Full course of Statistical Modelling
L3 MIDO

Course by Prof. Judith Rousseau

Transcribed by Ayda Atmani

Sponsorised by Bastien Marbaud et Antoine Nicolas–Lutfalla

Compiled by Samuel Lelouch and Arris Bouzouane with Gemini

December 10, 2025
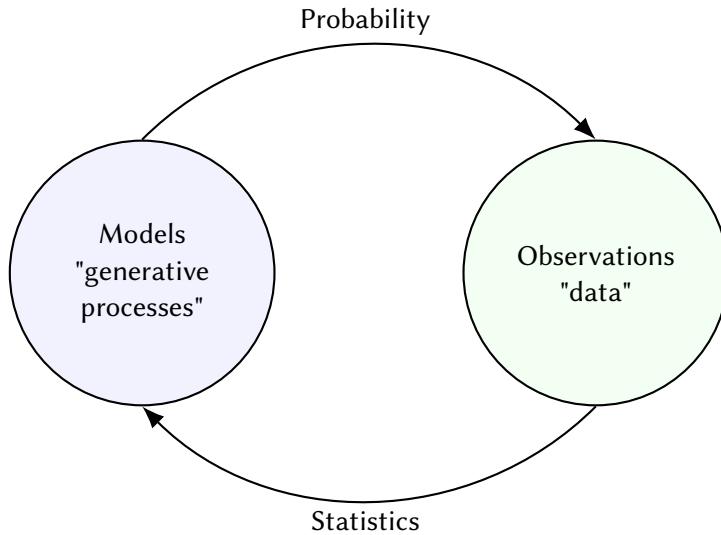
Contents

Chapter 1: Introduction to Statistical Modelling

Introduction Schema

I.   Data

1) $n = \# $ **L3 MIDO students**

- Were students in Dauphine in L2?

- Data $\in \{0, 1\}^n$:

    - 1: in L2 in Dauphine.
    - 0: not in L2 in Dauphine.

- $x = (0, 0, 1, 1, 0, \dots)$

- $\downarrow \text{sum}(x)$

$\implies$ x is a distribution of n independent experiences of Bernoulli of parameter p.

$$x \sim \text{Ber}(p)^{\otimes n} \quad \text{(i.i.d.)}$$

$$\implies \text{sum}(x) \sim \text{Bin}(n, p)$$
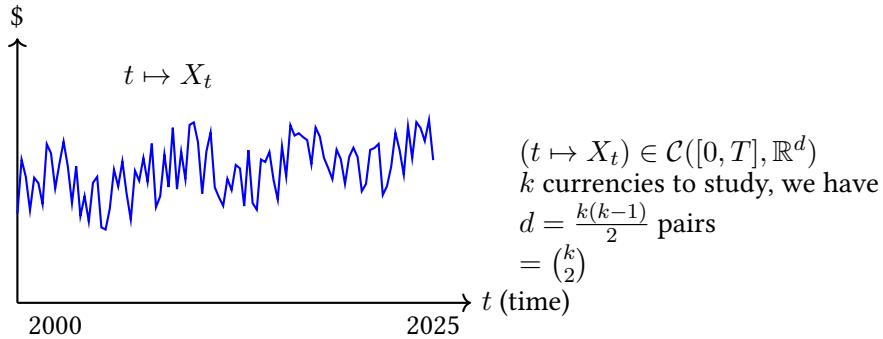
2) INSEE (economy)

Example: Employment rate between 2005 & 2025.

Male ♂	2005	2006	...	2025		Female ♀	2005	...
15-24 yo	α	α	α	α		15-24	α	α
25-49 yo	α	56%	α	α		25-49	α	α
50 +	α		α	α		50 +	α	α
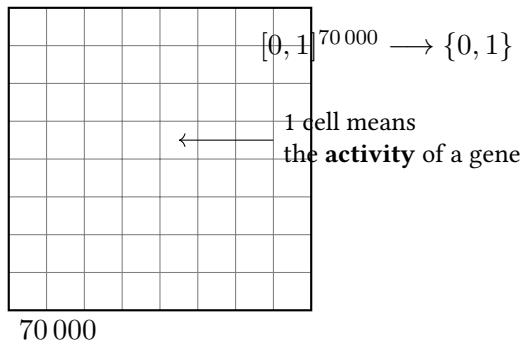
$\implies$ **Tensor:** multiple entries array.

3) Financial data

$(t \mapsto X_t) \in \mathcal{C}([0,T], \mathbb{R}^d)$
k currencies to study, we have
$d = \frac{k(k-1)}{2}$ pairs
$= \binom{k}{2}$

$\implies$ Brownian motion. $\implies$ Diffusion processes.

4) Microarrays

$[0,1]^{70\,000} \longrightarrow \{0,1\}$

1 cell means
the **activity** of a gene

$70\,000$

- Compute a big space into a smaller one: the cells of an individual translated to *healthy* or *not*.

- $[0,1]$ measures the activity of one cell. One person has $70\,000$ cells.

II.   <u>Statistical Modelling</u>

1) Stat models with words

To each set of data, we must associate a scientific question/objective.
    To answer that question, we must have at our disposition:

- A methodology.

- A quantitative mathematical (probabilistic) model that accounts for the properties of the data.

- Well suited methods (mathematical) that combine... and...

2) Abstract (simple) example

We toss a coin 18 times and observe ($n = 18, H = 0, T = 1$).

$$(0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0)$$

$\equiv$ Data (raw).

- **Statistical model:** We observe $n = 18$ random variables X_i; that are independent and that have the same distribution.

- $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = \theta$.

- $\text{Ber}(\theta)$ where $\theta \in \Theta = [0,1]$.

- θ is the **unknown parameter**.
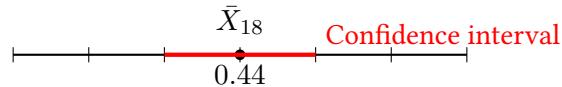
Questions:
1) What is a good (the best) estimation of θ?

- A good (?) estimator:

$$\underbrace{\bar{X}_n}_{\substack{\text{notation} \\ \text{of the mean}}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$= \frac{1}{18} \sum_{i=1}^{18} X_i = \frac{8}{18} = 0.44$$

- What is my accuracy of estimation?

$\implies$ We want:

1. Small length.

2. Good coverage.

2) Is the coin fair?
For instance, we compare $\bar{X}_{18}$ to 0.5. If $|\bar{X}_{18} - 0.5|$ is small, we accept the idea that the coin is fair, otherwise we reject that hypothesis.

Chapter 2: Models

I.   What is a statistical model?

> **Definition 1: Statistical Model**
>
> Let $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be a vector of n random variables, where each variable $X_i \in \mathcal{X}$ and $\mathcal{X}$ is a measurable space.
> A **model** for X is a set $\mathcal{P}$ of probability distributions on $\mathcal{X}^n$.
> Statistical inference will consist in estimating P (the distribution of X) or $F(P)$, with $P \in \mathcal{P}$.

Examples

i) Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

   Notation: $X_i \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ (Gaussian distribution with parameters μ and σ^2). The density is given by: $x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

   Here, $\mathcal{X} = \mathbb{R}$. The model is defined as:

   $$\mathcal{P} = \left\{ \mathcal{N}(\mu, \sigma^2)^{\otimes n}, \ (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty) \right\}$$

> **Remark**
>
> **Notations:** If (X_1, X_2) is a random vector of $\mathbb{R}^2$, with $X_1 \sim P_1$ and $X_2 \sim P_2$.
> (By the way: $X \sim P$ means that P is the distribution of $X \in \mathbb{R}$, meaning: $\forall A \in \mathcal{B}(\mathbb{R})$, $P(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$.)
>
> $$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$$
>
> P is defined via: $\forall A \in \mathcal{B}(\mathbb{R}), P(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega, X(\omega) \in A\})$.
> If $X_1 \sim P_1$, $X_2 \sim P_2$ and $X_1 \perp\!\!\!\perp X_2$ (X_1 and X_2 are independent):
>
> $$\mathbb{P}(X_1 \in A, X_2 \in B) = \mathbb{P}(X_1 \in A) \cdot \mathbb{P}(X_2 \in B)$$
>
> $\forall A, B \in \mathcal{B}(\mathbb{R})$, the distribution of (X_1, X_2) is defined on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ and is denoted by $P_1 \otimes P_2$.
> **Computation formula:**
>
> $$P_1 \otimes P_2(A \times B) = P_1(A) \cdot P_2(B) = \mathbb{P}(X_1 \in A) \cdot \mathbb{P}(X_2 \in B)$$

In our example, saying $(X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ exactly means:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

$$= \prod_{i=1}^{n} \int_{A_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} dx_i$$

$$= \int_{A_1} \cdots \int_{A_n} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) dx_1 \ldots dx_n$$

ii) $X_i \in \{0, 1\}, \forall i = 1, \ldots, n$, independently with same distribution $\text{Ber}(p)$, $p \in [0, 1]$.

$$\mathcal{P} = \left\{ \text{Ber}(p)^{\otimes n}, \ p \in [0, 1] \right\}, \quad \mathcal{X}^n = \{0, 1\}^n$$

Model Comparison and Types

Let us consider models of the form $\mathcal{P} = \{P^{\otimes n}, P \in \mathcal{P}_0\}$.

- Example (1): $\mathcal{P}_0 = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)\}$ (Parametric).

- Example (2): $\mathcal{P}_0 = \{\text{Ber}(p), p \in [0, 1]\}$.

- Example (3): $\mathcal{P}_0 = \{\text{All distributions on } (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}$ (Huge model).

Which is the best model? Generally, the best model is the one where the set of parameter values is the **smallest** (Parsimony principle).

> **Remark**
>
> In the 3 examples considered above, all observations are **i.i.d.** But, in many situations, this is not the case.
>
> **Example of a very standard non-i.i.d. model:**
>
> $$X_0 = x_0; \quad X_t = \rho X_{t-1} + \sigma \varepsilon_t, \quad t = 1, \ldots, T$$
>
> We observe $(X_1, \ldots, X_T)$. They are not i.i.d. Here $\rho \in \mathbb{R}$, $\sigma > 0$ (Volatility), and $\varepsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$. X_t could represent the log price of a financial asset.

In this course, we will mainly focus on the i.i.d. case.

II.   Parametric Model

> **Definition 2: Parametric Model**
>
> Let there be data $(X_1, \ldots, X_n) \in \mathcal{X}^n$. A model $\mathcal{P}$ for $(X_1, \ldots, X_n)$ is called **parametric** if:
>
> $$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$
>
> with $\Theta \subset \mathbb{R}^d$ for some $d \geq 1$. Here P_θ is a distribution on $\mathcal{X}^n$.

> **Definition 3: Nuisance Parameters**
>
> Let $X = (X_1, \ldots, X_n)$ i.i.d. with model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. If $\Theta = (\theta_1, \theta_2)$, and if we only are interested in estimating or predicting θ_1 and we don't care about θ_2, we say that:
>
> - θ_1 is the **parameter of interest**.
>
> - θ_2 is the **nuisance parameter**.

III.   Identifiability

Context: We have a statistical model for $X = (X_1, \ldots, X_n)$. $X_i \sim P_\theta$, $\theta \in \Theta \subset \mathbb{R}^d$. So $\mathcal{P} = \{P_\theta^{\otimes n}, \theta \in \Theta\}$.

Aim: Can we learn/estimate θ from X? This is only possible if we can learn θ from $P_\theta^{\otimes n}$ or P_θ.

Definition 4: Identifiability

A statistical model $\{P_\theta, \theta \in \Theta\}$ is **identifiable** (for θ) if:

$$P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$$

(i.e., the map $\theta \mapsto P_\theta$ is injective).

Examples

Example 1: Exponential Distribution

Let $(P_\theta, \theta \in \Theta) = (\mathcal{E}(\theta), \theta > 0)$. P_θ has density $f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{\{x > 0\}}$.
If $P_{\theta_1} = P_{\theta_2}$, then for every $x \in \mathbb{R}$:

$$f_{\theta_1}(x) = f_{\theta_2}(x) \implies \theta_1 e^{-\theta_1 x} = \theta_2 e^{-\theta_2 x} \quad \text{whenever } x > 0$$

Take $x \to 0$ (or $x = 0$ in the limit), we get $\theta_1 = \theta_2$. Thus, the model is identifiable.

Example 2: Gaussian Distribution

Let $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$.

$$P_\theta = \mathcal{N}(\mu, \sigma^2)$$

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbb{R}$$

Injectivity? Assume $P_{\theta_1} = P_{\theta_2}$. Then $\forall x \in \mathbb{R}$:

$$\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

Take $x = \mu_1$, then:

$$\frac{1}{\sigma_1} = \frac{1}{\sigma_2} e^{-\frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}}$$

Take $x = \mu_2$, then:

$$\frac{1}{\sigma_1} e^{-\frac{(\mu_2 - \mu_1)^2}{2\sigma_1^2}} = \frac{1}{\sigma_2}$$

This implies:

$$\frac{\sigma_2}{\sigma_1} = e^{-\frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}} \leq 1 \quad (\text{because } e^{-\cdot} \leq 1)$$

$$= e^{\frac{(\mu_1 - \mu_2)^2}{2\sigma_1^2}} \geq 1 \quad (\text{because } e^{\cdot} \geq 1)$$

Hence:

$$\frac{\sigma_2}{\sigma_1} = 1 \implies \sigma_1 = \sigma_2$$

And substituting back:

$$e^{-\frac{(\mu_1 - \mu_2)^2}{2\sigma^2}} = 1 \implies \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} = 0 \implies \mu_1 = \mu_2$$

Thus $\theta_1 = \theta_2$. The model is identifiable.

Proposition 1: Identifiability via CDF and Density

- If P_θ admits a cumulative distribution function (CDF) F_θ, the model is identifiable $\iff$ $(\forall \theta_1, \theta_2, F_{\theta_1} = F_{\theta_2} \implies \theta_1 = \theta_2)$.

- If P_θ admits a density f_θ, the model is identifiable $\iff$ $(\forall \theta_1, \theta_2, f_{\theta_1} = f_{\theta_2} \implies \theta_1 = \theta_2)$.

Example 3: Non-identifiable Model (Mixture)

Let $Y_1 \sim \text{Ber}(p_1)$, $Y_2 \sim \text{Ber}(p_2)$. $Y_1 \in \{0,1\}$. $\mathbb{P}(Y_1 = 1) = p_1$.
Let X be defined as:
$$X = \begin{cases} Y_1 & \text{with proba } \pi \\ Y_2 & \text{with proba } 1 - \pi \end{cases}$$

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 1 | X = Y_1)\mathbb{P}(X = Y_1) + \mathbb{P}(X = 1 | X = Y_2)\mathbb{P}(X = Y_2)$$
$$= \mathbb{P}(Y_1 = 1) \times \pi + \mathbb{P}(Y_2 = 1) \times (1 - \pi)$$
$$= p_1 \pi + p_2(1 - \pi)$$

We have a model with 3 parameters (p_1, p_2, π). $X \sim \text{Ber}(p)$ with $p = p_1 \pi + p_2(1 - \pi)$.
Knowing $\mathbb{P}(X = 1)$ gives limited information on the law of X.
Numeric Examples:

- $\pi = 1/2, p_1 = 0.6, p_2 = 0.2$. $\mathbb{P}(X = 1) = 0.6 \times 0.5 + 0.2 \times 0.5 = 0.4$.

- $\pi = 0.6, p_1 = 1/2, p_2 = 0.25$. $\mathbb{P}(X = 1) = 0.6 \times 0.5 + 0.4 \times 0.25 = 0.3 + 0.1 = 0.4$.

We have different parameters but the same distribution $\implies$ **Not identifiable**.

IV.   Empirical Distribution Function

Model: We observe $X_1, \ldots, X_n$ i.i.d. P. We set $F(t) = $ Cumulative distribution function of P at t

$$= \mathbb{P}(X_i \leq t), \quad \forall i$$

Remark

Reminder:

- F is non-decreasing.

- F is càdlàg (right-continuous with left limits).

- $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to +\infty} F(t) = 1$.

Definition 5: Empirical CDF

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$$

It is a good idea since when $n \to +\infty$: Fix t: $\hat{F}_n(t) \xrightarrow[n \to \infty]{} \mathbb{E}[Y_i]$ in probability (Weak Law of Large

Numbers).

$$\frac{1}{n} \sum_{i=1}^{n} Y_i$$

with $Y_i = \mathbb{1}_{\{X_i \leq t\}}$. The Y_i are i.i.d., $\mathbb{E}[|Y_i|] < +\infty$.

$$\mathbb{E}[Y_1] = \mathbb{P}(X_1 \leq t) = F(t)$$

So, we have: $\hat{F}_n(t) \to F(t)$ in probability.

Examples

1. $X_i = \begin{cases} 1 & \text{if I have an accident on day } i \\ 0 & \text{otherwise} \end{cases}$ $X_i \overset{i.i.d.}{\sim} \text{Ber}(p); p = \mathbb{P}(\text{avoir un accident}).$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n \to \infty]{\mathbb{P}} p \quad \text{(Law of Large Numbers)}$$

($\xrightarrow{\mathbb{P}}$ en probabilité).

2. $X_1, \ldots, X_n$ i.i.d. $X_i = $ Lifetime of computer i.

We want to estimate the function $t \mapsto F(t) = \mathbb{P}(X_i \leq t)$. To estimate F, we use at time t, $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$.

Chapter 3: Point estimators and MLE

I. Point Estimators

We consider a parametric model $(X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P_\theta$, with $\theta \in \Theta \subset \mathbb{R}^d$.
   Goal: Constructions of estimators for θ or for $g(\theta)$ with $g : \Theta \to \mathbb{R}^p$.

> **Definition 6: Point Estimator**
>
> A **point estimator** for θ is a quantity $\hat{\theta}$ which depends **only** on $X_1, \ldots, X_n$ (the data).

Examples:

- We observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Ber}(\theta)$.

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{is a point estimator of } \theta.$$

- We observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$.

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{is an estimator of } \theta.$$

- We observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}(\theta)$.

$$\hat{\theta}_n = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}$$

Recall: If $X \sim \mathcal{E}(\theta), \theta > 0$:

$$E[X] = \frac{1}{\theta}$$

$$\frac{1}{\hat{\theta}_n} \xrightarrow[n \to \infty]{} \frac{1}{\theta}$$

$$\frac{1}{\hat{\theta}_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

By the Weak Law of Large Numbers (WLLN), $\hat{\theta}_n \xrightarrow[n \to \infty]{\mathbb{P}} \theta$.

For this last example, if we want to estimate $g(\theta) = \frac{1}{\theta}$, in this case, $\hat{g} = \frac{1}{\hat{\theta}_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a "good estimator".

II. Quadratic Risk, Bias

Measure of error = Loss function. The most common loss function is the **quadratic loss**:

$$l(\hat{\theta}, \theta) = (\hat{\theta}_n - \theta)^2$$

We cannot compute $l(\hat{\theta}_n, \theta)$ since we do not know θ.

Definition 7: Quadratic Risk (Mean Squared Error)

We call **quadratic risk** (or mean squared error):

$$R(\hat{\theta}_n, \theta) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

Definition 8: Bias

Let $\hat{\theta}_n$ be an estimator of θ in a statistical model $(X_i)_{i \leq n} \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$.

  (i) We say that $\hat{\theta}_n$ is **unbiased** at θ iff $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$.

  (ii) We say that $\hat{\theta}_n$ is **unbiased** over Θ iff it is unbiased at all $\theta \in \Theta$.

The **bias** is defined as:

$$b(\theta) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta$$

Theorem 1: Bias-Variance Decomposition

In a statistical model $(X_i)_{i \leq n}$ i.i.d. $P_\theta, \theta \in \Theta$. If $\hat{\theta}_n$ is an estimator of θ, then:

$$R(\hat{\theta}_n, \theta) = b(\theta)^2 + \mathbb{V}_\theta(\hat{\theta}_n)$$

where $\mathbb{V}_\theta(\hat{\theta}_n)$ is the variance of $\hat{\theta}_n$ assuming that $X_i \overset{i.i.d.}{\sim} P_\theta$.

Proof.

$$R(\hat{\theta}_n, \theta) = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2]$$

Add and subtract $\mathbb{E}_\theta(\hat{\theta}_n)$:

$$= \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n) + \mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2]$$

Let $A = \hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n)$ and $B = \mathbb{E}_\theta(\hat{\theta}_n) - \theta = b(\theta)$.

$$= \mathbb{E}_\theta[(A + B)^2] = \mathbb{E}_\theta[A^2 + B^2 + 2AB]$$

$$= \underbrace{\mathbb{E}_\theta(A^2)}_{\mathbb{V}_\theta(\hat{\theta}_n)} + \underbrace{\mathbb{E}_\theta(B^2)}_{b(\theta)^2 \text{ (constant)}} + 2\underbrace{\mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n)) \cdot b(\theta)]}_{0}$$

The cross term is zero because:

$$2b(\theta)\mathbb{E}_\theta(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n)) = 2b(\theta)(\mathbb{E}_\theta(\hat{\theta}_n) - \mathbb{E}_\theta(\hat{\theta}_n)) = 0$$

Hence:

$$R(\hat{\theta}_n, \theta) = b(\theta)^2 + \mathbb{V}_\theta(\hat{\theta}_n)$$

(Q.E.D) □

Examples of Risk Calculation:

Example 4: Normal Mean

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$. Estimator $\hat{\theta}_n = \frac{1}{n} \sum X_i = \bar{X}_n$.

$$R(\hat{\theta}_n, \theta) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

$$= \mathbb{E}[(\bar{X}_n - \theta)^2]$$

$$= \text{Var}(\bar{X}_n) \quad (\text{as } \mathbb{E}[\bar{X}_n] = \theta, \text{ it is unbiased})$$

$$= \frac{1}{n}\text{Var}(X_1) = \frac{1}{n}$$

Example 5: Bernoulli Mean

We observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Ber}(\theta)$.

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$R(\hat{\theta}_n, \theta) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

$$= \mathbb{E}[(\bar{X}_n - \theta)^2]$$

$$= \text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n} = \frac{\theta(1 - \theta)}{n}$$

Remark

The function $x \in [0, 1] \mapsto x(1 - x)$ attains its maximum at $x = 1/2$.

$$\implies \forall \theta \in [0, 1], \quad \theta(1 - \theta) \leq \frac{1}{4}$$

$$\implies R(\hat{\theta}_n, \theta) \leq \frac{1}{4n}$$

Reminders on Convergence

Definitions

> ### Definition 9: Convergence in Probability
>
> $$X_n \xrightarrow{\mathbb{P}} X \iff \forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow[n \to \infty]{} 0.$$

> ### Definition 10: Convergence in L^p
>
> $$X_n \xrightarrow{L^p} X \iff \mathbb{E}(|X_n - X|^p) \xrightarrow[n \to \infty]{} 0.$$

> ### Definition 11: Convergence in Distribution
>
> $X_n \xrightarrow{\mathcal{D}} X$
>
> $$\iff \forall x \text{ (where } F_X \text{ is continuous)}, F_{X_n}(x) \xrightarrow[n \to \infty]{} F_X(x)$$
> $$\iff \forall g \text{ continuous and bounded: } \mathbb{E}(g(X_n)) \xrightarrow[n \to \infty]{} \mathbb{E}(g(X))$$

Main Theorems

> ### Theorem 2: Law of Large Numbers (Weak)
>
> If $(X_n)_n$ are i.i.d. random variables such that $\mathbb{E}(|X_n|) < \infty$, then:
>
> $$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n \to \infty]{\mathbb{P}} \mathbb{E}(X_1)$$

> ### Theorem 3: Central Limit Theorem (CLT)
>
> If $(X_n)_n$ are i.i.d. such that $\mathbb{E}(X_n^2) < \infty$, then:
>
> $$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$
>
> where $\bar{X}_n = \frac{1}{n} \sum X_i$, $\mu = \mathbb{E}(X_1)$, $\sigma^2 = \text{Var}(X_1)$.

> ### Theorem 4: Continuity Mapping Theorem
>
> i) If $X_n \xrightarrow{\mathbb{P}} X$ and if g is continuous, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.
>
> ii) If $X_n \xrightarrow{L^p} X$, $p \geq 1$, then $X_n \xrightarrow{L^q} X$ for all $1 \leq q \leq p$.
>
> iii) If $X_n \xrightarrow{\mathcal{D}} X$, and if g is continuous and bounded, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$. (Actually holds for any continuous g).

> **Theorem 5: Delta Method**
>
> If $(Y_n)_n$ are like $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$. If g is $\mathcal{C}^1$ at μ so that $|g'(\mu)| > 0$, then:
> $$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, g'(\mu)^2 \sigma^2)$$

Sketch of proof. Taylor expansion of $g(Y_n)$ around μ:
$$g(Y_n) \approx g(\mu) + g'(\mu)(Y_n - \mu)$$
$$\sqrt{n}(g(Y_n) - g(\mu)) \approx g'(\mu)\sqrt{n}(Y_n - \mu)$$

Since $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, the result follows. $\square$

> **Remark**
>
> If $\sqrt{n}(Y_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, then $Y_n \xrightarrow{\mathbb{P}} \mu$.

III.   Empirical Distribution Function

> **Definition 12: Dirac Mass**
>
> Dirac mass at $a \in \mathbb{R}$ is the distribution defined by:
> $$\forall A \subset \mathbb{R}, \quad \delta_{\{a\}}(A) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{if } a \notin A \end{cases}$$

Let $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\{X_i\}}$.

> **Definition 13: Empirical CDF**
>
> F_n = Empirical Cumulative Distribution Function (CDF) for $(X_i)_i$ i.i.d. with distribution P_X.
> $$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} = \mathbb{P}_n((-\infty, x])$$
>
> **Remark:** F_n is the CDF of $\mathbb{P}_n$.

> **Theorem 6: Asymptotic Normality of Empirical CDF**
>
> $$\forall x, \quad \sqrt{n}(F_n(x) - F_X(x)) \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, F_X(x)(1 - F_X(x)))$$
>
> Where F_X is the CDF of P_X.

Proof.
$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}_{\{X_i \leq x\}}}_{=Y_i}$$

The variables Y_i are i.i.d. Bernoulli variables:
$$Y_i \sim \text{Ber}(p) \quad \text{with } p = \mathbb{P}(X_i \leq x) = F_X(x)$$

$$\mathbb{E}[Y_i^2] < \infty$$

Applying CLT:

$$\sqrt{n}(\bar{Y}_n - \mathbb{E}(Y_1)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \operatorname{Var}(Y_1))$$

Here $\bar{Y}_n = F_n(x)$ and $\mathbb{E}(Y_1) = F_X(x)$.

$$\operatorname{Var}(Y_1) = F_X(x)(1 - F_X(x))$$

(Q.E.D) $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

More on Quadratic Risk

Expression of Risk using Density:

$$R(\hat{\theta}_n, \theta) = \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2]$$

Under the model $X_i \overset{i.i.d.}{\sim} P_\theta$:

$$= \begin{cases} \displaystyle\int (\hat{\theta}_n(x_1, \dots, x_n) - \theta)^2 \prod_{i=1}^n f_\theta(x_i)\, dx_i & \text{if continuous (density } f_\theta) \\ \displaystyle\sum_{x_1, \dots, x_n} (\hat{\theta}_n(x_1, \dots, x_n) - \theta)^2 \prod_{i=1}^n P_\theta(X = x_i) & \text{if discrete observations} \end{cases}$$

Example: Exponential Family

Example 2 (from beginning): $X_i \overset{i.i.d.}{\sim} \mathcal{E}(\theta), \theta > 0$.

$$X_i \text{ are continuous random variables with density } f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{\{x > 0\}}$$

Estimator:

$$\hat{\theta}_n = \frac{1}{\bar{X}_n}; \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

If $X \sim \mathcal{E}(\theta), \mathbb{E}_\theta(X) = \frac{1}{\theta}$. What is $R(\hat{\theta}_n, \theta)$? What is $b(\theta)$?
We need to calculate $\mathbb{E}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta\left(\frac{1}{\bar{X}_n}\right)$.
Recall:

- $\mathbb{E}(h(X)) = \int h(x) f_X(x) dx$ (if continuous).

- We write: $\mathbb{E}(h(X)) = \int h(x) dF_X(x) = \int h(x) dP_X(x)$.

- If $X_1, \dots, X_n \overset{i.i.d.}{\sim} \Gamma(a, b)$, then $\sum_{i=1}^n X_i \sim \Gamma(na, b)$.

Gamma distribution $\Gamma(a, b)$: $\mathcal{C}^0$ distribution on $\mathbb{R}^+$ with density:

$$f_{a,b}(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}_{\{x > 0\}}$$

Note: $\mathcal{E}(\theta) = \Gamma(1, \theta)$.
$\quad \implies \sum_{i=1}^n X_i \sim \Gamma(n, \theta)$.
Let $Z_n = \sum_{i=1}^n X_i$. Then $\hat{\theta}_n = \frac{n}{Z_n}$.
We want to compute $\mathbb{E}_\theta\left(\frac{n}{Z_n}\right) = n\mathbb{E}_\theta\left(\frac{1}{Z_n}\right)$.

$$\mathbb{E}_\theta\left(\frac{1}{Z_n}\right) = \int_0^{+\infty} \frac{1}{z} \cdot \frac{\theta^n}{\Gamma(n)} z^{n-1} e^{-\theta z} dz$$

$$= \frac{\theta^n}{\Gamma(n)} \int_0^{+\infty} z^{n-2} e^{-\theta z} dz$$

> **Remark**
>
> Let $f_{n-1,\theta}(z) = \frac{\theta^{n-1}}{\Gamma(n-1)} z^{n-2} e^{-\theta z}$ be the density of a $\Gamma(n-1,\theta)$.
>
> $$\implies \int_0^{+\infty} z^{n-2} e^{-\theta z} dz \times \frac{\theta^{n-1}}{\Gamma(n-1)} = 1$$
>
> $$\implies \int_0^{+\infty} z^{n-2} e^{-\theta z} dz = \frac{\Gamma(n-1)}{\theta^{n-1}}$$

Back to the expectation:

$$\mathbb{E}_\theta(\hat{\theta}_n) = \frac{n\theta^n}{\Gamma(n)} \times \frac{\Gamma(n-1)}{\theta^{n-1}}$$

And we have: $\Gamma(x+1) = x\Gamma(x)$ if $x > 0$. So $\Gamma(n) = (n-1)\Gamma(n-1)$.

$$\implies \mathbb{E}_\theta(\hat{\theta}_n) = \frac{n}{n-1}\theta$$

$$\implies b(\theta) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta = \frac{n}{n-1}\theta - \theta = \frac{\theta}{n-1}$$

Now, let's compute the quadratic risk. We need $\mathbb{E}_\theta(\hat{\theta}_n^2) = n^2 \mathbb{E}(\frac{1}{Z_n^2})$.

$$\mathbb{E}\left(\frac{1}{Z_n^2}\right) = \int y^{n-3} e^{-\theta y} dy \frac{\theta^n}{\Gamma(n)}$$

Using the same trick (identification with $\Gamma(n-2,\theta)$) for $n > 2$:

$$= \frac{\theta^n}{\Gamma(n)} \cdot \frac{\Gamma(n-2)}{\theta^{n-2}} = \frac{\theta^2\Gamma(n-2)}{(n-1)(n-2)\Gamma(n-2)} = \frac{\theta^2}{(n-1)(n-2)}$$

$$\implies \mathbb{E}_\theta(\hat{\theta}_n^2) = \frac{n^2\theta^2}{(n-1)(n-2)}$$

$$R(\theta, \hat{\theta}_n) = \underbrace{\frac{n^2\theta^2}{(n-1)(n-2)}}_{\mathbb{E}(\hat{\theta}_n^2)} - \underbrace{\left(\frac{n\theta}{n-1}\right)^2}_{\mathbb{E}(\hat{\theta}_n)^2} + \underbrace{\frac{\theta^2}{(n-1)^2}}_{b(\theta)^2}$$

(Using variance decomposition $\mathbb{V}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n^2) - \mathbb{E}(\hat{\theta}_n)^2$). Actually, let's calculate directly:

$$R(\theta, \hat{\theta}_n) = \frac{n^2\theta^2}{(n-1)(n-2)} - 2\theta\frac{n\theta}{n-1} + \theta^2$$

Simplifying (algebra):

$$= \frac{\theta^2 n^2}{n-1}\left(\frac{1}{n-2} - \frac{1}{n-1}\right) + \frac{\theta^2}{(n-1)^2}$$

$$= \frac{\theta^2}{(n-1)^2}\left[\frac{n^2}{n-2} + 1\right]$$

17

Unbiased estimator

We can construct an unbiased estimator $\tilde{\theta}_n$:

$$\tilde{\theta}_n = \hat{\theta}_n \times \left(\frac{n-1}{n}\right) \implies \mathbb{E}_\theta(\tilde{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) \times \frac{n-1}{n} = \theta.$$

Risk of the unbiased estimator:

$$R(\theta, \tilde{\theta}_n) = \mathbb{V}_\theta(\tilde{\theta}_n) = \left(\frac{n-1}{n}\right)^2 \mathbb{V}_\theta(\hat{\theta}_n)$$

$$= \frac{(n-1)^2}{n^2} \times \frac{n^2\theta^2}{(n-1)^2(n-2)} = \frac{\theta^2}{n-2}$$

We observe that:

$$\forall \theta, R(\theta, \tilde{\theta}_n) < R(\hat{\theta}_n, \theta) \implies \tilde{\theta}_n \text{ is better than } \hat{\theta}_n.$$

> **Remark**
>
> Unbiased estimators are not necessarily better than biased estimators.

> **Example 6: Empirical Variance**
>
> $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.
>
> $$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \quad \text{(Empirical variance)}$$
>
> We know that $S_n^2 = \frac{1}{n} \sum X_i^2 - \bar{X}_n^2$.
>
> $$\mathbb{E}_{\sigma^2}\left(\frac{1}{n} \sum X_i^2\right) = \sigma^2$$
>
> Since $\bar{X}_n \sim \mathcal{N}(0, \frac{\sigma^2}{n})$, $\mathbb{E}_{\sigma^2}(S_n^2) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{n-1}{n}$.
> Define $\tilde{\sigma}_n^2 = S_n^2 \cdot \frac{n}{n-1} = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$.
>
> $$\implies \mathbb{E}_{\sigma^2}(\tilde{\sigma}_n^2) = \sigma^2 \implies \text{Unbiased.}$$
>
> **Exercise:** Compute $R(\sigma^2, S_n^2)$ and $R(\sigma^2, \tilde{\sigma}_n^2)$ and show that $\forall \sigma > 0, R(\sigma^2, S_n^2) < R(\sigma^2, \tilde{\sigma}_n^2)$.
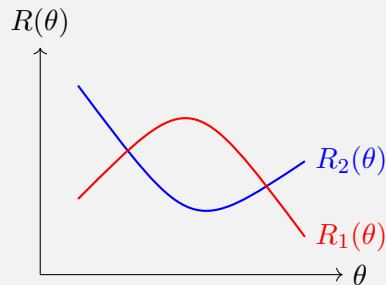
> **Remark**
>
> In most cases, if $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ are 2 estimators of θ, then risk functions $R_1(\theta) = R(\theta, \hat{\theta}_{n,1})$ and $R_2(\theta, \hat{\theta}_{n,2})$ cross.
>
>
>
> $\implies \hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ cannot be compared uniformly.

Example 7: Normal Mean Estimators

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1).$

$$\hat{\theta}_{n,1} = \bar{X}_n$$

$$\hat{\theta}_{n,2} = \frac{n\bar{X}_n}{n + \tau^2} \quad \text{with } \tau > 0$$

Compute their risks and show that they are not comparable.

Admissibility

Definition 14: Admissibility

An estimator $\hat{\theta}_n$ is **not admissible** iff $\exists \tilde{\theta}_n$, another estimator such that:

- $\forall \theta, R(\theta, \tilde{\theta}_n) \leq R(\theta, \hat{\theta}_n)$

- and $\exists \theta_0, R(\theta_0, \tilde{\theta}_n) < R(\theta_0, \hat{\theta}_n)$.

$\hat{\theta}_n$ is **admissible** iff it is not non-admissible.

IV. Consistency

Example 8: Bernoulli Consistency

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Ber}(p). \ \theta = p \in]0,1[. \ \bar{X}_n = \hat{\theta}_n$ is a possible estimator for θ.

$$\bar{X}_n \xrightarrow[n \to \infty]{P_\theta} \theta = \mathbb{E}_\theta(X_1) \quad \text{(LLN)}$$

Question: $R(\theta, \hat{\theta}_n) \xrightarrow[n \to \infty]{} 0, \forall \theta$?

$$R(\theta, \hat{\theta}_n) = \mathbb{E}_\theta[(\bar{X}_n - \theta)^2] = \mathbb{V}_\theta(\bar{X}_n) = \frac{1}{n^2} n \mathbb{V}_\theta(X_1)$$

$$= \frac{\theta(1 - \theta)}{n} \leq \frac{1}{4n}$$

So $\sup_{\theta \in (0,1)} R(\theta, \hat{\theta}_n) \xrightarrow[n \to \infty]{} 0.$
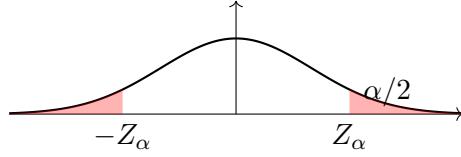
Can I say something like: $\hat{\theta}_{n,1} \leq \theta \leq \hat{\theta}_{n,2}$ with proba ≈ 0.95?

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, \theta(1 - \theta)) \quad \text{(CLT)}$$

$$\implies \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\theta(1 - \theta)}} \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

Since $\bar{X}_n \xrightarrow{\mathbb{P}} \theta$, by Slutsky's lemma (replacing θ with $\bar{X}_n$ in the denominator):

$$\frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

Let Z_α be such that $\mathbb{P}(|\mathcal{N}(0,1)| \le Z_\alpha) = 1 - \alpha$. (Usually $1 - \alpha = 0.95 \implies Z_\alpha \approx 1.96$).

$$\implies \mathbb{P}_\theta \left(-Z_\alpha \le \frac{\sqrt{n}(\bar{X}_n - \theta)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \le Z_\alpha \right) \xrightarrow[n \to \infty]{} 1 - \alpha$$

$$\implies \mathbb{P}_\theta \left(\underbrace{\bar{X}_n - \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)}Z_\alpha}{\sqrt{n}}}_{\hat{\theta}_{n,1}} \le \theta \le \underbrace{\bar{X}_n + \frac{Z_\alpha \sqrt{\bar{X}_n(1 - \bar{X}_n)}}{\sqrt{n}}}_{\hat{\theta}_{n,2}} \right) \to 1 - \alpha$$

With probability under $P_\theta \approx 1 - \alpha$, $\hat{\theta}_{n,1} \le \theta \le \hat{\theta}_{n,2}$.

Definition 15: Consistency

1. We say that an estimator $\hat{\theta}_n$ is **consistent in probability** at θ iff $\hat{\theta}_n \xrightarrow[n \to \infty]{P_\theta} \theta$.

2. $\hat{\theta}_n$ is consistent **over** Θ if it is consistent at $\theta, \forall \theta \in \Theta$.

3. $\hat{\theta}_n$ converges in **quadratic mean** at θ iff $R(\hat{\theta}_n, \theta) \xrightarrow[n \to \infty]{} 0$ (Notation: $\hat{\theta}_n \xrightarrow{q.m} \theta$).

4. $\hat{\theta}_n$ converges in quadratic mean over Θ iff $\hat{\theta}_n \xrightarrow{q.m} \theta, \forall \theta \in \Theta$.

5. $\hat{\theta}_n$ is **asymptotically normal** at θ with rate $\frac{1}{\sqrt{n}}$ iff $\exists V > 0$ s.t. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, V)$.

Estimation of $g(\theta)$

$X_1, \dots, X_n \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$.
  Examples:

1. $P_\theta \sim \text{Ber}(\theta)$. Interested in estimating $\eta = \log(\frac{\theta}{1 - \theta})$.

2. $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2)$. $g_1(\theta) = \mu$. $g_2(\theta) = \mathbb{P}_\theta(X > 1)$.

If $\hat{\theta}$ is an estimator of θ, then we can use $g(\hat{\theta})$ as an estimator of $g(\theta)$ (one possibility among others).

Theorem 7: Plug-in Consistency

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta \subset \mathbb{R}^d$. Let $\hat{\theta}_n$ be an estimator of θ and $g : \Theta \to \mathbb{R}^p$. Then:

  i) If $\hat{\theta}_n$ is consistent at θ in probability and g is $\mathcal{C}^0$ (continuous), then $\hat{\eta}_n = g(\hat{\theta}_n)$ is consistent in probability at $\eta = g(\theta)$.

  ii) If $\hat{\theta}_n \xrightarrow{L^2} \theta$ and g is $\mathcal{C}^0$ and **bounded**, then $g(\hat{\theta}_n) \xrightarrow{L^2(P_\theta)} g(\theta), \forall p \ge 1$.

iii) If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}(P_\theta)} \mathcal{N}(0, \sigma^2)$ and g is $\mathcal{C}^1$ with $g'(\theta) \neq 0$ (or $\nabla g(\theta)$ of full rank), then
$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n\to\infty]{\mathcal{L}(P_\theta)} \mathcal{N}(0, g'(\theta)^2 \sigma^2).$$

Remark

$\hat{\theta}_n$ is unbiased $\nRightarrow g(\hat{\theta}_n)$ is unbiased.

Proof for (ii). $\forall \varepsilon > 0$,
$$\mathbb{E}_\theta(|g(\hat{\theta}_n) - g(\theta)|^p) = \mathbb{E}_\theta(|g(\hat{\theta}_n) - g(\theta)|^p \mathbb{1}_{|g(\hat{\theta}_n) - g(\theta)| > \varepsilon}) + \mathbb{E}_\theta(|g(\hat{\theta}_n) - g(\theta)|^p \mathbb{1}_{|g(\hat{\theta}_n) - g(\theta)| \leq \varepsilon})$$

The second term is $\leq \varepsilon^p$. The first term:
$$\leq \mathbb{E}_\theta((2||g||_\infty)^p \mathbb{1}_{|g(\hat{\theta}_n) - g(\theta)| > \varepsilon})$$
$$= (2||g||_\infty)^p \mathbb{P}_\theta(|g(\hat{\theta}_n) - g(\theta)| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

Because $g(\hat{\theta}_n) \xrightarrow[n\to\infty]{\mathbb{P}} g(\theta)$ (from point (i), continuous mapping theorem).

Thus $\mathbb{E}_\theta(|g(\hat{\theta}_n) - g(\theta)|^p) \leq \varepsilon^p + 2||g||_\infty^p \times (\cdots \to 0)$. Taking limit $n \to \infty$ then $\varepsilon \to 0$ yields the result. $\square$

The Multivariate Case

$X_i \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta \subset \mathbb{R}^d. \theta = (\theta_1, \ldots, \theta_d)$.
$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta \iff \forall \varepsilon > 0, \mathbb{P}_\theta(||\hat{\theta}_n - \theta|| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$
$$\iff \forall j \leq d, \hat{\theta}_{n,j} \xrightarrow{\mathbb{P}_\theta} \theta_j$$

Because:

- If $\forall \varepsilon > 0, \mathbb{P}_\theta(||\hat{\theta}_n - \theta|| > \varepsilon) \to 0$. Let $j \leq d, \forall \varepsilon > 0, ||\hat{\theta}_n - \theta|| \geq |\hat{\theta}_{n,j} - \theta_j|$.
$$\mathbb{P}_\theta(|\hat{\theta}_{n,j} - \theta_j| > \varepsilon) \leq \mathbb{P}_\theta(||\hat{\theta}_n - \theta|| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

- **Reverse:** If $\forall j \leq d, \mathbb{P}_\theta(|\hat{\theta}_{n,j} - \theta_j| > \varepsilon) \to 0$.
$$||\hat{\theta}_n - \theta||^2 = \sum_j |\hat{\theta}_{n,j} - \theta_j|^2 \leq d \cdot \max_j |\hat{\theta}_{n,j} - \theta_j|^2$$
$$\mathbb{P}_\theta(||\hat{\theta}_n - \theta|| > \varepsilon) \leq \mathbb{P}_\theta\left(\max_j |\hat{\theta}_{n,j} - \theta_j| > \frac{\varepsilon}{\sqrt{d}}\right)$$
$$= \mathbb{P}_\theta\left(\bigcup_{j=1}^d \{|\hat{\theta}_{n,j} - \theta_j| > \frac{\varepsilon}{\sqrt{d}}\}\right)$$
$$\leq \sum_{j=1}^d \mathbb{P}_\theta\left(\{|\hat{\theta}_{n,j} - \theta_j| > \frac{\varepsilon}{\sqrt{d}}\}\right)$$
$$\xrightarrow[n\to\infty]{} 0 \quad \text{because } d \text{ is fixed.}$$

$\rightarrow$ Similarly, show that $\hat{\theta}_n \xrightarrow{q.m} \theta \iff \forall j, \hat{\theta}_{n,j} \xrightarrow{q.m} \theta_j$.

But, if $\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) \xrightarrow{\mathcal{L}(P_\theta)} \mathcal{N}(0, \sigma_j^2)$, it does **NOT** imply directly the joint normality.

$$\implies \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}(P_\theta)} \mathcal{N}(\mathbf{0}, \Sigma)$$

(Multivariate Central Limit Theorem required).

Chapter 4: Maximum Likelyhood Estimation

I.   The Likelihood

Model: $X_i \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$.

Observations: $(x_1, \ldots, x_n)$, where x_i is a realisation of X_i.

- If P_θ is **discrete**, then proba mass function of $X_1 \sim P_\theta$:

$$x \in \mathcal{X}, \quad f_\theta(x) = P_\theta(X_1 = x)$$

($\to$ density with respect to counting measure).

- If P_θ is a **continuous** distribution, then it has density f_θ on $\mathbb{R}$ (or $\mathbb{R}^d$) with respect to Lebesgue measure.

$$\int_A f_\theta(x)dx = P_\theta(X \in A)$$

In both cases, there is a density, with respect to either Lebesgue or counting measure, denoted f_θ.

Definition 16: Likelihood

Let $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$ with $\{P_\theta, \theta \in \Theta\}$ either discrete or continuous. Let f_θ be the density of P_θ (or probability mass function). We call the **likelihood** at $(x_1, \ldots, x_n)$ (realisations of $(X_1, \ldots, X_n)$) the function:

$$L_n : \Theta \to \mathbb{R}_+$$

$$\theta \mapsto \prod_{i=1}^n f_\theta(x_i)$$

Examples

1. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{P}(\lambda), \lambda > 0$ (Poisson). Observations $(x_1, \ldots, x_n), \forall x \in \mathbb{N}, f_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

$$L_n(\lambda) = \prod_{i=1}^n f_\lambda(x_i) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

2. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}(\theta), \theta > 0$. Observations $(x_1, \ldots, x_n), x_i > 0, \forall i$.

$$\forall x > 0, f_\theta(x) = \theta e^{-\theta x}$$

$$L_n(\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} \mathbb{1}_{\{x_i>0\}}$$

$$= \theta^n e^{-\theta \sum_{i=1}^n x_i} \mathbb{1}_{\{\min_i x_i > 0\}}$$

> **Remark**
>
> In the **discrete** case, $\forall \theta, L_n(\theta) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)$.
>
> But, in the **continuous** case, it is different:
>
> $$L_n(\theta) = \prod_{i=1}^n f_\theta(x_i) \neq \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = 0$$
>
> (Probability of exact points in continuous case is 0).

If $L_n(\theta)$ is large, it means that the observations are highly likely for θ. "There is a good fit between $(x_1, \ldots, x_n)$ and P_θ".

Exercise 1: Show that if $P_\theta \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, then:

$$L_n(\theta) = \frac{e^{-n\frac{(\bar{x}_n - \mu)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n \frac{(x_i - \bar{x}_n)^2}{2\sigma^2}}$$

II. Kullback-Leibler Divergence (K.L)

2nd interpretation: Kullback-Leibler divergence.

> **Definition 17: Kullback-Leibler Divergence**
>
> Let P and Q be 2 probabilities.
>
> - If P, Q are **discrete** on $\mathcal{X}$, then:
>
> $$KL(P, Q) = \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right)$$
>
> $$= \mathbb{E}_P\left(\ln\left(\frac{P(X)}{Q(X)}\right)\right)$$
>
> (where $P(x) = \mathbb{P}(X = x)$, $Q(x) = \mathbb{Q}(X = x)$).
>
> - If P and Q are **continuous** with density f_P and f_Q (resp), then:
>
> $$KL(P, Q) = \int_{\mathcal{X}} f_P(x) \ln\left(\frac{f_P(x)}{f_Q(x)}\right) dx$$
>
> $$= \mathbb{E}_P\left(\ln\left(\frac{f_P(X)}{f_Q(X)}\right)\right)$$

> **Proposition 2: Properties of KL**
>
>   i) $\forall P, Q, \quad KL(P, Q) \geq 0$.
>
>   ii) $KL(P, Q) = 0 \iff P = Q$.
>
> **But**, $KL(P, Q) \neq KL(Q, P)$ (Not symmetric).

Proof. Assume that P, Q are continuous.

i)

$$KL(P, Q) = \int_{\mathcal{X}} f_P(x) \ln\left(\frac{f_P(x)}{f_Q(x)}\right) dx$$

$$= \int_{\mathcal{X}} f_P(x) \left[-\ln\left(\frac{f_Q(x)}{f_P(x)}\right)\right] dx$$

By Jensen's inequality:

$$\geq -\ln\left[\int_{\mathcal{X}} f_P(x) \frac{f_Q(x)}{f_P(x)} dx\right]$$

$$= -\ln\left[\int_{\mathcal{X}} f_Q(x) dx\right] = -\ln(1) = 0$$

(Because $x \mapsto -\ln(x)$ is strictly convex on $\mathbb{R}$).

ii) Since $-\ln(\cdot)$ is strictly convex:

$$\mathbb{E}_P\left(-\ln(R(X))\right) > -\ln(\mathbb{E}_P(R(X))) \quad \text{with } R(x) = \frac{f_Q(x)}{f_P(x)}$$

if $\forall c, \mathbb{P}(R(X) = c) < 1$ ($R(X)$ is not constant).

$R(x) = \frac{f_Q(x)}{f_P(x)}$ is constant iff $\exists c \neq 0$ so that:

$$f_Q(x) = c f_P(x), \quad \forall x \in \mathcal{X}$$

$$\implies c = 1 \quad \left(\text{since } \int f_P = \int f_Q = 1\right)$$

$R(x) = \frac{f_Q(x)}{f_P(x)}$ is constant iff $f_Q = f_P \iff P = Q$. $\qquad\square$

Links with the likelihood

$\forall \theta, \theta_0 \in \Theta$:

$$\ln(L_n(\theta)) - \ln(L_n(\theta_0))$$

$$= \ln\left(\prod_{i=1}^{n} f_\theta(x_i)\right) - \ln\left(\prod_{i=1}^{n} f_{\theta_0}(x_i)\right)$$

$$= \sum_{i=1}^{n} \ln(f_\theta(x_i)) - \sum_{i=1}^{n} \ln(f_{\theta_0}(x_i))$$

$$= \sum_{i=1}^{n} \ln\left(\frac{f_\theta(x_i)}{f_{\theta_0}(x_i)}\right)$$

If $(x_i)_{i=1}^{n}$ are the realisations of $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_{\theta_0}$. Then when $X_i \overset{i.i.d.}{\sim} P_{\theta_0}$:

$$\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \underbrace{\text{converges under } P_{\theta_0}}_{n \to \infty} \text{ to } \mathbb{E}_{\theta_0}\left(\ln\left(\frac{f_\theta(X)}{f_{\theta_0}(X)}\right)\right)$$

as soon as $\mathbb{E}_{\theta_0}\left(\left|\ln\frac{f_\theta(X)}{f_{\theta_0}(X)}\right|\right) < \infty$.

$$\implies \text{if } \mathbb{E}_{\theta_0}(\ldots) < \infty, \text{ then we have:}$$

$$\frac{1}{n}\left(\ln(L_n(\theta)) - \ln(L_n(\theta_0))\right) \xrightarrow[n \to \infty]{P_{\theta_0}} \mathbb{E}_{\theta_0}\left[\ln\left(\frac{f_\theta(X)}{f_{\theta_0}(X)}\right)\right]$$

$$= -KL(P_{\theta_0}, P_\theta)$$

- θ_0 minimizes $KL(P_{\theta_0}, P_\theta)$ over θ.

- θ_0 maximizes $-KL(P_{\theta_0}, P_\theta)$ over θ.

We cannot maximize $-KL(P_{\theta_0}, P_\theta)$ because it depends on θ_0 unknown. $\implies$ We maximize:

$$\mathbb{E}_{P_n}\left[-\ln\left(\frac{f_\theta(X)}{f_{\theta_0}(X)}\right)\right] = \frac{1}{n}[\ln(L_n(\theta)) - \ln(L_n(\theta_0))]$$

(Empirical version of $-KL(P_{\theta_0}, P_\theta)$).

Maximizing in θ, $\ln(L_n(\theta)) - \ln(L_n(\theta_0))$ is equivalent to maximizing in θ, $\ln L_n(\theta)$.

Definition 18: Maximum Likelihood Estimator (MLE)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$, the model P_θ with density (or probability mass function) f_θ. We call **Maximum Likelihood Estimator** any selection $\hat\theta_n$ (when it exists) such that:

$$\hat\theta_n \in \underset{\theta \in \Theta}{\operatorname{argmax}}\{\ln(L_n(\theta))\}$$

$$\iff \log(L_n(\hat\theta_n)) \geq \log(L_n(\theta)), \forall \theta \in \Theta$$

Examples of Calculations

1) Bernoulli Model

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathrm{Ber}(\theta)$ $(\theta \in {]0, 1[})$.

$$L_n(\theta) = \prod_{i=1}^n P_\theta(X_i = x_i)$$

$$= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Define $l(\theta) = \log(L_n(\theta))$.

$$l(\theta) = \left(\sum_{i=1}^n x_i\right)\log(\theta) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\theta)$$

$\hat\theta_n$ maximizes $l(\theta)$ in θ. Let $\bar{x}_n = \frac{1}{n}\sum x_i$.

$$l(\theta) = n\bar{x}_n \log(\theta) + n(1-\bar{x}_n)\log(1-\theta)$$

Derivative:

$$l'(\theta) = \frac{n\bar{x}_n}{\theta} - \frac{n(1-\bar{x}_n)}{1-\theta}$$

Second derivative:

$$l''(\theta) = -\frac{n\bar{x}_n}{\theta^2} - \frac{n(1-\bar{x}_n)}{(1-\theta)^2} < 0$$

So $\theta \mapsto l(\theta)$ is **concave**. $\implies \hat\theta_n$ is the solution of $l'(\theta) = 0$.

$$l'(\theta) = 0 \iff \frac{n\bar{x}_n}{\theta} = \frac{n(1-\bar{x}_n)}{1-\theta} \iff \theta = \bar{x}_n$$

$$\implies \hat\theta_n = \bar{x}_n \text{ is the MLE (Maximum Likelihood Estimator)}$$

2) Exponential Model

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}(\theta), \theta > 0.$

$$X \sim \mathcal{E}(\theta) \iff f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{\{x>0\}}$$

$$L_n(\theta) = \theta^n e^{-n\theta \bar{x}_n} \prod_{i=1}^n \mathbb{1}_{\{x_i>0\}}$$

$$l(\theta) = \log(L_n(\theta)) = n\log(\theta) - n\bar{x}_n\theta \quad (\text{if } \forall i, x_i > 0)$$

Exercise: Show that $\theta \mapsto l(\theta)$ is concave and that $\hat{\theta}_n = \frac{1}{\bar{x}_n}$.

> **Remark**
>
> 1. The log likelihood as a function is not necessarily concave, therefore, $\hat{\theta}_n$ does not always exist and is not always unique.
>
> 2. In practice: how do we compute $\hat{\theta}_n$?
>
> $$\hat{\theta}_n \in \text{argmax}\{l(\theta), \theta \in \Theta\}$$
>
> Algorithms such as gradient descent, Newton-Raphson.

Summary:

- **Model:** $f_\theta(\cdot), \theta \in \Theta$ density or proba mass function. $X_i \overset{i.i.d.}{\sim} f_\theta$.

- **Log-likelihood:** $l(\theta) = \log(L_n(\theta)) = \sum_{i=1}^n \log(f_\theta(x_i))$.

- **MLE:** $\hat{\theta}_n$ such that $l(\hat{\theta}_n) \geq l(\theta), \forall \theta \in \Theta$.

> **Proposition 3: Invariance**
>
> If $\eta = g(\theta)$, where g is invertible ($\iff$ new parametrization).
> For instance: $X_i \overset{i.i.d.}{\sim} \text{Ber}(p), \theta = p \in]0, 1[$.
>
> $$\eta = \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$$
>
> Let $\tilde{l}(\eta) = \sum_{i=1}^n \log(f_{g^{-1}(\eta)}(x_i))$. Let $l(\theta) = \sum_{i=1}^n \log(f_\theta(x_i))$.
>
> - If $\hat{\eta}_n$ is the MLE of $\tilde{l}$ ($\iff \tilde{l}(\eta) \leq \tilde{l}(\hat{\eta}_n), \forall \eta$).
>
> - If $\hat{\theta}_n$ is the MLE of l ($\iff l(\theta) \leq l(\hat{\theta}_n), \forall \theta$).
>
> Then:
>
> $$\hat{\eta}_n = g(\hat{\theta}_n)$$

Proof. Note: $\hat{\eta}_n = g(\hat{\theta}_n)$. We verify that $\tilde{l}(\hat{\eta}_n) \geq \tilde{l}(\eta), \forall \eta$, that is, it will imply that $\hat{\eta}_n$ is an MLE of $\tilde{l}$.
Indeed:

$$\tilde{l}(\eta) = l(g^{-1}(\eta)) \leq l(\hat{\theta}_n), \forall \eta$$

And:

$$\tilde{l}(\hat{\eta}_n) = l(g^{-1}(g(\hat{\theta}_n))) = l(\hat{\theta}_n)$$

$$\implies \forall \eta, \tilde{l}(\eta) \leq \tilde{l}(\hat{\eta}_n). \quad \text{(Q.E.D)}$$

$\square$

Examples

MLE for $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Ber}(p)$.

$$l(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$x_1, \ldots, x_n$ are the observations.

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \to \text{MLE}$$

$$\eta = \log\left(\frac{p}{1-p}\right) \implies \hat{\eta}_n = \log\left(\frac{\hat{p}_n}{1-\hat{p}_n}\right) \text{ is the MLE of } \eta.$$

Asymptotic Normality of the MLE

Under which conditions on the model is the MLE asymptotically normal?

> **Definition 19: Regular Model**
>
> The model $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_\theta, \theta \in \Theta \subset \mathbb{R}^k$ (either density or probability mass function) is **regular** on Θ iff:
>
> (i) Θ is an open set and the model is identifiable.
>
> (ii) $\theta \mapsto \log(f_\theta(x))$ is $\mathcal{C}^2, \forall x \in \mathcal{X}$.
>
> (iii) $\forall \theta \in \Theta, \mathbb{E}_\theta(||\nabla_\theta \log(f_\theta(X))||^2) < +\infty$ and $\exists \delta > 0, \mathbb{E}_\theta(\sup_{|\theta'-\theta|<\delta} ||D_\theta^2 \log(f_{\theta'}(X))||) < +\infty$.
>
> (iv) $\forall \theta, I(\theta) = \mathbb{E}_\theta\left(\nabla_\theta \log(f_\theta(X)) \cdot \nabla_\theta \log(f_\theta(X))^T\right)$ is positive definite.

Fisher Information

$I(\theta)$ is called the **Fisher Information Matrix**.

 Notations: $D^2 \log(f_\theta(x))$ is the matrix:

$$(D^2 \log(f_\theta(x)))_{i,j} = \frac{\partial^2 \log(f_\theta(x))}{\partial \theta_i \partial \theta_j}$$

$$\nabla_\theta \log(f_\theta(x)) = \left(\frac{\partial}{\partial \theta_j}\log(f_\theta(x))\right)_{j=1,\ldots,k}$$

> **Lemma 1: Properties of Score and Fisher Information**
>
> If the model is regular:
>
> a) $\forall \theta \in \Theta, \mathbb{E}_\theta(\nabla_\theta \log(f_\theta(X))) = 0$. $S(\theta, x) = \nabla_\theta \log(f_\theta(x))$ is called the **score function**.
>
> b) $I(\theta) = -\mathbb{E}_\theta(D^2 \log(f_\theta(X)))$.

Proof. **a)** The model is regular $\implies \nabla_\theta \log(f_\theta(X))$ exists and is integrable. And:

$$\nabla_\theta \log(f_\theta(X)) = \frac{\nabla_\theta f_\theta(X)}{f_\theta(X)}$$

$$\mathbb{E}_\theta(\nabla_\theta \log(f_\theta(X))) = \begin{cases} \displaystyle\int_\mathcal{X} \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)dx & \text{if } X_i\text{'s are continuous} \\ \displaystyle\sum_{x \in \mathcal{X}} \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x) & \text{if } X_i\text{'s are discrete} \end{cases}$$

We have, if the X_i's are continuous:

$$\int_\mathcal{X} \frac{\partial}{\partial \theta_j} f_\theta(x)dx = \frac{\partial}{\partial \theta_j} \int_\mathcal{X} f_\theta(x)dx$$

Because $\mathbb{E}_\theta(||\nabla_\theta f_\theta(X)||) < +\infty$. Also $\forall \theta, \int_\mathcal{X} f_\theta(x)dx = 1 \implies \frac{\partial}{\partial \theta_j}(\int f_\theta(x)dx) = 0$.

So $\mathbb{E}_\theta\left(\frac{\partial}{\partial \theta_j} \frac{f_\theta(X)}{f_\theta(X)}\right) = 0$.

$$\implies \mathbb{E}_\theta(\nabla_\theta \log(f_\theta(X))) = 0$$

If f is discrete $\to$ same argument.

$$\sum_x \frac{\partial}{\partial \theta_j} f_\theta(x) = \frac{\partial}{\partial \theta_j} \sum_x f_\theta(x) = \frac{\partial}{\partial \theta_j}(1) = 0$$

(Q.E.D)

b) Show that $I(\theta) = -\mathbb{E}_\theta(D_\theta^2 \log(f_\theta(X)))$.

$$\frac{\partial^2 \log(f_\theta(x))}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_j}\left(\frac{\frac{\partial}{\partial \theta_i} f_\theta(x)}{f_\theta(x)}\right)$$

$$= \frac{\frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(x)}{f_\theta(x)} - \left(\frac{\frac{\partial}{\partial \theta_i} f_\theta(x)}{f_\theta(x)}\right)\left(\frac{\frac{\partial}{\partial \theta_j} f_\theta(x)}{f_\theta(x)}\right)$$

Similarly to before:

$$\mathbb{E}_\theta\left(\frac{\frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(X)}{f_\theta(X)}\right) = \int_\mathcal{X} \frac{\frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(x)}{f_\theta(x)} f_\theta(x)dx$$

$$= \int_\mathcal{X} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(x)dx = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_\mathcal{X} f_\theta(x)dx = 0$$

$$J(\theta)_{i,j} = -\mathbb{E}_\theta(D_\theta^2 \log(f_\theta(X)))$$

$$= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f_\theta(X))\right)$$

$$= -\left(-\mathbb{E}_\theta\left(\frac{\frac{\partial f}{\partial \theta_i}}{f} \cdot \frac{\frac{\partial f}{\partial \theta_j}}{f}\right)\right)$$

$$= I_{i,j}(\theta)$$

So $J(\theta) = I(\theta)$. (Q.E.D) $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

Theorem 8: Asymptotic Normality of MLE

Let $(x_1, \ldots, x_n)$ be observations from the model $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_\theta, \theta \in \Theta \subset \mathbb{R}^k$. Assume the model is **regular**.

(i) **Score Convergence:** For all $\theta \in \Theta$, let $S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta \log(f_\theta(X_i))$. Then:

$$S_n(\theta) \xrightarrow[n\to\infty]{\mathcal{L}_{P_\theta}} \mathcal{N}(0, I(\theta))$$

(ii) **MLE Convergence:** If in addition the MLE $\hat{\theta}_n$ is **consistent** over Θ (in probability), then:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}_{P_\theta}} \mathcal{N}(0, I^{-1}(\theta))$$

$\implies I^{-1}(\theta)$ is the asymptotical variance of $\sqrt{n}(\hat{\theta}_n - \theta)$.

Remark

Cramer-Rao Lower Bound: For the variance of a regular estimator is $\frac{I^{-1}(\theta)}{n}$.

Multivariate Gaussians Reminder

$X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^k$, $\mu \in \mathbb{R}^k$. Σ symmetric, positive semi-definite.
$\iff$ its density (w.r.t Lebesgue)

$$f_{\mu,\Sigma}(x) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{(2\pi)^{k/2}(\det(\Sigma))^{1/2}}$$

$\iff \forall a \in \mathbb{R}^d, a^T X \sim \mathcal{N}(a^T \mu, a^T \Sigma a)$.

<u>Proof of the Theorem</u>

Objective: Show that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}_{P_\theta}} \mathcal{N}(0, I^{-1}(\theta))$.

$\hat{\theta}_n$ is defined by $l(\hat{\theta}_n) \geq l(\theta)$ and verifies $\nabla_\theta l(\hat{\theta}_n) = 0$. The rest of the proof is done with $k = 1$.
$\theta \mapsto l(\theta) = \sum \log(f_\theta(x_i))$ is $\mathcal{C}^2$.

Taylor expansion:

$$l'(\theta) = l'(\hat{\theta}_n) + (\theta - \hat{\theta}_n)l''(\tilde{\theta}_n) \quad \text{with } \tilde{\theta}_n \in]\theta, \hat{\theta}_n[$$

$$\implies l'(\theta) = (\theta - \hat{\theta}_n)l''(\tilde{\theta}_n) \quad (\text{since } l'(\hat{\theta}_n) = 0)$$

$$\frac{l''(\tilde{\theta}_n)}{n} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log(f_{\tilde{\theta}_n}(x_i))$$

$$\frac{l''(\tilde{\theta}_n) - l''(\theta)}{n} = \Delta_n(\theta)$$

So:

$$l'(\theta) = (\theta - \hat{\theta}_n)(l''(\theta) + \Delta_n(\theta) \times n)$$

And $\frac{l'(\theta)}{\sqrt{n}} = -S_n(\theta)$.

$$S_n(\theta) = \sqrt{n}(\theta - \hat{\theta}_n)\left(\frac{l''(\theta)}{n} + \Delta_n(\theta)\right)$$

We examine the terms:

$$\frac{l''(\theta)}{n} = \frac{1}{n}\sum_{i=1}^{n} D^2\log(f_\theta(x_i)) \xrightarrow[n\to\infty]{P_\theta} \mathbb{E}_\theta(D^2\log(f_\theta(X))) = -I(\theta)$$

$$\implies S_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta)(-I(\theta) + \Delta_n(\theta) + B_n(\theta))$$

where $B_n(\theta) = \frac{l''(\theta)}{n} + I(\theta) \xrightarrow{P_\theta} 0$.

We need to handle $S_n(\theta)$ and $\Delta_n(\theta)$.

(i) Behavior of $S_n(\theta)$:

$$\forall \theta \in \Theta, S_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta\log(f_\theta(x_i)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta))$$

Proof for (i):

$$S_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_1(\theta, x_i)$$

And $\forall\theta, (s_1(\theta, x_i))_{i=1,\dots,n}$ are i.i.d. because (x_i) are i.i.d. And $\mathbb{E}_\theta(s_1(\theta, X_i)) = 0$ (Lemma). And $\mathbb{E}_\theta(||s_1(\theta, X)||^2) < +\infty$ (Assumption).

By the CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_1(\theta, X_i) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}_\theta(s_1(\theta, X)))$$

Also $\mathrm{Var}_\theta(s_1(\theta, X)) = \mathbb{E}_\theta(s_1(\theta, X)s_1(\theta, X)^T) = I(\theta)$.

(ii) Behavior of $\Delta_n(\theta)$: We show that $\Delta_n(\theta) \xrightarrow[n\to\infty]{P_\theta} 0$. Because when $|\hat{\theta}_n - \theta| \leq \delta$:

$$|\Delta_n(\theta)| \leq \frac{1}{n}\sum \sup_{|\theta'-\theta|\leq\delta} |D^2_{\theta'}\log(f_{\theta'}(x_i)) - D^2_\theta\log(f_\theta(x_i))|$$

LLN:

$$\frac{1}{n}\sum \sup_{|\theta'-\theta|\leq\delta} |D^2_{\theta'}\log(f_{\theta'}(x_i)) - D^2_\theta\log(f_\theta(x_i))|$$

$$\xrightarrow[P_\theta]{} \mathbb{E}_\theta\left(\sup_{|\theta'-\theta|\leq\delta} |D^2_{\theta'}\log(f_{\theta'}(x_i)) - D^2_\theta\log(f_\theta(x))|\right)$$

By the dominated convergence theorem:

$$\mathbb{E}_\theta\left(\sup_{|\theta'-\theta|\leq\delta} |D^2_{\theta'}\log(f_{\theta'}(x)) - D^2_\theta\log(f_\theta(x))|\right) \xrightarrow[\delta\to 0]{} 0$$

$\forall\varepsilon > 0, \exists\delta_\varepsilon > 0, \forall\delta \leq \delta_\varepsilon$, this expectation is $< \varepsilon$.

We know that $\hat{\theta}_n \xrightarrow{P_\theta} \theta$. $\forall\delta > 0, \mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \delta) \xrightarrow[n\to\infty]{} 0$.

Also $\theta \mapsto D^2_\theta\log(f_\theta(x))$ is continuous. $\implies \limsup_{\delta\to 0}\sup_{|\theta'-\theta|\leq\delta}|D^2_{\theta'}\log(f_{\theta'}(x)) - D^2_\theta\log(f_\theta(x))| = 0$ (almost surely). Let $\delta_p \to 0, \delta_p \downarrow$.

$$\sup_{|\theta'-\theta|\leq\delta_p} |D^2_{\theta'}\log(f_{\theta'}(x)) - D^2_\theta\log(f_\theta(x))|$$

$$\leq \sup_{|\theta'-\theta|\leq\delta_0} |D^2_{\theta'}\log(f_{\theta'}(x)) - D^2_\theta\log(f_\theta(x))|$$

And $\mathbb{E}_\theta(H(X)) < +\infty$ (by assumption).

31

So $\Delta_n(\theta) \xrightarrow{P_\theta} 0$.
Conclusion of Proof:

$$S_n(\theta) = \sqrt{n}(\hat{\theta}_n - \theta)(-I(\theta) + \Delta_n(\theta) + B_n(\theta))$$

$$B_n(\theta) = \frac{l''(\theta)}{n} + I(\theta) \xrightarrow{P_\theta} 0$$

$$\Delta_n(\theta) \xrightarrow{P_\theta} 0$$

$$S_n(\theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I(\theta))$$

$$\implies \sqrt{n}(\hat{\theta}_n - \theta) = \frac{S_n(\theta)}{I(\theta)(1 + \mu_n)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{I(\theta)}{I^2(\theta)}\right) = \mathcal{N}(0, I^{-1}(\theta))$$

With $\mu_n \xrightarrow{P_\theta} 0$.

III. <u>Exponential Families</u>

Definition 20: Exponential Family

Model $X \in \mathbb{R}^d \sim P_\theta, \theta \in \Theta \subset \mathbb{R}^k$ is an **exponential family** iff P_θ has density (or proba mass function) f_θ and:

- $\exists h : \mathbb{R}^d \to \mathbb{R}_+$

- $\exists A : \Theta \to \mathbb{R}^k$

- $\exists S : \mathbb{R}^d \to \mathbb{R}^k$

- $\exists C : \Theta \to \mathbb{R}$

Such that:
$$f_\theta(x) = h(x)\exp\left(A(\theta)^T S(x) - C(\theta)\right)$$

It's a **canonical exponential family** iff $A(\theta) = \theta$.

Remark

Cramer-Rao Lower Bound: For the variance of a regular estimator, the lower bound is $\frac{I^{-1}(\theta)}{n}$. What are regular models? Exponential families are a prime example.

Theorem 9: Properties of Exponential Families

Let P_θ be an exponential family. If $f_\theta(x) = h(x)\exp(A(\theta)^T S(x) - C(\theta))$ and if $\Theta = \{\theta, f_\theta(x) \text{ exists}\}$ is an open set, then:

(i) If $A(\theta) = \theta$, then the model is **regular**.

(ii) $\exp(C(\theta)) = \int_{\mathbb{R}^d} h(x)e^{A(\theta)^T S(x)}dx$ (if continuous) or $\sum_{x\in\mathcal{X}} h(x)e^{A(\theta)^T S(x)}$ (if discrete).

(iii) If $A(\theta) = \theta$, then $\theta \mapsto C(\theta)$ is $\mathcal{C}^\infty$ on Θ and:

$$\mathbb{E}_\theta(S(X)) = \nabla_\theta C(\theta)$$

$$\mathbb{V}_\theta(S(X)) = \frac{\partial^2 C(\theta)}{\partial \theta^2} = D^2 C(\theta)$$

(iv) If A is invertible and $\mathcal{C}^2$, and A^{-1} is $\mathcal{C}^2$, then the model is regular.

Remark

"f_θ exists" means:

- If P_θ is continuous: $\int_{\mathbb{R}^d} e^{A(\theta)^T S(x)} h(x) dx < +\infty$.

- If P_θ is discrete: $\sum_{x \in \mathcal{X}} e^{A(\theta)^T S(x)} h(x) < +\infty$.

Heuristic of the proof (Continuous case)

(ii) We know that $\int_{\mathbb{R}^d} f_\theta(x) dx = 1$.

$$\iff \int_{\mathbb{R}^d} h(x) e^{A(\theta)^T S(x) - C(\theta)} dx = 1$$

$$\iff \int_{\mathbb{R}^d} h(x) e^{A(\theta)^T S(x)} dx = e^{C(\theta)}$$

(iii) If $A(\theta) = \theta$.

$$e^{C(\theta)} = \int_{\mathbb{R}^d} e^{\theta^T S(x)} h(x) dx$$

And $\{\theta, \int_{\mathbb{R}^d} e^{\theta^T S(x)} h(x) dx < +\infty\}$ is open.

Differentiation:

$$\frac{\partial}{\partial \theta} e^{\theta^T S(x)} = S(x) e^{\theta^T S(x)}$$

If $k = 1$:

$$\frac{d}{d\theta} e^{C(\theta)} = C'(\theta) e^{C(\theta)} = \int_{\mathbb{R}^d} S(x) e^{\theta S(x)} h(x) dx$$

$$\implies C'(\theta) = e^{-C(\theta)} \int_{\mathbb{R}^d} S(x) e^{\theta S(x)} h(x) dx$$

$$= \int_{\mathbb{R}^d} S(x) \underbrace{e^{\theta S(x) - C(\theta)} h(x)}_{f_\theta(x)} dx = \mathbb{E}_\theta(S(X))$$

Second derivative:

$$\frac{d^2}{d\theta^2} e^{C(\theta)} = C''(\theta) e^{C(\theta)} + (C'(\theta))^2 e^{C(\theta)} = \int_{\mathbb{R}^d} S(x)^2 e^{\theta S(x)} h(x) dx$$

$$\implies C''(\theta) = - \underbrace{C'(\theta)^2}_{\mathbb{E}_\theta(S(X))^2} + \underbrace{\int_{\mathbb{R}^d} S(x)^2 e^{\theta S(x) - C(\theta)} h(x) dx}_{\mathbb{E}_\theta(S(X)^2)}$$

$$= \mathbb{E}_\theta(S(X)^2) - \mathbb{E}_\theta(S(X))^2 = \mathbb{V}_\theta(S(X))$$

> **Remark**
>
> Using properties of the log-likelihood:
>
> $$\frac{d^2}{d\theta^2} \log(f_\theta(x)) = -\frac{d^2}{d\theta^2} C(\theta)$$
>
> Or $I(\theta) = -\mathbb{E}_\theta \left(\frac{d^2}{d\theta^2} \log f_\theta(X) \right) = \frac{d^2}{d\theta^2} C(\theta)$.

Justification for differentiation under the integral sign (Proof details): If $\theta \in \Theta$, $\theta + \varepsilon \in \Theta$ and $\theta - \varepsilon \in \Theta$ (since open).

$$\implies \int e^{(\theta+\varepsilon)S(x)} h(x)dx < +\infty \quad \text{and} \quad \int e^{(\theta-\varepsilon)S(x)} h(x)dx < +\infty$$

$$|S(x)|e^{\theta S(x)} \leq e^{\theta S(x)} + \varepsilon|S(x)|e^{\theta S(x)} \dots$$

Wait, actually we use convexity arguments.

$$|S(x)|e^{\theta S(x)} \leq e^{(\theta+\varepsilon)S(x)} + e^{(\theta-\varepsilon)S(x)}$$

This bound (integrable function) allows the use of the Dominated Convergence Theorem.

$$\implies \int |S(x)|e^{\theta S(x)} h(x)dx < +\infty$$

> **Proposition 4: Stability of Exponential Families**
>
> If $\{P_\theta, \theta \in \Theta\}$ is an exponential family and $X_1, \dots, X_n \overset{i.i.d.}{\sim} P_\theta$, then the joint distribution is also an exponential family.
>
> $$f_\theta^{(n)}(x_1, \dots, x_n) = e^{A(\theta)^T \sum_{i=1}^{n} S(x_i) - nC(\theta)} \prod_{i=1}^{n} h(x_i)$$
>
> This is an exponential family with:
>
> $$A_n(\theta) = A(\theta), \quad S_n(x_1, \dots, x_n) = \sum_{i=1}^{n} S(x_i), \quad C_n(\theta) = nC(\theta)$$

Proof.

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

$$= \prod_{i=1}^{n} \left(h(x_i) e^{A(\theta)^T S(x_i) - C(\theta)} \right)$$

$$= \left(\prod_{i=1}^{n} h(x_i) \right) e^{A(\theta)^T \sum S(x_i) - nC(\theta)}$$

$\square$

Examples

Example 9: Binomial Distribution

$X \sim \text{Bin}(n, p), p = \theta$. Probability mass function:

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \dots, n\}$$

$$= \underbrace{\binom{n}{x}}_{h(x)} \exp \left(\underbrace{\log\left(\frac{p}{1-p}\right)}_{A(p)} x + \underbrace{n \log(1-p)}_{-C(p)} \right)$$

Here $S(x) = x$.

Canonical Parametrisation:

$$\eta = \log\left(\frac{p}{1-p}\right) \implies p = \frac{e^\eta}{1+e^\eta}$$

$$\tilde{f}_\eta(x) = \binom{n}{x} \exp\left(\eta x - n \log(1+e^\eta)\right)$$

With $C(\eta) = n \log(1 + e^\eta)$.

$$\Theta = \mathbb{R} \text{ (open)}$$

Moments check:

$$\mathbb{E}_\eta(X) = C'(\eta) = n\frac{e^\eta}{1+e^\eta} = np$$

$$\mathbb{V}_\eta(X) = C''(\eta) = n\frac{e^\eta(1+e^\eta) - e^\eta e^\eta}{(1+e^\eta)^2} = n\frac{e^\eta}{(1+e^\eta)^2} = np(1-p)$$

Remark: In a canonical exponential family, if Θ is an open set, then $I(\theta) = \frac{d^2 C(\theta)}{d\theta^2}$.

Proof:

$$\log(f_\theta(x)) = \theta^T S(x) - C(\theta) + \log(h(x))$$

$$\implies \frac{d}{d\theta} \log(f_\theta(x)) = S(x) - \nabla C(\theta)$$

$$\implies \frac{d^2}{d\theta^2} \log(f_\theta(x)) = -\frac{d^2}{d\theta^2} C(\theta)$$

$$\implies I(\theta) = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2} \log(f_\theta(X))\right] \quad \text{(see Lemma 1)}$$

$$= \frac{d^2}{d\theta^2} C(\theta)$$

Example 10: Exponential Distribution

$X \sim \mathcal{E}(\theta), \theta > 0$.

$$f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{x>0} = \mathbb{1}_{x>0} e^{-\theta x + \log(\theta)}$$

$h(x) = \mathbb{1}_{x>0}$. $A(\theta) = \theta$ (or $-\theta$). $S(x) = -x$. $C(\theta) = -\log(\theta)$. Canonical exponential family.

IV. Sufficient Condition for Consistency of MLE

> **Theorem 10: Consistency of MLE**
>
> If Θ is compact,
>
> If (i) $\displaystyle\sup_{\theta \in \Theta} \left| \frac{\ln(L_n(\theta)) - \ln(L_n(\theta_0))}{n} + KL(\theta_0, \theta) \right| \xrightarrow[n \to \infty]{P_{\theta_0}} 0$
>
> If (ii) $\forall \varepsilon > 0, \exists \delta > 0$ such that $\displaystyle\inf_{|\theta - \theta_0| > \varepsilon} KL(\theta_0, \theta) > \delta$.
>
> Then the MLE is consistent.
>
> **Where:**
> $$KL(\theta_0, \theta) = \mathbb{E}_{\theta_0}[\log f_{\theta_0}(X) - \log f_{\theta}(X)]$$
>
> (Kullback-Leibler divergence between f_{θ_0} and f_{θ}).

Proof.
$$KL(\theta_0, \theta) = \mathbb{E}_{\theta_0}[\log(f_{\theta_0}(X)) - \log(f_{\theta}(X))]$$

Idea of the proof: Prove that $\forall \varepsilon > 0, P_{\theta_0}\left[\sup_{|\theta - \theta_0| > \varepsilon} \ln(\theta) < \sup_{\theta \in \Theta} \ln(\theta)\right] \to 1$ as $n \to \infty$. (where $\ln(\theta)$ denotes the log-likelihood).

Because if $\sup_{|\theta - \theta_0| > \varepsilon} \ln(\theta) < \sup_{\theta \in \Theta} \ln(\theta)$, then $|\hat{\theta}_n - \theta_0| \leq \varepsilon$ (since $\hat{\theta}_n$ maximizes the likelihood).

Show that:
$$P_{\theta_0}\left(\sup_{|\theta - \theta_0| > \varepsilon} (\ln(\theta) - \ln(\theta_0)) < 0 \right) \xrightarrow{n \to \infty} 1$$

$$\sup_{|\theta - \theta_0| > \varepsilon} \left[\frac{\ln(\theta) - \ln(\theta_0)}{n} + KL(\theta_0, \theta) - KL(\theta_0, \theta) \right]$$

We have:
$$\frac{\ln(\theta) - \ln(\theta_0)}{n} = -\left[\frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{f_{\theta_0}(x_i)}{f_{\theta}(x_i)} \right) \right]$$

By LLN $\to -\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_{\theta_0}(X)}{f_{\theta}(X)} \right) \right] = -KL(\theta_0, \theta)$.

By assumption (i):
$$\sup_{|\theta - \theta_0| > \varepsilon} \left| \frac{\ln(\theta) - \ln(\theta_0)}{n} + KL(\theta_0, \theta) \right| \xrightarrow{P_{\theta_0}} 0$$

And using (ii):
$$\sup_{|\theta - \theta_0| > \varepsilon} -KL(\theta_0, \theta) = -\inf_{|\theta - \theta_0| > \varepsilon} KL(\theta_0, \theta) < -\delta$$

So:
$$\sup_{|\theta - \theta_0| > \varepsilon} \frac{\ln(\theta) - \ln(\theta_0)}{n} < -\delta + o(1)$$

Which is < 0 for n large enough. $\square$

V. Using Asymptotic Normality to Compute Confidence Regions

$\theta \in \Theta \subset \mathbb{R}$. Model $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$. $\hat{\theta}_n = $ MLE.

We know:
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \to \infty]{\mathcal{L}_{P_{\theta_0}}} \mathcal{N}(0, I^{-1}(\theta_0))$$

We want $I_n = [\hat{\theta}_{n,1}, \hat{\theta}_{n,2}]$ such that:

$$P_{\theta_0}(\theta_0 \in I_n) \geq 1 - \alpha$$

Then I_n is a $(1-\alpha)$ confidence interval for θ_0.

I_n is an **asymptotic** $(1-\alpha)$ confidence interval if $P_{\theta_0}(\theta_0 \in I_n) \xrightarrow[n\to\infty]{} 1 - \alpha$.

Exercise: Construct I_n.

VI. Delta method and confidence intervals

VI.1 Delta method

> **Remark**
>
> **Reminder:** If $\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_0)$, with $X_n \in \mathbb{R}^{d_1}$ and if $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is $\mathcal{C}^1$ and s.t (sub-tangent/standard) $(d_1 \leq d_2)$, $\nabla g(\mu) \in \mathbb{R}^{d_2 \times d_1}$ is of rank d_2
> Then:
> $$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla g(\mu) V_0 \nabla g(\mu)^\top)$$

Proof. Case $d_1 = d_2 = 1$.
Taylor expansion:

$$g(X_n) = g(X_n - \mu + \mu) = g(\mu) + (X_n - \mu)g'(\bar{\mu}_n)$$

where $\bar{\mu}_n \in]X_n, \mu[$.

This implies:

$$\sqrt{n}(g(X_n) - g(\mu)) = \sqrt{n}(X_n - \mu)g'(\bar{\mu}_n)$$
$$= \sqrt{n}(X_n - \mu)(g'(\mu) + o_p(1))$$

We know that $\sqrt{n}(X_n - \mu)g'(\mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, g'(\mu)^2 V_0)$.
And $\sqrt{n}(X_n - \mu)(g'(\bar{\mu}_n) - g'(\mu)) \xrightarrow{\mathbb{P}} 0$. $\qquad\square$

> **Remark**
> **Reminders:**
>
> - $X_n = O_p(1) \iff \lim_{C\to\infty} \limsup_n \mathbb{P}(|X_n| > C) = 0$.
>
> - $X_n = o_p(1) \iff X_n \xrightarrow{\mathbb{P}} 0$.
>
> - If $X_n \xrightarrow{\mathcal{L}} Q$ (where Q is a probability distribution), then $X_n = O_p(1)$.

Application to the MLE

If $\hat{\theta}_n$ is the MLE and if $\hat{\eta}_n = g(\hat{\theta}_n)$ with $g \in \mathcal{C}^1$, $\hat{\theta}_n \in \mathbb{R}^d$, $g : \mathbb{R}^d \to \mathbb{R}^k$.
If $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[\mathbb{P}_{\theta_0}]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta_0))$ with $\nabla g(\theta_0)$ of rank k.

Then:

$$\sqrt{n}(\hat{\eta}_n - g(\theta_0)) \xrightarrow[\mathbb{P}_{\theta_0}]{\mathcal{L}} \mathcal{N}(0, \nabla g(\theta_0) I^{-1}(\theta_0) \nabla g(\theta_0)^\top)$$

Example 11: Bernoulli

$X_1, ..., X_n \overset{i.i.d}{\sim} \text{Ber}(p)$ under $\mathbb{P}_{p_0}$.

MLE: $\hat{p}_n = \bar{X}_n$ and $\sqrt{n}(\hat{p}_n - p_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_0(1 - p_0))$.

Let $\eta = \log(\frac{p}{1-p})$ and $g(p) = \log(\frac{p}{1-p})$.

$g :]0, 1[\to \mathbb{R}$ is invertible. For $x \in \mathbb{R}$, $g^{-1}(x) = \frac{e^x}{1+e^x}$.

Derivative:
$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

$\forall p_0 \in]0, 1[, g'(p_0) \neq 0$.

$$\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow[\mathbb{P}_{p_0}]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{p_0(1 - p_0)}\right)$$

Note: $\frac{1}{p_0(1-p_0)} = \frac{(1+e^{\eta_0})^2}{e^{\eta_0}}$.

Example 12: Normal Distribution

$X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{+*}$.

MLE: $\hat{\theta}_n = (\bar{X}_n, S_n^2)$ where $S_n^2 = \frac{1}{n} \sum_{k=1}^{n}(X_k - \bar{X}_n)^2$.

Canonical parameters:
$$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} = \frac{e^{-\frac{x^2}{2\sigma^2} - \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}}}{\sqrt{2\pi}}$$

$$\implies \eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = \frac{-1}{2\sigma^2}$$

$$\implies \sqrt{n}(\hat{\eta}_n - \eta) \to \mathcal{N}(0, \tilde{V}_0)$$

Let $g(\eta) = \sigma^2$. We have $\sigma^2 = -\frac{1}{2\eta_2}$.

Gradient:
$$\nabla g(\eta) = \left(\frac{\partial g}{\partial \eta_1}, \frac{\partial g}{\partial \eta_2}\right) = \left(0, \frac{1}{2\eta_2^2}\right)$$

$$\implies \sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, (0, \frac{1}{2\eta_2^2})\tilde{V}_0 \begin{pmatrix} 0 \\ \frac{1}{2\eta_2^2} \end{pmatrix}\right)$$

Where $\tilde{V}_0 = \frac{d^2 c}{d^2 \eta}(\eta_0)$ and $c(\eta_0) = \frac{\eta_1^2}{8\eta_2} + \frac{1}{2}\log(-\frac{1}{2\eta_2})$.

$$\implies c'' = \dots$$

VI.2 Confidence intervals

> **Definition 21: Confidence Region**
>
> A confidence region of level α for an estimator of $g(\theta)$ is any region $C_\alpha(X_1, ..., X_n)$ verifying:
>
> $$\forall \theta \in \Theta, \quad \mathbb{P}_\theta[g(\theta) \in C_\alpha(X_1, ..., X_n)] \geq 1 - \alpha$$
>
> We can use asymptotic normality of an estimator to construct asymptotic confidence regions.

> **Example 13: Construction using MLE**
>
> $X_1, ..., X_n \overset{i.i.d}{\sim} \mathbb{P}_\theta, \theta \in \Theta \subset \mathbb{R}$.
> If the MLE verifies:
> $$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[\mathbb{P}_\theta]{\mathcal{L}} \mathcal{N}(0, I_\theta^{-1})$$
>
> We look for:
> $$C_\alpha(X_1, ..., X_n) = [\hat{\theta}_{n,1}, \hat{\theta}_{n,2}]$$
>
> s.t.
> $$\mathbb{P}_\theta[\theta \in [\hat{\theta}_{n,1}, \hat{\theta}_{n,2}]] \xrightarrow[n \to \infty]{} 1 - \alpha$$
>
> $$\hat{\theta}_{n,1} \leq \theta \leq \hat{\theta}_{n,2} \iff \sqrt{n}(\hat{\theta}_{n,1} - \hat{\theta}_n) \leq \sqrt{n}(\theta - \hat{\theta}_n) \leq \sqrt{n}(\hat{\theta}_{n,2} - \hat{\theta}_n)$$
>
> But:
> $$\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_\theta^{-1})$$
> $$\implies \sqrt{n} I_\theta^{1/2}(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$
>
> Since $\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta \implies I_{\hat{\theta}_n} \xrightarrow{\mathbb{P}_\theta} I_\theta$.
>
> $$\implies \sqrt{n} I_{\hat{\theta}_n}^{1/2}(\theta - \hat{\theta}_n) \xrightarrow[\mathbb{P}_\theta]{\mathcal{L}} \mathcal{N}(0, 1)$$
>
> If z_α is the $1 - \frac{\alpha}{2}$ quantile of a $\mathcal{N}(0, 1)$:
>
> $$\iff \Phi(z_\alpha) = 1 - \frac{\alpha}{2}, \quad \Phi = \text{CDF of } \mathcal{N}(0, 1)$$
>
>
>
> By symmetry: $\Phi(-z_\alpha) = \frac{\alpha}{2}$.
>
> $$\implies \lim_{n \to \infty} \mathbb{P}_\theta\left(-z_\alpha \leq \sqrt{n} I_{\hat{\theta}_n}^{1/2}(\theta - \hat{\theta}_n) \leq z_\alpha\right) = 1 - \alpha$$
>
> $$\iff \lim_{n \to \infty} \mathbb{P}_\theta\left(\frac{-z_\alpha}{I_{\hat{\theta}_n}^{1/2}} \leq \sqrt{n}(\theta - \hat{\theta}_n) \leq \frac{z_\alpha}{I_{\hat{\theta}_n}^{1/2}}\right) = 1 - \alpha$$
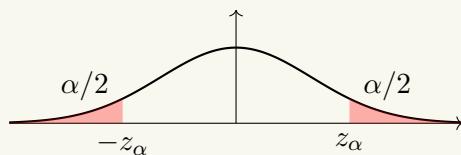
So if:
$$\hat{\theta}_{n,1} = \hat{\theta}_n - \frac{z_\alpha}{\sqrt{n}\sqrt{I_{\hat{\theta}_n}}} \quad ; \quad \hat{\theta}_{n,2} = \hat{\theta}_n + \frac{z_\alpha}{\sqrt{I_{\hat{\theta}_n}}\sqrt{n}}$$

Then:
$$\mathbb{P}_\theta(\hat{\theta}_{n,1} \le \theta \le \hat{\theta}_{n,2}) \xrightarrow[n\to\infty]{} 1 - \alpha$$

Chapter 5: Bayesian statistics

Model: $X_1, ..., X_n \overset{i.i.d}{\sim} \mathbb{P}_\theta$.
Estimator: $\hat{\theta}_n$.
Risk: $\mathcal{R}(\theta, \hat{\theta}_n) = \mathbb{E}[(\theta - \hat{\theta}_n)^2] = $ Quadratic risk.

I. In Bayesian statistics

The parameter θ is considered **unknown** $\to$ It is modelled as a random variable.

Definition 22: Prior Distribution

We call **prior distribution** on θ, a probability distribution Π over Θ.

Model:
$X_1, ..., X_n | \theta \overset{i.i.d}{\sim} \mathbb{P}_\theta$: Conditional distribution of $(X_1, ..., X_n)$ given θ.
Prior Marginal Probability over Θ:
$\theta \sim \Pi$ (Probability on Θ).
$\implies$ This defines a **Joint distribution** on $(X_1, ..., X_n, \theta)$.
Assume $\mathbb{P}_\theta$ is continuous, with density f_θ (with respect to Lebesgue measure).

Theorem 11: Bayes Theorem

The conditional distribution of θ given $(X_1, ..., X_n)$ has a density with respect to the measure ν (the reference measure of the prior):

$$\pi(\theta | X_1, ..., X_n) = \frac{\pi(\theta) \prod_{i=1}^n f_\theta(x_i)}{\int_\Theta \pi(\theta) \prod_{i=1}^n f_\theta(x_i) d\nu(\theta)}$$

Where $\pi(\theta)$ is the density of the prior Π with respect to ν. This conditional density is called the **posterior distribution** (or posterior density).

Definition 23: Bayesian Model

In Bayesian statistics, the Bayesian model is defined by:

i) The conditional law $[X|\theta] \sim \mathbb{P}_\theta$ with likelihood $L(\theta)$, for $\theta \in \Theta$.

ii) The Prior distribution Π on Θ.

Then the inference is made by the **posterior distribution** defined as the conditional distribution of θ given X. If π is the density of Π (wrt a measure ν), then the posterior distribution has a density $\pi(\theta|X)$ (wrt ν) given by:

$$\pi(\theta|X) = \frac{L(\theta)\pi(\theta)}{\int_\Theta L(\theta)\pi(\theta)d\nu(\theta)}$$

where the normalization constant is:

$$\int_\Theta L(\theta)\pi(\theta)d\nu(\theta) = \begin{cases} \int L(\theta)\pi(\theta)d\nu(\theta) & \text{if } \Pi \text{ is continuous} \\ \sum_{\theta \in \Theta} L(\theta)\pi(\theta) & \text{if } \Pi \text{ is discrete} \end{cases}$$

> **Remark**
>
> - If $\mathbb{P}_\theta$ is continuous:
> $L(\theta) = f_\theta(x) =$ density evaluated at the observations when θ is the parameter.
>
> - If $\mathbb{P}_\theta$ is discrete:
> $L(\theta) = f_\theta(x) = \mathbb{P}_\theta(X = x)$.
>
> The posterior density becomes:
>
> $$\pi(\theta|X = x) = \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_\theta(x)\pi(\theta)d\nu(\theta)}$$

Example 14: Poisson - Gamma

Let $X = (X_1, ..., X_n)$ where $X_i \overset{i.i.d}{\sim} \mathcal{P}(\theta)$, with $\theta > 0$.
Prior: $\theta \sim \Gamma(a, b)$. Here Π is a Gamma distribution $\Gamma(a, b)$.
The density of the prior is $\pi(\theta) \propto \theta^{a-1}e^{-b\theta}$.
Likelihood:

$$f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$$

Posterior density:

$$\implies \pi(\theta|X_1, ..., X_n) = \frac{\left(\prod_{i=1}^{n} f_\theta(x_i)\right)\theta^{a-1}e^{-b\theta}}{\int_\Theta \left(\prod_{i=1}^{n} f_\theta(x_i)\right)\theta^{a-1}e^{-b\theta}d\theta}$$

Looking at the numerator (and ignoring constants independent of θ):

$$\pi(\theta|X_1, ..., X_n) \propto e^{-n\theta}\prod_{i=1}^{n}\frac{\theta^{x_i}}{x_i!} \times \theta^{a-1}e^{-b\theta}$$

$$\propto e^{-(n+b)\theta}\theta^{a+\sum_{i=1}^{n} x_i - 1}$$

(Recall: for Poisson, $f_\theta(x_i) = e^{-\theta}\frac{\theta^{x_i}}{x_i!}$).

Since $\pi(\cdot|X)$ is a density, if $\frac{f(\theta)}{g(\theta)}$ is a constant (wrt θ), it implies $\exists c \neq 0, f(\theta) = cg(\theta)$.
Thus $\exists C > 0$ such that:

$$\pi(\theta|X_1, ..., X_n) = Ce^{-(b+n)\theta}\theta^{\sum_{i=1}^{n} x_i + a - 1}$$

We recognize the kernel of a Gamma distribution $\Gamma(\sum_{i=1}^{n} x_i + a, b + n)$.

$$\implies \pi(\theta|X_1, ..., X_n) = \frac{e^{-(b+n)\theta}\theta^{\sum_{i=1}^{n} x_i + a - 1}(b+n)^{a+\sum x_i}}{\Gamma(a + \sum_{i=1}^{n} x_i)}$$

<u>Conclusion:</u> The **Posterior distribution** is $\Gamma(\sum x_i + a, b + n)$.
Note regarding the constant calculation: If $g(\theta) = \theta^{a-1}e^{-b\theta}$, to make it a density we multiply by c:

$$\int_{\mathbb{R}^+} c \cdot g(\theta)d\theta = 1 \iff c = \frac{b^a}{\Gamma(a)}$$

Example 15: Normal - Normal

Bayesian model: $X_1, ..., X_n | \mu \overset{i.i.d}{\sim} \mathcal{N}(\mu, 1)$, with $\mu \in \mathbb{R}$.

Prior: $\mu \sim \mathcal{N}(a, b^2)$, with $a \in \mathbb{R}, b > 0$.

What is the Posterior distribution? It is continuous with density $\pi(\mu | X)$.

$$\pi(\mu | X_1, ..., X_n) \propto \left(\prod_{i=1}^{n} \frac{e^{-\frac{(x_i - \mu)^2}{2}}}{\sqrt{2\pi}} \right) \frac{e^{-\frac{(\mu - a)^2}{2b^2}}}{\sqrt{2\pi}b}$$

$$\propto \exp\left(-\frac{n}{2} \{\mu^2 + \bar{X}_n^2 - 2\mu\bar{X}_n\} - \frac{1}{2b^2} \{\mu^2 + a^2 - 2a\mu\} \right)$$

We used the decomposition of the sum of squares:

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

$$= n(\bar{X}_n - \mu)^2 + \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

(Note: The cross-term sums to zero).

Regrouping terms in μ:

$$\pi(\mu | X_1, ..., X_n) \propto \exp\left(-\frac{1}{2} \left[\mu^2 \left(n + \frac{1}{b^2} \right) - 2\mu \left(n\bar{X}_n + \frac{a}{b^2} \right) + \text{const} \right] \right)$$

Completing the square:

$$\pi(\mu | X_1, ..., X_n) \propto e^{-\frac{1}{2}(n + \frac{1}{b^2}) \left(\mu - \frac{\frac{a}{b^2} + n\bar{X}_n}{n + \frac{1}{b^2}} \right)^2}$$

$\implies$ The Posterior distribution is:

$$\mathcal{N}\left(\frac{\frac{a}{b^2} + n\bar{X}_n}{n + \frac{1}{b^2}}, \frac{1}{n + \frac{1}{b^2}} \right)$$

Definition 24: Marginal Likelihood

In a Bayesian model $X|\theta \sim f_\theta(x), \theta \in \Theta$ (Likelihood) and prior $\theta \sim \Pi$ with density π wrt ν.

The quantity:

$$m(x) = \int_\Theta f_\theta(x) \pi(\theta) d\nu(\theta)$$

is called the **marginal likelihood**.

II. Bayesian decision theory: Risks

II.1 Posterior and integrated risks

Recall: Quadratic Risk:

$$\mathcal{R}(\theta, \delta) = \begin{cases} \int (\theta - \delta(x))^2 f_\theta(x) dx \\ \sum_x (\theta - \delta(x))^2 \mathbb{P}_\theta(X = x) \end{cases}$$

Definition 25: Loss Function

A loss function is a function:
$$l : \Theta \times \mathcal{D} \to \mathbb{R}^+$$

where $\mathcal{D}$ is the set of decisions (i.e., estimators).

Example 16: Examples of Loss Functions

1. If $\mathcal{D} = \Theta$.

   a) The quadratic loss $l(\theta, \delta) = (\theta - \delta)^2$.

   b) The L_1 loss $l(\theta, \delta) = |\theta - \delta|$.

2. $\mathcal{D} = \{0, 1\}, \theta \in [0, 1]$.
   Aim is testing if $\theta > \frac{1}{2}$ (choosing 1) or $\theta \leq \frac{1}{2}$ (choosing 0).
   0-1 loss function defined by:

   $$l(\theta, \delta) = 1 \quad \text{if } \theta < \frac{1}{2} \text{ and } \delta = 1 \text{ (wrong decision)}$$

   $$\text{if } \theta > \frac{1}{2} \text{ and } \delta = 0 \text{ (wrong decision)}$$

   $$l(\theta, \delta) = 0 \quad \text{else (correct decision)}$$

Definition 26: Risks

In a Bayesian model $X|\theta \sim f_\theta(x), \theta \sim \Pi$ (with density π) and a loss function $l : \Theta \times \mathcal{D} \to \mathbb{R}^+$.

   a) We call the **posterior risk**:

   $$l(\Pi, \delta | X) = \int_\Theta l(\theta, \delta) \pi(\theta | X_1, ..., X_n) d\nu(\theta)$$

   b) We call the **integrated risk**:

   $$r(\Pi, \delta) = \begin{cases} \int_x l(\Pi, \delta | X = x) m(x) dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} l(\Pi, \delta | X = x) m(x) & \text{if } X \text{ is discrete} \end{cases}$$

Example 17: Calculation of Posterior Risk

1) Bayesian model, $X_1, ..., X_n | \theta \sim \mathcal{P}(\theta), \theta > 0$.
Prior: $\theta \sim \Gamma(a, b)$.

Loss function: Quadratic. $\mathcal{D} = \Theta = \mathbb{R}_+$.
$\forall \theta > 0, \delta > 0, \quad l(\theta, \delta) = (\theta - \delta)^2.$

Posterior risk:
Posterior distribution: $\Gamma(a + \sum_{i=1}^n X_i, b + n)$. Its density is:

$$\pi(\theta|X_1, ..., X_n) = \frac{(b+n)^{n\bar{X}_n + a} e^{-(b+n)\theta} \theta^{n\bar{X}_n + a - 1}}{\Gamma(a + n\bar{X}_n)}$$

With $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
$\forall \delta > 0$:

$$l(\Pi, \delta|X_1, ..., X_n) = \int_0^\infty (\theta - \delta)^2 \pi(\theta|X_1, ..., X_n) d\theta$$

$$= \int_0^\infty \theta^2 \pi(\theta|X_1, ..., X_n) d\theta + \delta^2 - 2\delta \int_0^\infty \theta \pi(\theta|X_1, ..., X_n) d\theta$$

Using the moments of the Gamma distribution ($E[Y] = \alpha/\beta, Var(Y) = \alpha/\beta^2, E[Y^2] = Var(Y) + E[Y]^2$):

$$\implies l(\Pi, \delta|X_1, ..., X_n) = \frac{a'(a'+1)}{b'^2} + \delta^2 - 2\delta \frac{a'}{b'}$$

where $a' = a + n\bar{X}_n$ and $b' = b + n$.

$\rightarrow$ This can be minimized in δ.

$$\delta(X_1, ..., X_n) = \text{minimizer of } l(\Pi, \delta|X_1, ..., X_n)$$

Differentiating with respect to δ and setting to 0, we can show that:

$$\delta(X_1, ..., X_n) = \frac{a'}{b'} = \frac{a + n\bar{X}_n}{b + n}$$

(which is the Posterior Mean).

Example 18: Hypothesis Testing (0-1 Loss)

Let $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$ be the loss function.

- Posterior risk:

$$\ell(\Pi, \delta|x_1, \ldots, x_n) = \int_\Theta \ell(\theta, \delta)\pi(\theta|x_1, \ldots, x_n) d\theta$$

Consider the model: $X_1, \ldots, X_n \sim \mathcal{P}(\theta)$ with prior $\theta \sim \Gamma(a, b)$.
The posterior density is (as seen previously):

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\theta^{a + n\bar{x}_n - 1} e^{-(b+n)\theta}(b+n)^{a+n\bar{x}_n}}{\Gamma(a + n\bar{x}_n)}$$

We set the hypotheses:

- $H_0 = \{\theta < 1\}$

- $H_1 = \{\theta \geq 1\}$

Decision rule δ:
$$\begin{cases} \delta = 1 & \text{if we choose } H_0 \\ \delta = 0 & \text{if we choose } H_1 \end{cases}$$

Let us use the 0-1 loss:
$$\ell(\theta, \delta) = \mathbb{1}_{\delta=1}\mathbb{1}_{\theta \geq 1} + \mathbb{1}_{\delta=0}\mathbb{1}_{\theta < 1}$$

The posterior risk becomes:

$$\ell(\Pi, \delta | x_1, \ldots, x_n) = \int_0^{+\infty} [\mathbb{1}_{\delta=1}\mathbb{1}_{\theta \geq 1} + \mathbb{1}_{\delta=0}\mathbb{1}_{\theta < 1}] \, \pi(\theta | x_1, \ldots, x_n) d\theta$$

$$= \mathbb{1}_{\delta=1} \underbrace{\int_1^\infty \pi(\theta | x_1, \ldots, x_n) d\theta}_{\Pi(\theta \geq 1 | X)} + \mathbb{1}_{\delta=0} \underbrace{\int_0^1 \pi(\theta | x_1, \ldots, x_n) d\theta}_{\Pi(\theta < 1 | X)}$$

$$= \mathbb{1}_{\delta=1}\Pi(\theta \geq 1 | x_1, \ldots, x_n) + \mathbb{1}_{\delta=0}\Pi(\theta < 1 | x_1, \ldots, x_n)$$

To minimize this risk, we choose $\delta = 1$ (choosing H_0) if $\Pi(\theta \geq 1 | X) < \Pi(\theta < 1 | X)$, i.e., if the posterior probability of H_0 is greater than 0.5.

Definition 27: Bayesian Estimator

Consider a Bayesian model $X | \theta \sim P_\theta, \theta \in \Theta$, with prior $\theta \sim \Pi$.
If $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$ is a loss function, we define the **Bayes estimators** δ^Π as (when they exist) the minimizers of the posterior risk $\ell(\Pi, \delta | X)$ in δ.

Proposition 5: Property

Bayesian estimators also minimize the integrated Risk.

Proof. Consider the Bayesian Model $P_\theta(x), \theta \in \Theta$ and Prior Π. Let $\ell(\theta, \delta)$ be the loss function with $\delta \in \mathcal{D}$.

The integrated risk is defined as $r(\Pi, \delta) = \int_\mathcal{X} \ell(\Pi, \delta | X = x) m(x) dx$.

Since for all x, the Bayes estimator $\delta^\Pi(x)$ verifies by definition:

$$\ell(\Pi, \delta^\Pi(x) | X = x) \leq \ell(\Pi, \delta | X = x) \quad \forall \delta \in \mathcal{D}$$

(Inequality on posterior risk).

Integrating this inequality with respect to the marginal density $m(x)$ (which is non-negative):

$$\implies \int_\mathcal{X} \ell(\Pi, \delta^\Pi(x) | X = x) m(x) dx \leq \int_\mathcal{X} \ell(\Pi, \delta(x) | X = x) m(x) dx$$

$$\implies r(\Pi, \delta^\Pi) \leq r(\Pi, \delta)$$

This holds for any estimator δ. Thus, δ^Π minimizes the integrated risk. $\qquad\square$

Why is it interesting?

<div style="border:1px solid #a00; padding:5px;">

Theorem 12: Statistical Model

Let $X|\theta \sim P_\theta$, $\theta \in \Theta$ be the Model, and Prior $\theta \sim \Pi$ with density π.

- <u>Loss function</u>: $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$

- The **Frequentist risk** is defined as the function $R : \Theta \times \mathcal{D} \to \mathbb{R}_+$ given by $R(\theta, \delta) = \mathbb{E}_\theta[\ell(\theta, \delta(X))]$.

$$= \begin{cases} \int_{\mathcal{X}} \ell(\theta, \delta(x)) f(x|\theta) dx & \text{if } X \text{ continuous} \\ \sum_{x \in \mathcal{X}} \ell(\theta, \delta(x)) f(x|\theta) & \text{if } X \text{ discrete} \end{cases}$$

<u>Then</u>:

$$\int_\Theta R(\theta, \delta) \pi(\theta) d\theta = r(\Pi, \delta)$$

The integrated risk is the average of the frequentist risk over the prior.

</div>

Example 19: Quadratic Risk

If $\ell(\theta, \delta) = (\theta - \delta)^2$, then $R(\theta, \delta) = \mathbb{E}_\theta[(\theta - \delta(X))^2]$ is the classical quadratic risk (MSE).

Proof of the Theorem. Recall that:

$$r(\Pi, \delta) = \int_{\mathcal{X}} \ell(\Pi, \delta | X = x) m(x) dx$$

Where the marginal density is $m(x) = \int_\Theta f(x|\theta) \pi(\theta) d\theta$.
And the posterior risk is $\ell(\Pi, \delta | X = x) = \int_\Theta \ell(\theta, \delta) \pi(\theta | X = x) d\theta$.
Substituting the expression of the posterior density $\pi(\theta | X = x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$:

$$r(\Pi, \delta) = \int_{\mathcal{X}} \left[\int_\Theta \ell(\theta, \delta(x)) \underbrace{\pi(\theta | X = x) m(x)}_{= f(x|\theta)\pi(\theta)} d\theta \right] dx$$

$$= \int_{\mathcal{X}} \int_\Theta \ell(\theta, \delta(x)) f(x|\theta) \pi(\theta) d\theta dx$$

By Fubini's theorem (assuming integrability), we swap the integrals:

$$= \int_\Theta \left[\int_{\mathcal{X}} \ell(\theta, \delta(x)) f(x|\theta) dx \right] \pi(\theta) d\theta$$

$$= \int_\Theta R(\theta, \delta) \pi(\theta) d\theta$$

$\square$

> **Example 20: Poisson - Gamma Calculation**
>
> $X_1, \ldots, X_n \sim \mathcal{P}(\theta)$. Prior density $\pi(\theta) = \frac{\theta^{a-1} e^{-b\theta} b^a}{\Gamma(a)}$, $\quad \theta > 0$.
> Case $a = 1, b = 1$: $\pi(\theta) = e^{-\theta} \mathbb{1}_{\theta > 0}$.
> The posterior density is:
>
> $$\pi(\theta | X_1, \ldots, X_n) = \frac{\theta^{\sum X_i} e^{-(n+1)\theta} (n+1)^{n\bar{X}_n + 1}}{\Gamma(n\bar{X}_n + 1)}$$
>
> This comes from the general formula:
>
> $$\pi(\theta | X) = \frac{\prod_{i=1}^{n} f(X_i | \theta) \pi(\theta)}{\int_{\Theta} \prod_{i=1}^{n} f(X_i | \theta) \pi(\theta) d\theta}$$

III.   Computation of Bayesian estimators

- Model: $X | \theta \sim P_\theta, \theta \in \Theta$

- Prior: $\theta \sim \Pi$

- Loss function: $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$

III.1   Quadratic loss

$$\ell(\theta, \delta) = (\theta - \delta)^2, \quad \mathcal{D} = \Theta$$

The posterior risk is:

$$\ell(\Pi, \delta | X = x) = \int_{\Theta} (\theta - \delta)^2 \pi(\theta | X = x) d\theta$$

This function is convex in δ. To find the minimum, we differentiate with respect to δ:

$$\frac{\partial}{\partial \delta} \ell(\Pi, \delta | X = x) = \int_{\Theta} \frac{\partial}{\partial \delta} (\theta - \delta)^2 \pi(\theta | X = x) d\theta$$

$$= \int_{\Theta} -2(\theta - \delta) \pi(\theta | X = x) d\theta$$

$$= 2\delta \underbrace{\int_{\Theta} \pi(\theta | X = x) d\theta}_{=1} - 2 \int_{\Theta} \theta \pi(\theta | X = x) d\theta$$

$$= 2\delta - 2\mathbb{E}^{\Pi}[\theta | X = x]$$

Setting the derivative to 0:

$$2\delta - 2\mathbb{E}^{\Pi}[\theta | X = x] = 0 \iff \delta(x) = \mathbb{E}^{\Pi}[\theta | X = x]$$

> **Proposition 6: Posterior Mean**
>
> The Bayesian estimator associated to the quadratic loss is the **posterior mean**.

III.2 L_1 loss

$$\ell(\theta, \delta) = |\theta - \delta|, \quad \mathcal{D} = \Theta = \mathbb{R}$$

Posterior risk:

$$\ell(\Pi, \delta | X = x) = \int_{\mathbb{R}} |\theta - \delta| \pi(\theta | X = x) d\theta$$

This function is convex in δ. We can split the integral:

$$\ell(\Pi, \delta | X = x) = \int_{-\infty}^{\delta} (\delta - \theta) \pi(\theta | X = x) d\theta + \int_{\delta}^{+\infty} (\theta - \delta) \pi(\theta | X = x) d\theta$$

Differentiating with respect to δ (using Leibniz integral rule):

$$\frac{\partial}{\partial \delta} \ell(\Pi, \delta | X = x) = \int_{-\infty}^{\delta} \pi(\theta | X = x) d\theta + [\delta - \delta] \pi(\delta | x)$$

$$- \int_{\delta}^{+\infty} \pi(\theta | X = x) d\theta - [\delta - \delta] \pi(\delta | x)$$

$$= \int_{-\infty}^{\delta} \pi(\theta | X = x) d\theta - \int_{\delta}^{+\infty} \pi(\theta | X = x) d\theta$$

$$= \Pi(\theta \le \delta | X = x) - \Pi(\theta > \delta | X = x)$$

Setting to 0:

$$\Pi(\theta \le \delta | X = x) = \Pi(\theta > \delta | X = x)$$

Since the sum of these probabilities is 1, this implies:

$$\Pi(\theta \le \delta | X = x) = \frac{1}{2}$$

Proposition 7: Posterior Median

The Bayesian estimator associated to the L_1 loss is the **median of the posterior distribution**.

IV.   Confidence intervals in Bayesian analysis: credible regions

Let $\alpha \in [0, 1]$ be the risk level, and Π be the prior distribution on $\Theta \subset \mathbb{R}^d$.

Definition 28: Credible Set

An α-credible set is a random set $C(\alpha, X_1, \ldots, X_n)$ such that:

$$\Pi(\theta \in C(\alpha, X_1, \ldots, X_n) | X_1, \ldots, X_n) = \int_{C(\alpha, X_1, \ldots, X_n)} \pi(\theta | X_1, \ldots, X_n) d\theta = 1 - \alpha$$

In particular, if $\Theta = \mathbb{R}$, an α-credible interval is denoted $[l_\alpha(X_1, \ldots, X_n), r_\alpha(X_1, \ldots, X_n)]$. l_α and r_α solve (satisfy):

$$\int_{-\infty}^{l_\alpha} \pi(\theta | X_1, \ldots, X_n) d\theta = \int_{r_\alpha}^{+\infty} \pi(\theta | X_1, \ldots, X_n) d\theta = \frac{\alpha}{2}$$

> **Remark**
>
> **Careful:** A Bayesian credible set $\neq$ Frequentist confidence set interval.
> In Bayesian analysis, we have:
>
> $$\Pi(\theta \in (l_\alpha, r_\alpha)|X_1, \ldots, X_n) \geq 1 - \alpha$$
>
> but we <u>do not have</u>:
>
> $$\mathbb{P}_\theta(\theta \in (l_\alpha, r_\alpha)) \geq 1 - \alpha$$
>
> However, if $f_\theta(x)$ is <u>nice enough</u> (regularity conditions), then credible sets can be turned into asymptotically optimal confidence intervals (as $n \to \infty$).

V. Choosing priors

V.1 Conjugate priors

Let $f(x|\theta) \in \mathcal{F}$ be a family of models (Poisson, Gamma, Normal, Bernoulli...).
Let $\pi(\theta) \in \widetilde{\mathcal{F}}$ be another family of distributions.

> **Definition 29: Conjugate Priors**
>
> The families $f(x|\theta) \in \mathcal{F}$ and $\pi(\theta) \in \widetilde{\mathcal{F}}$ are conjugate if:
>
> $$\pi(\theta|X_1, \ldots, X_n) \in \widetilde{\mathcal{F}}$$
>
> (The posterior distribution belongs to the same family as the prior).

> **Example 21: Uniform / Pareto**
>
> Consider the model:
>
> $$f(x|\theta) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x)$$
>
> And the prior (Pareto distribution):
>
> $$\pi(\theta) = \frac{C_{\alpha,\theta_{min}}}{\theta^{\alpha+1}} \mathbb{1}_{\{\theta > \theta_{min}\}}$$
>
> with $\alpha > 0$ and $\theta_{min} > 0$.
> **Posterior derivation:**
>
> $$\pi(\theta|X_1, \ldots, X_n) \propto L_n(\theta)\pi(\theta)$$
> $$= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \times \frac{1}{\theta^{\alpha+1}} \mathbb{1}_{\{\theta > \theta_{min}\}}$$
> $$= \frac{1}{\theta^n} \mathbb{1}_{\{\theta \geq \max(X_i)\}} \times \frac{1}{\theta^{\alpha+1}} \mathbb{1}_{\{\theta > \theta_{min}\}}$$
> $$= \frac{1}{\theta^{n+\alpha+1}} \times \mathbb{1}_{\{\theta > \max(\max(X_i),\theta_{min})\}}$$
>
> This is a Pareto distribution: $\text{Pareto}(\alpha + n, \theta_{min} \vee \max_{i=1\ldots n}(X_i))$.

V.2 Flat priors

The goal is to choose a prior that is non-informative.

- e.g. If $\Theta = [a, b]$, $\pi(\theta) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(\theta)$.

- What if $\Theta = \mathbb{R}$? We use an **Improper prior**.

Take $\pi(\theta) \propto c$, then $\int_{\mathbb{R}} \pi(\theta) d\theta = \infty$.
However, this is allowed as long as the posterior is proper:

$$\int_{\Theta} L_n(\theta'|X_1, \ldots, X_n)\pi(\theta')d\theta' < +\infty \quad \mathbb{P}_\theta - a.s.$$

> ### Example 22: Normal model with flat prior
>
> Let $f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
> Take $\pi(\theta) = 1$ on $\mathbb{R}$! (Improper).
> **Posterior derivation:**
>
> $$\pi(\theta|X_1, \ldots, X_n) \propto L_n(\theta|X_1, \ldots, X_n) \times 1$$
> $$\propto e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}$$
> $$= e^{-\frac{1}{2}\sum_{i=1}^{n}(\theta-X_i)^2}$$
>
> Let us expand the sum in the exponential:
>
> $$\sum_{i=1}^{n}(\theta - X_i)^2 = n\theta^2 - 2\sum_{i=1}^{n}\theta X_i + \sum_{i=1}^{n}X_i^2$$
> $$= n(\theta^2 - 2\theta\overline{X_n} + \ldots)$$
> $$= n(\theta - \overline{X_n})^2 + \text{const}$$
>
> (The terms not depending on θ are absorbed in the proportionality constant).
>
> Thus:
> $$\pi(\theta|X_1, \ldots, X_n) \propto e^{-\frac{1}{2}\frac{1}{1/n}(\theta-\overline{X_n})^2}$$
>
> We recognize a Normal distribution:
>
> $$\mathcal{N}\left(\overline{X_n}, \frac{1}{n}\right)$$
>
> [Image of standard normal distribution curve]

V.3 Jeffrey's prior

Jeffrey's idea: Choose a prior proportional to the square root of the Fisher information.

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}}$$

Where $I(\theta)$ is the Fisher information:

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

Result:

If $\pi(\theta) \propto \sqrt{I(\theta)}$, then under any reparameterisation $\eta = g(\theta)$, the corresponding prior $\tilde{\pi}(\eta) \propto \sqrt{\tilde{I}(\eta)}$ is consistent, where $\tilde{I}$ is the Fisher information for the model parametrized by η.