estimator replaces ("plugs in") the unknown distribution P by an empirical variant of this distribution

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

where the empirical distribution $\widehat{P}_n$ is defined below.

Definition 1.11. *The empirical distribution of* $X = (X_1, \dots, X_n)$ *is defined as*

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

where the Dirac measure in x *is defined as,* $\delta_x(A) = 1$ *if* $x \in A$, $\delta_x(A) = 0$ *otherwise, for all event* A.

$\widehat{P}_n$ *is a discrete distribution supported on* $\mathcal{S} = \{X_1, \dots, X_n\}$, *the set of distinct values in our sample, and for all* x,

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

The cdf of the distribution $\widehat{P}_n$ *can be shown to be the empirical cdf* $\widehat{F}_n$, *presented in Example 1.9.*

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

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

Remark 1.12. *When the functional $H(P)$ is defined as some expectation under P, the moment method and the plug-in method actually coincide. Indeed, if*

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

the plug-in method yields

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

Such estimators are also called "empirical estimators".

But plug-in estimator can be more general when H is not defined as some expectation. For example when we want to estimate some quantile of a distribution P, e.g. its median, the plug-in estimator, also called empirical quantile is the corresponding quantile of the empirical distribution $\widehat{P}_n$.

Example 1.13 (variance estimation)**.** *Given that the variance of a distribution can be written*

$$\mathrm{Var}[X] = \mathbb{E}_{X\sim P}[(X - \mathbb{E}_{X\sim P}[X])^2]$$

the plug-in method yields the estimator

$$\begin{aligned}\widehat{\sigma}_n^2 &= \mathbb{E}_{X\sim\widehat{P}_n}[(X - \mathbb{E}_{X\sim\widehat{P}_n}[X])^2]\\ &= \mathbb{E}_{X\sim\widehat{P}_n}[(X - \widehat{\mu}_n)^2]\\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2\end{aligned}$$

where we introduce the empirical mean $\widehat{\mu}_n = \mathbb{E}_{X\sim\widehat{P}_n}[X] = \frac{1}{n}\sum_{i=1}^{n}X_i$. We recover the same estimator as the one derived before the Gaussian distributions.

As for its properties, it is well known that the empirical variance is biased. Indeed one can check that $\mathbb{E}[\widehat{\sigma}_n^2] = \frac{n-1}{n}\sigma^2$ if $\sigma^2 = \mathrm{Var}[X_1]$. Hence an unbiased estimator of the variance is the adjusted empirical variance

$$\widetilde{\sigma}_n^2 = \frac{n}{n-1}\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2$$

which is often preferred in practise. Still, comparing $\widehat{\sigma}_n^2$ and $\widetilde{\sigma}_n^2$ in terms of mean-squared error for certain distributions (e.g. Gaussian) shows that the unadjusted estimator can have a smaller MSE.

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 1.14. *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 likelihood can be written

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

Example 1.15. *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

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

The likelihood can be interpreted as the probability of observing 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 interpretation 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 1.16. *A maximum likelihood estimator (MLE) of a parameter θ is an estimator satisfying*

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

As we will see in some exercises, the maximum likelihood is not always unique. From a computational perspective (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 1.17. *The log-likelihood of the observation X given a parameter θ is denoted by*

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

Example 1.18 (Bernoulli distributions). *As written above, the likelihood of an n-sample from a Bernoulli distribution with parameter θ is*

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

The log-likelihood takes the simple form

$$\ell(X;\theta) = \left(\sum_{i=1}^{n} X_i\right)\log(\theta) + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-\theta) := g(\theta)$$

In order to find the maximizer of $g(\theta)$, we compute the derivative

$$g'(\theta) = \frac{\sum_{i=1}^n X_i}{\theta} - \frac{n-\sum_{i=1}^n X_i}{1-\theta} = \frac{\sum_{i=1}^n X_i - \theta n}{\theta(1-\theta)}.$$

There is a unique solution to $g'(\theta) = 0$ given by $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Moreover $g'(\theta) > 0$ iff $\theta < \widehat{\theta}_n$, hence this solution is a maximizer, the MLE.

We remark that the same estimator could also have been obtained using the moment method by remarking that the parameter θ is also the mean of the Bernoulli distribution: $\theta = \mathbb{E}_{Z\sim\mathcal{B}(\theta)}[Z]$.

Example 1.19 (linear regression). *We collect pairs of independent samples (X_i, Y_i) such that $X_i \in \mathbb{R}^d$ comes from distribution with density f and $Y_i = \theta^\top X_i + \varepsilon_i$. Assuming that the noise ε_i is Gaussian with known variance σ^2 allows to write the likelihood of n independent observations:*

$$L((X_1,Y_1),\dots,(X_n,Y_n);\theta) = \prod_{i=1}^{n} f(X_i)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(Y_i-\theta^\top X_i)^2}{2\sigma^2}\right)$$

The MLE is a minimizer of the function $g : \mathbb{R}^d \to \mathbb{R}$ given by

$$g(\theta) = \sum_{i=1}^{n} (Y_i - \theta^\top X_i)^2,$$

also known as the least-squares estimate. The gradient of g is given by

$$\nabla g(\theta) = -2\sum_{i=1}^{n}(Y_i - \theta^\top X_i)X_i$$

and solving $\nabla g(\theta) = 0$ (and checking that the Hessian matrix is negative definite in this critical point) yields that the MLE is given by

$$\widehat{\theta}_n = \left(\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \sum_{i=1}^{n} Y_i X_i,$$

provided that the design matrix $\sum_{i=1}^n X_i X_i^\top$ is invertible.

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

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

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

1.4 MLE: Computational Considerations

In the examples of the previous sections, the Maximum Likelihood Estimator can be computed explicitly, by finding the critical point (for which the derivative, or the gradient is zero) and proving that it is indeed a maximizer (e.g., by checking that the second derivative, or the Hessian is negative in the critical point). In more complex cases, the maximizer in the definition of the MLE can only be approximated using some optimization algorithm converging towards the maximizer (e.g. a gradient ascent).

Example 1.20. *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*

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

We remark 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*

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

In this example, no closed-form expression exists for the MLE and we have to resort to an optimization algorithm.

In complex models involving latent variables, i.e. variable that are not actually observed (e.g. the membership of some individual in some cluster, which we also try to infer) as in mixture models, more fancy approximation scheme are needed, like the Expectation Maximization (EM) iterative algorithm. The next sections present some interesting particular case.

1.4.1 Gaussian Mixture Models

In this section we present the important Gaussian mixture models, that can be used for (model-based) clustering. We introduce the following notation for the probability simplex:

$$\Delta_k = \left\{ w \in [0, 1]^K : \sum_{k=1}^{K} w_i = 1 \right\}$$

which consists of all probability vectors over K elements.

Definition 1.21. *Let* $f_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)$ *be the density of a Gaussian distribution in* $\mathbb{R}^d$ *with mean* $\mu \in \mathbb{R}^d$ *and covariance matrix* $\Sigma \in \mathbb{R}^{d\times d}$*. A distribution is a Gaussian mixture with* K *components if it has a density that can be written*

$$f_\theta(x) = \sum_{k=1}^{K} \pi_k f_{\mu_k,\Sigma_k}(x)$$

The parameter $\theta = (\pi,\mu,\Sigma)$ *is made of the vector of mixing weights* $\pi = (\pi_1,\dots,\pi_K) \in \Delta_K$*, the mean vectors* $\mu = (\mu_1,\dots,\mu_K)$ *in* $(\mathbb{R}^d)^K$ *and the covariance matrices* $\Sigma = (\Sigma_1,\dots,\Sigma_K) \in (\mathbb{R}^{d\times d})^K$*.*

Figure 1.2 provides an illustration of $n = 1000$ iid samples from a Gaussian mixture in dimension 1 and in dimension 2. As we see, these distributions naturally generate clustered data, in which the means $\mu_1,\dots,\mu_K$ represent the centers of the different clusters. A natural question is therefore, can we estimate these parameters?

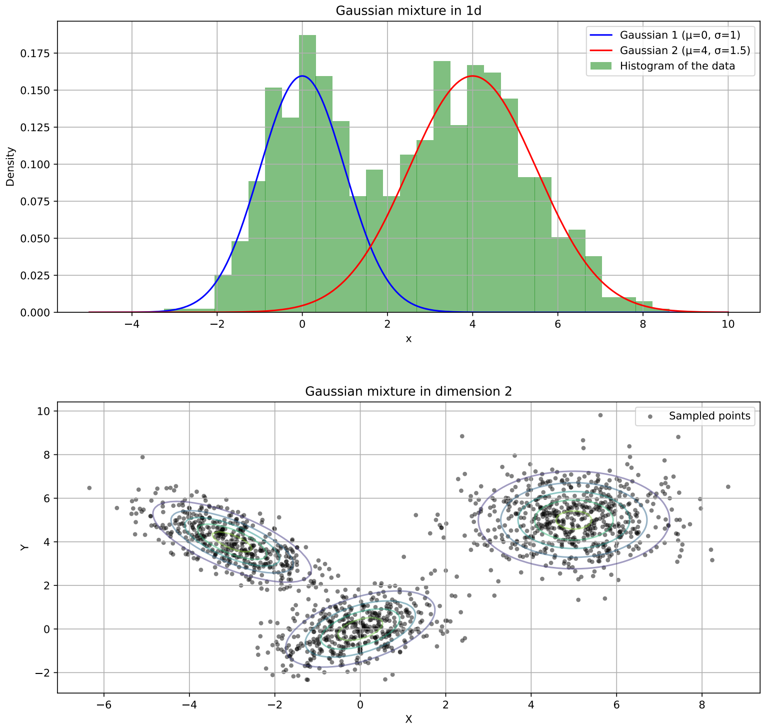

Figure 1.2: Top: histogram of the sampled points under a mixture of two Gaussian in dimension 1, with the densities of the two Gaussian distributions. Bottom: scatterplot of the sampled points under a mixture of 3 Gaussian in dimension 2, with the level sets of the densities of the 3 Gaussian distributions

Let $X_1,\dots,X_n$ be n iid sample from a Gaussian mixtures with K components, with parameters $\pi = (\pi_1,\dots,\pi_K)$, $\mu = (\mu_1,\dots,\mu_K)$ and $\Sigma = (\Sigma_1,\dots,\Sigma_K)$. The log-likelihood is

$$\ell(X;\pi,\mu,\Sigma) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_{\mu_k,\Sigma_k}(X_i)\right)$$

and it cannot be simplified much, unlike in previous examples.

To compute the MLE, this function should be maximized over $\Delta_K \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$, so we have to solve a constrained optimization problem (due to Δ_K), on a rather large space. The gradient of the log of the sum does not have a particularly nice form. Moreover, we note that there are multiple solutions to the MLE, as $\ell(X;\pi,\mu,\Sigma) = \ell(X;\pi_\sigma,\mu_\sigma,\Sigma_\sigma)$ for any permutation $\sigma \in \Sigma_K$ (with the notation $\pi_\sigma = (\pi_{\sigma(1)},\pi_{\sigma(2)},\dots,\pi_{\sigma(K)})$, that can be extended to μ_σ and Σ_σ). Hence, optimization algorithms will struggle to approximate an (exact) Maximum Likelihood Estimator in this model. This is why it is common to use another approach, which we now describe.

1.4.2 The Expectation Maximization (EM) Algorithm

The latent variable interpretation To introduce the EM algorithm, it is helpful to interpret the Gaussian mixture model as a model with some latent (i.e. un-observed) variable. Introducing a random variable Z with support $\{1,\dots,K\}$ such that $\mathbb{P}(Z=k)=\pi_k$, the random variable X specified by its conditional distribution given $(Z=k)$

$$X|(Z=k) \sim \mathcal{N}(\mu_k,\Sigma_k)$$

is a Gaussian mixture with parameters π,μ,Σ. The random variable Z can be viewed as the (hidden) identity of the Gaussian distribution that generated the data X.

Given $(X_i,Z_i)_{1\le i\le n}$ generated from this distribution (but for which only X_i is observed, as before), we can rewrite the above log-likelihood as

$$\ell(X;\pi,\mu,\Sigma) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \mathbb{P}(X_i,Z_i=k;\theta)\right) .$$

This is a generic writing for any mixture model, in which the distribution of X given Z could be arbitrary, and not necessary Gaussian.

A lower bound on the likelihood The idea is to find an (easy to optimize) *lower bound* on the log-likelihood. A first observation is that, for all i, for any distribution $q \in \Delta_K$ we can use the concavity of the logarithm to get the lower bound

$$\log\left(\sum_{k=1}^K \mathbb{P}(X_i,Z_i=k;\theta)\right) \ge \sum_{k=1}^K q_k \log\left(\frac{\mathbb{P}(X_i,Z_i=k;\theta)}{q_k}\right)$$

This trick allow to go from a log of sum to a sum of log. Now let θ^{old} be some particular reference value. If we pick $q_k^{(i)} = \mathbb{P}(Z_i=k|X_i,\theta^{\text{old}})$, the inequality becomes an equality for $\theta=\theta^{\text{old}}$, hence it must be a tight lower bound for θ close to θ^{old}. The lower bound is equal to

$$\mathbb{E}_{Z\sim\mathbb{P}(Z_i|X_i,\theta^{\text{old}})}\left[\log\left(\mathbb{P}(X_i,Z_i=Z;\theta)\right)\right] - \sum_{k=1}^K \mathbb{P}(Z_i=k|X_i,\theta^{\text{old}})\log\left(\mathbb{P}(Z_i=k|X_i,\theta^{\text{old}})\right)$$

Summing these inequalities, we obtain a lower bound on the log-likelihood that is of the form

$$\ell(X;\pi,\mu,\Sigma) \ge \sum_{i=1}^n \mathbb{E}_{Z\sim\mathbb{P}(Z_i|X_i,\theta^{\text{old}})}\left[\log\left(\mathbb{P}(X_i,Z_i=Z;\theta)\right)\right] - G(\theta^{\text{old}})$$

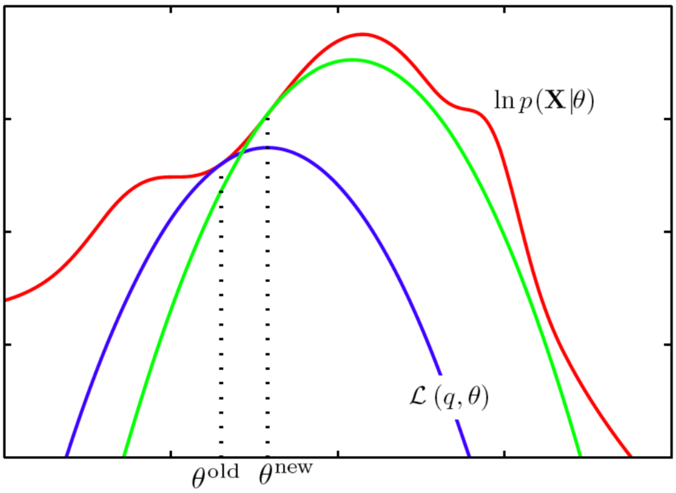

Figure 1.3: Two consecutive steps of the EM algorithm. Credit: some slides by Roger Grosse.

with an additive term $G(\theta^{\text{old}})$ that does not depend on θ.

The idea of the Expectation Maximization algorithm, illustrated above, is to update the current parameter θ^{old} by replacing it by a parameter θ^{new} that maximizes the above lower bound on the log-likelihood. Observe that the lower bound is an equality in $\theta = \theta^{\text{old}}$, so with this approach we are guaranteed to always increase the value of the likelihood. Yet the algorithm could get stuck in a local maximum.

The Expectation Maximization algorithm More formally (and to explain the name) EM is an iterative algorithm, that maintains a sequence of parameters (θ^m) that hopefully get close to a MLE. We start from a random value θ^0. In iteration $m+1$, the algorithm performs

- **The E (Expectation) Step**: compute the expectation term in the lower bound

$$\mathcal{L}_m(\pi,\mu,\Sigma) = \sum_{i=1}^{n} \mathbb{E}_{Z\sim\mathbb{P}(Z_i|X_i,\theta^m)} \left[\log\left(\mathbb{P}(X_i, Z_i = Z;\theta)\right)\right]$$

- **The M (Maximization) Step**: maximize the function $\mathcal{L}_m(\theta)$

$$\theta^{m+1} \in \underset{\pi,\mu,\Sigma}{\text{argmax}}\ \mathcal{L}_m(\pi,\mu,\Sigma)$$

The algorithm is guaranteed to converge to a local maximum of the likelihood. To get closer to a global maximum, it is common to run the algorithm multiple times, with different random initializations, and then to keep the value for which the likelihood is maximal.

EM for Gaussian Mixture Models The above formulation is actually quite general, and could apply to other latent variable models. For the particular case of Gaussian model, we now explain why both the E step and the M step are easy to perform.

We write $\theta^m = (\pi^m, \mu^m, \Sigma^m)$ as the current parameter maintained by EM. For the E-step, we need to compute

$$\gamma_{i,k}^m = \mathbb{P}(Z_i = k|X_i, \theta^m) = \frac{\pi_k^m f_{\mu_k^m,\Sigma_k^m}(X_i)}{\sum_{\ell=1}^K \pi_\ell^m f_{\mu_\ell^m,\Sigma_\ell^m}(X_i)}$$

where the equality follows from Bayes' formula. This quantity represents the probability that data i belongs to "cluster" k (i.e. was generated by the k-th mixture) given the current value of the parameter θ^m. The **E-Step** can be written

$$\begin{aligned}\mathcal{L}_m(\theta) &= \sum_{i=1}^n \sum_{k=1}^K \gamma_{i,k}^m \left(\log(\pi_k) + \log f_{\mu_k,\Sigma_k}(X_i)\right) \\ &= \sum_{k=1}^K \left(\sum_{i=1}^n \gamma_{i,k}^m\right) \log(\pi_k) + \sum_{k=1}^K \sum_{i=1}^n \gamma_{i,k}^m \log f_{\mu_k,\Sigma_k}(X_i)\end{aligned}$$

To perform the **M-Step** this time we can use further simplifications in $\log f_{\mu_k,\Sigma_k}$ (as we now have a sum of logs and not the log of a sum) and compute everything in closed form. For each k, solving for μ_k and Σ_k amounts to computing the MLE of μ and Σ in a multivariate Gaussian model in which each data point is weighted by $\gamma_{i,k}^m$. Letting $n_k^m = \sum_{i=1}^n \gamma_{i,k}^m$, we get

$$\begin{aligned}\mu_k^{m+1} &= \frac{1}{n_k^m} \sum_{i=1}^n \gamma_{i,k}^m X_i \, , \\ \Sigma_k^{m+1} &= \frac{1}{n_k^m} \sum_{i=1}^n \gamma_{i,k}^m (X_i - \mu_k^{m+1})(X_i - \mu_k^{m+1})^\top \, .\end{aligned}$$

For π^{m+1}, we optimize over the simplex to get

$$\pi_k^{m+1} = \frac{n_k^m}{n} \, .$$

EM for Clustering The variables $\gamma_{i,k}^m$ that are computed in the course of the EM algorithm are actually very useful for a clustering purpose: they can be interpreted as probability of the datapoint i to belong to the class k. This is a "soft" clustering assignment.

1.5 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 1.6). 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 1.22. *A (uni-dimensional) parametric 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, $S = \{x \in \mathcal{X} : f_\theta(x) > 0\}$ 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 1.23. *We can show that many classical parametric 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.*

1.5.1 The Fisher information

Definition 1.24. *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 1.25. *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 1.26. *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 $X_1 \sim P_\theta$.

1.5.2 Some properties of the Fisher information

Lemma 1.27. *Under a regular model, it holds that $I^X(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2\ell(X;\theta)}{\partial^2\theta}\right]$.*

Proof. Let us start by computing the right-hand side:

$$\begin{aligned}
\mathbb{E}_\theta\left[\tfrac{\partial^2 \ell(X;\theta)}{\partial^2\theta}\right] &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)}\right)\right] = \mathbb{E}_\theta\left[\frac{\left(\frac{\partial^2}{\partial^2\theta}f_\theta(X)\right)f_\theta(X) - \left(\frac{\partial}{\partial\theta}f_\theta(X)\right)^2}{(f_\theta(X))^2}\right] \\
&= \int_S \frac{\left(\frac{\partial^2}{\partial^2\theta}f_\theta(x)\right)f_\theta(x) - \left(\frac{\partial}{\partial\theta}f_\theta(x)\right)^2}{(f_\theta(x))^2} f_\theta(x)d\nu(x) \\
&= \int_S \left(\frac{\partial^2}{\partial^2\theta}f_\theta(x)\right)d\nu(x) - \int_S \left(\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}\right)^2 f_\theta(x)d\nu(x) \\
&= \frac{\partial^2}{\partial^2\theta}\underbrace{\int_S f_\theta(x)d\nu(x)}_{=1} - \mathbb{E}_\theta\left[(s(X;\theta))^2\right] \\
&= 0 - I^X(\theta)
\end{aligned}$$

which concludes the proof.

□

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 1.28. *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 1.29. *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 1.28, 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 1.27 can be proved:

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

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

The Fisher information will be shortly related to the minimal variance that a unbiased estimator can have. In this section, we give some elements of explanation as to why it can be called “information”.

First, due to its additivity property (Lemma 1.28), 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 1.30. *For any statistic $S = s(X)$ of an observation X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, we have $I^S(\theta) \le I^X(\theta)$.*

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 1.27 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

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

and the term between brackets is positive as it is the Fisher information of the conditional 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 1.31. *A statistic $S = s(X)$ is called sufficient for θ if the distribution of $X = (X_1, \dots, X_n)$ conditionally to S does not depend on θ.*

We admit the following characterization.

Theorem 1.32 (Neyman-Fisher)**.** *The statistic $S = s(X_1, \dots, X_n)$ is sufficient for θ is there exists two positive functions g and h such that the density of X can be written*

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

1.5.4 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 1.33. *For two probability measure P and Q that have a density 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 1.22 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 1.34. *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

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

Proposition 1.35. $\mathrm{KL}(P,Q) \geq 0$ *and* $\mathrm{KL}(P,Q) = 0$ *if and only if* $P = Q$.

1.6 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 1.36. *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 $\mathbb{E}_\theta[\widehat{g}] = g(\theta)$ (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 1.25.

Now we assume that $\mathbb{E}_\theta[h^2(X)] < \infty$ (otherwise, the inequality in Theorem 1.36 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 shows that there exist efficient estimators.

Example 1.37. *Combining Example 1.18 and Example 1.29, we can easily check that in the Bernoulli model $X_1, \ldots X_n \overset{iid}{\sim} \mathcal{B}(\theta)$ the MLE $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is an efficient estimator.*

1.7 Exponential families

Actually, the reason why we can find an efficient estimator in the Bernoulli model 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 1.38. *An exponential family is a set of probability distributions on some set $\mathcal{X}$ defined as*

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

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

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

Example 1.39. *The family of Bernoulli distributions $\mathcal{P} = \{\mathcal{B}(p), p \in (0,1)\}$ form an exponential family (of dimension 1). Indeed, its density with respect to the counting measure is*

$$\begin{aligned} f_p(x) &= p^x(1-p)^{1-x}\mathbb{1}(x \in \{0,1\}) \\ &= \exp(x\log(p) + (1-x)\log(1-p))\mathbb{1}(x \in \{0,1\}) \\ &= h(x)\exp\left(x\log\frac{p}{1-p} + \log(1-p)\right) \end{aligned}$$

with $h(x) = \mathbb{1}(x \in \{0,1\})$. Introducing the natural parameter $\theta = \log\frac{p}{1-p}$, we have $p = \frac{e^\theta}{1+e^\theta}$ and $\log(1-p) = -\log(1+e^\theta)$. Hence we have

$$f_p(x) = h(x)\exp\left(x\theta - b(\theta)\right)$$

with $b(\theta) = \log(1+e^\theta)$ and the family of Bernoulli distributions can be written as the family of densities

$$\{f_\theta(x) = h(x)\exp(a(\theta)T(x) - b(\theta)), \theta \in \mathbb{R}\}$$

where $a(\theta) = \theta$, $T(x) = x$ and $b(\theta) = \log(1+e^\theta)$ and the reference measure is the counting measure.

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 $X = (X_1, \ldots, X_n) \overset{iid}{\sim} P_\theta$. For each n, given an estimator $\widehat{g}_n = h(X_1, \ldots, X_n)$ 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 2.1. *Let $Z_1, Z_2, \ldots$ 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 $z = (z_1, \ldots, z_d)$ in which F is continuous,

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

Example 2.2. $Z_n \sim \mathcal{N}(0, \frac{1}{n})$. *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 2.3. *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 2.4 (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 2.5 (Slutsky lemma)**.** *If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{P} c$ 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 applies 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 2.6 (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 2.7 (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 2.8. *An estimator $\widehat{g}_n$ of $g(\theta)$ is consistent if for every $\theta \in \Theta$, $\widehat{g}_n \xrightarrow{P} g(\theta)$.*

When we further have an almost sure convergence instead of the convergence in probability, that is when $\widehat{g}_n \xrightarrow{a.s.} g(\theta)$, we shall say that $\widehat{g}_n$ is strongly consistent.

Proving consistency Consistency of estimators often follows directly from the law of large numbers (for estimators that are empirical averages). Lemma 2.3 and Lemma 2.4 also yield the following:

- 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 2.9. *Using the law of large number directly yields that the empirical mean $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ of any n sample with a finite expectation μ is a consistent estimator of μ. As for the empirical variance, which can be written*

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

the law of large numbers also gives the convergence almost surely to $\mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2$, the variance of the distribution, provided that X_1 has a second moment. Hence the consistency.

Given a consistent estimator $\widehat{g}_n$, we are now 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. We hope to prove a statement of the form

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

where $w(n)$ is some sequence that tends to infinity giving the convergence speed, and Z is some random variable. Equivalently we have

$$\widehat{g}_n = g(\theta) + \frac{1}{w(n)}Z_n \quad \text{where} \quad Z_n \rightsquigarrow Z$$

and we write

$$\widehat{g}_n \approx g(\theta) + \frac{1}{w(n)} Z \, .$$

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

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

Here, the limit distribution Z is Gaussian, and the convergence speed is $w(n) = \sqrt{n}$. 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 2.10. *An estimator is $\widehat{g}_n$ of $g(\theta)$ is asymptotically normal is it satisfies, for all $\theta \in \Theta$,*

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

where σ_θ^2 is called the asymptotic variance.

Proving asymptotic normality The first tool to prove asymptotic normality is obviously the CLT itself, as in the example below. The Slutsky lemma may also be of use when the asymptotic variance is further estimated (we will see some examples later). Finally, the counterpart of the continuous mapping lemma, which gives the asymptotic distribution of some continuous transformation of an estimator will be described in the next section.

Example 2.11. *Given a n sample $(X_1, \ldots, X_n)$ from some distribution with cdf F, we consider the value of the empirical cdf in x as an estimator for $F(x)$ (see Example 1.9). We have*

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

This estimator is the empirical mean of the random variables $Z_i = \mathbb{1}(X_i \leq x)$ whose mean is $F(x)$ and whose variance is $F(x)(1 - F(x))$. Using the central limit theorem, we can obtain the asymptotic distribution of $\widehat{F}_n(x)$:

$$\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) \rightsquigarrow \mathcal{N}\left(0, F(x)(1 - F(x))\right)$$

The mean of the estimator is $F(x)$ and it is asymptotically normal with variance $F(x)(1 - F(x))$.

It is worth mentioning that there exists estimators that are have a limiting distribution, but are not asymptotically normal. We provide an example below where $w(n) = n$ (hence a faster convergence speed than for asymptotically normal estimators) and Z is some exponential distribution.

Example 2.12. *As studied in exercise, in the model $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}([0, \theta])$, 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, 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. We illustrate below the asymptotic distribution of $\widehat{\theta}_n$ and $\widehat{\theta}_n$.

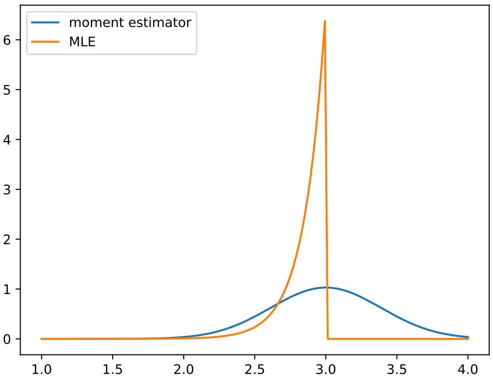

Figure 2.1: Densities of the asymptotic distribution of the moment estimator and the MLE in the uniform model for $\theta = 3$ and $n = 20$.

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.

Why asymptotic normality? Knowing the exact distribution of an estimator would be very useful to derive (exact) confidence region or tests with guaranteed type I error. Knowing it asymptotically allows to derive asymptotic confidence regions, and the corresponding tests, as we will recall at the beginning of the next chapter.

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 $\sigma_\theta^2 \le \widetilde{\sigma}_\theta^2$ 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 Delta method

We now present a useful tool to compute asymptotic distributions of some (continuous) transformation of an asymptotically normal estimator: the so-called Delta 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 2.13. *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

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

In other words,

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

Proof. The proof follows from using a Taylor expansion around μ. As g is differentiable, we have that for all n, there exists μ_n in (Z_n, μ) (if $Z_n < \mu$) or in (μ, Z_n) (if $Z_n \ge \mu$) such that

$$g(Z_n) = g(\mu) + g'(\mu_n)(Z_n - \mu)$$

hence

$$\sqrt{n}|g(Z_n) - g(\mu)| = g'(\mu_n)\sqrt{n}(Z_n - \mu) .$$

As $|\mu_n - \mu| \le |Z_n - \mu|$ and $Z_n \xrightarrow{P} \mu$, we have that $\mu_n \xrightarrow{P} \mu$. If we assume g' to be continuous[1], it follows from Lemma 2.4 that $g'(\mu_n) \xrightarrow{P} g'(\mu)$.

By assumption, we also have that $\sqrt{n}(Z_n - \mu) \rightsquigarrow Z$ where $Z \sim \mathcal{N}(0,1)$. It follows from Slutsky's lemma (Lemma 2.5) that

$$\sqrt{n}|g(Z_n) - g(\mu)| \rightsquigarrow g'(\mu)Z$$

whose distribution is $\mathcal{N}(0, (g'(\mu))^2\sigma^2)$.

□

There exists also a multi-variate version of the Delta method, stated below.

[1] A slightly more complicated proof can also be given when g is not continuous, see e.g. [Rivoirard and Stoltz, 2009]

Theorem 2.14. *Let $Z_n = (Z_{n,1}, \dots, Z_{n,d})$ be a sequence of random vectors in $\mathbb{R}^d$ such that*

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

where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and let $\nabla g(z) = (\frac{\partial g}{\partial z_1}(z), \dots, \frac{\partial g}{\partial z_d}(z))^\top$ be its gradient. If all the components of $\nabla g(\mu)$ are non-zero, then

$$\sqrt{n}\,(g(Z_n) - g(\mu)) \rightsquigarrow \mathcal{N}\left(0, (\nabla g(\mu))^\top \Sigma\, (\nabla g(\mu))\right)$$

Example 2.15. *In a clinical trials involving two treatments, we observe the outcome of treatment 1 (a placebo) on a pool of n patients. For $i \in \{1, \dots, n\}$, we record $X_i = 1$ if the treatment is a success for patient i, $X_i = 0$ is it is a failure. The outcome of treatment 2 (the new drug) is observed on another pool of n patients, with $Y_j \in \{0, 1\}$ indicating success of failure for patient $j \in \{1, \dots n\}$. We assume that all the X_i and Y_j are independent and that $X_i \sim \mathcal{B}(p_1)$ and $Y_j \sim \mathcal{B}(p_2)$ where p_1 and p_2 are the probability of efficacy of treatment 1 and 2, respectively. We are interested in estimating the treatment effect $\phi := p_2 - p_1$.*

First, we can derive the MLE estimator of the parameter $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \in \mathbb{R}^2$. The likelihood is

$$\ell(X_1, \dots, X_n, Y_1, \dots, Y_n; p_1, p_2) = p_1^{\sum_{i=1}^n X_i}(1 - p_1)^{n - \sum_{i=1}^n X_i} p_2^{\sum_{i=1}^n Y_i}(1 - p_2)^{n - \sum_{i=1}^n Y_i}$$

and it is maximized for $\widehat{p}_n = (\widehat{p}_{1,n}, \widehat{p}_{2,n})^\top$ where $\widehat{p}_{1,n} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\widehat{p}_{2,n} = \frac{1}{n}\sum_{i=1}^n Y_i$. Each of the estimator $\widehat{p}_{i,n}$ for $i = 1, 2$ is an unbiased estimator of p_i and the central limit theorem tells us that $\sqrt{n}(\widehat{p}_{i,n} - p_i) \rightsquigarrow \mathcal{N}(0, p_i(1 - p_i))$. As the estimators $\widehat{p}_{1,n}$ and $\widehat{p}_{2,n}$ are independent, we get that

$$\sqrt{n}\,(\widehat{p}_n - p) \sim \mathcal{N}\,(0, \Sigma) \quad \text{with } \Sigma = \begin{pmatrix} p_1(1 - p_1) & 0 \\ 0 & p_2(1 - p_2) \end{pmatrix}$$

As a treatment effect estimator, we propose $\widehat{\phi}_n = \widehat{p}_{2,n} - \widehat{p}_{1,n}$. It can be written $\widehat{\phi}_n = g(\widehat{p}_n)$ here $g : \mathbb{R}^2 \to \mathbb{R}$ is a simple linear function $g(p_1, p_2) = p_2 - p_1$ whose gradient is $\nabla g(p) = (-1, 1)^\top$. Using the multivariate Delta method, we get that

$$\sqrt{n}\left(\widehat{\phi}_n - \phi\right) \rightsquigarrow \mathcal{N}\left(0, p_1(1 - p_1) + p_2(1 - p_2)\right).$$

2.4 Asymptotic properties of the Maximum Likelihood Estimator

Given an iid sample $X_1, \dots, X_n \sim P_\theta$, we recall that the maximum likelihood estimator of the parameter θ is defined by

$$\widehat{\theta}_n \in \underset{\theta\in\Theta}{\text{argmax}}\; L(X_1, \dots, X_n; \theta).$$

Despite its implicit definition, as the maximizer of some function, we will see that this estimator enjoys strong asymptotic performance guarantees, when the model satisfies some assumptions. In particular, we will assume that the model is identifiable, that is

$$\forall(\theta, \theta') \in \Theta, P_\theta = P_{\theta'} \quad \text{if and only if} \quad \theta = \theta'.$$

2.4.1 Rationale

First, let us try to understand why the MLE is a good estimator. Let us denote by $\theta_\star$ the true parameter from which the data is generated. The maximum likelihood can be rewritten as follows, introducing artificially the likelihood under $\theta_\star$. Indeed, one can write

$$\begin{aligned}
\widehat{\theta}_n &\in \operatorname*{argmax}_{\theta\in\Theta} \frac{f_\theta(X_1)\dots f_\theta(X_n)}{f_{\theta_\star}(X_1)\dots f_{\theta_\star}(X_n)} \\
\widehat{\theta}_n &\in \operatorname*{argmin}_{\theta\in\Theta} \frac{f_{\theta_\star}(X_1)\dots f_{\theta_\star}(X_n)}{f_\theta(X_1)\dots f_\theta(X_n)} \\
\widehat{\theta}_n &\in \operatorname*{argmin}_{\theta\in\Theta} \log \frac{f_{\theta_\star}(X_1)\dots f_{\theta_\star}(X_n)}{f_\theta(X_1)\dots f_\theta(X_n)} \\
\widehat{\theta}_n &\in \operatorname*{argmin}_{\theta\in\Theta} \sum_{i=1}^n \log\left(\frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)}\right) \\
\widehat{\theta}_n &\in \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)}\right)
\end{aligned}$$

Hence, $\widehat{\theta}_n$ can be rewritten as the minimizer of some empirical average. Introducing the notation

$$M_n(\theta,\theta_\star) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_{\theta_\star}(X_i)}{f_\theta(X_i)}\right)$$

we know by the law of large number that, for all $\theta \in \Theta$,

$$M_n(\theta,\theta_\star) \xrightarrow{P} \mathbb{E}_{\theta_\star}\left[\log\left(\frac{f_{\theta_\star}(X_1)}{f_\theta(X_1)}\right)\right] = \mathrm{KL}\left(P_{\theta_\star}, P_\theta\right) \tag{2.1}$$

where $\mathrm{KL}(P,P')$ is the KL divergence between distributions, introduced in Definition 1.33. The KL divergence is not a distance but it still satisfies the following important property: $\mathrm{KL}(P,P') = 0$ if an only if $P = P'$. In particular, $\mathrm{KL}(P_{\theta_\star}, P_\theta) = 0$ if and only if $P_{\theta_\star} = P_\theta$, i.e. $\theta = \theta_\star$ as the model is identifiable. Thus we have

$$\operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}\left(P_{\theta_\star}, P_\theta\right) = \theta_\star.$$

Hence, our hope is to prove that, under the model $\mathbb{P}_{\theta_\star}$,

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta,\theta_\star) \xrightarrow{P} \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}\left(P_{\theta_\star}, P_\theta\right) = \theta_\star$$

This will require slightly more sophisticated arguments than the convergence of the objective function to minimize given in (2.1). We present them in the next section for more general M-estimators, that are also expressed as minimizer of empirical averages.

2.4.2 Consistency of M-estimators

A M-estimator is any estimator defined as a minimizer of some empirical average:

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta) \quad \text{with} \quad M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$$

Letting $M(\theta) = \mathbb{E}[m_\theta(X_1)]$, if this expectation is finite we have $M_n(\theta) \xrightarrow{P} M(\theta)$ for all $\theta \in \Theta$, and we hope that $\widehat{\theta}_n$, a minimizer of $M_n(\theta)$, converges to $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} M(\theta)$.

Example 2.16. *In supervised learning, we observe iid pairs of the form (X_i, Y_i) coming from some unknown distribution P on $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the feature space, often $\mathbb{R}^d$ and label space which is either finite (classification) or continuous (regression). The goal is to produce a predictor $\widehat{f}_n : \mathcal{X} \to \mathcal{Y}$, which is a data-dependent function mapping the feature to the label. Due to the generic empirical risk minimization principle, many predictor can be expressed as M-estimators.*

Given a class of function $\mathcal{F}$, and some loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, we can define

$$\widehat{f}_n \in \underset{f\in\mathcal{F}}{\mathit{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} L(f(X_i), Y_i).$$

In this general (non-parametric) setting, the "parameter" is a function f (possible predictor), and we have $m_f((X_i, Y_i)) = L(f(X_i), Y_i)$. We hope that $\widehat{f}_n$ converges to $f_0 \in \underset{f\in\mathcal{F}}{\mathit{argmin}}\, M(f)$ where $M(f) = \mathbb{E}_{(X,Y)\sim P}[L(f(X), Y)]$, that is to a predictor that minimizes the risk associated to the loss function L.

Sometimes, the class of function $\mathcal{F}$ can be described by a small set of parameters (e.g. a set of linear functions) and the regressor obtained by ridge regression can be defined as $\widehat{f}_n(x) = \widehat{\theta}_n^\top x$ for $x \in \mathbb{R}^d$ where

$$\widehat{\theta}_n = \underset{\theta\in\mathbb{R}^d : \|\theta\|_2 \le C}{\mathit{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} (Y_i - \theta^\top X_i)^2 .$$

In this case we hope that $\widehat{\theta}_n$ is close to $\theta_0 \in \underset{\theta\in\mathbb{R}^d, \|\theta\|\le C}{\mathit{argmin}}\ \mathbb{E}_{(X,Y)\sim P}[(Y - \theta^\top X)^2]$.

To establish the convergence of $\widehat{\theta}_n$ to θ_0, we need two properties. The first one is a property of the minimizer θ_0, which has to be a strict local minima, and the second is about the convergence from $M_n(\theta)$ to $M(\theta)$, which needs to be uniform.

Theorem 2.17. *Let $\widehat{\theta}_n = \underset{\theta\in\Theta}{\mathit{argmin}}\, M_n(\theta)$ and $\theta_0 = \underset{\theta\in\Theta}{\mathit{argmin}}\, M(\theta)$. For $\Theta \subseteq \mathbb{R}^d$, let d a distance on $\mathbb{R}^d$. Assume that the following two properties hold:*

1. *For all $\varepsilon > 0$, $\sup_{d(\theta,\theta_0)\ge\varepsilon} M(\theta) > M(\theta_0)$.*

2. $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$.

Then $\widehat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. From assumption 1., for every $\varepsilon > 0$, there exists η_ε such $d(\theta, \theta_0) \ge \varepsilon$ implies that $M(\theta) \ge M(\theta_0) + \eta_\varepsilon$. One can write

$$\begin{aligned}
\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \ge \varepsilon\right) &\le \mathbb{P}\left(M(\widehat{\theta}_n) \ge M(\theta_0) + \eta_\varepsilon\right)\\
&= \mathbb{P}\left(\eta_\varepsilon \le M(\widehat{\theta}_n) - M(\theta_0)\right)\\
&= \mathbb{P}\left(\eta_\varepsilon \le M(\widehat{\theta}_n) - M_n(\widehat{\theta}_n) + M_n(\widehat{\theta}_n) - M_n(\theta_0) + M_n(\theta_0) - M(\theta_0)\right)
\end{aligned}$$

As $\widehat{\theta}_n$ is a minimizer of M_n, we have $M_n(\widehat{\theta}_n) - M_n(\theta_0) \leq 0$ and

$$\begin{aligned}\mathbb{P}\left(d(\widehat{\theta}_n, \theta_0) \geq \varepsilon\right) &\leq \mathbb{P}\left(\eta_\varepsilon \leq M(\widehat{\theta}_n) - M_n(\widehat{\theta}_n) + M_n(\theta_0) - M(\theta_0)\right) \\ &\leq \mathbb{P}\left(\eta_\varepsilon \leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|\right) \\ &= \mathbb{P}\left(\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \frac{\eta_\varepsilon}{2}\right),\end{aligned}$$

and the right-hand side tends to zero by assumption 2., which concludes the proof.

□

Remark 2.18. *Consistency also holds if $\widehat{\theta}_n$ is not an exact minimizer of $M_n(\theta)$ (which can be hard to compute in some practical cases), as long as its approximation error converges to zero (in probability). A sufficient condition to obtain consistency for an approximate minimizer is to further assume that*

$$M_n(\widehat{\theta}_n) \leq M_n(\theta_0) + E_n$$

for some random variable $E_n \xrightarrow{P} 0$.

Application to the MLE estimator Using Theorem 2.17, we can propose some sufficient condition for the MLE to be a consistent estimator of $\theta_\star$ when $(X_1, \ldots, X_n) \sim P_{\theta_\star}$.

Theorem 2.19 (Consistency of the MLE)**.** *Assume that the model $\mathcal{M} = \{f_\theta, \theta \in \Theta\}$ satisfies the following properties:*

1. *$\mathcal{M}$ is identifiable, i.e., $f_\theta = f_{\theta'}$ implies $\theta = \theta'$ for all $(\theta, \theta') \in \Theta$.*
2. *Θ is compact and for all $x \in \mathcal{X}$, $\theta \mapsto f_\theta(x)$ is continuous.*
3. *For all $\theta \in \Theta$, $\mathbb{E}_\theta\left[\sup_{\theta' \in \Theta} |\log f_{\theta'}(X_1)|\right] < \infty$.*

Then for all $\theta_\star \in \Theta$, the MLE estimator built from a n sample $X_1, \ldots, X_n \sim f_{\theta_\star}$ satisfies $\widehat{\theta}_n \xrightarrow{P} \theta_\star$ (where the convergence is under the model $\mathbb{P}_{\theta_\star}$).

For theses assumption to be satisfied in simple models such as Bernoulli and Gaussian, we would need to restrict the set of possible values for the means (to $[p_0, 1 - p_0]$ for $p_0 > 0$ in the Bernoulli case, or to some bounded interval $[a, b]$ in the Gaussian case). But in these two cases, the consistency of the MLE (which coincides with the empirical means) can already easily be established directly using the law of large number. Still a result such as Theorem 2.19 provide some generic guarantees for the MLE in potentially more complex models, under some restriction on the parameter space.

2.4.3 Asymptotic normality of the MLE estimator

Under stronger assumptions, it is also possible to further exhibit the limiting distribution of the MLE estimator. We start by presenting the result and a sketch of proof for the estimation of a one-dimensional parameter $\theta \in \mathbb{R}$. Given a n sample $X_1, \ldots, X_n \sim P_\theta$, we recall that $I(\theta)$ denotes the Fisher information obtained from one sample X_1.

Theorem 2.20. *Let $\widehat{\theta}_n$ be the MLE of a parameter $\theta \in \mathbb{R}$ computed on a n sample $X_1, \dots, X_n \sim P_\theta$. If $\widehat{\theta}_n$ is consistent and if the model is regular (according to Definition 1.22) then if the Fisher information satisfies $I(\theta) > 0$, $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in distribution under $\mathbb{P}_\theta$ towards a Gaussian distribution:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{I(\theta)}\right).$$

Proof. Let $\ell(\theta) = \log f_\theta(X_1, \dots, X_n)$ be a simplified notation for the log-likelihood, and let $\ell'(\theta)$ be its derivative (in θ). As a minimizer of ℓ, the MLE estimator satisfies $\ell'(\widehat{\theta}_n) = 0$. Using a Taylor expansion of ℓ' in the true parameter θ, we can write

$$0 = \ell'(\widehat{\theta}_n) = \ell'(\theta) + (\widehat{\theta}_n - \theta)\ell''(\widetilde{\theta}_n)$$

for some $\widetilde{\theta}_n \in (\theta, \widehat{\theta}_n)$ (or $(\widehat{\theta}_n, \theta)$). Hence, one can write

$$\begin{aligned} \widehat{\theta}_n - \theta &= -\frac{\ell'(\theta)}{\ell''(\widetilde{\theta}_n)} \\ \sqrt{n}\left(\widehat{\theta}_n - \theta\right) &= \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\widetilde{\theta}_n)} \end{aligned} \tag{2.2}$$

The numerator in (2.2) can be written

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n s(X_i, \theta)\right).$$

Using that under a regular model $\mathbb{E}_\theta\left[s(X_1;\theta)\right] = 0$ and $\mathrm{Var}_\theta\left[s(X_1;\theta)\right] = I(\theta)$, one gets using the Central Limit Theorem that

$$\frac{1}{\sqrt{n}}\ell'(\theta) \rightsquigarrow \mathcal{N}(0, I(\theta))$$

From the consistency of $\widehat{\theta}_n$ we know that $\widehat{\theta}_n \xrightarrow{P} \theta$, from which we deduce that $\widetilde{\theta}_n \xrightarrow{P} \theta$. The denominator of (2.2) can be written (this is the part where the proof becomes approximately correct)

$$-\frac{1}{n}\ell''(\widetilde{\theta}_n) = \frac{1}{n}\sum_{i=1}^n -\left(\frac{\partial^2 \log f_\theta(X_i)}{\partial^2\theta}\right)_{\widetilde{\theta}_n} \simeq \frac{1}{n}\sum_{i=1}^n -\left(\frac{\partial^2 \log f_\theta(X_i)}{\partial^2\theta}\right)_{\theta}$$

By the law of large numbers, under the model $\mathbb{P}_\theta$, this empirical average converges in probability to $\mathbb{E}_\theta\left[-\frac{\partial^2 \log f_\theta(X_1)}{\partial^2\theta}\right]$ which is equal to the Fisher information $I(\theta)$ (using Lemma 1.27).

As $I(\theta) \neq 0$, we can use Slutsky's lemma to get that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \sim \frac{1}{I(\theta)}\mathcal{N}(0, I(\theta)) = \mathcal{N}\left(0, \frac{1}{I(\theta)}\right).$$

□

A useful consequence of Theorem 2.17 is that it allows us to build asymptotic confidence regions around the MLE estimator, by replacing the (unknown) quantity $I(\theta)$ by its empirical version $I(\widehat{\theta}_n)$.

Corollary 2.21. *Under the assumptions of Theorem 2.20, if $\theta \mapsto I(\theta)$ is continuous in θ then under the model $\mathbb{P}_\theta$,*

$$\sqrt{nI(\widehat{\theta}_n)}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0,1)\,.$$

Proof. Using the continuous mapping lemma and the consistency of $\widehat{\theta}_n$ yields that, under $\mathbb{P}_\theta$, $I(\widehat{\theta}_n) \xrightarrow{P} I(\theta)$. From Theorem 2.20 we have that under $\mathbb{P}_\theta$, $\sqrt{nI(\theta)}\left(\widehat{\theta}_n - \theta\right) \rightsquigarrow \mathcal{N}(0,1)$. Using Slutsky's lemma,

$$\sqrt{nI(\widehat{\theta}_n)}\left(\widehat{\theta}_n - \theta\right) = \sqrt{\frac{I(\widehat{\theta}_n)}{I(\theta)}} \times \sqrt{nI(\theta)}\left(\widehat{\theta}_n - \theta\right)$$

converges in distribution to $\mathcal{N}(0,1)$.

□

Hence, if our model is regular enough, we can build asymptotic confidence intervals of level $1-\alpha$ around the MLE estimator (and resulting tests, see the next chapter) that are of the form

$$\left[\widehat{\theta}_n - \sqrt{\frac{1}{nI(\widehat{\theta}_n)}}q_{1-\alpha/2} ; \widehat{\theta}_n - \sqrt{\frac{1}{nI(\widehat{\theta}_n)}}q_{1-\alpha/2}\right]$$

where $q_{1-\alpha}$ is such that $\mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(Z \le q_{1-\alpha}\right) = 1-\alpha$. We provide an example below.

Example 2.22. *Consider the Poisson model $X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{P}(\lambda)$ where we recall that*

$$f_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

for all $k \in \mathbb{N}$. The MLE is given by $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and the Fisher information (of one sample) satisfies

$$I(\lambda) = \mathbb{E}_\lambda\left[-\frac{\partial^2}{\partial^2\lambda}\log f_\lambda(X_1)\right]$$

We have

$$\begin{aligned}\frac{\partial \log f_\lambda(X_1)}{\partial\lambda} &= \frac{X_1}{\lambda} - 1\\ \frac{\partial \log f_\lambda(X_1)}{\partial\lambda} &= -\frac{X_1}{\lambda^2}\end{aligned}$$

hence $I(\lambda) = \mathbb{E}_\lambda\left[\frac{X_1}{\lambda^2}\right] = \frac{1}{\lambda}$ as the mean of a Poisson distribution with parameter λ is λ. Applying Corollary 2.21 yields that, under $\mathbb{P}_\lambda$,

$$\sqrt{\frac{n}{\widehat{\lambda}_n}}\left(\widehat{\lambda_n} - \lambda\right) \rightsquigarrow \mathcal{N}(0,1)\,.$$

Now, let's use this information to build an asymptotic confidence interval on λ. We have that

$$\begin{array}{rcl}\mathbb{P}_\lambda\left(-q_{1-\alpha/2} \leq \sqrt{\frac{n}{\widehat{\lambda}_n}}\left(\widehat{\lambda_n}-\lambda\right) \leq q_{1-\alpha/2}\right) & \underset{n\to\infty}{\longrightarrow} & \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(-q_{1-\alpha/2} \leq Z \leq q_{1-\alpha/2}\right) \\ & = & \mathbb{P}(Z \leq q_{1-\alpha/2}) - \mathbb{P}\left(Z \leq -q_{1-\alpha/2}\right) \\ & = & \mathbb{P}(Z \leq q_{1-\alpha/2}) - \mathbb{P}\left(Z > q_{1-\alpha/2}\right) \\ & = & 1-\frac{\alpha}{2}-\frac{\alpha}{2} \\ & = & 1-\alpha\end{array}$$

Putting λ in the center of the interval, we have

$$\mathbb{P}_\lambda\left(\widehat{\lambda}_n - \sqrt{\frac{\widehat{\lambda}_n}{n}}q_{1-\alpha/2} \leq \lambda \leq \widehat{\lambda}_n + \sqrt{\frac{\widehat{\lambda}_n}{n}}q_{1-\alpha/2}\right) \underset{n\to\infty}{\longrightarrow} 1-\alpha$$

which provides an asymptotic confidence interval of level $1-\alpha$.

Extensions of Theorem 2.20 First, this result is also true for the estimation of a multi-dimensional parameter $\theta \in \mathbb{R}^d$ using the MLE. Under similar assumptions, we obtain that under $\mathbb{P}_\theta$,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}\left(0, I(\theta)^{-1}\right)$$

but this time the Fisher information $I(\theta)$ is a $d \times d$ matrix, assumed to be invertible (see its definition in Section 1.5.2).

Then, while we presented the consistency results for the general family of M-estimators, we sticked to the MLE case for the asymptotic normality result. A counterpart of Theorem 2.20 also exists for M-estimators, see e.g. the book [Van der Vaart, 1998].

2.5 Asymptotic efficiency

In light of the Cramer-Rao lower bound given in Chapter 1, any estimator of a parameter $g(\theta) \in \mathbb{R}$ whose limit distribution satisfies

$$\widehat{g}_n \approx \mathcal{N}\left(g(\theta), \frac{(g'(\theta))^2}{nI(\theta)}\right)$$

is called *asymptotically efficient*. The reason is that, asymptotically, is is unbiased with a variance that is the minimal variance prescribed by the Cramer-Rao lower bound.

For estimating the parameter θ, (under appropriate regularity conditions) the MLE is an example of asymptotically efficient estimator, as we just saw that it satisfies

$$\widehat{g}_n \approx \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right).$$

However, we can find examples of MLE that are not efficient. Take for instance the MLE estimator of the variance of the Gaussian distribution, which is biased. We shall see other examples in exercises.

Chapter 3

(Optimal) Testing

We first give some reminder about the general formalism of (parametric) statistical test. Then we discuss two testing methodologies: one directly inspired by the previous chapter, based on an estimator of the parameter that the test is about. And then a more general method building again on the notion of likelihood: likelihood ratio tests.

3.1 Statistical tests

Given a parametric model $X \sim P_\theta$, where $\theta \in \Theta$, a statistical test is an answer about some question about the unknown parameter θ of the form: does θ belong to a certain subset $\Theta_0 \subset \Theta$? Given two disjoint subsets Θ_0 and Θ_1 (that do not have to form a partition of Θ), a testing problem is characterized by two *hypotheses*:

$$\mathcal{H}_0 : (\theta \in \Theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \in \Theta_1).$$

When Θ_i is reduced to a singleton, i.e. $\mathcal{H}_i = (\theta = \theta_0)$, the hypothesis to test is called simple, otherwise it is called composite. $\mathcal{H}_0$ is called the null hypothesis and $\mathcal{H}_1$ the alternative hypothese. As we shall see, they play different roles.

A test can be formalized as a decision function $D : \mathcal{X} \to \{0,1\}$, where $(D(X) = 1)$ means that given the observed data $X \in \mathcal{X}$ we tend to *reject* $\mathcal{H}_0$ and claim that $\mathcal{H}_1$ holds instead and $(D(X) = 0)$ means that we *do not reject* $\mathcal{H}_0$. Hence $(D(X) = 0)$ is not to be interpreted as a decision that $\mathcal{H}_0$ is true, rather that the observation X is still compatible with $\mathcal{H}_0$ being true.

The statistical guarantees that we can offer for a test are expressed in terms of type I error and type II error (or power).

- The type I error, denoted by α is often the primary concern of the statistician. It is defined for $\theta \in \Theta_0$ as
$$\alpha(\theta) = \mathbb{P}_\theta\left(D(X) = 1\right) .$$
A small type I error means that the probability to reject $\mathcal{H}_0$ when it is actually true is small.

- The type II error at a given alternative $\theta \in \Theta_1$ is denoted by
$$\beta(\theta) = \mathbb{P}_\theta\left(D(X) = 0\right)$$
and measure the probability to not detect that $\mathcal{H}_1$ in this alternative. Conversly, the *power* (in the alternative $\theta \in \mathcal{H}_1$) is the probability to detect that $\mathcal{H}_1$ holds, i.e. $1 - \beta(\theta)$.

In a statistical testing problem, we usually require to control the type I error for any possible $\theta \in \Theta_0$ and define the level of significance of the test to be $\alpha = \sup_{\theta\in\Theta_0} \alpha(\theta)$. Then, subject to this type I error control, we seek to provide guarantees on the type II error, or on the power, at least for θ in some part of the alternative.

Asymptotic test In the (common) particular case where $X = (X_1, \dots, X_n) \overset{iid}{\sim} P_\theta$ is a n samples, we can also consider the asymptotic quality of a test, when the sample size n goes large. In that case, indexing the test by its sample size, that is writing D_n instead of D, we define an asymptotic test of level α to be such that, for all $\theta \in \Theta_0$,

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(D_n(X_1, \dots, X_n) = 1\right) \leq \alpha.$$

Similarly, one can look at the asymptotic type II error or power of a test.

3.2 The testing methodology

Tests are often based on some *test statistic*, denoted by $T = t(X) \in \mathbb{R}$ (T_n in the n-sample case) and reject the null hypothesis when T belongs to some rejection region $\mathcal{R} \subseteq \mathbb{R}$ (which typically depends on the desired significance level α we seek). We write

$$D(X) = \mathbb{1}\left(T \in \mathcal{R}\right) .$$

The test statistic is chosen so that its distribution (or its asymptotic distribution) under the null hypothesis (that is, for all $\theta \in \Theta_0$) is known. This knowledge is used to find a rejection region $\mathcal{R}_\alpha$ such that

$$\forall \theta \in \Theta_0, \quad \mathbb{P}_\theta \left(T \in \mathcal{R}_\alpha\right) \leq \alpha$$

and our test at level α is $D_\alpha(X) = \mathbb{1}\left(T \in \mathcal{R}_\alpha\right)$.

For power computation, it can be useful to also know the distribution of the test statistic under the alternative. In the n-sample setting, the power will typically increase with the sample size, and it is common to select the sample size so as to guarantee a minimal power in some part of the alternative.

How to pick a test statistic? In many tests the null hypothesis can be described as "is a parameter of interest inside some region?".

In that case, a natural approach is to rely on some estimator of this parameter, and to define T_n as some function of this parameter whose distribution, or asymptotic distribution, is known under the null. We start with the simplest possible example, in which the distribution of the estimator is known exactly.

Example 3.1. *$X_1, \dots, X_n$ is a n sample from a $\mathcal{N}(\mu, 1)$ distribution and we are testing the hypothesis*

$$\mathcal{H}_0 : (\mu = \mu_0) \quad \textit{against} \quad \mathcal{H}_1 : (\mu = \mu_1)$$

for two distinct values $\mu_0 < \mu_1$.

The test is about the parameter μ so it is quite natural to base our test on an estimator of this parameter: we take the maximum likelihood estimator, which is the empirical mean $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$. In this simple Gaussian setting, $\widehat{\mu}_n$ also follows a Gaussian distribution: $\widehat{\mu}_n \sim \mathcal{N}\left(\mu_0, \frac{1}{n}\right)$. A natural test statistic is

$$T_n = \sqrt{n}\left(\widehat{\mu}_n - \mu_0\right)$$

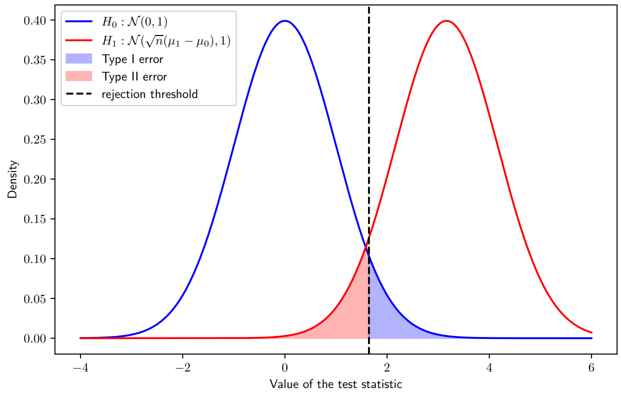

Figure 3.1: Distribution of the test statistic under $\mathcal{H}_0$ and $\mathcal{H}_1$ for $n = 10$, $\mu_0 = 0$, $\mu_1 = 1$.

as under $\mathcal{H}_0$, $T_n \sim \mathcal{N}(0,1)$. Note that we also know the distribution of T_n under $\mathcal{H}_1$. Indeed by writing $T_n = \sqrt{n}(\mu_1 - \mu_0) + \sqrt{n}\,(\widehat{\mu}_n - \mu_1)$ we have that under $\mathcal{H}_1$, $T_n \sim \mathcal{N}\left(\sqrt{n}(\mu_1 - \mu_0), 1\right)$.

Given these observations, we expect T_n to take values close to zero when $\mathcal{H}_0$ is true, and very large values when $\mathcal{H}_1$ is true. This motivates a simple rejection region $\mathcal{R} = [t, +\infty[$ for some threshold t, hence the resulting test is $D_n(X) = \mathbb{1}(T_n > t)$. Now we select the threshold in order to guarantee the desired type I error:

$$\mathbb{P}_{\theta_0}(D_n(X) = 1) = \mathbb{P}_{\theta_0}(T_n > t) = \alpha$$

if we choose $t = q_{1-\alpha}$, the $1-\alpha$ quantile of the standard normal distribution, where $\mathbb{P}_{Z \sim \mathcal{N}(0,1)}(Z \leq q_x) = x$ for all $x \in \mathbb{R}$. The type II error is

$$\mathbb{P}_{\theta_1}(D_n(X) = 0) = \mathbb{P}_{\theta_1}(T_n \leq q_{1-\alpha}) = \Phi(-\sqrt{n}(\mu_1 - \mu_0) + q_{1-\alpha}) = 1 - \Phi(\sqrt{n}(\mu_1 - \mu_0) - q_{1-\alpha})$$

where Φ is the standard Gaussian cdf.

In order to guarantee a desired type II error β (additionally to the type I error which is already guaranteed for any sample size n), we can choose the sample size n to satisfy

$$\sqrt{n}(\mu_1 - \mu_0) - q_{1-\alpha} = q_{1-\beta}, \quad i.e. \quad n = \frac{(q_{1-\alpha} + q_{1-\beta})^2}{(\mu_1 - \mu_0)^2}\ .$$

In this simple example we had access to the exact distribution of the test statistic (under both the null and the alternative !), which is rarely the case. In the sequel, we present two different approaches to find good test statistics. The first one leverages asymptotically normal estimators, and the second one is directly based on comparing the likelihood of different hypotheses.

3.3 Using Asymptotically Normal Estimators: Wald Test

In this section, we consider the particular testing problem

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{versus} \quad \mathcal{H}_1 : (\theta \neq \theta_0)$$

based on iid samples from the distribution P_θ, $\theta \in \Theta$. We assume that we have an asymptotically normal estimator of the parameter θ, for any $\theta \in \Theta$, that is

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\sigma_\theta^2}{n}}} \rightsquigarrow \mathcal{N}(0,1)$$

where the convergence in distribution is under P_θ.

Definition 3.2. *The Wald test of size α rejects $\mathcal{H}_0$ when $|W_n| > q_{1-\alpha/2}$ where*

$$W_n = \frac{\widehat{\theta}_n - \theta_0}{\sqrt{\frac{\sigma^2_{\widehat{\theta}_n}}{n}}} .$$

Using the asymptotic normality of $\widehat{g}_n$ together with Slutsky's lemma, and assuming further that the mapping $\theta \mapsto \sigma_\theta$ is continuous yields the following result.

Theorem 3.3. *The Wald test is asymptotically of size α, that is,*

$$\mathbb{P}_{\theta_0}\left(|W_n| > q_{1-\alpha/2}\right) \underset{n\to\infty}{\longrightarrow} \alpha$$

While the test statistic of the Wald test replaces σ_θ by an estimate $\sigma_{\widehat{\theta}_n}$, it is also possible to use the value of σ_θ in $\theta = \theta_0$ to get the same asymptotic size, that is, we can replace W_n by

$$\widetilde{W_n} = \frac{\widehat{\theta}_n - \theta_0}{\sqrt{\frac{\sigma^2_{\theta_0}}{n}}}.$$

Under the null hypothesis, the statistics $\widetilde{W_n}$ still converges in distribution to a standard normal, which allows for the same calibration as above.

Remark 3.4. *The same test can be used for different "one-sided" alternatives, of the form $\mathcal{H}_1 : (\theta > \theta_0)$, $(\theta < \theta_0)$ or $(\theta > \theta_1)$ for some $\theta_1 > \theta_0$. The asymptotic type I error will be the same, but the power may be improved by considering "one-sided" rejecting regions. For example $D_n(X) = \mathbb{1}(W_n > q_{1-\alpha})$ is still asymptotically of size α but is better suited (in terms of power) for $\mathcal{H}_1 : (\theta > \theta_0)$.*

Example 3.5. *Based on iid sample from a Bernoulli distribution with parameter p, we are interested in testing*

$$\mathcal{H}_0 : \left(p = \frac{1}{2}\right) \quad \textit{versus} \quad \mathcal{H}_1 : \left(p \neq \frac{1}{2}\right).$$

The previous chapters gave use an efficient estimator for p, which is further asymptotically normal. Letting $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$, we have, for all $p \in [0,1]$ that if p is the true parameter

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \rightsquigarrow \mathcal{N}(0,1)$$

Hence the Wald test can be written

$$D_n(X) = \mathbb{1}\left(\left|\frac{\hat{p}_n - \frac{1}{2}}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}\right| > q_{1-\frac{\alpha}{2}}\right)$$

and is asymptotically of size α.

Example 3.6 (treatment effect)**.** *Going back to Example 3.6 where we collect pairs of observations $(X_i, Y_i) \in \{0,1\}$ from two treatments 1 (placebo, with probability of efficacy p_1) and 2 (new drug, with probability of efficacy p_2). We gave an asymptotically normal estimator for the effect size $\phi = p_2 - p_1$*

$$\frac{\widehat{\phi}_n - \phi}{\sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{n}}} \rightsquigarrow \mathcal{N}(0,1)$$

where $\widehat{\phi}_n = \widehat{p}_{2,n} - \widehat{p}_{1,n}$. The Wald test statistic associated to the test

$$\mathcal{H}_0 : (\phi = 0) \quad \textit{versus} \quad \mathcal{H}_1 : (\phi \neq 0)$$

is

$$W_n = \frac{\widehat{\phi}_n}{\sqrt{\frac{\widehat{p}_{1,n}(1-\widehat{p}_{1,n})}{n} + \frac{\widehat{p}_{2,n}(1-\widehat{p}_{2,n})}{n}}}$$

and is asymptotically normal under the null hypothesis. Yet in this example, we may be more insterested to test whether or not the new treatment is better *(instead of different) than the placebo, thus considering the hypotheses*

$$\mathcal{H}_0 : (\phi = 0) \quad \textit{versus} \quad \mathcal{H}_1 : (\phi > 0)\ .$$

The test

$$D_n(X) = \mathbb{1}\left(W_n > q_{1-\alpha}\right)$$

is asymptotically of size α. Regarding its power, let us consider an alternative in which $p_2 = p_1 + \varepsilon$, in which there is an effect size of ε. Now consider (p_1, p_2) such that $p_2 > p_1$. The power in (p_1, p_2) is

$$\begin{aligned} \mathbb{P}_{(p_1,p_2)}\left(W_n > q_{1-\alpha}\right) &= \mathbb{P}_{(p_1,p_2)}\left(\frac{\widehat{\phi}_n - (p_2 - p_1)}{\sqrt{\frac{\widehat{p}_{1,n}(1-\widehat{p}_{1,n})}{n} + \frac{\widehat{p}_{2,n}(1-\widehat{p}_{2,n})}{n}}} > -\frac{(p_2 - p_1)}{\sqrt{\frac{\widehat{p}_{1,n}(1-\widehat{p}_{1,n})}{n} + \frac{\widehat{p}_{2,n}(1-\widehat{p}_{2,n})}{n}}} + q_{1-\alpha} \right) \\ &\simeq 1 - \Phi\left(\frac{(p_2 - p_1)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} - q_{1-\alpha} \right) \end{aligned}$$

If (p_1, p_2) is such that there is an effect of size ε, ie $p_2 = p_1 + \varepsilon$, to get (approximately) a power $1 - \beta$, we need to choose n such that

$$\frac{\sqrt{n}\varepsilon}{\sqrt{p_1(1-p_1)} + \sqrt{(p_1+\varepsilon)(1-p_1-\varepsilon)}} - q_{1-\alpha} \geq q_{1-\beta},$$

which requires further to have a good guess of p_1, the probability of efficacy of the placebo.

We now present a general receipe for (more general) tests, directly based on the likelihood.

3.4 Likelihood-Ratio Tests

Given two hypotheses

$$\mathcal{H}_0 : (\theta \in \Theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \in \Theta_1).$$

the idea of a likelihood ratio test is to compare the likelihood of the observation $L(X;\theta)$ for parameters θ that come from Θ_0 and from Θ_1.

A Likelihood-Ratio Test (LRT) reject $\mathcal{H}_0$ for large enough values of the test statistic

$$\Lambda(X) = \frac{\sup_{\theta\in\Theta_1} L(X;\theta)}{\sup_{\theta\in\Theta_0} L(X;\theta)}$$

or equivalently for large values of the log-likelihood ratio

$$\log\Lambda(X) = \log\left(\frac{\sup_{\theta\in\Theta_1} L(X;\theta)}{\sup_{\theta\in\Theta_0} L(X;\theta)}\right) .$$

More precisely, a LRT is of the form $D(X) = \mathbb{1}(\Lambda(X) > t)$ for some threshold $t > 1$ or $D(X) = \mathbb{1}(\log\Lambda(X) > u)$ for some $u > 0$. The idea of the test is to reject $\mathcal{H}_0$ when parameters in Θ_1 are significantly more likely that parameters in Θ_0.

Remark 3.7. *LRT are often used in settings where $\Theta_1 = \Theta\backslash\Theta_0$ where Θ is the entire set of possible parameters. In that case, one of the suprema will be equal to $L(X, \widehat{\theta}_{MLE})$. In particular, the likelihood-ratio test will reject $\mathcal{H}_0$ is the MLE belong to Θ_1 and satisfies*

$$\frac{L(X;\widehat{\theta}_{MLE})}{\sup_{\theta\in\Theta_0} L(X;\theta)} \geq t$$

for some well chosen threshold t.

Computational aspects In full generality, the computation of the likelihood ratio can be challenging, as unlike in the MLE case, we are required to solve at least one *constrained optimization* problem. In particular, when the hypotheses are composites (i.e. when neither Θ_0 nor Θ_1 are reduced to one element) the LRT is sometimes called the Generalized Likelihood Ratio Test (GLRT), the LRT being reserved to the simpler setting of testing two simple hypotheses where 1) the computation of the statistic is straightforward and 2) we have strong optimality guarantees, as will be explained in the next section.

Calibration of a LRT To use a LR test in practise, we need to find a threshold t_α such that if we define

$$D(X) = \mathbb{1}\left(\frac{\sup_{\theta\in\Theta_1} L(X;\theta)}{\sup_{\theta\in\Theta_0} L(X;\theta)} > t_\alpha\right)$$

the test has a type I error smaller than α for all $\theta \in \Theta_0$. This can be done case by case using arguments that are specific to the distribution at hand, but in Section 3.7 we will present a general result for calibrating (asymptotically) a GLRT.

Example 3.8. *Let's go back to the simple setting of Example 3.1. Letting f_μ be the density of a $\mathcal{N}(\mu, 1)$ distribution, the log-likelihood ratio in this case is*

$$\begin{aligned}
\log\frac{\prod_{i=1}^n f_{\mu_1}(X_i)}{\prod_{i=1}^n f_{\mu_0}(X_i)} &= \sum_{i=1}^n\left[-\frac{(X_i-\mu_1)^2}{2} + \frac{(X_i-\mu_0)^2}{2}\right] \\
&= \frac{1}{2}\sum_{i=1}^n(\mu_1-\mu_0)(2X_i-\mu_0-\mu_1) \\
&= (\mu_1-\mu_0)\left[n\widehat{\mu}_n - n\frac{\mu_0+\mu_1}{2}\right] .
\end{aligned}$$

Hence, we see that $\log \Lambda(X) > u$ *is equivalent to*

$$\widehat{\mu}_n > \frac{\mu_0 + \mu_1}{2} + \frac{u}{n(\mu_1 - \mu_0)} .$$

Hence the LRT is of the form $D_n(X) = \mathbb{1}(\widehat{\mu}_n > x)$ *for some threshold* x*. We remark that we this test is exactly the one proposed in Example 3.1, where we chose*

$$x = \mu_0 + \frac{q_{1-\alpha}}{\sqrt{n}}$$

to guarantee a type I error α*.*

3.5 The Neyman-Pearson lemma

The result of this section is a first motivation for the use of LR tests: in some simple settings, LRT can be better than other tests, according to the following definition.

Definition 3.9. *Let* $\alpha \in [0,1]$*. A statistical test* D *is called Uniformly More Powerful at level* α *(denoted by* $\mathrm{UMP}(\alpha)$*) if*

1. *the test* D *is of level* α*, i.e.,* $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(D(X) = 1) = \alpha$
2. *For all other test* D' *that is of level* α*,* $\forall \theta \in \Theta_1$*,* $\mathbb{P}_\theta(D(X) = 1) \geq \mathbb{P}_\theta(D'(X) = 1)$*.*

The Neyman-Pearson lemma shows that for testing two simple hypotheses

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta = \theta_1)$$

where θ_0 and θ_1 are two distinct points of Θ, likelihood ratio tests, that have the simple form

$$D_t(X) = \mathbb{1}\left(\frac{f_1(X)}{f_0(X)} > t\right) \tag{3.1}$$

can be UMP(α). Such simple likelihood-ratio tests are sometimes called Neyman-Pearson tests, due to the following result.

Theorem 3.10 (Neyman-Pearson lemma)**.** *For* $\alpha \in (0,1)$*, if there exists a treshold* $t_\alpha > 1$ *such that the likelihood ratio test* D_{t_α} *as defined in* (3.1) *satisfies* $\mathbb{P}_{\theta_0}(D_{t_\alpha}(X) = 1) = \alpha$*, then this test is* UMP*(*α*).*

Proof. To ease the notation we denote by $D(X) = \mathbb{1}\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} > t_\alpha\right)$ a LR test satisfying $\mathbb{P}_{\theta_0}(D(X) = 1) = \alpha$ and we let D' be another test satisfying $\mathbb{P}_{\theta_0}(D'(X) = 1) \leq \alpha$. In particular, we have

$$\mathbb{P}_{\theta_0}(D(X) = 1) - \mathbb{P}_{\theta_0}(D'(X) = 1) \geq 0 \quad \Leftrightarrow \quad \mathbb{E}_{\theta_0}[D(X) - D'(X)] \geq 0 .$$

Our goal is to prove that

$$\mathbb{P}_{\theta_1}(D(X) = 1) - \mathbb{P}_{\theta_1}(D'(X) = 1) \geq 0 \quad \Leftrightarrow \quad \mathbb{E}_{\theta_1}[D(X) - D'(X)] \geq 0 .$$

The proof consists in relating the expectation of $D(X) - D'(X)$ under θ_1 to that under θ_0. To do so, we introduce the function

$$g(x) = (D(x) - D'(x))(f_{\theta_1}(x) - t_\alpha f_{\theta_0}(x))$$

and we prove that this function is always non-negative by considering four cases:

1. If $f_{\theta_0}(x) = f_{\theta_1}(x) = 0$, then $g(x) = 0$.

2. If $f_{\theta_0}(x) = 0$ and $f_{\theta_1}(x) > 0$, then the value of the likelihood ratio $\frac{f_1(x)}{f_0(x)}$ is infinite, and as the threshold t_α is finite, we have $D(x) = 1$, which leads to $D(x) - D'(x) \geq 0$ for any D' and $g(x) = (D(x) - D'(x))f_{\theta_1}(x) \geq 0$.

3. If $f_{\theta_0}(x) > 0$ and $f_{\theta_1}(x) - t_\alpha f_{\theta_0}(x) > 0$, by definition of D, $D(x) = 1$ hence $D(x) - D'(x) \geq 0$ for any D' and $g(x) \geq 0$.

4. If $f_{\theta_0}(x) > 0$ and $f_{\theta_1}(x) - t_\alpha f_{\theta_0}(x) < 0$, by definition of D, $D(x) = 0$ hence $D(x) - D'(x) \leq 0$ for any D' and $g(x) \geq 0$.

We deduce that

$$\mathbb{E}_{\theta_1}[D(X) - D'(X)] - t_\alpha \mathbb{E}_{\theta_0}[D(X) - D'(X)] = \int_{\mathcal{X}} g(x) d\nu(x) \geq 0$$

hence $\mathbb{E}_{\theta_1}[D(X) - D(X')] \geq t_\alpha \mathbb{E}_{\theta_0}[D(X) - D'(X)] \geq 0$, which concludes the proof.

□

The statement of Theorem 3.10 suggests that for a given $\alpha \in (0,1)$, there does not always exist a LR test that has a type I error exactly equal to α. This has to be nuanced a bit.

For continuous distributions, that is when f_{θ_0} is a density with respect to the Lebesgue measure, there is actually no such issue: it is always possible to find t_α such that

$$\mathbb{P}_{\theta_0}\left(f_1(X) > t_\alpha f_0(X)\right) = \alpha.$$

Indeed, letting $\mathcal{E}(t) = (f_1(X) > t f_0(X))$, one can justify that $t \mapsto \mathbb{P}(\mathcal{E}(t))$ is continuous, non-increasing and satisfies $\mathbb{P}(\mathcal{E}(0)) = 1$ and $\lim_{t\to\infty} \mathbb{P}(\mathcal{E}(t)) = 0$. By the intermediate value theorem, for any $\alpha \in (0,1)$, there exists t_α such that $\mathbb{P}(\mathcal{E}(t_\alpha)) = \alpha$.

For discrete distribution however, it is not always possible to exactly match some level α with a LR test. We can look at a simple example to understand what happens: if we collect Bernoulli samples $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{B}(\theta)$ and we want to test $\mathcal{H}_0 : (\theta = \theta_0)$ and $\mathcal{H}_1 : (\theta = \theta_1)$ for $\theta_0 < \theta_1$, we can prove that a LR test is of the form $D_n(X) = \mathbb{1}\left(T(X) > t\right)$ where $T(X) = \sum_{i=1}^n X_i$ follows a binomial distribution $\mathcal{B}(n, \mu_0)$ under $\mathcal{H}_0$. As $T(X)$ only takes integer values in $\{0, 1, \dots, n\}$, there are only $n+1$ possible values of $\mathbb{P}_{\theta_0}(D_n(X) = 1)$, hence not all $\alpha \in (0,1)$ can be exactly attained. For example if $\theta_0 = 0.6$ and $n = 10$ we can show that $\mathbb{P}_{\theta_0}(T(X) > 7) \simeq 0.167$ while $\mathbb{P}_{\theta_0}(T(X) > 8) \simeq 0.046$, hence there is no threshold that provides a type I error exactly equal to $\alpha = 0.05$. In that case, it is common to choose the threshold giving the largest type I error that is smaller than α, that is $t = 8$ in this example.

Remark 3.11 (randomized test). *For discrete distributions, it is actually possible to exactly match a type I error α by considering the broader class of randomized tests. A randomized test is a mapping $\widetilde{D} : \mathcal{X} \to [0,1]$ and $\widetilde{D}(X) = \gamma$ with $\gamma \in (0,1)$ leads to rejecting $\mathcal{H}_0$ with probability γ. More concretely, the actual decision D to reject $\mathcal{H}_0$ or not reject it from the randomized test $\widetilde{D}$ can be written*

$$D(X) = \mathbb{1}(U \leq \widetilde{D}(X))$$

where $U \sim \mathcal{U}([0,1])$ is a uniform random variable that is independent from X.

A randomized Likelihood Ratio test can be defined by a threshold t and a parameter $\gamma \in [0,1)$ as follows:

$$\begin{aligned}
\widetilde{D}_{t,\gamma}(X) &= 1 \quad \textit{if} \quad \frac{L(X;\theta_1)}{L(X;\theta_0)} > t \\
\widetilde{D}_{t,\gamma}(X) &= \gamma \quad \textit{if} \quad \frac{L(X;\theta_1)}{L(X;\theta_0)} = t \\
\widetilde{D}_{t,\gamma}(X) &= 0 \quad \textit{if} \quad \frac{L(X;\theta_1)}{L(X;\theta_0)} < t
\end{aligned}$$

The type I error of this randomized test is

$$\mathbb{P}_{\theta_0}(D_{t,\gamma}(X) = 1) = \mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} > t\right) + \gamma \mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} = t\right).$$

Hence in the discrete setting by choosing $t_\alpha = \inf\{t : \mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} > t\right) \le \alpha\}$ and

$$\gamma_\alpha = \frac{\alpha - \mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} > t_\alpha\right)}{\mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} = t_\alpha\right)}$$

we end up with a test that has exactly a type I error α (and will have a larger power than when setting $\gamma = 0$).

By considering the more general class of randomized LRT, we can have a stronger version of Theorem 3.10 saying that for any $\alpha \in (0,1)$ there exists a (possibly randomized) LRT with level α and that any other test of level α has a smaller power.

3.6 Particular forms of the Neyman-Person test

We already mentioned two examples (Gaussian and Bernoulli distribution) for which the Neyman-Pearson test ends up having a simple form. First, we explain that this properties can be extended to exponential families. We recall that a family of distributions P_θ forms an exponential families if their densities with respect to some common reference measure ν is of the form

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

Proposition 3.12. *In an exponential family of the above form with $a(\theta_0) < a(\theta_1)$, the LR test for testing*

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \textit{against} \quad \mathcal{H}_1 : (\theta = \theta_1)$$

based on an n sample $X_1, \ldots, X_n$ from f_θ takes the form $D_n(X) = \mathbb{1}\left(\sum_{i=1}^n T(X_i) > t\right)$ where T is the canonical statistic of the exponential family.

Proof. The likelihood ratio can be written as follows:

$$\begin{aligned}
\Lambda(X) &= \frac{\prod_{i=1}^n h(X_i)\exp(T(X_i)a(\theta_1) - b(\theta_1))}{\prod_{i=1}^n h(X_i)\exp(T(X_i)a(\theta_0) - b(\theta_0))} = \frac{\exp(\sum_{i=1}^n T(X_i)a(\theta_0) - nb(\theta_0))}{\sum_{i=1}^n T(X_i)a(\theta_1) - nb(\theta_1))} \\
&= \exp\left((a(\theta_1) - a(\theta_0))\sum_{i=1}^n T(X_i) - n(b(\theta_1) - b(\theta_0))\right)
\end{aligned}$$

Hence $\log \Lambda(X) > u$ is equivalent to

$$(a(\theta_1) - a(\theta_0)) \sum_{i=1}^{n} T(X_i) - n(b(\theta_1) - b(\theta_0)) > u \, .$$

As $a(\theta_1) - a(\theta_0) > 0$ this is in turn equivalent to

$$\sum_{i=1}^{n} T(X_i) > \frac{u}{a(\theta_1) - a(\theta_0)} + n(b(\theta_1) - b(\theta_0)) \, .$$

□

This property extends to any family of distribution that possesses a sufficient statistic (see Section 1.5.3). Recall that the density of a n samples $X_1, \dots, X_n$ under such a distribution can be expressed as follows:

$$f_\theta(x_1, \dots, x_n) = g(x_1, \dots, x_n) h(S(x_1, \dots, x_n); \theta) \, .$$

In that case $\Lambda(X) = \frac{h(S(X);\theta_1)}{h(S(X);\theta_0)}$ and the LR is of the form

$$D_n(X) = \mathbb{1}(S(X_1, \dots, X_n) \in \mathcal{R})$$

where the rejection region is of the form $\mathcal{R} = \{s \in \mathcal{Y} : \frac{h(s;\theta_1)}{h(s;\theta_0)} > t\}$ for some $t > 0$ and $S : \mathcal{X}^n \to \mathcal{Y}$.

3.7 Testing composite hypotheses

Simple hypotheses are quite restrictive. As a first step towards generalization, a more useful test is to check whether the value of the parameter θ is larger than some reference value. This corresponds to a case in which $\mathcal{H}_0$ is a simple hypothesis, while $\mathcal{H}_1$ is composite.

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta > \theta_0) \tag{3.2}$$

Such a test could be useful to access the efficacy of a new treatment in a setting where the average efficacy of the standard of care (θ_0) is considered known.

Back to our Gaussian example In Example 3.1, we studied a LR test for testing

$$\mathcal{H}_0 : (\mu = \mu_0) \quad \text{against} \quad \mathcal{H}_1 : (\mu = \mu_1) \tag{3.3}$$

based on iid observations $X_1, \dots, X_n$ from $\mathcal{N}(\mu, 1)$, when $\mu_0 < \mu_1$. This test is given by

$$D_n(X) = \mathbb{1}\left(\widehat{\mu}_n > \mu_0 + \frac{q_{1-\alpha}}{\sqrt{n}}\right)$$

and has type I error α. As this test does not depend on μ_1, it is also a valid test for

$$\mathcal{H}_0 : (\mu = \mu_0) \quad \text{against} \quad \mathcal{H}_1 : (\mu > \mu_0) \, . \tag{3.4}$$

The type I error is still α (as $\mathcal{H}_0$ is the same). Now consider any other possible test D'_n for (3.4). For any $\mu_1 > \mu_0$, the Neyman-Pearson lemma applied to the test (3.3) tells us that $\mathbb{P}_{\mu_1}(D_n(X) = 1) \geq \mathbb{P}_{\mu_1}(D'_n(X) = 1)$. It follows that the test D_n is UMP(α) for (3.4).

In this simple example, one could investigate what a (Generalized) Likelihood Ratio would be for the composite hypothesis testing problem (3.4). The log-likelihood ratio is given by

$$\log \Lambda(X) = \sup_{\mu:\mu>\mu_0} \log \frac{L(X_1,\dots,X_n;\mu)}{L(X_1,\dots,X_n,\mu_0)}$$

and the same computations as in Example 3.8 further yield

$$\log \Lambda(X) = \sup_{\mu:\mu>\mu_0} (\mu - \mu_0)\left(n\widehat{\mu}_n - n\frac{\mu_0+\mu}{2}\right)$$

To compute this constrained maximization, one can consider two cases: either the MLE $\widehat{\mu}_n$ is larger than μ_0 and the supremum is attained for $\mu = \widehat{\mu}_n$, or $\widehat{\mu}_n \leq \mu_0$ in which case for all $\mu > \mu_0$, $\widehat{\mu}_n - \frac{\mu_0+\mu}{2} \leq 0$ and the function to maximize is always negative, and maximal at $\mu = \mu_0$, where it is zero. We obtain

$$\log \Lambda(X) = \frac{n}{2}(\widehat{\mu}_n - \mu_0)^2 \mathbb{1}(\widehat{\mu}_n \geq \mu_0) \, .$$

And $\log \Lambda(X) > u$ is equivalent to

$$\widehat{\mu}_n > \mu_0 + \sqrt{\frac{2u}{n}} \, ,$$

which is (again) the same form as the previous test. This provides an example of composite hypothesis testing problem in which the LRT is optimal, in that when calibrated to get a type I error α, it is UMP(α).

3.7.1 Testing complementary hypotheses

For composite hypotheses, Likelihood Ratio Tests are more common when the two hypotheses that are tested are complementary, that is when we have $\Theta_1 : \Theta\backslash\Theta_0$ and we are therefore testing

$$\mathcal{H}_0 : (\theta \in \Theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \in \Theta\backslash\Theta_0) \, . \tag{3.5}$$

In this context, the likehood ratio as we defined it should be written

$$\Lambda(X) = \frac{\sup_{\theta\in\Theta\backslash\Theta_0} L(X;\theta)}{\sup_{\theta\in\Theta_0} L(X;\theta)} \, .$$

But for hypotheses of the form (3.5) it is actually more common to define

$$\widetilde{\Lambda}(X) = \frac{\sup_{\theta\in\Theta} L(X;\theta)}{\sup_{\theta\in\Theta_0} L(X;\theta)} = \frac{L(X;\widehat{\theta}_{\text{MLE}})}{\sup_{\theta\in\Theta_0} L(X;\theta)} \, .$$

First, from a computational perspective, the latter is preferable as it only features one constrained optimization problem. Then, it can be observed that for any threshold $t > 1$,

$$\Lambda(X) > t \;\Leftrightarrow\; \widetilde{\Lambda}(X) > t$$

so the decision associated to comparing these two statistic to a threshold is the same. To justify this fact, we can observe that $\Lambda(X) > 1$ or $\widetilde{\Lambda}(X) > 1$ if an only if the MLE estimator $\widehat{\theta}_{\text{MLE}}$ belongs to $\Theta\backslash\Theta_0$, and in that case $\Lambda(X) = \widetilde{\Lambda}(X)$.

A testing problem for which the expression of the likelihood ratio is particularly simple is

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \neq \theta_0) \tag{3.6}$$

for which we have

$$\widetilde{\Lambda}(X) = \frac{L(X, \widehat{\theta}_{\mathrm{MLE}})}{L(X, \theta_0)}$$

As we see in that case there is no constrained optimization problem to be solved at all. Compared to the testing problem (3.2), the testing problem (3.6) is called two-sided (as the parameter in the null hypothesis have alternative parameters on both sides).

For a two-sided test, there is intuitively no hope to derive a *uniformly* more powerful test of level α: it is always possible to increase the power on one side of the alternative at the cost of decreasing it on the other side. For example the UMP(α) test that we derived in the previous Gaussian example for (3.4) has a very low power for any $\mu < \mu_0$ (as it rejects the null only when $\widehat{\mu}_n$ is significantly larger than μ_0). Conversely, rejecting when $\widehat{\mu}_n < \mu_0 - \frac{q_{1-\alpha}}{\sqrt{n}}$ would yield a good power for $\mu < \mu_0$ but a very small power for $\mu > \mu_0$. We would certainly prefer rejecting for $|\widehat{\mu}_n - \mu_0| > \frac{q_{1-\alpha/2}}{\sqrt{n}}$ which yields reasonable power on both sides of the alternative... but is not uniformly better than the two tests previously mentioned.

3.7.2 Optimality properties

Based on the previous example, one could hope that when UMP(α) exist, they are likelihood ratio tests. Sadly, there is no general result saying this. However, there exists some tests that are UMP(α) for particular (one-sided) composite hypothesis testing problems, under some conditions on the likelihood (and in some cases, those may coincide with LRTs, but there is no general rule about that). We present an example below.

Definition 3.13. *A family of distribution $\{P_\theta, \theta \in \Theta\}$ is a monotonic density ratio family if there exists a statistic $T(x)$ such that*

$$\forall \theta < \theta', \quad \frac{f_{\theta'}(x)}{f_\theta(x)} = g(T(x))$$

for some non-decreasing function g (that may depend on θ and θ').

Theorem 3.14 (Lehmann's theorem)**.** *Let $\theta_1 \geq \theta_0$. Consider the composite hypothesis testing*

$$\mathcal{H}_0 : (\theta \leq \theta_0) \quad \textit{against} \quad \mathcal{H}_1 : (\theta > \theta_1)$$

based on a n-sample $X_1, \dots, X_n$ whose (joint) distribution P_θ belongs to a monotonic density ratio family with statistic $T(x)$. Then the test given by $D_n(X) = \mathbb{1}(T(X_1, \dots, X_n) > t)$ is UMP(α) where $\alpha = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta(D_n(X) = 1)$.

Example 3.15. *The computations of Example 3.8 shows that the Gaussian n-sample model (with a fixed variance) is a monotonic density ratio family, with the statistic $T(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^n X_i$. The same holds for exponential families, if $\theta \mapsto a(\theta)$ is increasing (see the proof of Proposition 3.12).*

3.7.3 Asymptotic calibration of a Likelihood-Ratio Test

Even if there are no general optimality properties (expressed with the UMP(α) property introduced in this chapter), we show below a general result providing a way to calibrate a Generalized Likelihood Ratio test using asymptotic considerations. This results holds for particular forms of testing problems that generalize the two-sided test

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \neq \theta_0)$$

to possibly higher dimension of the parameter space.

Theorem 3.16 (Wilk's theorem)**.** *Consider a n-samples $X_1, \dots, X_n \sim \mathbb{P}_\theta$ coming from a parameteric model $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Assume that Θ_0 defines a linear sub-hypothesis of Θ with $\dim(\Theta) = p$ and $\dim(\Theta_0) = q$. Assume that the MLE estimator satisfy the conditions of Theorem 2.20 to be asymptotically normal. Then, for any $\theta \in \Theta_0$,*

$$2 \log \widetilde{\Lambda}(X_1, \dots, X_n) \rightsquigarrow \chi^2(p - q)$$

Hence the test $D_n(X) = \mathbb{1}\left(2 \log \widetilde{\Lambda}(X_1, \dots, X_n) > t_\alpha\right)$ with t_α equal to the $1 - \alpha$ quantile of the chi-square distribution with $p - q$ degrees of freedom is asymptotically of level α.

The notion of "linear sub-hypothesis" is a bit vague. It means that Θ_0 puts some constraints on the parameters $(\theta_1, \dots, \theta_p) \in \Theta$, and that these constraints are linear (or actually affine). Here are a few examples of such constraints:

- a subset (say m) of the p variable are set to fixed values: the dimension of Θ_0 is $q = p - m$
- a subset of the variables are equal: $\theta_1 = \theta_2 = \cdots = \theta_m$: the dimension of Θ_0 is $q = p - m + 1$

A particular two-dimensional example of interest is the two sample test

$$\mathcal{H}_0 : (\theta_1 = \theta_2) \quad \text{against} \quad \mathcal{H}_1 : (\theta_1 \neq \theta_2)$$

when we collect data from two distributions P_{θ_1} and P_{θ_2} (which generalizes the test for treatment effect discussed in Example 3.6). In that case $\dim(\Theta_0) = 1$ while $\dim(\Theta) = 2$.

Example 3.17. *Consider the two sample test*

$$\mathcal{H}_0 : (\mu_1 = \mu_2) \quad \textit{versus} \quad \mathcal{H}_1 : (\mu_1 \neq \mu_2)$$

based on n_1 iid samples $X_1, \dots, X_{n_1}$ from $\mathcal{N}(\mu_1, \sigma^2)$ and n_2 iid samples $Y_1, \dots, Y_{n_2}$ from $\mathcal{N}(\mu_2, \sigma^2)$, with X_i independent from Y_j.

Assuming that σ^2 is known, the likelihood of $Z = (X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2})$ is

$$L(Z; \mu_1, \mu_2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n} \prod_{i=1}^{n_1} \exp\left(-\frac{(X_i - \mu_1)^2}{2\sigma^2}\right) \prod_{i=1}^{n_2} \exp\left(-\frac{(Y_i - \mu_2)^2}{2\sigma^2}\right)$$

The MLE is given by $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$ whereas the maximizer of the likelihood under the constraint that $\mu_1 = \mu_2$ is $(\widetilde{\mu}, \widetilde{\mu})$ where

$$\widetilde{\mu} = \frac{1}{n_1 + n_2} \left(\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i\right) = \frac{n_1}{n_1 + n_2} \widehat{\mu}_1 + \frac{n_2}{n_1 + n_2} \widehat{\mu}_2.$$

The (log) Generalized Likelihood Ratio statistic for this test is therefore

$$\begin{aligned}
\log \widetilde{\Lambda}(t) &= \log \frac{\prod_{i=1}^{n_1} \exp\left(-\frac{(X_i-\widehat{\mu}_1)^2}{2\sigma^2}\right) \prod_{i=1}^{n_2} \exp\left(-\frac{(Y_i-\widehat{\mu}_2)^2}{2\sigma^2}\right)}{\prod_{i=1}^{n_1} \exp\left(-\frac{(X_i-\widetilde{\mu})^2}{2\sigma^2}\right) \prod_{i=1}^{n_2} \exp\left(-\frac{(Y_i-\widetilde{\mu})^2}{2\sigma^2}\right)} \\
&= \sum_{i=1}^{n_1}\left[\frac{(X_i-\widetilde{\mu})^2}{2\sigma^2} - \frac{(X_i-\widehat{\mu}_1)^2}{2\sigma^2}\right] + \sum_{i=1}^{n_2}\left[\frac{(Y_i-\widetilde{\mu})^2}{2\sigma^2} - \frac{(Y_i-\widehat{\mu}_2)^2}{2\sigma^2}\right] \\
&= \frac{n_1}{2\sigma^2}(\widetilde{\mu}-\widehat{\mu}_1)^2 + \frac{n_2}{2\sigma^2}(\widetilde{\mu}-\widehat{\mu}_2)^2 \\
&= \frac{n_1}{2\sigma^2}\left(\frac{n_2}{n_1+n_2}(\hat{\mu}_1-\hat{\mu}_2)\right)^2 + \frac{n_2}{2\sigma^2}\left(\frac{n_1}{n_1+n_2}(\hat{\mu}_1-\hat{\mu}_2)\right)^2 \\
&= \frac{n_1 n_2}{2\sigma^2(n_1+n_2)}(\hat{\mu}_1-\hat{\mu}_2)^2 \\
&= \frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{\frac{2\sigma^2}{n_1}+\frac{2\sigma^2}{n_2}}
\end{aligned}$$

Under $\mathcal{H}_0$, this statistic is the square of a Gaussian, so it exactly follows a χ_2 distribution with 1 degree of freedom. The asymptotic calibration of Wilk's theorem is therefore exact in that case.

If σ^2 is unknown, the derivations are a bit more tedious, but the LRT statistic takes the form

$$\log \widetilde{\Lambda}(t) = \frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{\frac{2\widehat{\sigma}_p^2}{n_1}+\frac{2\widehat{\sigma}_p^2}{n_2}}$$

with the pooled variance $\widehat{\sigma}_p^2 = \frac{(n_1-1)\widehat{\sigma}_1+(n_2-1)\widehat{\sigma}_2}{n_1+n_2-2}$. Here also instead of relying on a asymptotic calibration, we can derive the distribution of the GLR statistic, which is the square of a Student (t) distribution. The LRT test is a t-test.

3.8 Wald Test versus LRT

Is the LRT better than other tests? In the composite setting, this question is complicated, as there are no strong optimality property for the LRT. In many cases, it actually ends up being close to tests that we could have derived using another approaches.

If we go back to the test

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \neq \theta_0)$$

and consider as an estimator the MLE, we can try to draw some similarities between the Wald Test and a GLR test. Under appropriate regularity assumptions, we saw in Chapter 2 (see Corollary 2.21) that under $\mathcal{H}_0$,

$$\sqrt{nI(\widehat{\theta}_n)}\left(\widehat{\theta}_n - \theta_0\right) \rightsquigarrow \mathcal{N}(0,1)$$

where $I(\theta)$ is the Fisher information, so the Wald test statistic is

$$W_n = \frac{\widehat{\theta}_n - \theta_0}{\sqrt{\frac{I(\widehat{\theta}_n)}{n}}}\,.$$

For this particular case, we can argue that the Wald test and the LRT test are actually close, by proposing an approximation of the log-likelihood ratio. Using a Taylor expansion in $\widehat{\theta}_n$, we have

$$\begin{aligned}
\log \widetilde{\Lambda}(X) &= \log L(X;\widehat{\theta}_n) - \log L(X;\theta_0) \simeq -\underbrace{\frac{\partial \ell(X;\widehat{\theta}_n)}{\partial \theta}}_{=0}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}\frac{\partial^2 \ell(X;\widehat{\theta}_n)}{\partial^2 \theta}\left(\widehat{\theta}_n - \theta_0\right)^2 \\
2\log \widetilde{\Lambda}(X) &\simeq -\frac{1}{n}\frac{\partial^2 \ell(X;\widehat{\theta}_n)}{\partial^2 \theta}\left(\sqrt{n}(\widehat{\theta}_n - \theta_0)\right)^2 \\
2\log \widetilde{\Lambda}(X) &\simeq I(\theta_0)\left(\sqrt{n}(\widehat{\theta}_n - \theta_0)\right)^2 \\
2\log \widetilde{\Lambda}(X) &\simeq \frac{I(\theta_0)}{I(\widehat{\theta}_n)} W_n^2
\end{aligned}$$

where the last but one step uses the law of large number and the definition of the Fisher information, together with the consistency of $\widehat{\theta}_n$. Despite their closeness in the asymptotic regime, for moderate values of the sample size, the LRT and the Wald test could still give different results, and both can be worth trying.

3.9 A glimpse into sequential testing

So far we have looked at *batch* statistical tests: given n samples from our distribution, we should decide whether or not we want to reject $\mathcal{H}_0$. When $\mathcal{H}_1$ is correct and we don't reject, it can of course be because we've been unlucky (we got atypical realization of our test statistic) or it can be because we did not have enough samples (our test statistic is still not very concentrated, and the type II error is large). In sequential testing, we are allowed to *adaptively* stop the data collection until we are actually able to make a *decision* confidently.

In a sequential test, we collect a stream $(X_i)_{i\in\mathbb{N}}$ of iid samples from some distribution P_θ. There are two hypotheses

$$\mathcal{H}_1 : (\theta \in \Theta_1) \quad \text{and} \quad \mathcal{H}_2 : (\theta \in \Theta_2)$$

and a test consists of a stopping rule τ (that is a stopping rule with respect to the filtration $\mathcal{F}_t = \sigma(X_1,\dots,X_t)$) and a decision rule $\widehat{\delta} \in \{1,2\}$ which indicates the chosen hypothesis and depends on the data collected upon stopping, $X_1,\dots,X_\tau$.

The goal is to control the type I and type II errors with prescribed levels α and β:

$$\forall \theta \in \Theta_1, \mathbb{P}_\theta\left(\widehat{\delta} = 2\right) \le \alpha \quad \text{and} \quad \forall \theta \in \Theta_2, \mathbb{P}_\theta\left(\widehat{\delta} = 1\right) \le \beta\,.$$

The Sequential Probability Ratio Test In 1945, Wald proposed the Sequential Probability Ratio Test for the case of simple hypothesis testing [Wald, 1945]

$$\mathcal{H}_1 : (\theta = \theta_1) \quad \text{and} \quad \mathcal{H}_2 : (\theta = \theta_2)\,.$$

It is a natural extension of the Likelihood Ratio Testing principle to the sequential testing. The tests takes as input two thresholds $A < 1$ and $B > 1$ and stops when either the likelihood ratio becomes significantly larger than one, or significantly smaller than 1.

$$\tau_{A,B} = \inf\left\{ t \in \mathbb{N} : \frac{\prod_{i=1}^t f_{\theta_2}(X_i)}{\prod_{i=1}^t f_{\theta_1}(X_i)} \notin [A,B] \right\}$$

and the decision is $\widehat{\delta}_{A,B} = 2$ if and only if

$$\frac{\prod_{i=1}^{\tau_{A,B}} f_{\theta_2}(X_i)}{\prod_{i=1}^{\tau_{A,B}} f_{\theta_1}(X_i)} > B\ .$$

The SPRT enjoys a very strong optimality property, which can be seen as a counterpart of the notion of UMP(α) test in the sequential setting.

Theorem 3.18. *Let $(\tau, \widehat{\delta})$ be a SPRT and let*

$$\alpha = \mathbb{P}_{\theta_1}\left(\widehat{\delta} = 2\right) \quad \textit{and} \quad \beta = \mathbb{P}_{\theta_2}\left(\widehat{\delta} = 1\right).$$

Let $(\widetilde{\tau}, \widetilde{\delta})$ be another sequential test such that

$$\mathbb{P}_{\theta_1}\left(\widetilde{\delta} = 2\right) \leq \alpha \quad \textit{and} \quad \mathbb{P}_{\theta_2}\left(\widetilde{\delta} = 1\right) \leq \beta.$$

Then this tests requires on average more samples than the SPRT: for all $i \in \{1, 2\}$, $\mathbb{E}_{\theta_i}[\widetilde{\tau}] \geq \mathbb{E}_{\theta_i}[\tau]$.

Some properties How to calibrate the SPRT? A simple choice of A and B to get desired type I and type II errors is provided below.

Lemma 3.19. *Choosing $A = \beta$ and $B = \frac{1}{\alpha}$ yields*

$$\mathbb{P}_{\theta_1}\left(\widehat{\delta}_{A,B} = 2\right) \leq \alpha \quad \textit{and} \quad \mathbb{P}_{\theta_2}\left(\widehat{\delta}_{A,B} = 1\right) \leq \beta.$$

Proof. The argument relies on a change of distribution.

$$\begin{aligned}
\mathbb{P}_{\theta_1}\left(\widehat{\delta}_{A,B} = 2\right) &= \mathbb{E}_{\theta_1}\left[\sum_{n=1}^{\infty} \mathbb{1}(\tau = n)\mathbb{1}\left(\widehat{\delta}_{A,B} = 2\right)\right] \\
&\leq \mathbb{E}_{\theta_1}\left[\sum_{n=1}^{\infty} \mathbb{1}(\tau = n)\mathbb{1}\left(\widehat{\delta}_{A,B} = 2\right)\frac{1}{B}\frac{\prod_{i=1}^{n} f_{\theta_2}(X_i)}{\prod_{i=1}^{n} f_{\theta_1}(X_i)}\right] \\
&= \frac{1}{B}\mathbb{E}_{\theta_2}\left[\sum_{n=1}^{\infty} \mathbb{1}(\tau = n)\mathbb{1}\left(\widehat{\delta}_{A,B} = 2\right)\right] \\
&= \frac{1}{B}\mathbb{P}_{\theta_2}\left(\widehat{\delta}_{A,B} = 2\right) \\
&\leq \alpha
\end{aligned}$$

for $B = \frac{1}{\alpha}$. Similarly, with $A = \beta$, we get

$$\begin{aligned}
\mathbb{P}_{\theta_2}\left(\widehat{\delta}_{A,B} = 1\right) &= \mathbb{E}_{\theta_2}\left[\sum_{n=1}^{\infty} \mathbb{1}(\tau = n)\mathbb{1}\left(\widehat{\delta}_{A,B} = 1\right)\right] \\
&\leq \mathbb{E}_{\theta_2}\left[\sum_{n=1}^{\infty} \mathbb{1}(\tau = n)\mathbb{1}\left(\widehat{\delta}_{A,B} = 1\right)A\frac{\prod_{i=1}^{n} f_{\theta_1}(X_i)}{\prod_{i=1}^{n} f_{\theta_2}(X_i)}\right] \\
&= A\mathbb{P}_{\theta_1}\left(\widehat{\delta}_{A,B} = 1\right) \\
&\leq \beta\ .
\end{aligned}$$

□

We note that the calibration is quite different from that of a batch test: we give as an input α and β and the sample size adjusts automatically to match those. In the limit of small α and β of the same order of magnitude, it can be shown that the expected number of samples used by the SPRT is

$$\mathbb{E}_{\theta_1}[\tau] \simeq \frac{\log(1/\alpha)}{\mathrm{KL}(P_{\theta_1}, P_{\theta_2})} \quad \text{and} \quad \mathbb{E}_{\theta_2}[\tau] \simeq \frac{\log(1/\alpha)}{\mathrm{KL}(P_{\theta_2}, P_{\theta_1})}$$

In the Gaussian case, you can check that this expectation is twice less than the number of samples we would need to set up in a batch test to get the same type I and type II error guarantees.

Bibliography

[Rivoirard and Stoltz, 2009] Rivoirard, V. and Stoltz, G. (2009). *Statistique en Action*. Vuibert.

[Van der Vaart, 1998] Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[Wald, 1945] Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186.