

# Advanced Machine Learning

Emilie Chouzenoux<sup>(1)</sup>, L. Omar Chehab<sup>(2)</sup> and Frédéric Pascal<sup>(3)</sup>

<sup>(1)</sup> Center for Computer Vision (CVN), CentraleSupélec / Opis Team, Inria

<sup>(2)</sup> Parietal Team, Inria <sup>(3)</sup> Laboratory of Signals and Systems (L2S), CentraleSupélec,  
University Paris-Saclay

{emilie.chouzenoux, frederic.pascal}@centralesupelec.fr, l-emir-omar.chehab@inria.fr

<http://www-syscom.univ-mlv.fr/~chouzeno/>

<http://fredericpascal.blogspot.fr>

**MDS**

Sept. - Dec., 2020



CentraleSupélec

# Contents

- 1 Introduction - Reminders of probability theory and mathematical statistics (Bayes, estimation, tests) - FP
- 2 Robust regression approaches - EC / OC
- 3 Hierarchical clustering - FP / OC
- 4 Stochastic approximation algorithms - EC / OC
- 5 Nonnegative matrix factorization (NMF) - EC / OC
- 6 Mixture models fitting / Model Order Selection - FP / OC
- 7 Inference on graphical models - EC / VR
- 8 Exam

## Key references for this course

- Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer, 2009.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*. Springer, 2013

+ many many references...

# Course 1

Introduction - Reminders of probability theory and  
mathematical statistics

I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

IV. Statistical modelling

V. Theory of Point Estimation

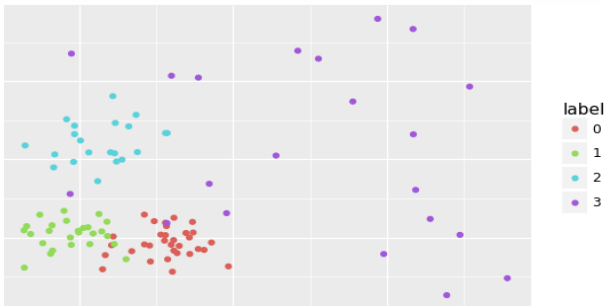
VI. Hypothesis testing - Decision theory

# What is Machine Learning?

Statistical machine learning is concerned with the development of algorithms and techniques that learn from observed data by constructing stochastic models that can be used for making predictions and decisions.

Topics covered include Bayesian inference and maximum likelihood modeling; regression, classification, density estimation, clustering, principal component analysis; parametric, semi-parametric, and non-parametric models; basis functions, neural networks, kernel methods, and graphical models; deterministic and stochastic optimization; overfitting, regularization, and validation.

# From data to processing - robustness, dimension...



## Big Picture

Data driven

Model driven



$$n > p$$

$$\begin{aligned} (n > p) \\ n < p \end{aligned}$$

$$\begin{aligned} (n > p) \\ (n < p) \end{aligned}$$

$$R < n, p$$

Classical  
Processing

Regularization

Structure  
a priori

# General context

## Statistical Signal Processing

*Signals*  $\mathbf{z}$  : multivariate random complex observations (vectors).

*Example* :  $\mathbf{z} \in \mathbb{C}^p$  Signal corrupted by an additive noise:

$$\mathbf{z} = \beta \mathbf{d}(\theta) + \mathbf{n}$$

with  $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ ,  $\theta$  and  $\beta$  unknown.

## Several processes

- PCA and dimension reduction
- Parameter estimation
- Detection / Filtering
- Clustering / Classification
- ...



# Covariance & Subspace

Two quantities common to all these processes

“Optimal” processes rely on the second order statistics of  $\mathbf{z}$ , notably on:

**The covariance matrix** (assume circularity):

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{z}\mathbf{z}^H]$$

Information on the variance and correlations between elements of  $\mathbf{z}$ .

**The principal subspace** (of rank  $R$ )

$$\mathbf{\Pi}_R = \mathcal{P}_R \{ \mathbb{E}[\mathbf{z}\mathbf{z}^H] \}$$

Rank  $R$  orthogonal subspace where most of the information lies in.

# Examples

## Estimation (MLE, GMM...)

Parameter  $\theta$  of the signal  $\mathbf{d}(\theta)$  to be estimated from observations

*Example* : Maximum Likelihood Estimator (MLE)

$$\min_{\theta} (\mathbf{d}(\theta) - \mathbf{z})^H \boldsymbol{\Sigma}^{-1} (\mathbf{d}(\theta) - \mathbf{z})$$

Low rank version (e.g. MUSIC): replace  $\boldsymbol{\Sigma}^{-1}$  by  $\boldsymbol{\Pi}^{\perp}$

*Applications*: DoA, inverse problems, source separation...

## Detection (ACE, GLRT, ANMF, MSD...)

Binary hypothesis test: is  $\mathbf{d}(\theta_0)$  present?

*Example* : Adaptive Cosine Estimator (ACE, or ANMF):

$$\Lambda_{ACE} = \frac{|\mathbf{d}(\theta_0)^H \boldsymbol{\Sigma}^{-1} \mathbf{z}|^2}{(\mathbf{d}(\theta_0)^H \boldsymbol{\Sigma}^{-1} \mathbf{d}(\theta_0)) (\mathbf{z}^H \boldsymbol{\Sigma}^{-1} \mathbf{z})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Low rank version: replace  $\boldsymbol{\Sigma}^{-1}$  by  $\boldsymbol{\Pi}^{\perp}$

*Applications* : RADAR, imaging, audio...

## Filtering (MF, AMF, Projection...)

Maximizing the output signal to noise ratio (SNR):

*Example* : Adaptive Matched Filter

$$y = \frac{|\mathbf{d}^H(\theta) \boldsymbol{\Sigma}^{-1} \mathbf{z}|^2}{\mathbf{d}(\theta_0)^H \boldsymbol{\Sigma}^{-1} \mathbf{d}(\theta_0)}$$

Low rank version: replace  $\boldsymbol{\Sigma}^{-1}$  by  $\boldsymbol{\Pi}^\perp$

*Applications* : De-noising, interference cancellation (telecom)...

## Classification (SVM, K-means, KL divergence...)

Select a class for the observations: covariance and subspace are descriptors

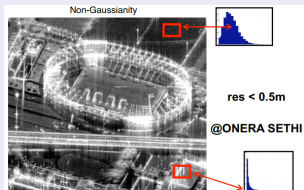
*Example* : KL divergence between two distributions (or other divergences, Wasserstein, Riemannian ...)

$$\begin{aligned} KL(\mathbf{Z}^1, \mathbf{Z}^2) &= \frac{1}{2} [\text{Tr}(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1) + \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_2) - 2k] \\ W_2^2(\mathbf{Z}^1, \mathbf{Z}^2) &= \text{Tr}(\boldsymbol{\Sigma}_1) + \text{Tr}(\boldsymbol{\Sigma}_2) - 2 \text{Tr} \left( (\boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{1/2})^{1/2} \right) \end{aligned}$$

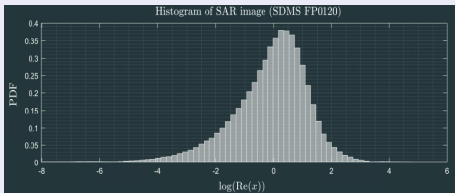
*Applications* : Machine learning, segmentation, profile determination...

# Example of non Gaussianity (1/3): High Resolution SAR images

## HR SAR images

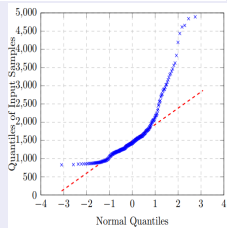
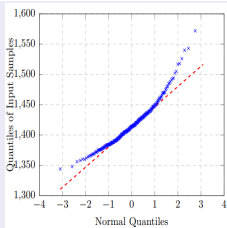
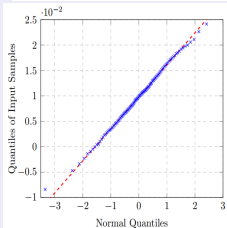
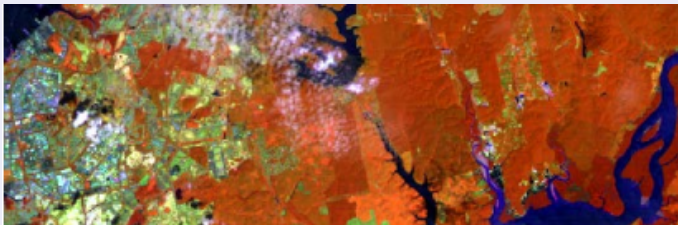


## SMDS Data



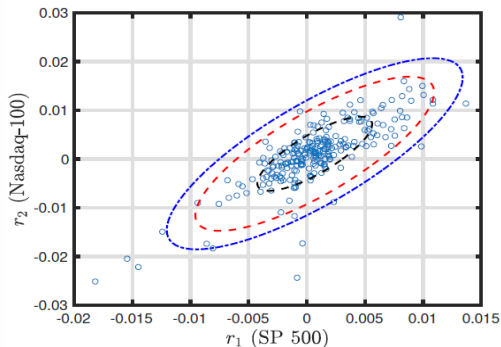
# Example of non Gaussianity (2/3): Hyperspectral data

NASA Hyperion sensor

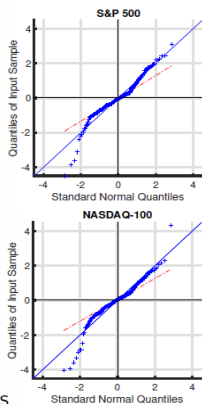


# Example of non Gaussianity (3/3): Financial data

Nasdaq-100, SP 500



inside the 50% ellipse: 65.6% of returns  
inside the 95 and 99% ellipses = 93.2% 95.6% of returns



Courtesy of E. Ollila [Ollila18]

I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

IV. Statistical modelling

V. Theory of Point Estimation

VI. Hypothesis testing - Decision theory

# Menu - Probabilities and statistics basics

Example: Fair Six-Sided Die:

- **Sample space:**  $\Omega = \{1, 2, 3, 4, 5, 6\}$
- **Events:** Even =  $\{2, 4, 6\}$ , Odd =  $\{1, 3, 5\} \subseteq \Omega$
- **Probability:**  $P(6) = \frac{1}{6}$ ,  $P(\text{Even}) = P(\text{Odd}) = \frac{1}{2}$
- **Outcome:**  $6 \in E$ .
- **Conditional probability:**  $P(6|\text{Even}) = \frac{P(6 \cap \text{Even})}{P(\text{Even})} = \frac{1/6}{1/2} = \frac{1}{3}$

General Axioms:

- $P(\emptyset) = 0 \leq P(A) \leq 1 = P(\Omega)$ ,
- $P(A \cup B) + P(A \cap B) = P(A) + P(B)$ ,
- $P(A \cap B) = P(A|B)P(B)$ .



# Menu - Probabilities and statistics basics

Example: (Un)fair coin:  $\Omega = \{\text{Tail}, \text{Head}\} \simeq \{0, 1\}$  with  $P(1) = \theta \in [0, 1]$ :

- **Likelihood:**  $P(1101|\theta) = \theta \times \theta \times (1 - \theta) \times \theta$
- **Maximum Likelihood (ML) estimate:**  $\hat{\theta} = \operatorname{argmax}_{\theta} P(1101|\theta) = \frac{3}{4}$
- **Prior:** If we are indifferent, then  $P(\theta) = \text{const.}$
- **Evidence:**  $P(1101) = \sum_{\theta} P(1101|\theta)P(\theta) = \frac{1}{20}$  (actually  $f$ )
- **Posterior:**  $P(\theta|1101) = \frac{P(1101|\theta)P(\theta)}{P(1101)} \propto \theta^3(1 - \theta)$  (Bayes rule).
- **Maximum a Posterior (MAP) estimate:**  $\hat{\theta} = \operatorname{argmax}_{\theta} P(\theta|1101) = \frac{3}{4}$
- **Predictive distribution:**  $P(1|1101) = \frac{P(11011)}{P(1101)} = \frac{2}{3}$
- **Expectation:**  $E[f|\dots] = \sum_{\theta} f(\theta)P(\theta|\dots)$ , e.g.  $E[\theta|1101] = \frac{2}{3}$
- **Variance:**  $V(\theta|1101) = E[(\theta - E[\theta])^2|1101] = \frac{2}{63}$
- **Probability density:**  $P(\theta) = \frac{1}{\varepsilon}P([\theta, \theta + \varepsilon])$  for  $\varepsilon \rightarrow 0$

# Random Variables (r.v.) / Vectors (r.V.)

## Notations

Let  $X$  (resp.  $\mathbf{x}$ ) a random variable (resp. vectors). Denote by  $P$  or  $P_\theta$  its probability :

- $P(X = x)$  or  $P_\theta(X = x)$  for the discrete case
- $f(x)$  or  $f_\theta(x)$  for the continuous case (with PDF)

Some other notations:

- $E[\cdot]$  or  $E_\theta[\cdot]$  (resp.  $V[\cdot]$  /  $V_\theta[\cdot]$ ) stands for the statistical expectation (resp. the variance)
- i.i.d.  $\rightarrow$  Independent (denoted  $\perp$ ) and Identically Distributed, i.e. same distribution and  $X \perp Y \iff$  for any measurable functions  $h$  and  $g$ ,  $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$  .
- $n$ -sample  $(X_1, \dots, X_n) \iff X_1, \dots, X_n$  are i.i.d.
- PDF, CDF and iff resp. means Probability Density Function, Cumulative Distribution Function and "if and only if"

# Convergences

## Multivariate case

Let  $(\mathbf{x})_{n \in \mathbb{N}} \in \mathbb{R}^d$  a sequence of r.V. and  $(\mathbf{x}) \in \mathbb{R}^d$  defined on the same probability space  $(\Omega, \mathcal{A}, P)$ , then

- **Almost Sure CV:**  $\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{a.s.} \mathbf{x} \iff \exists N \in \mathcal{A}$  such that  $P(N) = 0$  and  $\forall \omega \in N^c, \lim_{n \rightarrow \infty} \mathbf{x}_n(\omega) = \mathbf{x}(\omega)$
- **CV in probability:**  $\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{P} \mathbf{x} \iff \forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(\|\mathbf{x}_n - \mathbf{x}\| \geq \varepsilon) = 0$  where  $\|\mathbf{x}\| = (\sum_{i=1}^d x_i^2)^{1/2}$  for  $\mathbf{x} \in \mathbb{R}^d$ .  
 $\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{P} \mathbf{x} \iff$  each component converges in probability.
- **CV in  $\mathcal{L}^p$ :** Let  $p \in \mathbb{N}^*, \mathbf{x}_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}^p} \mathbf{x} \iff (\mathbf{x})_{n \in \mathbb{N}}, \mathbf{x} \in \mathcal{L}^p$  and  $E[\|\mathbf{x}_n - \mathbf{x}\|_{\mathcal{L}^p}^p] \xrightarrow[n \rightarrow \infty]{} 0$ .

# Convergence in distribution

- **CV in distribution:**  $\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{dist.} \mathbf{x}$  if for any continuous and bounded function  $g$ , one has  $\lim_{n \rightarrow \infty} E[g(\mathbf{x}_n)] = E[g(\mathbf{x})]$ .

⚠ The CV in distribution of a sequence of r.V. is stronger than the CV of each component!

How to characterise the CV in distribution?

## Theorem (Levy continuity Theorem)

Let  $\varphi_n(u) = E[\exp(iu^t \mathbf{x}_n)]$  and  $\varphi(u) = E[\exp(iu^t \mathbf{x})]$  the characteristic functions of  $\mathbf{x}_n$  and  $\mathbf{x}$ . Then,

$$\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{dist.} \mathbf{x} \iff \forall u \in \mathbb{R}^d, \varphi_n(u) \xrightarrow[n \rightarrow \infty]{} \varphi(u).$$

## Proposition (a.s., $P$ , dist. convergences)

$\mathbf{x}_n \xrightarrow[n \rightarrow \infty]{} \mathbf{x} \implies h(\mathbf{x}_n) \xrightarrow[n \rightarrow \infty]{} h(\mathbf{x})$ , if  $h$  is a continuous function

Discussion on the cv hierarchy...

I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

- SLLN and CLT
- Slutsky theorem and the Delta-method
- Gaussian-related distributions

IV. Statistical modelling

V. Theory of Point Estimation

VI. Hypothesis testing - Decision theory

# SLLN and CLT

## Theorem (Strong (Weak) Law of Large Numbers)

Let  $(\mathbf{x}_n)_{n \in \mathbb{N}^*}$  a sequence of i.i.d. r.V. in  $\mathbb{R}^d$  s.t.  $E[|\mathbf{x}_1|] < +\infty$ . Let  $\mu = E[\mathbf{x}_1]$  the expectation of  $\mathbf{x}_1$ . Then,

$$\bar{\mathbf{x}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \xrightarrow[n \rightarrow \infty]{a.s., P} \mu.$$

## Theorem (Central Limit Theorem)

Let  $(\mathbf{x}_n)_{n \in \mathbb{N}^*}$  a sequence of i.i.d. r.V. in  $\mathbb{R}^d$  s.t.  $E[|\mathbf{x}_1|^2] < +\infty$ . Let  $\mu = E[\mathbf{x}_1]$  and  $\Sigma = E[\mathbf{x}_1 \mathbf{x}_1^t] - E[\mathbf{x}_1]E[\mathbf{x}_1]^t$  the covariance matrix of  $\mathbf{x}_1$ . Let

$\bar{\mathbf{x}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  the empirical mean. Then,

$$\sqrt{n}(\bar{\mathbf{x}}_n - \mu) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}(\mathbf{0}, \Sigma).$$

# Slutsky theorem

## Theorem (Slutsky theorem)

Let  $(\mathbf{x}_n)_{n \in \mathbb{N}^*}$  a sequence of r.V. in  $\mathbb{R}^d$  that cv in dist. to  $\mathbf{x}$ . Let  $(\mathbf{y}_n)_{n \in \mathbb{N}^*}$  a sequence of r.V. in  $\mathbb{R}^m$  (defined on the same proba. space as  $(\mathbf{x}_n)_{n \in \mathbb{N}^*}$ ) that cv a.s. (or in  $P$ , or in dist.) towards a constant  $\mathbf{a}$ . Thus, the sequence  $(\mathbf{x}_n, \mathbf{y}_n)_{n \in \mathbb{N}^*}$  cv *in dist.* towards  $(\mathbf{x}, \mathbf{a})$ ,  $(\mathbf{x}_n, \mathbf{y}_n) \xrightarrow[n \rightarrow \infty]{\text{dist.}} (\mathbf{x}, \mathbf{a})$

## Remark (Important Applications of Slutsky (IAS))

Under previous assumptions, one has:

- 1  $\mathbf{x}_n + \mathbf{y}_n \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathbf{x} + \mathbf{a}$  if  $m = d$
- 2  $\mathbf{x}_n \mathbf{y}_n \xrightarrow[n \rightarrow \infty]{\text{dist.}} a \mathbf{x}$  if  $m = 1$
- 3  $\mathbf{x}_n / \mathbf{y}_n \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathbf{x} / a$  if  $m = 1, a \neq 0$

# Delta-method

## Theorem (Delta-method)

Let  $(\mathbf{x}_n)_{n \in \mathbb{N}^*}$  a sequence of r.V. in  $\mathbb{R}^d$  and  $\theta$  a (deterministic) vector of  $\mathbb{R}^d$ .  
Let  $h: \mathbb{R}^d \mapsto \mathbb{R}^m$  a function that is differentiable (at least) at point  $\theta$ .

Let us denote  $\frac{\partial h}{\partial \theta^t}(\theta)$  the  $m \times d$  matrix s.t.  $\left( \frac{\partial h_i}{\partial \theta_j}(\theta) \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq d}}$  and

$\frac{\partial h^t}{\partial \theta}(\theta) = \left( \frac{\partial h}{\partial \theta^t}(\theta) \right)^t$  its transpose. Assume that  $\sqrt{n}(\mathbf{x}_n - \theta) \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathbf{x}$ . Then

$$\sqrt{n}(h(\mathbf{x}_n) - h(\theta)) \xrightarrow[n \rightarrow \infty]{\text{dist.}} \frac{\partial h}{\partial \theta^t}(\theta) \mathbf{x}.$$

Particular case:

If  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ , then  $\sqrt{n}(h(\mathbf{x}_n) - h(\theta)) \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathcal{N}\left(\mathbf{0}, \frac{\partial h}{\partial \theta^t}(\theta) \Sigma \frac{\partial h^t}{\partial \theta}(\theta)\right)$



# Gamma and Beta distributions

## Definition (Gamma distribution)

Let  $p > 0$  et  $\lambda > 0$ . A real-valued r.v.  $X \sim \Gamma(p, \lambda)$  if its PDF is defined as

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^+}(x),$$

where  $\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) dt$  for  $x \in \mathbb{C}$  s.t.  $\Re(x) > 0$ . Also

$\Gamma(x+1) = x\Gamma(x)$  ( $n \in \mathbb{N}^*$ ,  $\Gamma(n) = (n-1)!$ ).

If  $X \sim \Gamma(p, \lambda)$  and  $a > 0$ , then  $aX \sim \Gamma(p, \lambda/a)$

## Proposition (Beta distributions)

1 Let  $Y \sim \Gamma(q, \lambda)$  and  $X \sim \Gamma(p, \lambda)$  2 independent r.v. Thus,

- $X + Y \sim \Gamma(p + q, \lambda)$ ,
- $X + Y$  and  $\frac{X}{X+Y}$  (resp.  $X + Y$  and  $\frac{X}{Y}$ ) are independent
- Distributions of  $\frac{X}{X+Y}$  and  $\frac{X}{Y}$  do **NOT** depend on  $\lambda$ . It resp. corresponds to **Beta distributions of 1<sup>st</sup> and 2<sup>nd</sup> kind**, denoted  $\beta^1(p, q)$  and  $\beta^2(p, q)$ . PDF...

# Gamma and Beta distributions

## Definition (Beta PDFs)

$$\left\{ \begin{array}{l} \beta^1(p, q) : f(x) = \frac{x^{p-1}(1-x)^{q-1}}{\beta(p, q)} \mathbb{1}_{[0,1]}(x), \\ \beta^2(p, q) : f(x) = \frac{x^{p-1}}{(1+x)^{p+q}\beta(p, q)} \mathbb{1}_{\mathbb{R}^+}(x), \end{array} \right.$$

with  $\beta(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ .

## Proposition

- If  $U \sim \beta^1(p, q)$ ,  $\frac{U}{1-U} \sim \beta^2(p, q)$ ,
- If  $V \sim \beta^2(p, q)$ ,  $\frac{V}{1+V} \sim \beta^1(p, q)$ ,
- If  $V \sim \beta^2(p, q)$ ,  $\frac{1}{V} \sim \beta^2(q, p)$ .

# $\chi^2$ , Student- $t$ and Fisher (or $F$ ) distributions

## Definition ( $\chi^2$ dist.)

Let  $(X_n)_{n \in \mathbb{N}^*}$  a sequence of i.i.d. real-valued r.v.  $\sim \mathcal{N}(0, 1)$ . Thus,

- $\sum_{i=1}^k X_i^2$  follows a  $\chi^2$ -distribution with  $k$  d.o.f., (denoted  $\chi^2(k)$ ).
- $X_1^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$  and  $\sum_{i=1}^k X_i^2 \sim \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$

## Definition (Student- $t$ and $F$ - distributions)

- 1 If  $X \sim \mathcal{N}(0, 1)$ ,  $Y \sim \chi^2(k)$ , and  $X, Y$  independent, then  $T = \frac{X}{\sqrt{Y/k}}$  follows a Student- $t$  dist. with  $k$  d.o.f. (denoted  $t(k)$ ).
- 2 If  $p$  and  $q$  are integers, if  $X \sim \chi^2(p)$ ,  $Y \sim \chi^2(q)$ , and  $X, Y$  are independent, then  $F = \frac{X/p}{Y/q}$  follows a  $F$ -dist. with  $p$  and  $q$  d.o.f., (denoted  $F(p, q)$ ).

# Student Theorem

## Theorem (Student theorem)

Let  $(X_n)_{n \in \mathbb{N}^*}$  a sequence a real-valued i.i.d. r.v.  $\sim \mathcal{N}(\mu, \sigma^2)$ . Then, one has:

1  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$

2  $R_n = \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2 \chi^2(n-1).$

3  $\bar{X}_n$  and  $R_n$  are independent.

4 If  $S_n = \sqrt{\frac{R_n}{n-1}}$ , then  $T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t(n-1).$

## Proof

Some elements...

# Some applications

## Estimate unknown parameters??

A1 Mean estimation:  $(X_1, \dots, X_n) \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

- $\sigma^2$  known
- $\sigma^2$  unknown

A2 Variance estimation:  $(X_1, \dots, X_n) \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

- $\mu$  known
- $\mu$  unknown

A3 Variance comparison (test) between two independent samples:

$(X_1, \dots, X_n) \stackrel{iid}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$  and  $(Y_1, \dots, Y_n) \stackrel{iid}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$

- $\mu_X$  and  $\mu_Y$  known
- $\mu_X$  and  $\mu_Y$  unknown

# Possible answers with confidence intervals

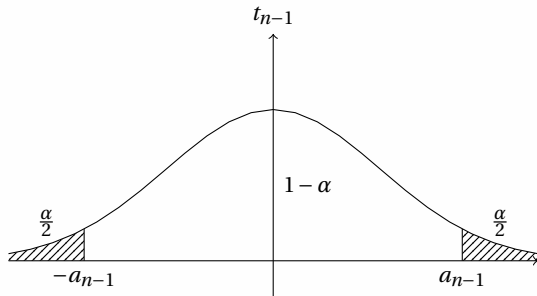
A1 Based on  $\hat{\mu} = \bar{X}_n \dots$

- $I_n = \left[ \bar{X}_n \pm \frac{1,96\sigma}{\sqrt{n}} \right]$  is an **exact** 95%-confidence interval
- $\tilde{I}_n = \left[ \bar{X}_n \pm \frac{1,96\hat{\sigma}_n}{\sqrt{n}} \right]$  is an asymptotic 95%-confidence interval.

OR use

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t(n-1) \Rightarrow \hat{I}_n = \left[ \bar{X}_n \pm \frac{a_{n-1}S_n}{\sqrt{n}} \right]$$

is an **exact** 95%-confidence interval



# Possible answers with confidence intervals

A2 Based on ...



$$R_n^* = \sum_{i=1}^n (X_i - \mu)^2 \sim \sigma^2 \chi^2(n) \Rightarrow I_n = \left[ \frac{n\hat{\sigma}_n^2}{b_n}, \frac{n\hat{\sigma}_n^2}{a_n} \right]$$

is an **exact** 95%-confidence interval with  $\hat{\sigma}_n^2 = R_n^*/n$ .



$$R_n = \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2 \chi^2(n-1) \Rightarrow \hat{I}_n = \left[ \frac{(n-1)\hat{\sigma}_n^2}{b_{n-1}}, \frac{(n-1)\hat{\sigma}_n^2}{a_{n-1}} \right]$$

is an **exact** 95%-confidence interval with  $\hat{\sigma}_n^2 = R_n/(n-1)$

↪ Loss when unknowns are present..., i.e. length of CI increases...

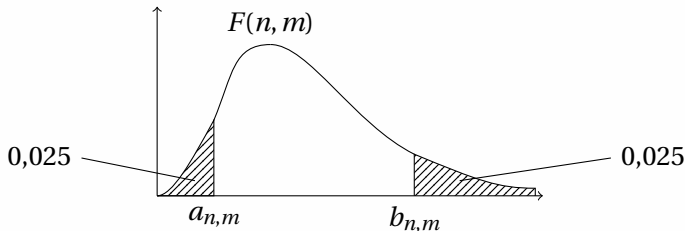
# Possible answers with confidence intervals

A3 Based on ...

$$\blacksquare R_{n,X}^* = \sum_{i=1}^n (X_i - \mu_X)^2 \sim \sigma_X^2 \chi^2(n), R_{m,Y}^* = \sum_{i=1}^m (Y_i - \mu_Y)^2 \sim \sigma_Y^2 \chi^2(m)$$

$$\frac{R_{n,X}^*}{R_{m,Y}^*} \sim F(n, m) \Rightarrow \frac{\sigma_X^2}{\sigma_Y^2} \in \left[ \frac{1}{b_{n,m}} \frac{\hat{\sigma}_{n,X}^2}{\hat{\sigma}_{m,Y}^2}, \frac{1}{a_{n,m}} \frac{\hat{\sigma}_{n,X}^2}{\hat{\sigma}_{m,Y}^2} \right]$$

with  $\hat{\sigma}_{n,X}^2 = R_{n,X}^*/n$  and  $\hat{\sigma}_{m,Y}^2 = R_{m,Y}^*/m$



**Same thing for  $\mu_X$  and  $\mu_Y$  unknown...**



I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

IV. Statistical modelling

- Generalities
- Sufficiency
- Exponential family
- Fisher information
- Optimality
- Cramer-Rao bound

V. Theory of Point Estimation

# Statistical modelling

## Generalities

- $n$ -sample  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$
- **Dominated models**  $\rightsquigarrow$  **Likelihood Function (LF)**, denoted  $L(\mathbf{x}, \theta)$
- Parametric models, i.e.  $\theta \in \Theta \subset \mathbb{R}^d$

### Definition (**Identifiability conditions**)

A model  $(\mathcal{X}, \mathcal{A}, \{P_\theta, \theta \in \Theta\})$  is said **identifiable** if the mapping from  $\Theta$  onto the probabilities space  $(\mathcal{X}, \mathcal{A})$ , which to  $\theta$  gives  $P_\theta$  is injective.

### Definition (**Statistic**)

In a statistical model  $\{\mathcal{X}, \mathcal{A}, \{P_\theta, \theta \in \Theta\}\}$ , one said **statistic**, for any (**measurable or  $\sigma$ -finite**) mapping  $S$  from  $(\mathcal{X}, \mathcal{A})$  onto an arbitrary space. Let's say a **statistic is a function of r.V.**  $S(\mathbf{x}_1, \dots, \mathbf{x}_n)$ .

e.g.,  $\bar{X}_n, R_{n,X}, \hat{\sigma}_n^2$ , or even  $X_1, \dots$

# Statistical modelling

## Sufficient statistics

**Very important concept!** for high-dimensional data, **dimension reduction without reducing the information brought by the data.**

Main idea: Where is contained the information of interest (i.e. related to the unknowns) in the data?

Example: Coin toss  $\rightarrow$  Head and Tails - One wants to know the probability of Head or if the coin is biased ... No need to keep the whole dataset...

### Definition (**Sufficient statistic**)

A statistic  $S$  is said to be sufficient iff the conditional distribution  $\mathcal{L}_\theta(X|S(X))$  does not depend on  $\theta$ .

### Remark (Pros and cons)

- *Difficulty to use the definition*
- *Dimension of  $S$  has to be minimal!  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  is always a sufficient stat.*

# Statistical modelling

## Sufficient statistics characterization

### Theorem (**Factorisation Criterion (FC)**)

A statistic  $S$  is sufficient iff the likelihood function can be written as:

$$L(x; \theta) = \psi(S(x); \theta) \lambda(x).$$

This is a sort of separability theorem...

Example: let  $(X_1, \dots, X_n)$  i.i.d following a non-centred exponential dist., i.e. with PDF

$$f(x_i, \theta) = \frac{1}{\theta_2} \exp\left(-\frac{1}{\theta_2}(x_i - \theta_1)\right) \mathbb{1}_{\{x_i \geq \theta_1\}} \quad \text{with} \quad \theta = (\theta_1, \theta_2)^t.$$

$$\Rightarrow S(X) = \left( \min_{i=1, \dots, n} (X_i), \sum_{i=1}^n X_i \right) \text{ is sufficient!}$$

# Exponential family

## Definition (**Complete statistics**)

A statistic  $S$  is said to be complete if for any measurable real-valued function  $\phi$ , one has

$$\{\forall \theta \in \Theta, E_{\theta} [\phi \circ S(X)] = 0\} \Rightarrow \{\forall \theta \in \Theta, \phi \circ S(X) = 0 \text{ a.s. } [P_{\theta}]\}.$$

Purely theoretical... for optimal unbiased estimation...

## Definition (**Exponential family**)

A model is said to be exponential iff its LF can be written as:

$$L(x; \theta) = h(x)\phi(\theta) \exp \left\{ \sum_{i=1}^r Q_i(\theta) S_i(x) \right\}. \quad (1)$$

where  $S(\cdot) = (S_1(\cdot), \dots, S_r(\cdot))$  is the **canonical statistic**.

Discussion:  $r$ , large family (discrete and continuous models),...

# Exponential family

Some very useful properties in the class of models...

## Proposition

*The canonical statistic is sufficient.*

trivial with FC...

## Proposition

*For exponential family, if the  $S_i(\cdot)$  are linearly independent (affine sense), i.e.,*

$$\forall x \in \mathcal{X}, \sum_{i=1}^r a_i S_i(x) = a_0 \implies a_0 = a_j = 0 \forall j$$

*Thus  $P_{\theta_1} = P_{\theta_2} \iff Q_j(\theta_1) = Q_j(\theta_2)$ .*

## Corollary

*For exponential family, if the  $S_i(\cdot)$  are linearly independent,  $\theta$  is identifiable  $\iff \theta \mapsto Q(\theta)$  is injective.*

# Exponential family

Some very useful properties in the class of models...

## Theorem

*If  $Q(\Theta)$  contains a non-empty set of  $\mathbb{R}^r$ , the canonical statistic is complete.*

## Proposition

*Of course, the canonical statistic follows an exponential model.*

Models examples:

- Exponential dist.!
- Gaussian
- Poisson
- Binomial dist.
- ...
- Exhaustive list on Wikipedia

# Fisher Information (FI) Matrix (FIM)

## Definition (Score)

The *score function* is the r.V.  $s_\theta(\mathbf{x})$  defined by:

$$s_\theta(\mathbf{x}) = \frac{\partial}{\partial \theta} l(\mathbf{x}; \theta),$$

where  $l(x; \theta) = \log(L(x; \theta))$  is the log-likelihood function.

## Proposition

The score is zero-mean, i.e.  $E[s_\theta(\mathbf{x})] = 0$ .

## Definition (FIM)

If one has (A<sub>5</sub>) the score is square-integrable, the FIM is the variance (covariance matrix in multidimensional case) of the score:

$$I(\theta) = \text{var}_\theta(s_\theta(\mathbf{x})) = E_\theta [s_\theta(\mathbf{x}) s_\theta(\mathbf{x})^t].$$



# FIM

## Remark

In case of a  $n$ -sample,  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ , the score can be written as:

$$s_{n,\theta}(\mathbf{x}) = \frac{\partial}{\partial \theta} l_n(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} l(\mathbf{x}_i; \theta),$$

where  $l_n(x_1, \dots, x_n; \theta)$  is the log-likelihood function of the  $n$ -sample. In such case, the FIM,  $I_n(\theta)$  can be written (by independence) as

$$I_n(\theta) = nI(\theta).$$

## Proposition

Let's assume a regular model, plus  $(A_5)$ , then for a real  $\theta$ , one has:

$$I(\theta) = -E_{\theta} \left[ \frac{\partial^2}{\partial \theta \partial \theta^t} l(\mathbf{x}; \theta) \right].$$

# FIM

Some examples...

Let us consider a  $n$ -sample of r.v. Prove the following results:

1 If  $P_\theta \sim B(\theta, 1), \theta \in ]0, 1[$ , thus  $I_n(\theta) = \frac{n}{\theta(1-\theta)}$ .

2 If  $P_\theta \sim \text{Poisson}(\theta), \theta > 0$ , thus  $I_n(\theta) = \frac{n}{\theta}$ .

3 If  $P_\theta \sim \mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ , thus:

$$I_n(\theta) = n \begin{pmatrix} 1 & 0 \\ \sigma^2 & 1 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

# Unbiased estimation - Decision theory

Main idea: give an answer  $d$  regarding the data...

Define a loss function  $\rho(d, \theta)$  between  $d$  and the (true) value of the unknowns  $\theta$  or  $g(\theta)$ . Generally,

Definition (quadratic loss)

$$\rho(d, \theta) = (d - g(\theta))^t \mathbf{A}(\theta) (d - g(\theta))$$

where  $\mathbf{A}(\cdot)$  is positive-definite

Use  $\mathbf{A}(\theta) = \mathbf{I}$  leads to  $\rho(d, \theta) = (d - g(\theta))^2 \dots$

Definition (**Estimator**)

An *estimator* of  $g(\theta)$  is a statistic  $\delta(\mathbf{x})$  mapping  $\mathcal{X}$  into  $\mathcal{D} = g(\Theta)$ .

Definition (**Mean Square Error (MSE)**)

$$R_\delta(\theta) = E_\theta [\rho(\theta, \delta(\mathbf{x}))] = E_\theta [(g(\theta) - \delta(\mathbf{x}))^2].$$

# Cramer-Rao lower bound

## Theorem (Cramer-Rao lower Bound (CRB) - FDCR inequality)

Let  $\delta$  an unbiased, regular estimator of  $g(\theta) \in \mathbb{R}^k$  where  $\theta \in \Theta \subset \mathbb{R}^p$ . The function  $g$  is of class  $C^1$ . Let's also assume that  $I(\theta)$  is positive-definite. Thus, for a  $n$ -sample, and for all  $\theta \in \Theta$ , one has:

$$R_{\delta}(\theta) = \text{var}_{\theta}(\delta) \geq \frac{1}{n} \frac{\partial g}{\partial \theta^t}(\theta) I(\theta)^{-1} \frac{\partial g^t}{\partial \theta}(\theta),$$

with  $\frac{\partial g}{\partial \theta^t}(\theta)$  the  $p \times k$ -matrix defined by  $\left( \frac{\partial g_i}{\partial \theta_j}(\theta) \right)_{1 \leq i \leq p, 1 \leq j \leq k}$  and

$\frac{\partial g^t}{\partial \theta}(\theta) = \left( \frac{\partial g}{\partial \theta^t}(\theta) \right)^t$  its transpose.

# Cramer-Rao lower bound

## Definition (Efficiency)

An unbiased estimator is said to be *efficient* iff its variance is the CRB.

## Proposition

If  $T$  is an efficient estimator of  $g(\theta)$ , then the affine transform  $\mathbf{A}T + \mathbf{b}$  is an efficient estimator of  $\mathbf{A}g(\theta) + \mathbf{b}$  (for  $\mathbf{A}$  and  $\mathbf{b}$  with appropriate dimensions)

## Proposition

*An efficient estimator is optimal.*  
*The converse is (obviously) wrong.*

Think about the students grades in a given course

## Link with exponential family

Consider an exponential model (1),  $L(\mathbf{x}; \theta) = h(\mathbf{x})\phi(\theta) \exp \left\{ \sum_{i=1}^r Q_i(\theta) S_i(\mathbf{x}) \right\}$   
and make the change of variable  $\lambda_j = Q_j(\theta)$ . Then, one obtains:

**Definition (Exponential model under a natural form...)**

... when the LR is

$$L(\mathbf{x}, \lambda) = K(\lambda) h(\mathbf{x}) \exp \left[ \sum_{j=1}^r \lambda_j S_j(\mathbf{x}) \right] \quad (2)$$

The new parameters  $(\lambda_1, \dots, \lambda_r) \in \Lambda = Q(\Theta) \subset \mathbb{R}^r$

**Theorem (Regularity)**

Let an exponential model (2). If  $\Lambda$  is a non-empty open set of  $\mathbb{R}^r$ , then the model is regular and  $(A_5)$  is verified,  $\Rightarrow I(\lambda)$  exists. Furthermore

$$I(\lambda) = -E_{\lambda} \left[ \frac{\partial^2 \ln L(\mathbf{x}, \lambda)}{\partial \lambda \partial \lambda^t} \right]$$

# Link with exponential family

## Theorem (**Identifiability**)

Let us consider the exponential model (2) where  $\Lambda$  is a (non-empty) open set of  $\mathbb{R}^r$ . Then, the model is identifiable, i.e.,  $(P_{\lambda_1} = P_{\lambda_2} \implies \lambda_1 = \lambda_2)$  *iff* the FIM  $I(\lambda)$  is invertible  $\forall \lambda \in \Lambda$ .

## Theorem (**Necessary condition**)

Let us consider the exponential model (1). Let us assume that the model is regular et let  $\delta$  an unbiased regular estimator of  $g(\theta)$ . Moreover, let us assume that  $g$  is of class  $C^1$  and that  $I(\theta)$  is invertible  $\forall \theta \in \Theta$ . Thus, if  $\delta$  is efficient, it is necessary an affine function of  $S(\mathbf{x}) = (S_1(\mathbf{x}), \dots, S_r(\mathbf{x}))^t$ .

## Remark

*Previous theorem is useful for proving the NON efficiency of an estimator...*

## Theorem (Converse of the CRB - Equality)

Given a regular model where  $\Theta \subset \mathbb{R}^d$  is a non-empty open set, let  $g: \Theta \rightarrow \mathbb{R}^p$  of class  $C^1$  s.t.  $\frac{\partial g}{\partial \theta^t}(\theta)$  is a **square** invertible matrix  $\forall \theta \in \Theta$  so that  $p = d$ . Assume that  $I(\theta)$  exists and is invertible  $\forall \theta \in \Theta$ .

Thus  $\delta(\mathbf{x})$  is a regular and EFFICIENT (unbiased) estimator of  $g(\theta)$  iff  $L(x, \theta)$  can be written as:

$$L(x, \theta) = C(\theta) h(x) \exp \left[ \sum_{j=1}^d Q_j(\theta) S_j(x) \right]$$

where functions  $Q$  and  $C$  are s.t.

- $Q$  and  $C$  are differentiable  $\forall \theta \in \Theta$
- $\frac{\partial Q}{\partial \theta^t}(\theta)$  is invertible  $\forall \theta \in \Theta$
- $g(\theta) = - \left( \frac{\partial Q}{\partial \theta^t}(\theta) \right)^{-1} \frac{\partial \ln C}{\partial \theta^t}(\theta)$ .



I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

IV. Statistical modelling

V. Theory of Point Estimation

- Basics
- Method of Moment
- Method of Maximum Likelihood
- Bayesian estimation - MAP and MMSE

VI. Hypothesis testing - Decision theory

## Basics

Let us denote  $T_n(\mathbf{x}_1, \dots, \mathbf{x}_n)$  or  $\hat{\theta}_n$  an estimator of  $\theta$  (or the true value  $\theta_0$  if needed).

### Definition (**Consistency**)

An estimator  $\hat{\theta}_n$  of  $g(\theta)$  is strongly (resp. weakly) consistent if it  $P_{\theta_0}$ -almost surely (resp. in proba.) converges towards  $g(\theta_0)$ , with  $g: \Theta \rightarrow \mathbb{R}^p$ .

### Definition (**Asymptotically unbiased**)

An estimator  $\hat{\theta}_n$  of  $g(\theta)$  is **asymptotically unbiased** if its limiting distribution is zero-mean, i.e.,

$$\exists c_n \rightarrow \infty \text{ s.t. } c_n(\hat{\theta}_n - g(\theta_0)) \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathbf{z} \text{ with } E_{\theta_0}[\mathbf{z}] = 0.$$

Remark: Different from “unbiased at the limit”:  $E_{\theta_0}[\hat{\theta}_n] \xrightarrow[n \rightarrow \infty]{} g(\theta_0)$ .

# Basics

## Definition (Asymptotically normal)

$\hat{\theta}_n$  is asymptotically normal if

$$\sqrt{n}(\hat{\theta}_n - g(\theta_0)) \xrightarrow[n \rightarrow \infty]{\text{dist.}} \mathcal{N}(\mathbf{0}, \Sigma(\theta_0))$$

where  $\Sigma(\theta_0)$  (PDS) is the asymptotic CM of  $\hat{\theta}_n$ .

Remark: This implies that  $\hat{\theta}_n$  is asymptotically unbiased.

## Definition (Asymptotically efficient)

An estimator is asymptotically efficient if it is asymptotically normal and if:

$$\Sigma(\theta_0) = \frac{\partial g}{\partial \theta^t}(\theta_0) I(\theta_0)^{-1} \frac{\partial g^t}{\partial \theta}(\theta_0)$$

## Method of Moment

Let a  $n$ -sample  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  i.i.d. with  $\mathbf{x}_1 \sim P_\theta$  where  $\theta \in \Theta \subset \mathbb{R}^d$  s.t.  $E[|\mathbf{x}_1|^d] < \infty$ . Let us assume that:

$$\mathbf{m} = \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix} = \begin{pmatrix} \phi_1(\theta_1, \dots, \theta_d) \\ \vdots \\ \phi_d(\theta_1, \dots, \theta_d) \end{pmatrix} = \phi \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$$

where  $m_k = E[\mathbf{x}^k]$ . If function  $\phi$  is invertible (with inverse  $\psi$ ), one has:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} = \begin{pmatrix} \psi_1(m_1, \dots, m_d) \\ \vdots \\ \psi_d(m_1, \dots, m_d) \end{pmatrix} = \psi \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix}$$

### Theorem

- $U_p \xrightarrow[n \rightarrow \infty]{a.s.} m_p$  where  $\forall p, U_p = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^p$
- $\sqrt{n}(\mathbf{U} - \mathbf{m}) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}(\mathbf{0}, \mathbf{Z})$  where  $\mathbf{U} = (U_1, \dots, U_p)^t$ ,  $\mathbf{m} = (m_1, \dots, m_p)^t$ .

# Method of Moment

The estimator of the Method of Moment (MME) is defined as

$$\hat{\theta}_n = \begin{pmatrix} \hat{\theta}_{n1} \\ \vdots \\ \hat{\theta}_{nd} \end{pmatrix} = \begin{pmatrix} \psi_1(U_1, \dots, U_d) \\ \vdots \\ \psi_d(U_1, \dots, U_d) \end{pmatrix} = \psi \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}$$

where  $\forall p, U_p = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^p$  with  $\mathbf{x}_i$  are i.i.d.

## Theorem (Asymptotics of the MM estimator)

If function  $\psi$  is differentiable, then

- $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{a.s.} \theta$
- $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}(\mathbf{0}, \mathbf{A}(\theta))$  where  $\mathbf{A}(\theta) = \frac{\partial \psi}{\partial \theta^t}(m) \Sigma(\theta) \frac{\partial \psi^t}{\partial \theta}(m)$  with  $m = \phi(\theta)$ .

**MME strongly consistent, asymptotically normal BUT generally NOT asymptotically efficient!**

# Method of Maximum Likelihood

Assume a regular model +  $(A_5)$  +

$(A_6)$   $\forall x \in \Delta$ , for  $\theta$  close to  $\theta_0$ ,  $\log(f(x; \theta))$  is 3 $\times$  differentiable w.r.t.  $\theta$  and

$$\left| \frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log(f(x; \theta)) \right| \leq M(x)$$

with  $E_{\theta_0} [M(x)] < +\infty$ .

## Proposition

Assume the model is identifiable, then  $\forall \theta \neq \theta_0$ , one has

$$P_{\theta_0} (L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta_0) > L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta)) \xrightarrow[n \rightarrow \infty]{} 1$$

where  $L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta)$  is the LF.

The LF is maximum at the point  $\theta_0$ ...

# Method of Maximum Likelihood

## Definition (Maximum Likelihood Estimator (MLE))

The MLE is defined by

$$T: (\mathbf{x}_1, \dots, \mathbf{x}_n) \rightarrow \hat{\theta}_n \in \arg \max_{\theta \in \Theta} L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta).$$


The MLE has to verified the following likelihood equations!

$$\begin{cases} \frac{\partial}{\partial \theta} l(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) = 0 \\ \frac{\partial^2}{\partial \theta \partial \theta^t} l(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) \leq 0, \end{cases}$$

where  $l(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) = \log(L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta))$

## Definition

Let  $g: \Theta \rightarrow \mathbb{R}^p$ . If  $\hat{\theta}_n$  is a MLE of  $\theta$ , then  $g(\hat{\theta}_n)$  is also a MLE of  $g(\theta)$ .

 the MLE is not necessary unique...

# MLE asymptotics

## Theorem

Assume: identifiable model,  $(A_1)$ ,  $(A_2)$ ,  $\theta_0 \in \Theta \neq \emptyset$ , compact, and

- $x_1 \mapsto L(x_1, \theta)$  is bounded  $\forall \theta \in \Theta$ ;
- $\theta \mapsto L(x_1, \theta)$  is continuous  $\forall x_1 \in \Delta$ ;

Thus,  $\hat{\theta}_n^{ML} \xrightarrow[n \rightarrow \infty]{a.s.} \theta_0$  (Existence from a given  $n_0$ )

## Theorem (Classical asymptotics)

Assume: identifiable model,  $\Theta$  open set of  $\mathbb{R}^d$  and  $(A_1) - (A_6)$ .

Thus,  $\exists \hat{\theta}_n^{ML}$  (from a given  $n_0$ ) solution to the likelihood equations s.t.

$$\left\{ \begin{array}{l} \hat{\theta}_n^{ML} \xrightarrow[n \rightarrow \infty]{a.s.} \theta_0 \\ \sqrt{n}(\hat{\theta}_n^{ML} - \theta_0) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}(\mathbf{0}, I_1(\theta_0)^{-1}) \end{array} \right.$$



# MLE asymptotics

## Theorem (Classical asymptotics)

Assume: identifiable model,  $\Theta$  open set of  $\mathbb{R}^d$  and  $(A_1) - (A_6)$  AND  $g: \mathbb{R}^d \rightarrow \mathbb{R}^p$  differentiable

Thus,  $\exists \hat{\theta}_n^{ML}$  (from a given  $n_0$ ) solution to the likelihood equations s.t.

$$\left\{ \begin{array}{l} g(\hat{\theta}_n^{ML}) \xrightarrow[n \rightarrow \infty]{a.s.} g(\theta_0) \\ \sqrt{n}(g(\hat{\theta}_n^{ML}) - g(\theta_0)) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}\left(\mathbf{0}, \frac{\partial g}{\partial \theta^t}(\theta_0) I_1(\theta_0)^{-1} \frac{\partial g^t}{\partial \theta}(\theta_0)\right) \end{array} \right.$$

## Conclusions

The MLE is strongly consistent, asymptotically normal and asymptotically efficient.

# Come back on exponential models

## Theorem

Let an exponential model (2) (under natural form)

$$L(x, \lambda) = K(\lambda) h(x) \exp\left(\sum_{j=1}^r \lambda_j S_j(x)\right)$$

where  $\lambda \in \Lambda$  and  $\Lambda$  is a non-empty open-set of  $\mathbb{R}^r$ . Moreover, let us assume that  $I(\lambda)$  is invertible  $\forall \lambda \in \Lambda$  (identifiable model).

Thus, the MLE exists (from a given  $n_0$ ), is unique, strongly consistant and asymptotically efficient (which includes asymptotically normal).

## Proof

Up to you ...

# Bayesian estimation

- **Principles:** Philosophy is different from previous MM/ML estimation approaches (frequentist methods). **The purpose is the same: estimating an unknown parameter  $\theta \in \mathbb{R}$  or  $\mathbb{R}^p$  thanks to the sample  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  likelihood (parameterized by  $\theta$ ) and an a priori distribution  $p(\theta)$ .**

So,  $\theta$  is assumed to random...

- **Ideas:** To that end, one has to minimize a cost function  $c(\theta, \hat{\theta})$  that represents the error between  $\theta$  and its estimator  $\hat{\theta}$ .
- **Reminders:** A posteriori distribution / posterior distribution

$$p(\theta | \mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) p(\theta)}{f(\mathbf{x}_1, \dots, \mathbf{x}_n)} = \frac{L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) p(\theta)}{\int_{\mathbb{R}^p} L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) p(\theta) d\theta}$$
$$\propto L(\mathbf{x}_1, \dots, \mathbf{x}_n; \theta) p(\theta)$$

## MMSE estimator

**MMSE estimator** (mean of the posterior PDF) is the estimator that minimizes the MSE as the cost function:  $c(\theta, \hat{\theta}) = E[(\theta - \hat{\theta})^2]$ .

$$\theta \in \mathbb{R}$$

$$E[(\theta - \hat{\theta}_{MMSE}(\mathbf{x}))^2] = \min_{\pi} E[(\theta - \pi(\mathbf{x}))^2]$$

with  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ , hence the **MMSE estimator** is

$$\hat{\theta}_{MMSE}(\mathbf{x}) = E[\theta|\mathbf{x}]$$

$$\theta \in \mathbb{R}^p$$

The **MMSE estimator**  $\hat{\theta}_{MMSE}(\mathbf{x}) = E[\theta|\mathbf{x}]$  minimizes the quadratic cost

$$E[(\theta - \pi(\mathbf{x}))^t \mathbf{Q} (\theta - \pi(\mathbf{x}))]$$

for any symmetric definite positive matrix  $\mathbf{Q}$  (and in particular for  $\mathbf{Q} = \mathbf{I}_p$ , the identity matrix).

# MAP estimator

$\theta \in \mathbb{R}$

The **MAP estimator**  $\hat{\theta}_{MAP}(\mathbf{x})$  minimizes the average of a “uniform” cost function

$$c((\theta - \pi(\mathbf{x}))) = \begin{cases} 0 & \text{if } |\theta - \pi(\mathbf{x})| \leq \Lambda/2 \\ 1 & \text{if } |\theta - \pi(\mathbf{x})| > \Lambda/2 \end{cases}$$

and is defined by

$$c((\theta - \hat{\theta}_{MAP}(\mathbf{x}))) = \min_{\pi} c((\theta - \pi(\mathbf{x})))$$

If  $\Lambda$  is arbitrary small,  $\hat{\theta}_{MAP}(\mathbf{x})$  is the value of  $\pi(\mathbf{x})$  which maximizes the posterior  $p(\theta|\mathbf{x})$  hence its name **MAP estimator**.  $\hat{\theta}_{MAP}(\mathbf{x})$  is computed by setting to zero the derivative of  $p(\theta|\mathbf{x})$  (or of its log) with respect to  $\theta$ .

$\theta \in \mathbb{R}^p$

Determine the values of  $\theta_i$  which make the partial derivatives of  $p(\theta|\mathbf{x})$  (or of its logarithm) with respect to  $\theta_i$  equal to zero.

I. Introduction in stat. signal processing

II. Random Variables / Vectors / CV

III. Essential theorems

IV. Statistical modelling

V. Theory of Point Estimation

VI. Hypothesis testing - Decision theory

- Generalities
- UMP tests
- Student- $t$  test
- Asymptotic Tests

# Generalities

Let a  $n$ -sample  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$  i.i.d.  $\sim P_\theta, \theta \in \Theta$ . Let  $H_0$  and  $H_1$ , 2 non-empty disjoint subsets of  $\Theta$  s.t.  $H_0 \cup H_1 = \Theta$ .

$H_0$  is the **null hypothesis** while  $H_1$  is called the **alternative hypothesis**.

**Remember: no symmetry!**

**Goal: To find a procedure that allows to decide whether  $\theta$  belongs to  $H_0$  or not, regarding the datasets  $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ .**

## Definition

*An hypothesis is said **simple** if it is reduced to a single element. Else, it is called **composite**.*

## Definition

*A (**pure**) test is a mapping  $\delta$  from  $\mathcal{X}^n$  onto  $\{0, 1\}$  s.t.:*

*If  $\delta(x) = 0$ , one decides  $H_0$ , while if  $\delta(x) = 1$ , one rejects  $H_0$ .*

*The region  $W = \{x \in \mathcal{X}^n \mid \delta(x) = 1\}$  is called the **rejection region** or the **critical region**. Its complement is called the **acceptance region**.*

# Generalities

## Remark

*A test is characterized (and will be identified) by its rejection region  $W$ .*

## Definition (Different errors)

*For a test, there are two possible errors:*

- *rejecting  $H_0$  when it is true: **type-I error** or error of 1<sup>st</sup> kind.*
- *accepting  $H_0$  when it is false: **type-II error** or error of 2<sup>nd</sup> kind.*

## Definition (Type-I and Type-II errors)

*For a test  $\delta$  with critical region  $W$ , one has*

- **Type-I error:**  $\alpha_W: \begin{cases} H_0 \rightarrow [0, 1] \\ \theta \mapsto P_\theta(W); \end{cases}$
- **Type-II error:**  $\beta_W: \begin{cases} H_1 \rightarrow [0, 1] \\ \theta \mapsto P_\theta(W^c) = 1 - P_\theta(W). \end{cases}$



# Generalities

## Definition (**Power of the test**)

The *power* of a test  $W$  is defined as:

$$\rho_W : \begin{cases} H_1 \rightarrow [0, 1] \\ \theta \mapsto P_\theta(W) = 1 - \beta_W(\theta). \end{cases}$$

## Definition (**Randomized test (more general)**)

A random test is a mapping  $\varphi$  from  $\mathcal{X}^n$  into  $[0, 1]$  where  $\varphi(x)$  is the probability of rejecting  $H_0$  for the dataset  $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ .

## Remark

For  $\varphi = \mathbb{1}_W$ , one retrieves the simple test!

# Generalities

## Definition (Type-I and Type-II errors, power for a test $\varphi$ )

- **Type-I error:**  $\alpha_\varphi : \begin{cases} H_0 \rightarrow [0, 1] \\ \theta \mapsto E_\theta [\varphi(\mathbf{x})]; \end{cases}$
- **Type-II error:**  $\beta_\varphi : \begin{cases} H_1 \rightarrow [0, 1] \\ \theta \mapsto 1 - E_\theta [\varphi(\mathbf{x})]; \end{cases}$
- **Power of the test:**  $\rho_\varphi = 1 - \beta_\varphi = E_{H_1} [\varphi(\mathbf{x})].$

## Definition (Level of significance (ls))

The **level of significance**  $\alpha$  (typically 0.01 or 0.05 as for the IC) for a test  $\varphi$  is:

$$\alpha = \sup_{\theta \in H_0} \alpha_\varphi(\theta) = \sup_{\theta \in H_0} E_\theta [\varphi(\mathbf{x})].$$

# Neyman Principle

Goal: one wants to control (or fix) the type-I error, i.e. the probability of rejecting  $H_0$  when it is true.

The Neyman principle consists in considering all tests with a  $l_s \leq$  to a fixed  $\alpha$ , and then, in finding (among these tests) the one with the smallest Type-II error.

Since  $\rho_\varphi = 1 - \beta_\varphi$ , such test will be said to be UMP.

Definition (**Uniformly Most Powerful (UMP)**)

$\varphi$  is UMP at the threshold  $\alpha$  if its  $l_s \leq \alpha$  and if  $\forall \varphi'$  with a  $l_s \leq \alpha$ , one has:

$$\forall \theta \in H_1, E_\theta [\varphi(\mathbf{x})] \geq E_\theta [\varphi'(\mathbf{x})].$$

# Simple hypothesis testing

In this part, for the  $n$ -sample  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ , one considers,

$$H_0 : \{\theta = \theta_0\} \text{ versus } H_1 : \{\theta = \theta_1\},$$

which means that  $\Theta = \{\theta_0, \theta_1\}$ .

So, 2 probabilities  $P_{\theta_0}$  (or  $P_0$ ) and  $P_{\theta_1}$  (or  $P_1$ ), that implies 2 LF  $L_0(x) = L(x; \theta_0)$  and  $L_1(x) = L(x; \theta_1)$ , for  $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ .

## Definition (Neyman test or Likelihood Ratio Test (LRT))

A Neyman test is a test  $\varphi$  s.t.  $\exists k \in \mathbb{R}_+^*$ , and

$$\varphi(x) = \begin{cases} 1 & \text{if } L(x; \theta_1) > kL(x; \theta_0) \\ 0 & \text{if } L(x; \theta_1) < kL(x; \theta_0) \end{cases}$$

The value of  $\varphi$  is not specified for  $\{x \in \mathcal{X}^n \mid L_1(x) = kL_0(x)\}$ .

# Neyman-Pearson Lemma

## Remark

$L_1(x)/L_0(x)$  is called the **Likelihood Ratio (LR)**. The Neyman test consists in accepting the most likely hypothesis for a given observation  $x$ .

## Proposition (Neyman-Pearson Lemma)

- 1 Existence**  $\forall \alpha \in (0, 1)$ , it exists a Neyman test s.t.  $E_{\theta_0}(\varphi) = \alpha$ .  
Moreover,  $k$  is the quantile of order  $(1 - \alpha)$  of the LR distribution  $\frac{L_1(x)}{L_0(x)}$  under  $P_0$  and one can impose that  $\varphi$  is constant for  $x \in \mathcal{X}^n$  s.t.  $\underline{L_1(x) = kL_0(x)}$ . If the LR CDF under  $P_0$  evaluated in  $k$  is  $(1 - \alpha)$  (**continuous CDF**), thus one can choose this constant = 0 (pure test).
- 2 S. cond.**  $\forall \alpha \in (0, 1)$ , a Neyman test s.t.  $E_{\theta_0}(\varphi) = \alpha$  is **UMP** at level  $\alpha$ .
- 3 N. cond.**  $\forall \alpha \in (0, 1)$ , a UMP test at level  $\alpha$  is **necessarily a Neyman test**.

## Proof

*Essential to built the Neyman test...*

# Neyman-Pearson Lemma

## Remark

- 1 *Conclusion: the only UMP tests at level  $\alpha$  are the Neyman tests of level of significance  $\alpha$ .*
- 2 *If the LR CDF under  $H_0$  is continuous, one obtains the test of critical region  $W = \{x \in \mathcal{X}^n \mid L_1(x) > kL_0(x)\}$  where  $k$  is defined by  $P_0(L_1(X) > kL_0(X)) = \alpha$ .*
- 3 *The power  $E_1(\varphi)$  of a UMP test at level  $\alpha$  is necessarily  $\geq \alpha$ . Indeed,  $\varphi$  is preferable to the constant test  $\psi = \alpha$  (which is of ls  $\alpha$ ), thus  $E_1(\varphi) \geq E_1(\psi) = \alpha$ .*

# Neyman-Pearson Lemma

**Example 1:** Let us consider the exponential model (1)

$$L(x, \theta) = C(\theta) h(x) \exp \left[ \sum_{j=1}^d Q_j(\theta) S_j(x) \right]$$

where  $\theta \in \{\theta_0, \theta_1\}$ , with  $\theta_1 > \theta_0$ . Assume an identifiable model:  
 $Q(\theta_0) \neq Q(\theta_1)$  (e.g.,  $Q(\theta_1) > Q(\theta_0)$ ).

Goal: test  $H_0 : \{\theta = \theta_0\}$  versus  $H_1 : \{\theta = \theta_1\}$ .

**Example 2:** Let us consider  $(X_1, \dots, X_n) \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known.

Goal: test  $H_0 : \{\mu = \mu_0\}$  versus  $H_1 : \{\mu = \mu_1\}$ , with  $\mu_0 < \mu_1$ .

**Example 3:** Let us consider  $(X_1, \dots, X_n) \stackrel{iid}{\sim} \text{Poisson}(\theta)$ .

Goal: test  $H_0 : \{\theta = \theta_0\}$  versus  $H_1 : \{\theta = \theta_1\}$ , with  $\theta_0 < \theta_1$ .

## Composite tests - One-sided hypotheses

Now, let us consider a model with only 1 parameter and where  $\Theta$  is an interval of  $\mathbb{R}$ . One assume  $L(x, \theta) > 0, \forall x \in \mathcal{X}^n, \forall \theta \in \Theta$ .

**Goal: test  $H_0 : \{\theta \leq \theta_0\}$  versus  $H_1 : \{\theta > \theta_0\}$ .**

More general problem!

Let us consider the family having **monotone likelihood ratio**:

### Definition (**Monotone LR**)

The family  $\{P_\theta^{\otimes n}, \theta \in \Theta\}$  is said to have **monotone likelihood ratio** if it exists a real-valued statistic  $U(x)$  s.t.  $\forall \theta' < \theta'', \frac{L(x, \theta'')}{L(x, \theta')}$  is a strictly increasing (or decreasing) function of  $U$ .

### Remark

By changing  $U$  into  $-U$ , one can always assume strictly increasing in previous definition.



# Lehman Theorem

## Theorem (**Lehman theorem**)

Let  $\alpha \in (0, 1)$ . If the family  $(P_\theta, \theta \in \Theta)$  has **monotone (increasing) likelihood ratio**, there exists a UMP test at level  $\alpha$  for testing  $H_0 : \{\theta \leq \theta_0\}$  versus  $H_1 = \{\theta > \theta_0\}$ . This test is defined by:

$$\begin{cases} \varphi(x) = 1 & \text{if } U(x) > c \\ \varphi(x) = \gamma & \text{if } U(x) = c \\ \varphi(x) = 0 & \text{if } U(x) < c \end{cases}$$

where  $c$  and  $\gamma$  are obtained with  $E_{\theta_0}[\varphi] = \alpha$ . The same test is UMP at level  $\alpha$  for testing:

1  $H_0 : \{\theta = \theta_0\}$  versus  $H_1 : \{\theta > \theta_0\}$

2  $H_0 : \{\theta = \theta_0\}$  versus  $H_1 : \{\theta = \theta_1\}$

where  $\theta_1 > \theta_0$ .

# Lehman Theorem

## Remark

If the inequalities are reversed in the test, i.e.  $H_0 : \{\theta \geq \theta_0\}$  and  $H_1 : \{\theta < \theta_0\}$ , then the UMP test is obtained by reversing the inequalities (in the test).

**Example:** The exponential model with LF  $L(x, \theta) = C(\theta)h(x) \exp(Q(\theta)S(x))$  where  $Q(\theta)$  is strictly increasing, has increasing LR with  $U(X) = S(X)$ .

## Remark (Important)

In general, *it does NOT exist* UMP test for testing  $H_0 : \{\theta = \theta_0\}$  versus  $H_1 : \{\theta \neq \theta_0\}$  (even for monotone LR).

For instance, let's consider the Gaussian model,  $\sigma^2$  known. The UMP test

for  $H_0 : \{\mu = \mu_0\}$  versus  $H_1 : \{\mu > \mu_0\}$  is  $\begin{cases} \rho(x) = 1 & \text{if } \sum x_i > c \\ \rho(x) = 0 & \text{if } \sum x_i \leq c \end{cases}$  while the

UMP test for  $H_0 : \{\mu = \mu_0\}$  versus  $H_1 : \{\mu < \mu_0\}$  is  $\begin{cases} \rho(x) = 1 & \text{if } \sum x_i < c \\ \rho(x) = 0 & \text{if } \sum x_i \geq c \end{cases}$

$\Rightarrow$  **no UMP test for testing  $\mu = \mu_0$  versus  $\mu \neq \mu_0$ .**

# Student test

Let  $(X_1, \dots, X_n) \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\mu$  and  $\sigma^2$  unknown.

**Goal:** test  $H_0 : \{\mu = \mu_0\}$  versus  $H_1 : \{\mu \neq \mu_0\}$  at level  $\alpha \in (0, 1)$ .

## General methodology

- 1 From the *Student theorem*, one has

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t(n-1)$$

where  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  and  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ .

- 2 Under  $H_0$ :

$$\xi_n = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} \sim t(n-1)$$

- 3 Under  $H_1$ : From the SLLN,  $\bar{X}_n - \mu_0 \xrightarrow[n \rightarrow \infty]{a.s.} \mu - \mu_0$  and  $S_n \xrightarrow[n \rightarrow \infty]{a.s.} \sigma$ .

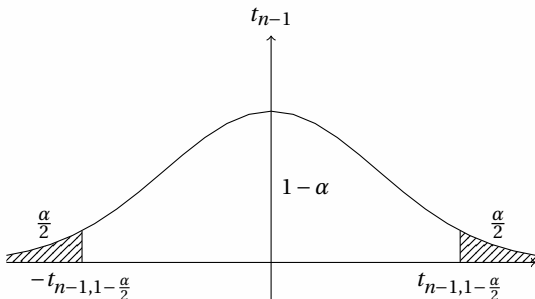
Thus  $\xi \xrightarrow[n \rightarrow \infty]{a.s.} +\infty$  if  $\mu > \mu_0$  and  $\xi \xrightarrow[n \rightarrow \infty]{a.s.} -\infty$  if  $\mu < \mu_0$

- 4 **Critical region:**

$$\mathbf{W}_n = \{|\xi_n| > \mathbf{a}\}$$

# Student test

- Let  $t_{n-1,r}$  the quantile of order  $r$  of the  $t$ -distribution  $t_{n-1}$ :



Thus, under  $H_0$ ,  $P(|\xi_n| > t_{n-1,1-\frac{\alpha}{2}}) = \alpha$ .

Previously, one have seen that  $I_n = \left[ \bar{X}_n - \frac{t_{n-1,1-\alpha/2} S_n}{\sqrt{n}}, \bar{X}_n + \frac{t_{n-1,1-\alpha/2} S_n}{\sqrt{n}} \right]$  is a  $(1 - \alpha)$ -CI for  $\mu_0$ . Here is the link between CI and Student (bilateral) test  $\mu_0 \in I_n$  iff  $|\xi_n| \leq t_{n-1,1-\frac{\alpha}{2}}$ . Finally, the associated  $p$ -value is  $p = P(|T| > |\xi_n^{obs}|)$  where  $T \sim t(n-1)$  and  $\xi_n^{obs}$  is the observed value of  $\xi_n$ .

# Generalities

As for estimators, in many situations, one CANNOT find the distribution of the LR (or the statistic of the monotone LR). As a consequence, one cannot set the parameters  $k$  and  $\gamma$  for the test.

A solution (like in point estimation theory) is to rely on asymptotic properties!

Now, instead of considering a test  $W$ , we will consider a sequence of tests  $(W_n)_{n \in \mathbb{N}^*}$ .

## Definition (Asymptotic level)

An asymptotic test  $W_n$  is at asymptotic level  $\alpha$  if

$$\lim_{n \rightarrow \infty} \sup_{\theta \in H_0} P_\theta(W_n) = \alpha.$$

# Generalities

## Definition (**Uniform asymptotic level**)

An asymptotic test  $W_n$  is at **uniform asymptotic level**  $\alpha$  if

$$\sup_{\theta \in H_0} \lim_{n \rightarrow \infty} P_{\theta}(W_n) = \alpha.$$

## Definition (**Consistant (or convergent) test**)

An asymptotic test  $W_n$  is said to be **consistant (or convergent)** if its power tends towards 1, i.e.,

$$\forall \theta \in H_1, \lim_{n \rightarrow \infty} P_{\theta}(W_n) = 1.$$

***This means that the Type-II error tends to 0!***

**Example:** the  $t$ -test is consistant...

# Asymptotic tests

**Implicit constraint:**  $H_0 : \{\theta | g(\theta) = 0\}$

where  $g$  a mapping from  $\mathbb{R}^d$  into  $\mathbb{R}^r$ , of class  $C^1$  s.t. the  $r \times d$  matrix

$$\frac{\partial g}{\partial \theta^t} = \left( \frac{\partial g_i}{\partial \theta_j} \right)_{1 \leq i \leq r, 1 \leq j \leq d} \text{ is of rank } r \text{ (so } r \leq d \text{)}.$$

Goal: test  $H_0 : \{\theta \in \Theta, g(\theta) = 0\}$  versus the alternative hypothesis  
 $H_1 : \{\theta \in \Theta, g(\theta) \neq 0\}$

**More general than**  $H_0 : \{\theta = \theta_0\}$  **versus**  $H_1 : \{\theta \neq \theta_0\}$

To answer such problems, there exist (at least) 3 asymptotic tests:

- Wald test
- Rao (score) test
- Likelihood Ratio Test (LRT)

# Wald test

## Proposition (Wald test)

Let  $\hat{\theta}_n^{ML}$  the MLE of  $\theta$ . Under  $H_0$ , the sequence of r.V., one has:  
 $(\sqrt{n}g(\hat{\theta}_n^{ML})) \xrightarrow[n \rightarrow \infty]{dist.} \mathcal{N}(\mathbf{0}, \Sigma(\theta_0))$ , where  $\theta_0 \in H_0$  is the true value of the parameter  $\theta$  and where  $\Sigma(\theta_0) = \frac{\partial g}{\partial \theta^t}(\theta_0) I_1(\theta_0)^{-1} \frac{\partial g^t}{\partial \theta}(\theta_0)$ .

Furthermore, the test statistic  $\xi_n^W = n g(\hat{\theta}_n^{ML})^t \Sigma(\hat{\theta}_n^{ML})^{-1} g(\hat{\theta}_n^{ML})$  converges in distribution under  $H_0$  towards a  $\chi^2$ -distribution with  $r$  d.o.f.:

$$\xi_n^W \xrightarrow[n \rightarrow \infty]{dist.} \chi^2(r)$$

The Wald tests are defined by the following critical region:

$$W_n = \{\xi_n^W > q_r(1 - \alpha)\}$$

where  $q_r(1 - \alpha)$  is the quantile of order  $(1 - \alpha)$  of the  $\chi^2$ -distribution with  $r$  d.o.f. This test is strongly convergent at asymptotic level  $\alpha = P(\chi^2(r) > q_r(1 - \alpha))$ .



# Wald test

## Definition (*p-value*)

The asymptotic *p*-value of the Wald test is defined by

$$p = P(\chi^2(r) > \xi_n^W(x_1, \dots, x_n))$$

where  $\chi^2(r)$  is a r.v. following a  $\chi^2$ -dist. with  $r$  d.o.f. and  $\xi_n^W(x_1, \dots, x_n)$  is the observed test statistic. *One rejects  $H_0$  if  $p < \alpha$ ...*

## Remark

If one cannot compute  $I_1(\theta)$ . One can estimate  $I_1(\theta)$  by the MM and replace it in the Wald test *WITHOUT changing the results!*

$$\hat{I}_1(\cdot) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln L(x_i, \cdot)}{\partial \theta^t} \frac{\partial \ln L(x_i, \cdot)}{\partial \theta} \quad \text{ou} \quad \hat{I}_1(\cdot) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ln L(x_i, \cdot)}{\partial \theta \partial \theta^t}.$$

## Proof (Wald test)

*Allows to understand the methodology...*

## Wald test

**Example:** Let a Gaussian  $n$ -sample  $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}_{i \in \{1, \dots, n\}} \sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right)$  with  $\sigma_1$  and  $\sigma_2$  known. Let  $\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ .

**Goal:** test  $\mu_1 = \mu_2$ , i.e.,  $H_0 : \{\mu_1 - \mu_2 = 0\}$  versus  $H_1 : \{\mu_1 - \mu_2 \neq 0\}$ .

Let us set  $g(\theta) = \mu_2 - \mu_1$  and show that the Wald test statistic is

$$\xi_n^W = \frac{n(\hat{\mu}_1 - \hat{\mu}_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where  $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i$  and  $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n Y_i$ . One has

$$\xi_n^W \xrightarrow[n \rightarrow \infty]{dist.} \chi^2(1)$$

# Rao-score test and Likelihood Ratio test (LRT)

Let  $\hat{\theta}_n^c$  the MLE of  $\theta$  under the constraint  $g(\theta) = 0$ , i.e. under  $H_0$ .

## Theorem (Rao test and LRT)

The test statistics are defined by:

$$\xi_n^R = \frac{1}{n} \frac{\partial \ln L(x_i, \dots, x_n; \hat{\theta}_n^c)}{\partial \theta^t} I_1(\hat{\theta}_n^c)^{-1} \frac{\partial \ln L(x_i, \dots, x_n; \hat{\theta}_n^c)^t}{\partial \theta}$$
$$\xi_n^{LR} = 2(\ln L(x_i, \dots, x_n; \hat{\theta}_n) - \ln L(x_i, \dots, x_n; \hat{\theta}_n^c))$$

Rao test and the LRT are defined by the following critical region

$$W_n = \{\xi_n^i > q_r(1 - \alpha)\}$$

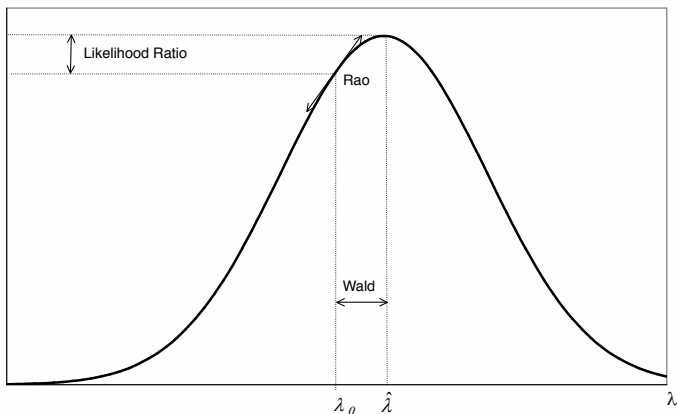
where  $q_r(1 - \alpha)$  is the quantile of order  $(1 - \alpha)$  of the  $\chi^2$ -distribution with  $r$  d.o.f. These tests are strongly convergent at asymptotic level  $\alpha = P(\chi^2(r) > q_r(1 - \alpha))$ . Furthermore, *under  $H_0$* , one has:

$$\xi_n^W - \xi_n^R \xrightarrow[n \rightarrow \infty]{P} 0 \quad \text{and} \quad \xi_n^W - \xi_n^{LR} \xrightarrow[n \rightarrow \infty]{P} 0$$

# Rao-score test and Likelihood Ratio test (LRT)

**Example** Testing  $H_0 : \{\lambda = \lambda_0\}$  versus  $H_1 : \{\lambda \neq \lambda_0\}$  in case of a Poisson distribution with parameter  $\lambda$ ...

...



## $\chi^2$ test: Goodness-of-Fit to a given distribution

Goal: test the goodness of fit of r.V. to a discrete and finite distribution (e.g., binomial, ...)

Quite restrictive but it CAN be extended to all distributions!

Let the  $n$ -sample  $(X_1, \dots, X_n)$  i.i.d. with values in  $\{a_1, \dots, a_m\}$  and distribution  $P$ , where  $P$  is characterized by its weights  $P = (p_1, \dots, p_m)$  (it is a PMF) with  $\sum_{i=1}^m p_i = 1$  and  $\forall j = 1, \dots, n, \forall i = 1, \dots, m, p_i = P(X_j = a_i)$ .

One wants to test  $H_0 : \{P = P_{p_0}\}$ , where  $p_0 = (p_1^0, \dots, p_m^0)$  is given (no unknown parameter) with  $\sum_{i=1}^m p_i^0 = 1, p_i^0 > 0, \forall i = 1, \dots, m$ .

Let  $N_i$  the counting statistic and  $p_i$  is the empirical frequency of  $\{X_k = a_i\}$ :

$$N_i = \sum_{k=1}^n \mathbb{1}_{\{X_k = a_i\}} \quad \text{and} \quad \hat{p}_i = \frac{N_i}{n}$$

# $\chi^2$ test: Goodness-of-Fit to a given distribution

## Theorem ( $\chi^2$ - test)

*Under  $H_0$*

$$\xi_n = \sum_{i=1}^m \frac{(N_i - np_i^0)^2}{np_i^0} = n \sum_{i=1}^m \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}$$

And  $\xi_n$  converges in distribution towards a  $\chi^2$ -distribution with  $(m-1)$  d.o.f. when  $n \rightarrow +\infty$ .

The test is defined by the critical region:

$$W_n = \{\xi_n > q_{m-1}(1-\alpha)\}$$

where  $q_{m-1}(1-\alpha)$  is the quantile of order  $(1-\alpha)$  of the  $\chi^2$ -distribution with  $(m-1)$  d.o.f. This test is strongly convergent at asymptotic level  $\alpha = P(\chi^2(m-1) > q_{m-1}(1-\alpha))$ .

**Example:** Toss a coin...

## $\chi^2$ test: Goodness-of-Fit to a given distribution

Now, let us test  $H_0 : \{p = p(\theta)\}$  versus  $H_1 : \{p \neq p(\theta)\}$  where  $\theta \in \Theta \subset \mathbb{R}^d$ ,  $\Theta$  open-set and  $\theta$  is unknown!

### Theorem (General $\chi^2$ - test)

Under  $H_0$

$$\xi_n = \sum_{i=1}^m \frac{(N_i - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)} = n \sum_{i=1}^m \frac{(\hat{p}_i - p_i(\hat{\theta}_n))^2}{p_i(\hat{\theta}_n)}$$

where  $\hat{\theta}_n$  is the MLE of  $\theta$ .

And  $\xi_n$  converges in distribution towards a  $\chi^2$ -distribution with  $(m-1-d)$  d.o.f. when  $n \rightarrow +\infty$ .

The test is defined by the critical region:

$$W_n = \{\xi_n > q_{m-1-d}(1-\alpha)\}$$

where  $q_{m-1-d}(1-\alpha)$  is the quantile of order  $(1-\alpha)$  of the  $\chi^2$ -distribution with  $(m-1-d)$  d.o.f. This test is strongly convergent at asymptotic level  $\alpha = P(\chi^2(m-1-d) > q_{m-1-d}(1-\alpha))$ .

# $\chi^2$ test: Goodness-of-Fit to a given distribution

How to generalize those  $\chi^2$  tests to continuous distribution or infinite discrete distribution?

## Remark (On the use of $\chi^2$ tests!)

- It is an *asymptotic* test. In practice, it works if  $np_i(\hat{\theta}_n) > 5, \forall i$  and if  $N_i \geq 5, \forall i$ . Else, one regroups classes (cf exercise in the problems).
- In case of continuous r.v. with unknown distribution, one wants to test if it belongs to the family  $\{P_\theta, \theta \in \Theta\}$ . The idea is to partition  $\mathbb{R}$  into  $m$  intervals  $(A_i)_{i=1, \dots, m}$ . The choice of  $m$  is a tradeoff:
  - $m$  should be sufficiently large so that the discrete dist.  $\{\pi_i = \pi(A_i)\}$  and  $\{p_{\theta,i} = P_\theta(A_i)\}$  be sufficiently close to  $\pi$  and  $P_\theta$  (if  $m$  is small, the test will be less powerful).
  - On the other hand,  $m$  should not be too large so that the  $p_{\theta,i}$  be sufficiently large to satisfy  $np_i(\hat{\theta}_n) > 5$ .



## $\chi^2$ test for independence

Let  $(X_k, Y_k), k = 1, \dots, n$  i.i.d. with values in  $\{a_1, \dots, a_l\} \times \{b_1, \dots, b_r\}$ . Let us denote  $p_{i,j} = P(X_1 = a_i, Y_1 = b_j)$  and

$$p_{i,\cdot} = P(X_1 = a_i) = \sum_{j=1}^r p_{i,j} \text{ and } p_{\cdot,j} = P(Y_1 = b_j) = \sum_{i=1}^l p_{i,j}$$

One wants to know if  $X_1$  and  $Y_1$  are independent, i.e. if

$$H_0 : \{p_{i,j} = p_{i,\cdot} p_{\cdot,j}, \forall i, j\}$$

Let  $N_{i,j} = \sum_{k=1}^n \mathbb{1}_{\{X_k = a_i, Y_k = b_j\}}$  the counting statistic and

$$N_{i,\cdot} = \sum_{k=1}^n \mathbb{1}_{\{X_k = a_i\}} \text{ and } N_{\cdot,j} = \sum_{k=1}^n \mathbb{1}_{\{Y_k = b_j\}}$$

# $\chi^2$ test for independence

## Theorem ( $\chi^2$ - test for independence)

*Under  $H_0$*

$$\xi_n = \sum_{i=1}^l \sum_{j=1}^r \frac{\left( N_{i,j} - \frac{N_{i.} N_{.j}}{n} \right)^2}{\frac{N_{i.} N_{.j}}{n}}$$

*And  $\xi_n$  converges in distribution towards a  $\chi^2$ -distribution with  $(r-1)(l-1)$  d.o.f.*

*The test is defined by the critical region:*

$$W_n = \{ \xi_n > q_{(r-1)(l-1)}(1 - \alpha) \}$$

*where  $q_{(r-1)(l-1)}(1 - \alpha)$  is the quantile of order  $(1 - \alpha)$  of the  $\chi^2$ -distribution with  $(r-1)(l-1)$  d.o.f. This test is strongly convergent at asymptotic level  $\alpha = P(\chi^2((r-1)(l-1)) > q_{(r-1)(l-1)}(1 - \alpha))$ .*

## $\chi^2$ test for independence

**Example** A study on 592 women: is there a correlation between eyes color and hairs color?

Eyes \ Hairs	Dark	Light-brown	Red	Blond
Black	68	119	26	7
Brown	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

One obtains  $\xi_n = 138,29$ ,  $dof = 9$ ,  $P(\chi^2_q \leq 16,91) = 0,95$ . Since  $138,29 \gg 16,91$ , one rejects  $H_0$ .