Basics of Probability and Statistics

Clément Pellegrini

clement.pellegrini@math.univ-toulouse.fr

Institut de Mathématiques de Toulouse,
Statistics and Probability team,
Bureau 220 Bâtiment 1R1

Content

- **Statistical Model**
- **Probability Background**
- **Law of Large Numbers, Central Limit Theorem**
- **Gaussian Vectors**

Content

- **Conditioning**
- **Estimation**
- **Confidence Set**
- **Basic of Regression**
- **Component Principal Analysis: Introduction**

Acknowledgments: Part of theses lectures are based on materials provided by Thierry Klein

Statistical Model

Definition

Let Ω be a set

Definition

$\mathcal{A} \subset \mathcal{P}(\Omega)$ is a σ-algebra on Ω if the following conditions are satisfied

1. $\Omega \in \mathcal{A}$
2. $\mathcal{A}$ is stable by the complementary operation i.e if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
3. $\mathcal{A}$ is stable by countable union i.e if $(A_n)_n$ is a countable family of elements of $\mathcal{A}$ i.e $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ then $\bigcup_n A_n \in \mathcal{A}$

1. $\{\emptyset, \Omega\}$ is the smallest $\sigma - algebra$
2. $\mathcal{P}(\Omega)$ is called the trivial $\sigma - algebra$, usually considered when Ω is discrete
3. When Ω is a topologic space equipped with a family of open sets, the smallest $\sigma-$ algebra which contains all these open is called the **Borel $\sigma-$algebra**. We denote it by $\mathcal{B}(\Omega)$. Why does it always exists?

Let Ω be a set

Definition

$\mathcal{A} \subset \mathcal{P}(\Omega)$ is a σ-algebra on Ω if the following conditions are satisfied

1. $\Omega \in \mathcal{A}$
2. $\mathcal{A}$ is stable by the complementary operation i.e if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
3. $\mathcal{A}$ is stable by countable union i.e if $(A_n)_n$ is a countable family of elements of $\mathcal{A}$ i.e $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ then $\bigcup_n A_n \in \mathcal{A}$

1. $\{\emptyset, \Omega\}$ is the smallest $\sigma - algebra$
2. $\mathcal{P}(\Omega)$ is called the trivial $\sigma - algebra$, usually considered when Ω is discrete
3. When Ω is a topologic space equipped with a family of open sets, the smallest $\sigma-$ algebra which contains all these open is called the **Borel $\sigma-$algebra**. We denote it by $\mathcal{B}(\Omega)$. Why does it always exists?

Definition

A set Ω equipped with a $\sigma-$algebra $\mathcal{A}$ is called a measurable space and we denote it by $(\Omega, \mathcal{A})$

Definition

A measure μ on $(\Omega, \mathcal{A})$ is an application from $\mathcal{A} \to [0, +\infty]$ such that

1. $\mu(\emptyset) = 0$
2. If $(A_n)_n$ is a countable family of elements of $\mathcal{A}$ mutally disjoints i.e $A_i \cap A_j = \emptyset$ if $i \neq j$ then

$$\mu(\bigcup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mu(A_n)$$

- Dirac measure δ_a. Counting measure $\sum_{n \in \mathbb{N}} \delta_n$.
- Lebesgue measure $\lambda([a, b]) = \lambda(]a, b]) = \lambda([a, b[) = \lambda(]a, b[) = b - a$

Definition

1. The triplet $(\Omega, \mathcal{A}, \mu)$ is called a measured set.
2. **When μ is of mass** 1 that is $\mu(\Omega) = 1$ we speak about **probability measure**. In this case we denote μ by $\mathbb{P}$.
3. **A probability space** is then a measurable space $(\Omega, \mathcal{A})$ equipped with a probability measure $\mathbb{P}$: $(\Omega, \mathcal{A}, \mathbb{P})$
4. One important situation in statistics is when the probability measure $\mathbb{P}$ depends on a **unknown parameter** θ^*. We usually denote $\mathbb{P}_{\theta^*}$ this probability.
5. We shall assume that the probability $\mathbb{P}_{\theta^*}$ belongs to a class of probability measure that we shall denote $\mathcal{P}$.
6. One of the aim of statistics is to find how can we obtain information on this parameter?

Definition

1. The triplet $(\Omega, \mathcal{A}, \mu)$ is called a measured set.
2. **When** μ **is of mass** 1 that is $\mu(\Omega) = 1$ we speak about **probability measure**. In this case we denote μ by $\mathbb{P}$.
3. **A probability space** is then a measurable space $(\Omega, \mathcal{A})$ equipped with a probability measure $\mathbb{P}$: $(\Omega, \mathcal{A}, \mathbb{P})$
4. One important situation in statistics is when the probability measure $\mathbb{P}$ depends on a **unknown parameter** θ^*. We usually denote $\mathbb{P}_{\theta^*}$ this probability.
5. We shall assume that the probability $\mathbb{P}_{\theta^*}$ belongs to a class of probability measure that we shall denote $\mathcal{P}$.
6. One of the aim of statistics is to find how can we obtain information on this parameter?

Definition

Definition

Let E and F be two sets equipped with $\sigma-$algebras $\mathcal{A}$ for E and $\mathcal{B}$ for F. An application $f : (E, \mathcal{A}) \to (E, \mathcal{B})$ is called measurable if

$$\forall B \in \mathcal{B}, f^{-1}(B) \in \mathcal{A}$$

- Recall that a random variable X is a measurable function from Ω to $\mathbb{R}$ or a discrete or countable space
- Let us throw two dices and compute the sum $S : \{1, \ldots, 6\}^2 \to \{2, \ldots, 12\} : S(i, j) = i + j$ is a r.v
- When is X is valued on $\mathbb{R}^k$, $k > 1$, we usually speak of random vectors

Statistical Model

Definition

A statistical model is a triplet $(\Omega, \mathcal{A}, \mathcal{P})$ where

1. Ω is called the space of realizations
2. $\mathcal{A}$ is a σ–algebra
3. $\mathcal{P}$ is a family of probability measure defined on $\mathcal{A}$

- Family of Gaussian laws:

$$\mathcal{P} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

Recall that the density of $\mathcal{N}(m, \sigma^2)$ is given by

$$f_{\mathcal{N}(m,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

- Family of Bernoulli laws:

$$\mathcal{P} = \{\mathcal{B}(\theta), \theta \in [0, 1]\}$$

Statistical Model

Definition

A statistical model is a triplet $(\Omega, \mathcal{A}, \mathcal{P})$ where

1. Ω is called the space of realizations
2. $\mathcal{A}$ is a $\sigma-$algebra
3. $\mathcal{P}$ is a family of probability measure defined on $\mathcal{A}$

- Family of Gaussian laws:

$$\mathcal{P} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

  Recall that the density of $\mathcal{N}(m, \sigma^2)$ is given by

$$f_{\mathcal{N}(m,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

- Family of Bernoulli laws:

$$\mathcal{P} = \{\mathcal{B}(\theta), \theta \in [0, 1]\}$$

Definition

- The examples
  - Family of Gaussian laws:

$$\mathcal{P} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

  - Family of Bernoulli laws:

$$\mathcal{P} = \{\mathcal{B}(\theta), \theta \in [0, 1]\}$$

  are usually associated with a random variable X whose law is either Gaussian or Bernoulli.

- Assume you want to extract information on m, σ or θ (these are unknown parameters). You can easily guess that one realization (one observation) of the value of X is not enough.

- Usually we are faced to n independent realizations of the same random variable. This way we consider $X_1, \ldots, X_n$ n r.v independent and identically distributed such as $X_i \sim X$ for all $i \in \{1, \ldots, n\}$

Definition

- The examples
  - Family of Gaussian laws:

$$\mathcal{P} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

  - Family of Bernoulli laws:

$$\mathcal{P} = \{\mathcal{B}(\theta), \theta \in [0, 1]\}$$

  are usually associated with a random variable X whose law is either Gaussian or Bernoulli.

- Assume you want to extract information on m, σ or θ (these are unknown parameters). You can easily guess that one realization (one observation) of the value of X is not enough.

- Usually we are faced to n independent realizations of the same random variable. This way we consider $X_1, \dots, X_n$ n r.v independent and identically distributed such as $X_i \sim X$ for all $i \in \{1, \dots, n\}$

Definition

- The examples
  - Family of Gaussian laws:

$$\mathcal{P} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

  - Family of Bernoulli laws:

$$\mathcal{P} = \{\mathcal{B}(\theta), \theta \in [0, 1]\}$$

  are usually associated with a random variable X whose law is either Gaussian or Bernoulli.

- Assume you want to extract information on m, σ or θ (these are unknown parameters). You can easily guess that one realization (one observation) of the value of X is not enough.

- Usually we are faced to n independent realizations of the same random variable. This way we consider $X_1, \ldots, X_n$ n r.v independent and identically distributed such as $X_i \sim X$ for all $i \in \{1, \ldots, n\}$

Definition

In the situation where you have n observations i.i.d $X_1, \ldots, X_n$, the statical models can be described by

- Gaussian: $\Omega = \mathbb{R}^n = \mathbb{R} \times \ldots \times \mathbb{R}$ (n times), $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$,

$$\mathcal{P} = \{\mathcal{N}^{\otimes n}(m, \sigma), m \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$$

- Bernoulli: $\Omega = \{0, 1\}^n$, $\mathcal{A} = \mathcal{P}(\Omega)$

$$\mathcal{P} = \{\mathcal{B}^{\otimes n}(\theta), \theta \in [0, 1]\}$$

the notation $\otimes n$ means that we consider the product of measure on the cartesian product $\mathbb{R}^n$ or $\{0, 1\}^n$. This corresponds to the fact that we consider independent situation.

- Exercise: describe the statistical model where you throw 100 times 10 dices and you just look at the sum of each result.

1. Other situations. Assume you observe n realizations of random variables X_i valued in $\mathbb{R}$ such that

$$\mathbb{E}[X_i] = i\theta$$

where θ is an unknown parameter and the law of X_i are unknown (you do not know the forme of the density for example). Your focus is on θ! only and not on the distribution of X_i

   - $\Omega = \mathbb{R}^n$
   - $\mathcal{P} = \left\{\mathbb{P}_{X_1} \otimes \ldots \otimes \mathbb{P}_{X_n}, \int_{\mathbb{R}} x dP_{X_i}(x) = i\theta, \theta \in \mathbb{R}\right\}$

2. Assume simply that you observe n independent and identical realizations of X. What can you say?

Definition

1. Other situations. Assume you observe n realizations of random variables X_i valued in $\mathbb{R}$ such that

$$\mathbb{E}[X_i] = i\theta$$

   where θ is an unknown parameter and the law of X_i are unknown (you do not know the forme of the density for example). Your focus is on θ! only and not on the distribution of X_i

   - $\Omega = \mathbb{R}^n$
   - $\mathcal{P} = \left\{\mathbb{P}_{X_1} \otimes \ldots \otimes \mathbb{P}_{X_n}, \int_{\mathbb{R}} x dP_{X_i}(x) = i\theta, \theta \in \mathbb{R}\right\}$

2. Assume simply that you observe n independent and identical realizations of X. What can you say?

Definition

1. **Parametric Model**: the family law is parametrized by a subset of $\mathbf{R}^d$.
2. **Semi- parametric Model**: the family laws is not parametrized by a subset of $\mathbf{R}^d$ but the quantity of interest is.
3. **Non parametric models**: all the other cases.

Definition

Definition

1. **Parametric Model**: the family law is parametrized by a subset of $\mathbf{R}^d$.
2. **Semi- parametric Model**: the family laws is not parametrized by a subset of $\mathbf{R}^d$ but the quantity of interest is.
3. **Non parametric models**: all the other cases.

Definition

Definition

1. **Parametric Model**: the family law is parametrized by a subset of $\mathbf{R}^d$.
2. **Semi- parametric Model**: the family laws is not parametrized by a subset of $\mathbf{R}^d$ but the quantity of interest is.
3. **Non parametric models**: all the other cases.

Definition

1. Now we have clearly defined what is a statistical model and what kind of different model we can address let us come back to the main statistical questions.
2. Estimation
3. Hypothesis testing

Definition

1. Now we have clearly defined what is a statistical model and what kind of different model we can address let us come back to the main statistical questions.
2. Estimation
3. Hypothesis testing

Definition

1. Estimation: Assume you want to estimate an unknown parameter θ or a function $g(\theta)$. This estimation has to be based only on the observations; this is done by the **notion of estimator**. We shall concentrate only the i.i.d situation

Definition

Let $X_1, \ldots, X_n$ be a sample that is the r.v are independent and identically distributed. An estimator is a measurable function of the observations.

2. An estimator can not be defined with unknown parameters
3. Usual estimator take the form $T = f(X_1, \ldots, X_n)$. An estimator is a r.v. When you have an observations $(x_1, \ldots, x_n)$, the quantity $t = f(x_1, \ldots, x_n)$ is a realization of T and is called an estimation
4. Examples:

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T = max(X_1, \ldots, X_n)$$

can be considered as estimators

Definition

1. Estimation: Assume you want to estimate an unknown parameter θ or a function $g(\theta)$. This estimation has to be based only on the observations; this is done by the **notion of estimator**. We shall concentrate only the i.i.d situation

Definition

Let $X_1, \ldots, X_n$ be a sample that is the r.v are independent and identically distributed. An estimator is a measurable function of the observations.

2. An estimator can not be defined with unknown parameters
3. Usual estimator take the form $T = f(X_1, \ldots, X_n)$. An estimator is a r.v. When you have an observations $(x_1, \ldots, x_n)$, the quantity $t = f(x_1, \ldots, x_n)$ is a realization of T and is called an estimation
4. Examples:

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T = max(X_1, \ldots, X_n)$$

can be considered as estimators

Definition

1. Estimation: Assume you want to estimate an unknown parameter θ or a function $g(\theta)$. This estimation has to be based only on the observations; this is done by the **notion of estimator**. We shall concentrate only the i.i.d situation

Definition

Let $X_1, \ldots, X_n$ be a sample that is the r.v are independent and identically distributed. An estimator is a measurable function of the observations.

2. An estimator can not be defined with unknown parameters
3. Usual estimator take the form $T = f(X_1, \ldots, X_n)$. An estimator is a r.v. When you have an observations $(x_1, \ldots, x_n)$, the quantity $t = f(x_1, \ldots, x_n)$ is a realization of T and is called an estimation
4. Examples:

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T = max(X_1, \ldots, X_n)$$

can be considered as estimators

Definition

1. Estimation: Assume you want to estimate an unknown parameter θ or a function $g(\theta)$. This estimation has to be based only on the observations; this is done by the **notion of estimator**. We shall concentrate only the i.i.d situation

Definition

Let $X_1, \ldots, X_n$ be a sample that is the r.v are independent and identically distributed. An estimator is a measurable function of the observations.

2. An estimator can not be defined with unknown parameters
3. Usual estimator take the form $T = f(X_1, \ldots, X_n)$. An estimator is a r.v. When you have an observations $(x_1, \ldots, x_n)$, the quantity $t = f(x_1, \ldots, x_n)$ is a realization of T and is called an estimation
4. Examples:

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T = max(X_1, \ldots, X_n)$$

can be considered as estimators

Definition

1. Hypothesis testing: Assume that your unknown parameter $\theta^* \in \Theta = \Theta_1 \cup \Theta_2$ where the union is disjoint.
2. Within the observations you want to take a decision: the parameter θ^* belongs either to Θ_1 or to Θ_2
3. Again this decision has to be made in a measurable way with respect to the observations. A test is a measurable function of $(X_1, \ldots, X_n)$
4. We won't study the theory of hypothesis testing in this course and we shall concentrate on estimation

Definition

1. Hypothesis testing: Assume that your unknown parameter $\theta^* \in \Theta = \Theta_1 \cup \Theta_2$ where the union is disjoint.
2. Within the observations you want to take a decision: the parameter θ^* belongs either to Θ_1 or to Θ_2
3. Again this decision has to be made in a measurable way with respect to the observations. A test is a measurable function of $(X_1, \ldots, X_n)$
4. We won't study the theory of hypothesis testing in this course and we shall concentrate on estimation

Definition

1. Before going further: **Important point:** making statistic is assuming that you are going to make mistakes, errors.
2. Indeed you won't be able, in general, to be sure having founded the unknown parameter only with a finite number of observations
3. Statisticians are Mathematicians who are able to control the error they will make by establishing qualitative analysis of their estimators or tests.
4. Before going into the details, we shall recall some basic probability result.

Definition

1. Before going further: **Important point:** making statistic is assuming that you are going to make mistakes, errors.
2. Indeed you won't be able, in general, to be sure having founded the unknown parameter only with a finite number of observations
3. Statisticians are Mathematicians who are able to control the error they will make by establishing qualitative analysis of their estimators or tests.
4. Before going into the details, we shall recall some basic probability result.

Probability background

First concentration inequality

- This part will be a glossary of notions of probability we shall need in the sequel
- Let us start with two useful concentration inequalities. Let us consider a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
- If $X \in L^1$, the mean, average, expectation is denoted by $\mathbb{E}[X]$
- If $X \in L^2$, the variance is denoted by $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- If X is L^1: **Markov inequality**

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}(|X|)}{t}$$

- If X is L^2: **Bienaymé-Tchebychev inequality**

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant t) \leqslant \frac{Var(X)}{t^2}$$

First concentration inequality

- This part will be a glossary of notions of probability we shall need in the sequel
- Let us start with two useful concentration inequalities. Let us consider a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
- If $X \in L^1$, the mean, average, expectation is denoted by $\mathbb{E}[X]$
- If $X \in L^2$, the variance is denoted by $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- If X is L^1: **Markov inequality**

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}(|X|)}{t}$$

- If X is L^2: **Bienaymé-Tchebychev inequality**

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant t) \leqslant \frac{Var(X)}{t^2}$$

Characteristic function

Definition

The characteristic function of a r.v X is defined by

$$\phi_X(t) = \mathbb{E}[e^{itX}], \forall t \in \mathbb{R}$$

The characeristic function of a random vector is

$$\phi_X(u) = \mathbb{E}[e^{i<u,X>}], \forall u \in \mathbb{R}^d,$$

where $<,>$ denote the scalar product on $\mathbb{R}^d$.

characteristic function

- $X \sim \mathcal{B}(p)$ then $\phi_X(t) = 1 - p + pe^{it}$
- $X \sim \mathcal{B}(n, p)$ then $\phi_X(t) = (1 - p + pe^{it})^n$
- $X \sim \mathcal{P}(\lambda)$ then $\phi_X(t) = exp(\lambda(e^{it} - 1))$

- $X \sim \mathcal{U}([a, b])$ then $\phi_X(t) = \frac{e^{ibt} - e^{iat}}{(b-a)it}$
- $X \sim \mathcal{E}(\lambda)$ then $\phi_X(t) = \frac{\lambda}{\lambda - it}$
- $X \sim C(a)$ then $\phi_X(t) = exp(-a|t|)$
- $X \sim \mathcal{N}(m, \sigma^2)$ then $\phi_X(t) = exp(imt - \frac{\sigma^2 t^2}{2})$

characteristic function

- $X \sim \mathcal{B}(p)$ then $\phi_X(t) = 1 - p + pe^{it}$
- $X \sim \mathcal{B}(n,p)$ then $\phi_X(t) = (1 - p + pe^{it})^n$
- $X \sim \mathcal{P}(\lambda)$ then $\phi_X(t) = exp(\lambda(e^{it} - 1))$

- $X \sim \mathcal{U}([a,b])$ then $\phi_X(t) = \frac{e^{ibt} - e^{iat}}{(b-a)it}$
- $X \sim \mathcal{E}(\lambda)$ then $\phi_X(t) = \frac{\lambda}{\lambda - it}$
- $X \sim C(a)$ then $\phi_X(t) = exp(-a|t|)$
- $X \sim \mathcal{N}(m,\sigma^2)$ then $\phi_X(t) = exp(imt - \frac{\sigma^2 t^2}{2})$

Proposition

Let X be a r.v which admits a moment of order p then its characteristic function is p times differentiable and we have

$$\phi_X^{(p)}(0) = i^p \mathbb{E}[X^p]$$

Other transformation

- The moment generator function of a r.v X with values in $S(X) \subset \mathbb{N}$ and $p_k = \mathbb{P}(X = k)$ is

$$G_X(t) = \mathbb{E}[t^X] = \sum_k p_k t^k$$

  This function is C^∞ on $[0, 1[$ and p times differentiable on 1 if $\mathbb{E}[X^p] < +\infty$

$$G_X^{(k)}(0) = k!p_k, k \in \mathbb{N}$$

  If the mean exists, we have $G_X'(1) = \mathbb{E}(X)$
- Laplace transform. For a r.v X, we call its Laplace transform

$$\phi_X(t) = \mathbb{E}[e^{tX}]$$

Other transformation

1. As we shall see in the sequel, we shall be interested in limits of estimator when the number of observations n goes to infinity.
2. This asks for convergence of random variables.

Definition

Let (X_n) be a sequence of r.v and X be a r.v. We say that (X_n) converge towards X

- **Almost surely a.s** if $\mathbb{P}(\lim X_n = X) = 1$ we note $X_n \xrightarrow{a.s} X$
- **In L^p norm** if $\lim_{n\to+\infty} \mathbb{E}[|X_n - X|^p] = 0$ we note $X_n \xrightarrow{L^p} X$
- **In probability** if $\forall\epsilon > 0, \lim_{n\to+\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0$ we note $X_n \xrightarrow{\mathbb{P}} X$
- **In law** if for all continuous and bounded functions f we have $\lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ we note $X_n \xrightarrow{\mathcal{L}} X$

When the law of X depends on a unknown parameter θ we make appear this dependency.

Definition

Definition

Let (X_n) be a sequence of r.v and X be a r.v. We say that (X_n) converge towards X

- **Almost surely a.s** if $\mathbb{P}(\lim X_n = X) = 1$ we note $X_n \xrightarrow{a.s} X$
- **In L^p norm** if $\lim\limits_{n \to +\infty} \mathbb{E}[|X_n - X|^p] = 0$ we note $X_n \xrightarrow{L^p} X$
- **In probability** if $\forall \epsilon > 0, \lim\limits_{n \to +\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0$ we note $X_n \xrightarrow{\mathbb{P}} X$
- **In law** if for all continuous and bounded functions f we have $\lim\limits_{n \to +\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ we note $X_n \xrightarrow{\mathcal{L}} X$

When the law of X depends on a unknown parameter θ we make appear this dependency.

Definition

Let (X_n) be a sequence of r.v and X be a r.v. We say that (X_n) converge towards X

- **Almost surely a.s** if $\mathbb{P}(\lim X_n = X) = 1$ we note $X_n \xrightarrow{a.s} X$
- **In L^p norm** if $\lim\limits_{n \to +\infty} \mathbb{E}[|X_n - X|^p] = 0$ we note $X_n \xrightarrow{L^p} X$
- **In probability** if $\forall \epsilon > 0, \lim\limits_{n \to +\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0$ we note $X_n \xrightarrow{\mathbb{P}} X$
- **In law** if for all continuous and bounded functions f we have $\lim\limits_{n \to +\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ we note $X_n \xrightarrow{\mathcal{L}} X$

When the law of X depends on a unknown parameter θ we make appear this dependency.

Definition

Let (X_n) be a sequence of r.v and X be a r.v. We say that (X_n) converge towards X

- **Almost surely a.s** if $\mathbb{P}(\lim X_n = X) = 1$ we note $X_n \xrightarrow{a.s} X$
- **In L^p norm** if $\lim_{n\to+\infty} \mathbb{E}[|X_n - X|^p] = 0$ we note $X_n \xrightarrow{L^p} X$
- **In probability** if $\forall\epsilon > 0, \lim_{n\to+\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0$ we note $X_n \xrightarrow{\mathbb{P}} X$
- **In law** if for all continuous and bounded functions f we have $\lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ we note $X_n \xrightarrow{\mathcal{L}} X$

When the law of X depends on a unknown parameter θ we make appear this dependency.

Convergence en loi

For a r.v we denote its partition function F_X and recall that ϕ_X denotes its characteristic function

Theorem

(X_n) converge in law towards X if and only if

$$F_{X_n}(t) \to F_X(t)$$

in all points where F_X is continuous i.e in all points t such that $\mathbb{P}(X = t) = 0$

Theorem

(X_n) converges in law towards X if and only if

$$\phi_{X_n}(t) \to \phi_X(t)$$

for all $t \in \mathbb{R}$.

Usual Convergence mode

In order to finish let us recall the usual convergence mode

Theorem

- ***Beppo Levy Theorem:*** *let* (X_n) *be a non decreasing sequence of non negative numbers then if* $\lim_n X_n = X$ *we have*

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$$

- ***Fatou Lemma:*** *let* (X_n) *be a sequence of non negative numbers then*

$$\mathbb{E}[\liminf_n X_n] \leqslant \liminf_n \mathbb{E}[X_n]$$

- ***Lebesgue dominated convergence Theorem:*** *let* (X_n) *be a sequence such that* X_n *converges a.s to* X*. Let* Y *such that* $\mathbb{E}[|Y|] < \infty$ *and* $|X_n| < Y|$ *then*

Usual Convergence mode

In order to finish let us recall the usual convergence mode

Theorem

- ***Beppo Levy Theorem:*** *let* (X_n) *be a non decreasing sequence of non negative numbers then if* $\lim_n X_n = X$ *we have*

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$$

- ***Fatou Lemma:*** *let* (X_n) *be a sequence of non negative numbers then*

$$\mathbb{E}[\liminf_n X_n] \leqslant \liminf_n \mathbb{E}[X_n]$$

- ***Lebesgue dominated convergence Theorem:*** *let* (X_n) *be a sequence such that* X_n *converges a.s to* X*. Let* Y *such that* $\mathbb{E}[|Y|] < \infty$ *and* $|X_n| < Y|$ *then*

Usual Convergence mode

In order to finish let us recall the usual convergence mode

Theorem

- ***Beppo Levy Theorem:*** *let (X_n) be a non decreasing sequence of non negative numbers then if $\lim_n X_n = X$ we have*

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$$

- ***Fatou Lemma:*** *let (X_n) be a sequence of non negative numbers then*

$$\mathbb{E}[\liminf_n X_n] \leqslant \liminf_n \mathbb{E}[X_n]$$

- ***Lebesgue dominated convergence Theorem:*** *let (X_n) be a sequence such that X_n converges a.s to X. Let Y such that $\mathbb{E}[|Y|] < \infty$ and $|X_n| < Y|$ then*

Links between convergence modes

Recall the usual links

- Almost sure convergence $\Longrightarrow$ Convergence in probability
- L^p Convergence $p \geqslant 1 \Longrightarrow L^1$ Convergence $\Longrightarrow$ Convergence in probability
- All convergence modes $\Longrightarrow$ Convergence in law
- Almost surely convergence + domination $\Longrightarrow L^1$ convergence
- L^1 convergence $\Longrightarrow$ Almost sure convergence for a sub-sequence

Links between convergence modes

Recall the usual links

- Almost sure convergence $\Longrightarrow$ Convergence in probability
- L^p Convergence $p \geqslant 1 \Longrightarrow L^1$ Convergence $\Longrightarrow$ Convergence in probability
- All convergence modes $\Longrightarrow$ Convergence in law
- Almost surely convergence + domination $\Longrightarrow L^1$ convergence
- L^1 convergence $\Longrightarrow$ Almost sure convergence for a sub-sequence

Links between convergence modes

Recall the usual links

- Almost sure convergence $\Longrightarrow$ Convergence in probability
- L^p Convergence $p \geqslant 1 \Longrightarrow L^1$ Convergence $\Longrightarrow$ Convergence in probability
- All convergence modes $\Longrightarrow$ Convergence in law
- Almost surely convergence + domination $\Longrightarrow L^1$ convergence
- L^1 convergence $\Longrightarrow$ Almost sure convergence for a sub-sequence

Links between convergence modes

Recall the usual links

- Almost sure convergence $\Longrightarrow$ Convergence in probability
- L^p Convergence $p \geqslant 1 \Longrightarrow L^1$ Convergence $\Longrightarrow$ Convergence in probability
- All convergence modes $\Longrightarrow$ Convergence in law
- Almost surely convergence + domination $\Longrightarrow L^1$ convergence
- L^1 convergence $\Longrightarrow$ Almost sure convergence for a sub-sequence

Links between convergence modes

Recall the usual links

- Almost sure convergence $\Longrightarrow$ Convergence in probability
- L^p Convergence $p \geqslant 1 \Longrightarrow L^1$ Convergence $\Longrightarrow$ Convergence in probability
- All convergence modes $\Longrightarrow$ Convergence in law
- Almost surely convergence + domination $\Longrightarrow L^1$ convergence
- L^1 convergence $\Longrightarrow$ Almost sure convergence for a sub-sequence

- When (X_n) converges in law to X and (Y_n) converges in law to Y this does not implies in general that (X_n, Y_n) converges in law to (X, Y). But we have this useful result:

Proposition

(Slutsky)

- *If (X_n) converges in law to X and (Y_n) converges in law to c then (X_n, Y_n) converges in law to (X, c)*

Slutsky

- When (X_n) converges in law to X and (Y_n) converges in law to Y this does not implies in general that (X_n, Y_n) converges in law to (X, Y). But we have this useful result:

Proposition

(Slutsky)

- *If (X_n) converges in law to X and (Y_n) converges in law to c then (X_n, Y_n) converges in law to (X, c)*

Slutsky

- In the sequel we shall also need the notion of $\circ_{\mathbb{P}}$
- We say that $X_n = \circ_{\mathbb{P}}(Y_n)$ if

$$\frac{X_n}{Y_n} \xrightarrow{\mathbb{P}} 0$$

- Note that if R is a continuous function such that $R(h) = \circ(||h||^p)$ and (X_n) is a sequence which converges in probability to 0 then

$$R(X_n) = \circ_{\mathbb{P}}(||X_n||^p)$$

Here we shall use the fact that if $X_n \xrightarrow{\mathbb{P}} X$ then for all continuous function $f(X_n) \xrightarrow{\mathbb{P}} f(X)$

Slutsky

- In the sequel we shall also need the notion of $\circ_{\mathbb{P}}$
- We say that $X_n = \circ_{\mathbb{P}}(Y_n)$ if

$$\frac{X_n}{Y_n} \xrightarrow{\mathbb{P}} 0$$

- Note that if R is a continuous function such that $R(h) = \circ(\|h\|^p)$ and (X_n) is a sequence which converges in probability to 0 then

$$R(X_n) = \circ_{\mathbb{P}}(\|X_n\|^p)$$

Here we shall use the fact that if $X_n \xrightarrow{\mathbb{P}} X$ then for all continuous function $f(X_n) \xrightarrow{\mathbb{P}} f(X)$

Slutsky

- In the sequel we shall also need the notion of $\circ_{\mathbb{P}}$
- We say that $X_n = \circ_{\mathbb{P}}(Y_n)$ if

$$\frac{X_n}{Y_n} \xrightarrow{\mathbb{P}} 0$$

- Note that if R is a continuous function such that $R(h) = \circ(\|h\|^p)$ and (X_n) is a sequence which converges in probability to 0 then

$$R(X_n) = \circ_{\mathbb{P}}(\|X_n\|^p)$$

Here we shall use the fact that if $X_n \xrightarrow{\mathbb{P}} X$ then for all continuous function $f(X_n) \xrightarrow{\mathbb{P}} f(X)$

Law of Large Numbers (LLN) and Central Limit Theorem (CLT)

Objectif

- The objective of this section is to understand the convergence of

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

$$\sqrt{n}(\bar{X}_n - m) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m)$$

  when (X_n) is a sequence of i.i.d random variables where $m = \mathbb{E}[X_1]$.

- As we shall see the first quantity is a good estimator of m and the second quantity allows to control the error we make when making estimation

Weak Law of Large Numbers L^2 and L^1

Theorem

Let (X_n) be a sequence of i.i.d r.v which are L^2 then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$$

- Let (X_n) be a sequence of i.i.d r.v $\mathcal{B}(p)$ then $M_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} p$
- First step towards estimation of an unknown proportion

Theorem

Let (X_n) be a sequence of i.i.d r.v which are L^1 then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$$

Weak Law of Large Numbers L^2 and L^1

Theorem

Let (X_n) be a sequence of i.i.d r.v which are L^2 then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$$

- Let (X_n) be a sequence of i.i.d r.v $\mathcal{B}(p)$ then $M_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} p$
- First step towards estimation of an unknown proportion

Theorem

Let (X_n) be a sequence of i.i.d r.v which are L^1 then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$$

Law of Large Numbers

Theorem

Law of Large Numbers: *Let (X_n) be a sequence of i.i.d r.v and L^1 then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s}{\to} \mathbb{E}[X_1]$$

- Application: Monte Carlo Method. Let f be a measurable function such that $f(X_1)$ Let L^1

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \overset{a.s}{\to} \mathbb{E}[f(X_1)]$$

  Rq: note that the advantage of this method is that we do not require any regularity property of f.

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$ are estimators of the mean and of the variance

Law of Large Numbers

Theorem

Law of Large Numbers: *Let (X_n) be a sequence of i.i.d r.v and L^1 then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s}{\to} \mathbb{E}[X_1]$$

- Application: Monte Carlo Method. Let f be a measurable function such that $f(X_1)$ Let L^1

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \overset{a.s}{\to} \mathbb{E}[f(X_1)]$$

  Rq: note that the advantage of this method is that we do not require any regularity property of f.
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$ are estimators of the mean and of the variance

Theorem

Central Limit Theorem: *Let* (X_n) *be a sequence of i.i.d r.v which are* L^2*. Let* m *be the common mean and* σ^2 *the common variance. We put*

$$S_n = \sum_{i=1}^{n} X_i = n\bar{X}_n$$

then

$$\frac{1}{\sqrt{n\sigma^2}} \sum_{i=1}^{n} (X_i - m) = \frac{S_n - nm}{\sqrt{n\sigma^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Central Limit Theorem

- This is a strong refinement of the LLN: somehow it gives the rate of convergence of the empirical mean towards the mean.
- As we shall see later, this allows to construct confidence interval
- Sometimes we need to consider $f(\bar{X}_n)$ for f sufficiently smooth. It is easy to see that

$$f(\bar{X}_n) \xrightarrow{a.s} f(\mathbb{E}[X_1])$$

  using the continuity of f
- Concerning extension of CLT one is interested in convergence in law of

$$\sqrt{n}(f(\bar{X}_n) - f(\mathbb{E}[X_1]))$$

- This asks for the so called Delta method which will be exposed at the end of the next part concerning Gaussian laws.

Central Limit Theorem

- This is a strong refinement of the LLN: somehow it gives the rate of convergence of the empirical mean towards the mean.
- As we shall see later, this allows to construct confidence interval
- Sometimes we need to consider $f(\bar{X}_n)$ for f sufficiently smooth. It is easy to see that

$$f(\bar{X}_n) \overset{a.s}{\to} f(\mathbb{E}[X_1])$$

using the continuity of f
- Concerning extension of CLT one is interested in convergence in law of

$$\sqrt{n}(f(\bar{X}_n) - f(\mathbb{E}[X_1]))$$

- This asks for the so called Delta method which will be exposed at the end of the next part concerning Gaussian laws.

Central Limit Theorem

- This is a strong refinement of the LLN: somehow it gives the rate of convergence of the empirical mean towards the mean.
- As we shall see later, this allows to construct confidence interval
- Sometimes we need to consider $f(\bar{X}_n)$ for f sufficiently smooth. It is easy to see that

$$f(\bar{X}_n) \overset{a.s}{\to} f(\mathbb{E}[X_1])$$

  using the continuity of f
- Concerning extension of CLT one is interested in convergence in law of

$$\sqrt{n}(f(\bar{X}_n) - f(\mathbb{E}[X_1]))$$

- This asks for the so called Delta method which will be exposed at the end of the next part concerning Gaussian laws.

Gaussian Vectors

Definition

Definition

A random vector $X = (X_1, \dots, X_d)^t$ is called Gaussian vector if all linear combination of its coordinates are Gaussian, that is for all $a \in \mathbb{R}^d$ the r.v

$$< a, X >= \sum_{i=1}^{d} a_i X_i$$

is a Gaussian r.v.

- If X is a Gaussian vector then for all matrices A the vector AX is still a Gaussian vector

Definition

A random vector $X = (X_1, \ldots, X_d)^t$ is called Gaussian vector if all linear combination of its coordinates are Gaussian, that is for all $a \in \mathbb{R}^d$ the r.v

$$< a, X >= \sum_{i=1}^{d} a_i X_i$$

is a Gaussian r.v.

- If X is a Gaussian vector then for all matrices A the vector AX is still a Gaussian vector

Definition

Let $X = (X_1, \ldots, X_d)^t$ be a Gaussian vector we note K its covariance matrix defined by

$$K_{i,j} = Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j],$$

for all $i, j = 1, \ldots, d$. We shall also note

$$m = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^t$$

the vector of mean. We shall note $X \sim \mathcal{N}_d(m, K)$

- The matrix K is semi-definite positive
- $\mathbb{E}[< a, X >] =< a, \mathbb{E}[X] >$
- $Var(< a, X >) = Var\left(\sum_{i=1}^{d} a_i X_i\right) = \sum_{i,j=1}^{d} a_i a_j Cov(X_i, X_j) = a^t K a =< a, Ka >$

matrix de covaroiance

Definition

Let $X = (X_1, \ldots, X_d)^t$ be a Gaussian vector we note K its covariance matrix defined by

$$K_{i,j} = Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j],$$

for all $i, j = 1, \ldots, d$. We shall also note

$$m = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^t$$

the vector of mean. We shall note $X \sim \mathcal{N}_d(m, K)$

- The matrix K is semi-definite positive
- $\mathbb{E}[< a, X >] =< a, \mathbb{E}[X] >$
- $Var(< a, X >) = Var\left(\sum_{i=1}^{d} a_i X_i\right) = \sum_{i,j=1}^{d} a_i a_j Cov(X_i, X_j) = a^t K a =< a, Ka >$

- One can check that

$$\phi_{<a,X>}(t) = \exp\left(i < a, m > t - \frac{1}{2} a^t K a\, t^2\right)$$

- $\phi_X(x) = \mathbb{E}[e^{i<x,X>}] = \phi_{<x,X>}(1)$

Proposition

The characteristic function of a Gaussian vector is given by

$$\phi_X(x) = \exp\left(i < x, m > -\frac{1}{2} x^t K x\right)$$

- The coordinates of a Gaussian vector are independent if and only if its covariance matrix is diagonal

Transformation linéaire

Proposition

Let $X \sim \mathcal{N}_d(m, K)$ *then for all matrices* $A \in \mathbb{M}_{p,d}(\mathbb{R})$ *then*

$$AX \sim \mathcal{N}_p(AX, AKA^t)$$

- If $X \sim \mathcal{N}_d(0, I_d)$ then the law of X is invariant by all rotation.

Centrer and réduire un Gaussian vector

- We shall say that a Gaussian vector X is degenerate if its covariance matrix K is non invertible
- In the degenerate case, there exists a such that $Ka = 0$ which implies that

$$Var(< a, X >) = 0$$

and then $< a, X >= b$ a.s. Then X leaves in the affine space

$$\{< a, x >= b, x \in \mathbb{R}^d\}$$

- If K is invertible then $\sqrt{K}^{-1}(X - m) \sim \mathcal{N}(0, I_d)$
- If $N \sim \mathcal{N}(0, I_d)$ then $X = \sqrt{K}N + m \in \mathcal{N}(m, K)$

Density

- If $X \sim \mathcal{N}_d(0, I_d)$ then the coordinates $(X_i)_{i=1,\dots,d}$ are i.i.d and $X_1 \sim \mathcal{N}(0, 1)$. Then the density of X is given by the product of densities i.e

$$f_X(x_1, \dots, x_d) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{1}{2}\sum_{i=1}^{d} x_i^2\right)$$

- In the case where K is invertible we have

$$f_X(x_1, \dots, x_d) = \frac{1}{\sqrt{(2\pi)^d \det K}} \exp(-\frac{1}{2} < (x-m), K^{-1}(x-m) >)$$

- Rq: if X is Gaussian all its coordinates are Gaussian, the converse is not true in general.

CLT multidimensional

Theorem

Let $X^{(n)}$ be a sequence of random vectors of $\mathbb{R}^d$ which are i.i.d and L^2 of mean vector m and of covariance matrix K. We put $S^{(n)} = \sum_{i=1}^{n} X^{(i)}$ then we have

$$n^{-1/2}\sqrt{K}^{-1}(S^{(n)} - nm) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, I_d)$$

or

$$n^{-1/2}(S^{(n)} - nm) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, K)$$

- Let $X \sim \mathcal{N}(0,1)$ and consider $Z = X^2$. Let f be a continuous and bounded function

$$
\begin{aligned}
\mathbb{E}[f(Z)] &= \mathbb{E}[f(X^2)] \\
&= \int_{\mathbb{R}} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= 2\int_0^{+\infty} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_0^{+\infty} f(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} dz
\end{aligned}
$$

- Then $Z \sim \chi^2(1)$ where $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} \mathbf{1}_{z \geqslant 0}$

- Let $X \sim \mathcal{N}(0,1)$ and consider $Z = X^2$. Let f be a continuous and bounded function

$$
\begin{aligned}
\mathbb{E}[f(Z)] &= \mathbb{E}[f(X^2)] \\
&= \int_{\mathbb{R}} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= 2 \int_0^{+\infty} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_0^{+\infty} f(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} dz
\end{aligned}
$$

- Then $Z \sim \chi^2(1)$ where $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} \mathbf{1}_{z \geqslant 0}$

Transformation of Gaussian law

- Let $X \sim \mathcal{N}(0,1)$ and consider $Z = X^2$. Let f be a continuous and bounded function

$$
\begin{aligned}
\mathbb{E}[f(Z)] &= \mathbb{E}[f(X^2)] \\
&= \int_{\mathbb{R}} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= 2 \int_0^{+\infty} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_0^{+\infty} f(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} dz
\end{aligned}
$$

- Then $Z \sim \chi^2(1)$ where $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} \mathbf{1}_{z \geqslant 0}$

Transformation of Gaussian law

- Let $X \sim \mathcal{N}(0,1)$ and consider $Z = X^2$. Let f be a continuous and bounded function

$$
\begin{aligned}
\mathbb{E}[f(Z)] &= \mathbb{E}[f(X^2)] \\
&= \int_{\mathbb{R}} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= 2 \int_0^{+\infty} f(x^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_0^{+\infty} f(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} dz
\end{aligned}
$$

- Then $Z \sim \chi^2(1)$ where $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} (\sqrt{z})^{-1} \mathbf{1}_{z \geqslant 0}$

Transformation of Gaussian law

- Let $X = (X_1, \ldots, X_d)$ a Gaussian random vector where (X_i) are i.i.d of law $\mathcal{N}(0, 1)$ then

$$Z = \sum_{i=1}^{d} X_i^2$$

  is a random variable whose law is $\chi^2(d)$ where d is called the degree of freedom

- The density of this r.v is

$$f_Z(z) = \frac{1}{2\Gamma(k/2)} z^{\frac{k}{2}-1} e^{-\frac{z}{2}} \mathbf{1}_{z \geqslant 0}$$

  where Γ is the Gamma function

Transformation of Gaussian law

- Let $X \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2(k)$ then the r.v

$$T = \frac{X}{\sqrt{Z/k}}$$

  is said to be distributed as the Student law of degree k
- The density is given by

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} (1 + \frac{t^2}{2})^{-\frac{k+1}{2}}$$

Transformation of Gaussian law

- Let $X \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2(k)$ then the r.v

$$T = \frac{X}{\sqrt{Z/k}}$$

is said to be distributed as the Student law of degree k

- The density is given by

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} (1 + \frac{t^2}{2})^{-\frac{k+1}{2}}$$

Cochran Theorem

Proposition

Let $X \sim \mathcal{N}_d(0, I_d)$ and let $\mathbb{R}^d = F_1 \oplus \ldots \oplus F_k$ a decomposition in orthogonal space with $\dim(F_i) = d_i$. We note $P_{F_i}, i = 1, \ldots, k$ the orthogonal projectors associated with space $F_i, i = 1 \ldots, k$. In this case the vectors $P_{F_1}(X), \ldots, P_{F_k}(X)$ are independent Gaussian vectors. We have also

$$\|P_{F_i}(X)\|^2 \sim \chi^2(d_i), i = 1, \ldots, k$$

- This is linear algebra
- We can express a more general result $X \sim \mathcal{N}(0, K)$ with non degenerate K by introducing a scalar product with respect to K i.e $< a, b >_K = < a, Kb >$.

chi^2 Test

Test of adequation χ^2:

- We observe a random variable X where the set of values $S(X) = \{a_1, \dots, a_r\}$ and $p_j = \mathbb{P}(X = a_j) = Q(\{a_j\}), j = 1, \dots, r$ unknown. We note $p = (p_1, \dots, p_r)$ the corresponding vector of probability.
- We consider a reference probability $Q_0 = \sum_i \pi_i \delta_i$ with same support but with a known vector $\pi = (\pi_1, \dots, \pi_r)$ where $\pi_i > 0$
- The Hypothesis testing is $H_0 : Q = Q_0$ against $H_1 : Q \neq Q_0$.
- Let (X_n) be a sequence of i.i.d.r.v of law Q. For $n \in \mathbb{N}$, we put

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = a_j}$$

- The random vector $N = (N_1, N_2, \dots, N_r)^t$ follows a multinomial law $\mathcal{M}(n, p_1, \dots, p_r)$ i.e

$$\mathbb{P}(N_1 = n_1, \dots, N_r = n_r) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}, \quad n_1 + \dots + n_r = n$$

chi^2 Test

Test of adequation χ^2:

- We observe a random variable X where the set of values $S(X) = \{a_1, \dots, a_r\}$ and $p_j = \mathbb{P}(X = a_j) = Q(\{a_j\}), j = 1, \dots, r$ unknown. We note $p = (p_1, \dots, p_r)$ the corresponding vector of probability.
- We consider a reference probability $Q_0 = \sum_i \pi_i \delta_i$ with same support but with a known vector $\pi = (\pi_1, \dots, \pi_r)$ where $\pi_i > 0$
- The Hypothesis testing is $H_0 : Q = Q_0$ against $H_1 : Q \neq Q_0$.
- Let (X_n) be a sequence of i.i.d.r.v of law Q. For $n \in \mathbb{N}$, we put

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = a_j}$$

- The random vector $N = (N_1, N_2, \dots, N_r)^t$ follows a multinomial law $\mathcal{M}(n, p_1, \dots, p_r)$ i.e

$$\mathbb{P}(N_1 = n_1, \dots, N_r = n_r) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}, \quad n_1 + \dots + n_r = n$$

chi^2 Test

Test of adequation χ^2:

- We observe a random variable X where the set of values $S(X) = \{a_1, \ldots, a_r\}$ and $p_j = \mathbb{P}(X = a_j) = Q(\{a_j\}), j = 1, \ldots, r$ unknown. We note $p = (p_1, \ldots, p_r)$ the corresponding vector of probability.
- We consider a reference probability $Q_0 = \sum_i \pi_i \delta_i$ with same support but with a known vector $\pi = (\pi_1, \ldots, \pi_r)$ where $\pi_i > 0$
- The Hypothesis testing is $H_0 : Q = Q_0$ against $H_1 : Q \neq Q_0$.
- Let (X_n) be a sequence of i.i.d.r.v of law Q. For $n \in \mathbb{N}$, we put

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = a_j}$$

- The random vector $N = (N_1, N_2, \ldots, N_r)^t$ follows a multinomial law $\mathcal{M}(n, p_1, \ldots, p_r)$ i.e

$$\mathbb{P}(N_1 = n_1, \ldots, N_r = n_r) = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r}, \quad n_1 + \ldots + n_r = n$$

chi^2 Test

Test of adequation χ^2:

- We observe a random variable X where the set of values $S(X) = \{a_1, \ldots, a_r\}$ and $p_j = \mathbb{P}(X = a_j) = Q(\{a_j\}), j = 1, \ldots, r$ unknown. We note $p = (p_1, \ldots, p_r)$ the corresponding vector of probability.
- We consider a reference probability $Q_0 = \sum_i \pi_i \delta_i$ with same support but with a known vector $\pi = (\pi_1, \ldots, \pi_r)$ where $\pi_i > 0$
- The Hypothesis testing is $H_0 : Q = Q_0$ against $H_1 : Q \neq Q_0$.
- Let (X_n) be a sequence of i.i.d.r.v of law Q. For $n \in \mathbb{N}$, we put

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = a_j}$$

- The random vector $N = (N_1, N_2, \ldots, N_r)^t$ follows a multinomial law $\mathcal{M}(n, p_1, \ldots, p_r)$ i.e

$$\mathbb{P}(N_1 = n_1, \ldots, N_r = n_r) = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r}, \quad n_1 + \ldots + n_r = n$$

Test of adequation χ^2:

- We observe a random variable X where the set of values $S(X) = \{a_1, \dots, a_r\}$ and $p_j = \mathbb{P}(X = a_j) = Q(\{a_j\}), j = 1, \dots, r$ unknown. We note $p = (p_1, \dots, p_r)$ the corresponding vector of probability.
- We consider a reference probability $Q_0 = \sum_i \pi_i \delta_i$ with same support but with a known vector $\pi = (\pi_1, \dots, \pi_r)$ where $\pi_i > 0$
- The Hypothesis testing is $H_0 : Q = Q_0$ against $H_1 : Q \neq Q_0$.
- Let (X_n) be a sequence of i.i.d.r.v of law Q. For $n \in \mathbb{N}$, we put

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{X_i = a_j}$$

- The random vector $N = (N_1, N_2, \dots, N_r)^t$ follows a multinomial law $\mathcal{M}(n, p_1, \dots, p_r)$ i.e

$$\mathbb{P}(N_1 = n_1, \dots, N_r = n_r) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}, \quad n_1 + \dots + n_r = n$$

- We put

$$T_n = \sum_{j=1}^{r} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$

- Under H_0 this quantity is close to 0 whereas under H_1 this quantity is big.

Theorem

- *Under H_0 we have*

$$T_n \xrightarrow{\mathcal{L}} \chi^2(r-1)$$

- *Under H_1 we have*

$$T_n \xrightarrow{a.s} +\infty$$

- Homogeneity Test, Independency Test

Delta method

- Let (X_n) be a sequence of i.i.d r.v L^2. Denote $\theta = \mathbb{E}[X_1]$ and $\sigma^2 = Var(X_1)$. Recall
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- Recall that the CLT says
$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$
- As already announced, for a particular class of f we would like to understand the convergence of
$$\sqrt{n}\left(f(\bar{X}_n) - f(\theta)\right)$$
- To this end we use the delta method

Delta method

- Let (X_n) be a sequence of i.i.d r.v L^2. Denote $\theta = \mathbb{E}[X_1]$ and $\sigma^2 = Var(X_1)$. Recall

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

- Recall that the CLT says

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- As already announced, for a particular class of f we would like to understand the convergence of

$$\sqrt{n}\left(f(\bar{X}_n) - f(\theta)\right)$$

- To this end we use the delta method

Delta method

- Keep in mind the CLT

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- First let us consider $f(x) = ax + b$ then we have

$$\sqrt{n}\left(a\bar{X}_n - a\theta\right) \xrightarrow{\mathcal{L}} a\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, a^2\sigma^2)$$

- Now suppose that f is differentiable in θ you can write $f(x) = f(\theta) + f'(\theta)(x - \theta) + \circ(|x - \theta|)$. Since $\bar{X}_n - \theta$ converges to 0 almost surely it converges to 0 in probability which allows to write

$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + \circ_{\mathbb{P}}(|\bar{X}_n - \theta|)$$

Delta method

- Keep in mind the CLT

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- First let us consider $f(x) = ax + b$ then we have

$$\sqrt{n}\left(a\bar{X}_n - a\theta)\right) \xrightarrow{\mathcal{L}} a\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, a^2\sigma^2)$$

- Now suppose that f is differentiable in θ you can write $f(x) = f(\theta) + f'(\theta)(x - \theta) + \circ(|x - \theta|)$. Since $\bar{X}_n - \theta$ converges to 0 almost surely it converges to 0 in probability which allows to write

$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + \circ_{\mathbb{P}}(|\bar{X}_n - \theta|)$$

Delta method

- Keep in mind the CLT

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

- First let us consider $f(x) = ax + b$ then we have

$$\sqrt{n}\left(a\bar{X}_n - a\theta\right) \xrightarrow{\mathcal{L}} a\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, a^2\sigma^2)$$

- Now suppose that f is differentiable in θ you can write $f(x) = f(\theta) + f'(\theta)(x - \theta) + \circ(|x - \theta|)$. Since $\bar{X}_n - \theta$ converges to 0 almost surely it converges to 0 in probability which allows to write

$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + \circ_{\mathbb{P}}(|\bar{X}_n - \theta|)$$

- Plugging
$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + o_{\mathbb{P}}(|\bar{X}_n - \theta|)$$
into $\sqrt{n}(f(\bar{X}_n) - f(\theta))$, we get
$$\sqrt{n}(f(\bar{X}_n) - \theta) = \sqrt{n} f'(\theta)(\sqrt{n}(\bar{X}_n - \theta))(1 + o_{\mathbb{P}}(1))$$

- Now the term $1 + o_{\mathbb{P}}(1)$ converges towards 1 in probability and then in Law (since the limit is a constant). Using the Slutsky Lemma allows to conclude that
$$\sqrt{n}(f(\bar{X}_n) - f(\theta)) \xrightarrow{\mathcal{L}} f'(\theta)\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, f'(\theta)^2\sigma^2)$$

- Note that it is easy to extend such result to situation where (T_n) satisfy that there exist a sequence (r_n) and a r.v T (non necessary Gaussian) such that
$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

Delta method

- Plugging
$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + o_{\mathbb{P}}(|\bar{X}_n - \theta|)$$
into $\sqrt{n}(f(\bar{X}_n) - f(\theta))$, we get
$$\sqrt{n}(f(\bar{X}_n) - \theta) = \sqrt{n}f'(\theta)(\sqrt{n}(\bar{X}_n - \theta))(1 + o_{\mathbb{P}}(1))$$

- Now the term $1 + o_{\mathbb{P}}(1)$ converges towards 1 in probability and then in Law (since the limit is a constant). Using the Slutsky Lemma allows to conclude that
$$\sqrt{n}(f(\bar{X}_n) - f(\theta)) \xrightarrow{\mathcal{L}} f'(\theta)\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, f'(\theta)^2\sigma^2)$$

- Note that it is easy to extend such result to situation where (T_n) satisfy that there exist a sequence (r_n) and a r.v T (non necessary Gaussian) such that
$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

Delta method

- Plugging
$$f(\bar{X}_n) = f(\theta) + f'(\theta)(\bar{X}_n - \theta) + \circ_{\mathbb{P}}(|\bar{X}_n - \theta|)$$
into $\sqrt{n}(f(\bar{X}_n) - f(\theta))$, we get
$$\sqrt{n}(f(\bar{X}_n) - \theta) = \sqrt{n}f'(\theta)(\sqrt{n}(\bar{X}_n - \theta))(1 + \circ_{\mathbb{P}}(1))$$

- Now the term $1 + \circ_{\mathbb{P}}(1)$ converges towards 1 in probability and then in Law (since the limit is a constant). Using the Slutsky Lemma allows to conclude that
$$\sqrt{n}(f(\bar{X}_n) - f(\theta)) \xrightarrow{\mathcal{L}} f'(\theta)\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, f'(\theta)^2\sigma^2)$$

- Note that it is easy to extend such result to situation where (T_n) satisfy that there exist a sequence (r_n) and a r.v T (non necessary Gaussian) such that
$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

Delta method: General version

Theorem

Let θ in $\mathbb{R}^k$. Let ϕ be an application from $\mathbb{R}^k$ to $\mathbb{R}^m$ differentiable in θ. We denote $D_\theta\phi(.)$ the corresponding differential application. Let (T_n) be a sequence of random vectors of $\mathbb{R}^k$ such that there exists a sequence (r_n) and a random vector T such that

$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

then we have

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} D_\theta\phi(T)$$

- In the Gaussian case if $Z \sim \mathcal{N}(0, K)$ where K is the covariance matrix and Z a Gaussian vector, then we have

$$D_\theta\phi(Z) \sim \mathcal{N}(0, J_\theta\phi K J_\theta\phi^t),$$

where $J_\theta\phi$ is the Jacobian matrix of ϕ.

Delta method: General version

Theorem

Let θ in $\mathbb{R}^k$. Let ϕ be an application from $\mathbb{R}^k$ to $\mathbb{R}^m$ differentiable in θ. We denote $D_\theta\phi(.)$ the corresponding differential application. Let (T_n) be a sequence of random vectors of $\mathbb{R}^k$ such that there exists a sequence (r_n) and a random vector T such that

$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

then we have

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} D_\theta\phi(T)$$

- In the Gaussian case if $Z \sim \mathcal{N}(0, K)$ where K is the covariance matrix and Z a Gaussian vector, then we have

$$D_\theta\phi(Z) \sim \mathcal{N}(0, J_\theta\phi K J_\theta\phi^t),$$

where $J_\theta\phi$ is the Jacobian matrix of ϕ.

Delta method: General version

Theorem

Let θ in $\mathbb{R}^k$. Let ϕ be an application from $\mathbb{R}^k$ to $\mathbb{R}^m$ differentiable in θ. We denote $D_\theta\phi(.)$ the corresponding differential application. Let (T_n) be a sequence of random vectors of $\mathbb{R}^k$ such that there exists a sequence (r_n) and a random vector T such that

$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

then we have

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} D_\theta\phi(T)$$

- In the Gaussian case if $Z \sim \mathcal{N}(0, K)$ where K is the covariance matrix and Z a Gaussian vector, then we have

$$D_\theta\phi(Z) \sim \mathcal{N}(0, J_\theta\phi K J_\theta\phi^t),$$

where $J_\theta\phi$ is the Jacobian matrix of ϕ.

Delta method: General version

Theorem

Let θ in $\mathbb{R}^k$. Let ϕ be an application from $\mathbb{R}^k$ to $\mathbb{R}^m$ differentiable in θ. We denote $D_\theta\phi(.)$ the corresponding differential application. Let (T_n) be a sequence of random vectors of $\mathbb{R}^k$ such that there exists a sequence (r_n) and a random vector T such that

$$r_n(T_n - \theta) \xrightarrow{\mathcal{L}} T$$

then we have

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} D_\theta\phi(T)$$

- In the Gaussian case if $Z \sim \mathcal{N}(0, K)$ where K is the covariance matrix and Z a Gaussian vector, then we have

$$D_\theta\phi(Z) \sim \mathcal{N}(0, J_\theta\phi K J_\theta\phi^t),$$

where $J_\theta\phi$ is the Jacobian matrix of ϕ.

Conditioning

Definition

Definition

Let B be a event of non zero probability i.e $\mathbb{P}(B) \neq 0$. For all events A we define the conditional probability A knowing B by

$$\mathbb{P}_B(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$
- The application $\mathbb{P}(\cdot|B)$ defines a measure on $(\Omega, \mathcal{A})$
- If $A \perp\!\!\!\perp B$ then $\mathbb{P}(A|B) = \mathbb{P}(A)$

Definition

Let B be a event of non zero probability i.e $\mathbb{P}(B) \neq 0$. For all events A we define the conditional probability A knowing B by

$$\mathbb{P}_B(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$
- The application $\mathbb{P}(\cdot|B)$ defines a measure on $(\Omega, \mathcal{A})$
- If $A \perp\!\!\!\perp B$ then $\mathbb{P}(A|B) = \mathbb{P}(A)$

Total probability law formula and Bayes formula

- Total probability law:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

- Two players A and B owns respectively a and b euros. They throw a dice where a odd number apear with probability p. The player B gives 1 euro to A if a odd number appear and the converse if a even number appears. We define u_a the probability that A bankrupt. We have

$$u_a = pu_{a+1} + (1-p)u_{a-1}$$

- Bayes law:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

In the practice, the total probability law is used to compute $\mathbb{P}(A)$.

Total probability law formula and Bayes formula

- Total probability law:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

- Two players A and B owns respectively a and b euros. They throw a dice where a odd number apear with probability p. The player B gives 1 euro to A if a odd number appear and the converse if a even number appears. We define u_a the probability that A bankrupt. We have

$$u_a = pu_{a+1} + (1-p)u_{a-1}$$

- Bayes law:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

In the practice, the total probability law is used to compute $\mathbb{P}(A)$.

Total probability law formula and Bayes formula

Proposition

Let $A_1, \ldots, A_N$ a partition of Ω then

$$\mathbb{P}(A) = \sum_{i=1}^{N} \mathbb{P}(A|A_i)\mathbb{P}(A_i)$$

$$\mathbb{P}(A_i|A) = \frac{\mathbb{P}(A|A_i)\mathbb{P}(A_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{N} \mathbb{P}(A|A_i)\mathbb{P}(A_i)}$$

Conditional Law

- Let X and Y two random variables. One can write

$$\mathbb{P}(Y \in A, X \in B) = \int \mathbb{P}(Y \in A|X = x)\mathbb{P}_X(dx) = \mathbb{E}[\mathbf{1}_B \mathbb{P}(Y \in A|X)]$$

- The quantity $\mathbb{P}(Y \in A|X = x)$ is a notation which corresponds to the Radon Nykodym derivative
- The family $(\mathbb{P}(Y \in \cdot|X = x)_{x \in \mathbb{R}}$ is called conditional probability law family of Y knowing X.
- The conditional law of Y knowing X is denoted by $\mathbb{P}(Y \in \cdot|X)$

Conditional Law

- In the discrete case, the conditional probability law family is easy to obtain. In particular

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$$

we have then

$$\mathbb{P}(Y \in \cdot|X) = \sum_{x \in S(X)} \mathbb{P}(Y = y|X = x)\mathbf{1}_{X=x}$$

- In the continuous case, we speak about conditional density. To this end, we put

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}\mathbf{1}_{f_X(x)>0}$$

with

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

Conditional Law

- In the discrete case, the conditional probability law family is easy to obtain. In particular

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$$

we have then

$$\mathbb{P}(Y \in \cdot|X) = \sum_{x \in S(X)} \mathbb{P}(Y = y|X = x)\mathbf{1}_{X=x}$$

- In the continuous case, we speak about conditional density. To this end, we put

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}\mathbf{1}_{f_X(x)>0}$$

with

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

Conditional expectation

- So far we have addressed conditional probability. We want to construct a notion of conditional expectation. Let us consider the following

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \mathbb{E}[\mathbf{1}_A|B]$$

- Then one is tempting to define the conditional expectation of a r.v knowing an event by

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}$$

-

- Now we aim to extend this notation to the conditional expectation to a r.v knowing a σ-algebra $\mathcal{B}$:

$$\mathbb{E}[X|\mathcal{B}]???$$

Conditional expectation

- So far we have addressed conditional probability. We want to construct a notion of conditional expectation. Let us consider the following
$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \mathbb{E}[\mathbf{1}_A|B]$$

- Then one is tempting to define the conditional expectation of a r.v knowing an event by
$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}$$

-

- Now we aim to extend this notation to the conditional expectation to a r.v knowing a σ-algebra $\mathcal{B}$:

$$\mathbb{E}[X|\mathcal{B}]???$$

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let B such that $0 < \mathbb{P}[B] < 1$. Consider $\mathcal{B} = \sigma(B)$ the σ-algebra generated by B.

$$\mathcal{B} = \{\emptyset, B, B^c, \Omega\},$$

- We put for X a L^1 r.v

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- This is a random variable called conditional expectation of X knowing $\mathcal{B}$.
- Note that this r.v is measurable with respect to $\mathcal{B}$

Conditional expectation

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let B such that $0 < \mathbb{P}[B] < 1$. Consider $\mathcal{B} = \sigma(B)$ the σ-algebra generated by B.

$$\mathcal{B} = \{\emptyset, B, B^c, \Omega\},$$

- We put for X a L^1 r.v

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- This is a random variable called conditional expectation of X knowing $\mathcal{B}$.
- Note that this r.v is measurable with respect to $\mathcal{B}$

Conditional expectation

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let B such that $0 < \mathbb{P}[B] < 1$. Consider $\mathcal{B} = \sigma(B)$ the σ-algebra generated by B.

$$\mathcal{B} = \{\emptyset, B, B^c, \Omega\},$$

- We put for X a L^1 r.v

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- This is a random variable called conditional expectation of X knowing $\mathcal{B}$.
- Note that this r.v is measurable with respect to $\mathcal{B}$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- Let us investigate the property of this random variable

$$Y = \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c}$$

- First note that

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}_B] &= \mathbb{E}[(\mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c})\,\mathbf{1}_B] \\
&= \mathbb{E}[(\mathbb{E}[X|B])\,\mathbf{1}_B] \\
&= \mathbb{E}[X|B]\mathbb{E}[\mathbf{1}_B] \\
&= \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}[B]}\mathbb{P}[B] \\
&= \mathbb{E}[X\mathbf{1}_B] \\
\mathbb{E}[Y\mathbf{1}_{B^c}] &= \mathbb{E}[X\mathbf{1}_{B^c}]
\end{aligned}$$

- We easy see also that $\mathbb{E}[Y\mathbf{1}_\emptyset] = \mathbb{E}[X\mathbf{1}_\emptyset]$ and $\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_\Omega] = \mathbb{E}[X\mathbf{1}_\Omega] = \mathbb{E}[X]$

Conditional expectation

- As a conclusion we can see that for all event $G \in \mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we have

$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \tag{1}$$

- The r.v $Y = \mathbb{E}[X|\mathcal{B}]$ is the only r.v $\mathcal{B}$ mesurable satisfying the above property.
- Indeed a $\mathcal{B}$ mesurable r.v Z can be written in form of

$$Z = \alpha\mathbf{1}_B + \beta\mathbf{1}_{B^c}$$

then asking (1) implies $\alpha = \mathbb{E}[X|B]$ and $\beta = \mathbb{E}[X|B^c]$

Conditional expectation

- As a conclusion we can see that for all event $G \in \mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we have
$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \tag{1}$$
- The r.v $Y = \mathbb{E}[X|\mathcal{B}]$ is the only r.v $\mathcal{B}$ mesurable satisfying the above property.
- Indeed a $\mathcal{B}$ mesurable r.v Z can be written in form of
$$Z = \alpha\mathbf{1}_B + \beta\mathbf{1}_{B^c}$$
then asking (1) implies $\alpha = \mathbb{E}[X|B]$ and $\beta = \mathbb{E}[X|B^c]$

Conditional expectation

- As a conclusion we can see that for all event $G \in \mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we have
$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \tag{1}$$

- The r.v $Y = \mathbb{E}[X|\mathcal{B}]$ is the only r.v $\mathcal{B}$ mesurable satisfying the above property.
- Indeed a $\mathcal{B}$ mesurable r.v Z can be written in form of
$$Z = \alpha\mathbf{1}_B + \beta\mathbf{1}_{B^c}$$
then asking (1) implies $\alpha = \mathbb{E}[X|B]$ and $\beta = \mathbb{E}[X|B^c]$

Conditional expectation

- Let us go further and consider $\mathcal{B} = \sigma\{B_i, i = 1, \ldots, N\}$, where B_i is a partition of Ω, that is

$$\Omega = \bigcup_{i=1}^{N} B_i, \quad B_i \cap B_j = \emptyset, i \neq j$$

- We define

$$\mathbb{E}[X|\mathcal{B}] = \sum_{i=1}^{N} \mathbb{E}[X|B_i] \mathbf{1}_{B_i}$$

- One can verify that for all $G \in \mathcal{B}$

$$\mathbb{E}\left[\mathbb{E}[X|\mathcal{B}]\mathbf{1}_G\right] = \mathbb{E}[X\mathbf{1}_G]$$

and this is the only $\mathcal{B}$ mesurable r.v satisfying such a property.

Conditional expectation

- We have the following theorem

Theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{B} \subset \mathcal{A}$. Let X be a L^1 r.v. There exists a unique r.v Y with is $\mathcal{B}$ mesurable such that

$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G],$$

for all $G \in \mathcal{B}$. We denote this r.v

$$\mathbb{E}[X|\mathcal{B}]$$

the conditional expectation knowing $\mathcal{B}$

Conditional expectation

- We have the following theorem

Theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{B} \subset \mathcal{A}$. Let X be a L^1 r.v. There exists a unique r.v Y with is $\mathcal{B}$ mesurable such that

$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G],$$

for all $G \in \mathcal{B}$. We denote this r.v

$$\mathbb{E}[X|\mathcal{B}]$$

the conditional expectation knowing $\mathcal{B}$

- The conditioning calls for partial information and as we shall see the r.v $\mathbb{E}[X|\mathcal{B}]$ is somehow best "approximation" of X knowing only the information included in $\mathcal{B}$.
- Come back to $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we recall that

$$
\begin{aligned}
\mathbb{E}[X|\mathcal{B}] &= \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c} && (2)\\
&= \sqrt{\mathbb{P}[B]}\mathbb{E}[X|B]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \sqrt{\mathbb{P}[B^c]}\mathbb{E}[X|B^c]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} && (3)\\
&= \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} && (4)
\end{aligned}
$$

Conditional expectation

- The conditioning calls for partial information and as we shall see the r.v $\mathbb{E}[X|\mathcal{B}]$ is somehow best "approximation" of X knowing only the information included in $\mathcal{B}$.
- Come back to $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we recall that

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c} \tag{2}$$

$$= \sqrt{\mathbb{P}[B]}\mathbb{E}[X|B]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \sqrt{\mathbb{P}[B^c]}\mathbb{E}[X|B^c]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \tag{3}$$

$$= \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \tag{4}$$

Conditional expectation

- The conditioning calls for partial information and as we shall see the r.v $\mathbb{E}[X|\mathcal{B}]$ is somehow best "approximation" of X knowing only the information included in $\mathcal{B}$.
- Come back to $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we recall that

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c} \tag{2}$$

$$= \sqrt{\mathbb{P}[B]}\mathbb{E}[X|B]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \sqrt{\mathbb{P}[B^c]}\mathbb{E}[X|B^c]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \tag{3}$$

$$= \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \tag{4}$$

- The conditioning calls for partial information and as we shall see the r.v $\mathbb{E}[X|\mathcal{B}]$ is somehow best "approximation" of X knowing only the information included in $\mathcal{B}$.
- Come back to $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ we recall that

$$
\begin{aligned}
& \mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X|B]\mathbf{1}_B + \mathbb{E}[X|B^c]\mathbf{1}_{B^c} && (2)\\
= \quad & \sqrt{\mathbb{P}[B]}\mathbb{E}[X|B]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \sqrt{\mathbb{P}[B^c]}\mathbb{E}[X|B^c]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} && (3)\\
= \quad & \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} && (4)
\end{aligned}
$$

Conditional expectation

- If X is L^2 one can write

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \qquad (5)$$

in the form

$$\mathbb{E}[X|\mathcal{B}] = \left\langle X, \frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right\rangle\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \left\langle X, \frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right\rangle\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \qquad (6)$$

where

$$\langle X, Y\rangle = \mathbb{E}[XY],$$

is the scalar product in L^2

- Note that one can easily check that $\left\{\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}, \frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right\}$ is an orthonormal basis of $L^2((\Omega, \mathcal{B}, \mathbb{P}))$
- $\mathbb{E}[X|\mathcal{B}]$ is then just the L^2 orthonormal projection of X onto $L^2((\Omega, \mathcal{B}, \mathbb{P}))$.

Conditional expectation

- If X is L^2 one can write

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}\left[X\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right]\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \mathbb{E}\left[X\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right]\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \quad (5)$$

in the form

$$\mathbb{E}[X|\mathcal{B}] = \left\langle X, \frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}\right\rangle\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}} + \left\langle X, \frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right\rangle\frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}} \quad (6)$$

where

$$\langle X, Y\rangle = \mathbb{E}[XY],$$

is the scalar product in L^2

- Note that one can easily check that $\left\{\frac{\mathbf{1}_B}{\sqrt{\mathbb{P}[B]}}, \frac{\mathbf{1}_{B^c}}{\sqrt{\mathbb{P}[B^c]}}\right\}$ is an orthonormal basis of $L^2((\Omega, \mathcal{B}, \mathbb{P}))$
- $\mathbb{E}[X|\mathcal{B}]$ is then just the L^2 orthonormal projection of X onto $L^2((\Omega, \mathcal{B}, \mathbb{P}))$.

Conditional expectation

- In fact, in the case where X is L^2, the property

$$\mathbb{E}[\mathbb{E}[X|\mathcal{B}]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G]$$

  for all $G \in \mathcal{B}$ means that $\mathbb{E}[X|\mathcal{B}]$ is the orthogonal projection of X onto $L^2((\Omega, \mathcal{B}, \mathbb{P}))$

- We can then express the following result which is useful in some situation (for example in the Gaussian context)

Theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{B} \subset \mathcal{A}$. Let X be a L^2 r.v.

The conditional expectation of X knowing $\mathcal{B}$ is the orthogonal projection of X onto $L^2((\Omega, \mathcal{B}, \mathbb{P}))$

Conditional expectation

- Recall that the conditional law of Y knowing X was given by

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}\mathbf{1}_{f_X(x)>0}, \quad f_{Y|X}(y) = \frac{f_{X,Y}(X,y)}{f_X(X)}\mathbf{1}_{f_X(x)>0}$$

with

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

- Let denote $\mathbb{E}[h(Y)|X] = \mathbb{E}[h(Y)|\sigma(X)]$, where $\sigma(X)$ is the σ-algebra generated by X
- We have

$$\mathbb{E}[h(Y)|X] = \int h(y)f_{Y|X}(X,y)dy$$

Conditional expectation

- Recall that the conditional law of Y knowing X was given by

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \mathbf{1}_{f_X(x)>0}, \quad f_{Y|X}(y) = \frac{f_{X,Y}(X,y)}{f_X(X)} \mathbf{1}_{f_X(x)>0}$$

  with

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

- Let denote $\mathbb{E}[h(Y)|X] = \mathbb{E}[h(Y)|\sigma(X)]$, where $\sigma(X)$ is the σ-algebra generated by X
- We have

$$\mathbb{E}[h(Y)|X] = \int h(y) f_{Y|X}(X,y)dy$$

Conditional expectation

- Some useful properties

$$\mathbb{E}[\mathbb{E}[X|\mathcal{B}]] = \mathbb{E}[X]$$

- if X is independent of $\mathcal{B}$

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X]$$

- If X is $\mathcal{B}$ mesurable

$$\mathbb{E}[X|\mathcal{B}] = X$$

- If Z is $\mathcal{B}$ mesurable

$$\mathbb{E}[X\,Z|\mathcal{B}] = \mathbb{E}[X|\mathcal{B}]Z$$

Estimation

Generality

- Let us consider a parametric model where θ is an unknown parameter valued in $\Theta \subset \mathbb{R}^d$
- Recall that an estimator of θ is a r.v which is measurable with respect to a n sample $X_1, \ldots, X_n$

Definition

- An estimator T is said to be unbiased if for all $\theta \in \Theta$

$$\mathbb{E}_\theta[T] = \theta$$

- T is said to be consistent if for all $\theta \in \Theta$

$$T(X_1, \ldots, X_n) \to_{n\to\infty} \theta$$

  in probability or almost surely (with respect to $\mathbb{P}_\theta$)
- T is said asymptotically normal if there exists a sequence (a_n) converging to ∞ such that

$$a_n \left(T(X_1, \ldots, X_n) - \theta\right) \to \mathcal{N}(0, 1)$$

Generality

- Let us consider a parametric model where θ is an unknown parameter valued in $\Theta \subset \mathbb{R}^d$
- Recall that an estimator of θ is a r.v which is measurable with respect to a n sample $X_1, \ldots, X_n$

Definition

- An estimator T is said to be unbiased if for all $\theta \in \Theta$

$$\mathbb{E}_\theta[T] = \theta$$

- T is said to be consistent if for all $\theta \in \Theta$

$$T(X_1, \ldots, X_n) \to_{n \to \infty} \theta$$

  in probability or almost surely (with respect to $\mathbb{P}_\theta$)
- T is said asymptotically normal if there exists a sequence (a_n) converging to ∞ such that

$$a_n \left(T(X_1, \ldots, X_n) - \theta \right) \to \mathcal{N}(0, 1)$$

Generality

- Let us consider a parametric model where θ is an unknown parameter valued in $\Theta \subset \mathbb{R}^d$
- Recall that an estimator of θ is a r.v which is measurable with respect to a n sample $X_1, \ldots, X_n$

Definition

- An estimator T is said to be unbiased if for all $\theta \in \Theta$

$$\mathbb{E}_\theta[T] = \theta$$

- T is said to be consistent if for all $\theta \in \Theta$

$$T(X_1, \ldots, X_n) \to_{n \to \infty} \theta$$

  in probability or almost surely (with respect to $\mathbb{P}_\theta$)
- T is said asymptotically normal if there exists a sequence (a_n) converging to ∞ such that

$$a_n \left(T(X_1, \ldots, X_n) - \theta \right) \to \mathcal{N}(0, 1)$$

Moment estimation

- Let $(X_1, \ldots, X_n)$ a sample
- Recall that the moment of order k for a r.v is

$$\mathbb{E}[X_1^k] = \mathbb{E}[X_i^k], i = 1, \ldots, k$$

- We can replace these moments by their empirical version that is

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The centered version

$$\mathbb{E}\left[(X_1 - E[X_1])^k\right]$$

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^k$$

Moment estimation

- Let $(X_1, \dots, X_n)$ a sample
- Recall that the moment of order k for a r.v is

$$\mathbb{E}[X_1^k] = \mathbb{E}[X_i^k], i = 1, \dots, k$$

- We can replace these moments by their empirical version that is

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The centered version

$$\mathbb{E}\left[(X_1 - E[X_1])^k\right]$$

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^k$$

Moment estimation

- Let $(X_1, \ldots, X_n)$ a sample
- Recall that the moment of order k for a r.v is

$$\mathbb{E}[X_1^k] = \mathbb{E}[X_i^k], i = 1, \ldots, k$$

- We can replace these moments by their empirical version that is

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The centered version

$$\mathbb{E}\left[(X_1 - E[X_1])^k\right]$$

$$\bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^k$$

Moment estimation

- Method principle
- Assuming that you can apply the Law of large numbers we have

$$\bar{X}_n^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{a.s} \mathbb{E}X_1^k$$

- Assume that $X = (X_1, \ldots, X_n)$ is distributed along $\mathbb{P}_\theta$ where $\theta \in \Theta$ is unknown.
- Hope: extract information on θ by knowing the moment

Moment estimation

- Example
- Bernoulli of parameter θ: $\mathcal{B}(\theta)$

$$\mathbb{E}[X_1] = \theta$$

we can use the first moment

$$T = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \theta$$

- We also have

$$\mathbb{E}[X_1^2] = \theta$$

we can use the second moment

$$\bar{X}_n \xrightarrow{a.s} \theta, \quad T = \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{a.s} \theta$$

Moment estimation

- Example
- Bernoulli of parameter θ: $\mathcal{B}(\theta)$

$$\mathbb{E}[X_1] = \theta$$

we can use the first moment

$$T = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \theta$$

- We also have

$$\mathbb{E}[X_1^2] = \theta$$

we can use the second moment

$$\bar{X}_n \xrightarrow{a.s} \theta, \quad T = \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{a.s} \theta$$

Moment estimation

- Example
- Binomial of parameter (k, θ). Assume you know k and just want to estimate θ

$$\mathbb{E}[X_1] = k\theta$$

we can use the first moment

$$T = \frac{1}{k}\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \theta$$

- Assume you do not know k and need to estimate k and θ you should use also the second moment

$$Var(X_1) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))^2] = k\theta(1-\theta) = \mathbb{E}[X_1](1-\theta)$$

- Then

$$\theta = 1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}, \quad k = \frac{\mathbb{E}[X_1]}{1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}}$$

Moment estimation

- Example
- Binomial of parameter (k, θ). Assume you know k and just want to estimate θ

$$\mathbb{E}[X_1] = k\theta$$

we can use the first moment

$$T = \frac{1}{k}\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s} \theta$$

- Assume you do not know k and need to estimate k and θ you should use also the second moment

$$Var(X_1) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))^2] = k\theta(1-\theta) = \mathbb{E}[X_1](1-\theta)$$

- Then

$$\theta = 1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}, \quad k = \frac{\mathbb{E}[X_1]}{1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}}$$

- Then

$$\theta = 1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}, \quad k = \frac{\mathbb{E}[X_1]}{1 - \frac{Var(X_1)}{\mathbb{E}[X_1]}}$$

- Then we can estimate k and θ by putting

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

and defining

$$\hat{\theta}(X_1, \ldots, X_n) = 1 - \frac{\hat{\sigma}_n^2}{\bar{X}_n}, \quad \hat{k}(X_1, \ldots, X_n) = \frac{\bar{X}_n}{1 - \frac{\hat{\sigma}_n^2}{\bar{X}_n}}$$

Moment estimation

- Case of a sample $(X_1, \ldots, X_n)$ whose density is $f_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{\mathbb{R}^+}(x)$
- simple computation shows that

$$\mathbb{E}[X_1] = \frac{1}{\theta}$$

- Then our estimator of θ can be chosen as

$$\hat{\theta} = \frac{1}{\bar{X}_n}$$

- Exercise: do the same job for $(X_1, \ldots, X_n)$ distributed along $\mathcal{N}(\mu, \sigma^2)$

Moment estimation

- In order to summarize. Assume you want to estimate $g(\theta)$. First you should find h such that

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

- Determine the number p of moments you shall need to recover $g(\theta)$
- Then compute the p moments you need and connect them to the quantity you aim to estimate
- Replace these p moments by their empirical version.
- Unbiaised, asymptotic normality, Delta method

Moment estimation

- In order to summarize. Assume you want to estimate $g(\theta)$. First you should find h such that

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

- Determine the number p of moments you shall need to recover $g(\theta)$
- Then compute the p moments you need and connect them to the quantity you aim to estimate
- Replace these p moments by their empirical version.
- Unbiaised, asymptotic normality, Delta method

Moment estimation

- In order to summarize. Assume you want to estimate $g(\theta)$. First you should find h such that

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

- Determine the number p of moments you shall need to recover $g(\theta)$
- Then compute the p moments you need and connect them to the quantity you aim to estimate
- Replace these p moments by their empirical version.
- Unbiaised, asymptotic normality, Delta method

Moment estimation

- In order to summarize. Assume you want to estimate $g(\theta)$. First you should find h such that

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

- Determine the number p of moments you shall need to recover $g(\theta)$
- Then compute the p moments you need and connect them to the quantity you aim to estimate
- Replace these p moments by their empirical version.
- Unbiaised, asymptotic normality, Delta method

Moment estimation

- Comme back to the initial question with the notion of bias and asymtptotic normality.
- If you have found h such that $\mathbb{E}[h(X_1)] = g(\theta)$ then using

$$T = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s} g(\theta)$$

  T is an unbiaised estimator of $g(\theta)$
- Assume that $Var(h(X_1)) = \sigma^2(\theta)$ then we have

$$\sqrt{n}\left(\frac{T - g(\theta)}{\sigma(\theta)}\right) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0,1)$$

Moment estimation

- One can see that the moment method has weakness
- First you can see that in the study of asymptotically normality one see that it depends on $\sigma(\theta)$ which is also unknown.
- You can avoid this obstacle using Slutsky Lemma, you look at

$$\sqrt{n}\left(\frac{T - g(\theta)}{\hat{\sigma}_n^2}\right) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0, 1)$$

Moment estimation

- It is not evident to find h such that $\mathbb{E}[h(X_1)] = g(\theta)$. For example the density case where $f_\theta(x) = \theta e^{-\theta x}\mathbf{1}_{\mathbb{R}^+}(x)$, the estimator of θ was

$$T = \frac{n}{X_1 + \ldots + X_n}$$

and it is not even easy to compute $\mathbb{E}[T]$ which makes the study of bias not straightforward.

- Concerning the asymptotically normality property you have to use delta method to get

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta}\right) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0, 1/\theta^2), \text{then} \quad \sqrt{n}\,(T-\theta) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0,\theta^2)$$

Moment estimation

- It is not evident to find h such that $\mathbb{E}[h(X_1)] = g(\theta)$. For example the density case where $f_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{\mathbb{R}^+}(x)$, the estimator of θ was

$$T = \frac{n}{X_1 + \ldots + X_n}$$

and it is not even easy to compute $\mathbb{E}[T]$ which makes the study of bias not straightforward.

- Concerning the asymptotically normality property you have to use delta method to get

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta}\right) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0, 1/\theta^2), \text{then} \quad \sqrt{n}\,(T - \theta) \xrightarrow{\mathcal{L}_\theta} \mathcal{N}(0, \theta^2)$$

Maximum likelihood

- The framework is the following, we consider a parametric model $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ and we consider that the model is dominated in the sense that for all θ there exists f_θ such that for all $A \in \mathcal{A}$:

$$\mathbb{P}_\theta(A) = \int_A f_\theta(x) d\mu(x)$$

Definition (Vraissemblance)

Let $(X_1, \ldots, X_n)$ be a n-sample of probability $\mathbf{P}_\theta$, we call likelihood of this sample, the joint density of this sample with respect to μ. We denote it as

$$L(x_1, \ldots, x_n; \theta;).$$

In general this can be expressed as

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Maximum likelihood

- The framework is the following, we consider a parametric model $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ and we consider that the model is dominated in the sense that for all θ there exists f_θ such that for all $A \in \mathcal{A}$:

$$\mathbb{P}_\theta(A) = \int_A f_\theta(x) d\mu(x)$$

Definition (Vraissemblance)

Let $(X_1, \ldots, X_n)$ be a n-sample of probability $\mathbf{P}_\theta$, we call likelihood of this sample, the joint density of this sample with respect to μ. We denote it as

$$L\left(x_1, \ldots, x_n; \theta;\right).$$

In general this can be expressed as

$$L\left(x_1, \ldots, x_n; \theta\right) = \prod_{i=1}^{n} f_\theta(x_i).$$

- In the discrete case it takes the form

$$L_n(x_1, \dots, x_n, \theta) = \mathbb{P}_\theta(X_1 = x_1) \dots \mathbb{P}_\theta(X_n = x_n)$$

- In the continuous case

$$L_n(x_1, \dots, x_n, \theta) = f_\theta(x_1) \dots f_\theta(x_n)$$

where f_θ corresponds to the density of X_1 with respect to the Lebesgue measure.

- In the discrete case it takes the form

$$L_n(x_1, \dots, x_n, \theta) = \mathbb{P}_\theta(X_1 = x_1) \dots \mathbb{P}_\theta(X_n = x_n)$$

- In the continuous case

$$L_n(x_1, \dots, x_n, \theta) = f_\theta(x_1) \dots f_\theta(x_n)$$

  where f_θ corresponds to the density of X_1 with respect to the Lebesgue measure.

- Example
- Let $(X_1, \ldots, X_n)$ be a n-sample of law $\mathcal{N}(m, \sigma^2)$. Assume that the unknown parameters are $\theta = (m, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+$.

$$L\left(x_1, \ldots, x_n; \theta\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2 \prod \sigma^2}} e^{-\frac{(x_i - m)^2}{2\sigma^2}} = \frac{1}{\left(2 \prod \sigma^2\right)^{n/2}} e^{-\frac{\sum_{i=1}^{n} (x_i - m)^2}{2\sigma^2}}.$$

- Let $(X_1, \ldots, X_n)$ be a n-sample of law $\mathcal{P}(\theta)$. Assume that the unknown parameter $\theta \in \mathbf{R}$.

$$L\left(x_1, \ldots, x_n; \theta\right) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} = e^{-n\theta} \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

MLE

Definition

Let consider a statistical model dominated by a measure μ and let $L(X, \theta)$ be its likelihood function. All statistic $\hat{\theta}_n^{MV} = \hat{\theta}_n^{MV}(X_1, \ldots, X_n)$ such that

$$L(X_1, \ldots, X_n, \hat{\theta}_n^{MV}) = \max_{\theta} L(X_1, \ldots, X_n, \theta)$$

is called estimator of the maximum likelihood. We shall denote

$$\hat{\theta}_n^{MV} = argmax \ L(X_1, \ldots, X_n, \theta)$$

if there are several point where the maximum is reached, we can replace $=$ by $\in$

In the sequel, we shall denote the so-called log likelihood

$$l_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \ln L(X_i, \theta).$$

Definition

Let consider a statistical model dominated by a measure μ and let $L(X,\theta)$ be its likelihood function. All statistic $\hat{\theta}_n^{MV} = \hat{\theta}_n^{MV}(X_1,\ldots,X_n)$ such that

$$L(X_1,\ldots,X_n,\hat{\theta}_n^{MV}) = \max_{\theta} L(X_1,\ldots,X_n,\theta)$$

is called estimator of the maximum likelihood. We shall denote

$$\hat{\theta}_n^{MV} = argmax\ L(X_1,\ldots,X_n,\theta)$$

if there are several point where the maximum is reached, we can replace $=$ by $\in$

In the sequel, we shall denote the so-called log likelihood

$$l_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln L(X_i,\theta).$$

MLE

- Example
- Laplace model $f(x, \theta) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\theta|}{\sigma}\right)$, $\theta \in \mathbf{R}$, unknown and σ known.

$$l_n(\theta) = \ln(2\sigma) + \frac{1}{n\sigma} \sum_{i=1}^{n} |X_i - \theta|.$$

- We shall need to find the minium of $\sum |X_i - \theta|$. Note that this function is almost surely differentiable and its differential h is given by

$$-\sum_{i=1}^{n} sign(X_i - \theta) = h(\theta).$$

if n is even the differential vanishes on every point of $[X_{(n/2)}, X_{(n/2+1)}]$ and then any point of this interval is an MLE. If n is odd a unique MLE is the mediane but there is no point where the differential vanishes.

MLE

- Example
- Laplace model $f(x,\theta) = \frac{1}{2\sigma}\exp\left(-\frac{|x-\theta|}{\sigma}\right)$, $\theta \in \mathbf{R}$, unknown and σ known.

$$l_n(\theta) = \ln(2\sigma) + \frac{1}{n\sigma}\sum_{i=1}^{n} |X_i - \theta|.$$

- We shall need to find the minium of $\sum |X_i - \theta|$. Note that this function is almost surely differentiable and its differential h is given by

$$-\sum_{i=1}^{n} sign(X_i - \theta) = h(\theta).$$

if n is even the differential vanishes on every point of $[X_{(n/2)}, X_{(n/2+1)}]$ and then any point of this interval is an MLE. If n is odd a unique MLE is the mediane but there is no point where the differential vanishes.

- Cauchy law $f(x) = \frac{1}{\pi(1+(x-\theta)^2)}$.
  The critical point study is not explicit, in general there exists many critical point and then many MLE.
- Consider a model of the form

$$f(x,\theta) = f_0(x-\theta)$$

  with

$$f_0(x) = \frac{e^{-|x|/2}}{2\sqrt{2\pi|x|}}.$$

  then the likelihood converges towards $+\infty$ when $\theta \to X_i$ for all i then tehre is no MLE.

MLE

- Normal case $\mathcal{N}(\mu, \sigma^2)$
- Bernoulli case: $\mathcal{B}(\theta)$
- Uniform law case: $\mathcal{U}([0, \theta])$

MLE

- What can we say about the asymptotic behaviour of the MLE
- First we shall address the consistency
- To this end we introduce an assumption

$$\int |\ln f_\theta(x)| f_{\theta^*}(x) d\mu(x) < \infty, \ \forall \theta \in \Theta. \tag{7}$$

- This means that the r.v

$$-\ln(f_\theta(X_1)) \in L^1$$

and then we can applied the LLN to get that

$$l_n(\theta) \xrightarrow{\mathbb{P}_{\theta^*} a.s} J(\theta) := -\int f(x, \theta^*) \ln f(x, \theta) d\mu$$

MLE

- What can we say about the asymptotic behaviour of the MLE
- First we shall address the consistency
- To this end we introduce an assumption

$$\int |\ln f_\theta(x)| f_{\theta^*}(x) d\mu(x) < \infty, \ \forall \theta \in \Theta. \tag{7}$$

- This means that the r.v

$$-\ln(f_\theta(X_1)) \in L^1$$

and then we can applied the LLN to get that

$$l_n(\theta) \overset{\mathbb{P}_{\theta^*} a.s}{\longrightarrow} J(\theta) := -\int f(x, \theta^*) \ln f(x, \theta) d\mu$$

1. We have $J(\theta) \geqslant J(\theta^*)$.
2. If moreover the model is identifiable the inequality is strict as soon as $\theta \neq \theta^*$.
3. Now we know that $I_n(\theta)$ converges towards $J(\theta)$ we can hope that the argmin of $I_n(\theta)$ converges towards the argmin of $J(\theta)$ which appears to be θ^* under the hypotheses of identifiability.

MLE

1. We have $J(\theta) \geqslant J(\theta^*)$.
2. If moreover the model is identifiable the inequality is strict as soon as $\theta \neq \theta^*$.
3. Now we know that $l_n(\theta)$ converges towards $J(\theta)$ we can hope that the argmin of $l_n(\theta)$ converges towards the argmin of $J(\theta)$ which appears to be θ^* under the hypotheses of identifiability.

Theorem

Suppose that Θ is an open set of $\mathbf{R}$ *and*

1. *that for all x the density $f(x, \theta)$ is continuous in θ,*
2. *that the model is identifiable*
3. *that the Hypothesis (7) is satisfied*
4. *that for all n $\hat{\theta}_n^{MV}$ exists and that the set of local minima of $l_n(\theta)$ is a bounded closed interval include in θ.*

then $\hat{\theta}_n^{MV}$ is a consistant estimator (which converges in probability with respect to $\mathbb{P}_{\theta^*}$).

MLE

Weibull Model of density $f(x,\theta) = \theta x^{\theta-1}\exp(-x^{\theta})\mathbf{1}_{x>0}$. We then obtain

$$l_n(\theta) = -\ln\theta - (\theta-1)\frac{1}{n}\sum_{i=1}^{n}\ln X_i + \frac{1}{n}\sum_{i=1}^{n} X_i^{\theta}$$

$$l_n'(\theta) = -\frac{1}{\theta} - \frac{1}{n}\sum_{i=1}^{n}\ln X_i + \frac{1}{n}\sum X_i^{\theta}\ln X_i$$

$$l_n''(\theta) = \frac{1}{\theta^2} + \frac{1}{n}\sum X_i^{\theta}(\ln X_i)^2 > 0.$$

a study of the function shows that there exists only one critival point which is then a global minimum, we have then existence and uniqueness $\hat{\theta}_n^{MV}$. It remains just to verify that

$$\mathbf{E}_{\theta^*}\left(\left|\ln f_{\theta}(X)\right|\right) < +\infty.$$

and then we conclude that $\hat{\theta}_n^{MV}$ is consistent.

MLE

We shall say that a model is ML regular if

1. The model is dominated
2. Θ is an open set of $\mathbf{R}$ and $f(x,\theta) > 0 \iff f(x,\theta') > 0$
3. The functions f and $l = \ln f$ are C^2 in θ.
4. $\forall\theta^*$ there exists a neighborhood of θ^* denoted by U and a function $\Lambda(x)$ such that
   $|l''(x,\theta)| \leqslant \Lambda(x),\ |l'(x,\theta)| \leqslant \Lambda(x),\ |l'(x,\theta)|^2 \leqslant \Lambda(x)$ for all $\theta \in U$ and μ almost surely in x and
   $$\int \Lambda(x) \sup_{\theta\in U} f(x,\theta) d\mu < \infty.$$
5. $I(\theta) := \mathbf{E}_{\theta^*}\left[l'(X,\theta^*)l'(X,\theta^*)^t\right] = -\mathbf{E}_{\theta^*}\left[l''(X,\theta^*)\right] > 0,\ \forall\theta \in \Theta.$

Theorem (T.C.L pour $\hat{\theta}_n^{MV}$)

Suppose that the model M.V. is regular and Let $\hat{\theta}_n^{MV}$ be a sequence of consistent de square root of $l_n'(\theta) = 0$. Then $\forall \theta^* \in \theta$

$$\sqrt{n}(\hat{\theta}_n^{MV} - \theta^*) \to \mathcal{N}(0, 1/I(\theta^*)).$$

The quantity

$$I(\theta) := \mathbf{E}_{\theta^*}\left[l'(X,\theta^*)l'(X,\theta^*)^t\right] = -\mathbf{E}_{\theta^*}\left[l''(X,\theta^*)\right]$$

is usually called the Fisher information

MLE

- Why are we interested by unbiased estimator?
- Let (T_n) an estimator of θ, we have the quadratic risk defined by

$$\mathbb{E}((T_n - \theta)^2)$$

which corresponds to the L^2 distance between our estimator T_n and the target θ

- One can write

$$\begin{aligned}
& \mathbb{E}((T_n - \theta)^2) \\
= \ & \mathbb{E}((T_n - \mathbb{E}(T_n) + \mathbb{E}(T_n) - \theta)^2) \\
= \ & \mathbb{E}((T_n - \mathbb{E}(T_n))^2 + 2\mathbb{E}((T_n - \mathbb{T}_n)(\mathbb{E}(T_n) - \theta)) + (\mathbb{E}(T_n) - \theta)^2 \\
= \ & \color{red}{\mathbb{E}((T_n - \mathbb{E}(T_n))^2} + \color{blue}{(\mathbb{E}(T_n) - \theta)^2}
\end{aligned}$$

which is called the variance-bias decomposition. The bias makes the distance larger.

Confidence set

Confidence set

- In this section we shall follow an example to make clear the idea behind the confidence set
- Essentially when we make an estimation we are forced to make an error. Confidence set are here to control this error.
- The idea is to construct a random interval (or set in higher dimension) who contains the true parameter with high probability.
- For example if $\bar{\mu}$ is an estimation we want to determine ϵ such that a true parameter satisfies

$$\mathbb{P}[\mu \in [-\epsilon + \bar{\mu}, \epsilon + \bar{\mu}]] = 1 - \alpha$$

where α is small (such that $\mathbb{P}[\mu \in [-\epsilon + \mu, \epsilon + \mu]]$ is close to 1)

Confidence set

- In this section we shall follow an example to make clear the idea behind the confidence set
- Essentially when we make an estimation we are forced to make an error. Confidence set are here to control this error.
- The idea is to construct a random interval (or set in higher dimension) who contains the true parameter with high probability.
- For example if $\bar{\mu}$ is an estimation we want to determine ϵ such that a true parameter satisfies

$$\mathbb{P}[\mu \in [-\epsilon + \bar{\mu}, \epsilon + \bar{\mu}]] = 1 - \alpha$$

where α is small (such that $\mathbb{P}[\mu \in [-\epsilon + \mu, \epsilon + \mu]]$ is close to 1)

Confidence set

- Let consider the guiding example of $(X_1, \dots, X_n)$ a n-sample of Bernoulli law of parameter θ^*: $\mathcal{B}(\theta^*)$
- As we have seen a good estimator is

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

- We know that

$$\bar{X}_n \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*$$

- Let us try to estimate

$$\mathbb{P}[\theta^* \in [\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]] = \mathbb{P}_{\theta^*}[|\bar{X}_n - \theta^*| \leqslant \epsilon]$$

Confidence set

- Let consider the guiding example of $(X_1, \ldots, X_n)$ a n-sample of Bernoulli law of parameter θ^*: $\mathcal{B}(\theta^*)$
- As we have seen a good estimator is

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

- We know that

$$\bar{X}_n \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*$$

- Let us try to estimate

$$\mathbb{P}[\theta^* \in [\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]] = \mathbb{P}_{\theta^*}[|\bar{X}_n - \theta^*| \leqslant \epsilon]$$

Confidence set

- First let us check that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1] = \theta^*$$

and

$$Var_\theta\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\theta^*(1-\theta^*)}{n}\ .$$

- Then we can apply Bienaymé Chebyschev

$$\begin{aligned}\mathbb{P}[|\bar{X}_n - \theta^*| > \epsilon] &\leqslant \frac{Var(\bar{X}_n)}{\epsilon^2} && (8)\\ &= \frac{\theta^*(1-\theta^*)}{n\epsilon^2} && (9)\end{aligned}$$

Confidence set

- First let us check that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1] = \theta^*$$

and

$$Var_\theta\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\theta^*(1-\theta^*)}{n}.$$

- Then we can apply Bienaymé Chebyschev

$$\begin{aligned}\mathbb{P}[|\bar{X}_n - \theta^*| > \epsilon] &\leqslant \frac{Var(\bar{X}_n)}{\epsilon^2} &(8)\\ &= \frac{\theta^*(1-\theta^*)}{n\epsilon^2} &(9)\end{aligned}$$

Confidence set

- Now one can see that for all $x \in [0, 1]$

$$x(1-x) \leqslant \frac{1}{4}$$

then

$$\mathbb{P}[|\bar{X}_n - \theta^*| > \epsilon] \leqslant \frac{1}{4n\epsilon^2}$$

- Fixing

$$\alpha = \frac{1}{4n\epsilon^2}$$

which imposes

$$\epsilon = \frac{1}{\sqrt{4n\alpha}}$$

Confidence set

- Now one can see that for all $x \in [0, 1]$

$$x(1 - x) \leqslant \frac{1}{4}$$

then

$$\mathbb{P}[|\bar{X}_n - \theta^*| > \epsilon] \leqslant \frac{1}{4n\epsilon^2}$$

- Fixing

$$\alpha = \frac{1}{4n\epsilon^2}$$

which imposes

$$\epsilon = \frac{1}{\sqrt{4n\alpha}}$$

Confidence set

- We can then conclude that

$$\mathbb{P}[|\bar{X}_n - \theta^*| > \frac{1}{\sqrt{4n\alpha}}] \leqslant \alpha$$

which finally yields

$$\mathbb{P}[\theta^* \in [\bar{X}_n - \frac{1}{\sqrt{4n\alpha}}, \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}]] \geqslant 1 - \alpha$$

- As we can see through this approach we can adjust the parameter α to make the above probability close to 1. This parameter represents a risk.
- Often we choose $\alpha = 0,05 = 5.10^{-2}$

Confidence set

- We can then conclude that

$$\mathbb{P}[|\bar{X}_n - \theta^*| > \frac{1}{\sqrt{4n\alpha}}] \leqslant \alpha$$

which finally yields

$$\mathbb{P}[\theta^* \in [\bar{X}_n - \frac{1}{\sqrt{4n\alpha}}, \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}]] \geqslant 1 - \alpha$$

- As we can see through this approach we can adjust the parameter α to make the above probability close to 1. This parameter represents a risk.
- Often we choose $\alpha = 0,05 = 5.10^{-2}$

Confidence set

- The confidence interval is then

$$[\bar{X}_n - \frac{1}{\sqrt{4n\alpha}}, \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}]$$

- Assume you want a small interval this imposes

$$\frac{1}{\sqrt{4n\alpha}}$$

  to be small
- For example for $\alpha = 0,05$ if you want $\frac{1}{\sqrt{4n\alpha}} = 0,1$ you need $n =$
- For example for $\alpha = 0,05$ if you want $\frac{1}{\sqrt{4n\alpha}} = 0,01$ you need $n =$
- Note that since this is $\sqrt{n}$ which is involved, when you want to obtain a smallr interval (gaining a significative number you need a sample 100 times bigger).

Confidence set

- Using this approach you can see that you can need a large number n. But when n is large enough you can use the Central Limit Theorem.
- Recall that
$$\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta^*)}}\right) \xrightarrow{\mathcal{L}_{\theta^*}} \mathcal{N}(0,1)$$
- Since
$$\bar{X}_n \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*,$$
then by Slutsky we have
$$\sqrt{n}\left(\frac{\bar{X}_n - \theta^*}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right) = \frac{\sqrt{\theta(1-\theta^*)}}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta^*)}}\right) \xrightarrow{\mathcal{L}_{\theta^*}} \mathcal{N}(0,1)$$

Confidence set

- Using this approach you can see that you can need a large number n. But when n is large enough you can use the Central Limit Theorem.
- Recall that

$$\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta^*)}}\right) \xrightarrow{\mathcal{L}_{\theta^*}} \mathcal{N}(0,1)$$

- Since

$$\bar{X}_n \xrightarrow{\mathbb{P}_{\theta^*}} \theta^*,$$

then by Slutsky we have

$$\sqrt{n}\left(\frac{\bar{X}_n - \theta^*}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right) = \frac{\sqrt{\theta(1-\theta^*)}}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta^*)}}\right) \xrightarrow{\mathcal{L}_{\theta^*}} \mathcal{N}(0,1)$$

Confidence set

- Keep in mind that for n large enough we have

$$\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right) \overset{\mathcal{L}_{\theta^*}}{\simeq} \mathcal{N}(0,1)$$

- We can say that

$$\mathbb{P}_{\theta^*}\left(\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ \right] \ni \theta^*\right)$$

$$= \mathbb{P}_{\theta^*}\left(\left|\bar{X}_n - \theta^*\right| \leqslant q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right)$$

$$= \mathbb{P}_{\theta^*}\left(\left|\sqrt{n}\,\frac{\hat{\theta}^*{}_n - \theta^*}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right| \leqslant q_{1-\alpha/2}\right) \simeq \mathbb{P}[|X| \leqslant q_{1-\alpha/2}], \qquad (10)$$

where $X \sim \mathcal{N}(0,1)$.

Confidence set

- Keep in mind that for n large enough we have

$$\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right) \overset{\mathcal{L}_{\theta^*}}{\simeq} \mathcal{N}(0,1)$$

- We can say that

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ \right] \ni \theta^*\right) \\
&= \mathbb{P}_{\theta^*}\left(\left|\bar{X}_n - \theta^*\right| \leqslant q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right) \\
&= \mathbb{P}_{\theta^*}\left(\left|\sqrt{n}\,\frac{\hat{\theta}^*{}_n - \theta^*}{\sqrt{\bar{X}_n^*(1-\bar{X}_n^*)}}\right| \leqslant q_{1-\alpha/2}\right) \simeq \mathbb{P}[|X| \leqslant q_{1-\alpha/2}], \qquad (10)
\end{aligned}$$

where $X \sim \mathcal{N}(0,1)$.

Confidence set

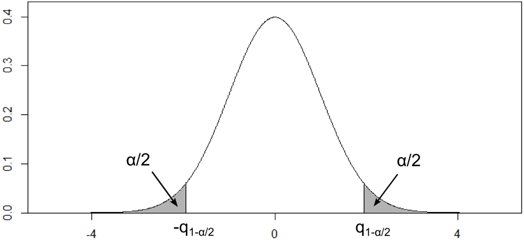
- So far we have

$$\mathbb{P}_{\theta^*}\left(\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ \right] \ni \theta^*\right) \tag{11}$$

$$\simeq \mathbb{P}[|X| \leqslant q_{1-\alpha/2}], \tag{12}$$

- Now we can say what is $q_{1-\alpha/2}$,

$$\mathbb{P}(|X| \leqslant q_{1-\alpha/2}) = 1 - (\alpha/2 + \alpha/2) = 1 - \alpha$$



Confidence set

- This way we have construct a confidence interval

$$\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ \right]$$

- For example for $\alpha = 0,05$, we get $q_{1-\alpha/2} = 1,96$. This can be read on table of the $\mathcal{N}(0,1)$ law.

Confidence set

- Can we compare the two interval that we have constructed. In fact we can show that

$$\lim_{n\to\infty} \mathbb{P}_{\theta^*}\left(\left[\bar{X}_n - \frac{1}{\sqrt{4n\alpha}}\ ;\ \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}\right] \ni \theta\right) \geqslant 1 - \exp\left(-\frac{1}{2\alpha}\right) = 1 - \circ(\alpha)$$

- Essentially this means that for large n, we have

$$\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right] \quad (13)$$

$$\subset \left[\bar{X}_n - \frac{1}{\sqrt{4n\alpha}}\ ;\ \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}\right] \quad (14)$$

  then for large n the confidence interval obtained via the CLT is better than the one obtained by Bienaymé Chebychev

- The interest of Bienaymé Tchebychev is that it is true for all n. This can give information for small sample.

Confidence set

- Can we compare the two interval that we have constructed. In fact we can show that

$$\lim_{n\to\infty} \mathbb{P}_{\theta^*}\left(\left[\bar{X}_n - \frac{1}{\sqrt{4n\alpha}} \; ; \; \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}\right] \ni \theta\right) \geqslant 1 - \exp\left(-\frac{1}{2\alpha}\right) = 1 - \circ(\alpha)$$

- Essentially this means that for large n, we have

$$\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \; ; \; \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right] \tag{13}$$

$$\subset \left[\bar{X}_n - \frac{1}{\sqrt{4n\alpha}} \; ; \; \bar{X}_n + \frac{1}{\sqrt{4n\alpha}}\right] \tag{14}$$

  then for large n the confidence interval obtained via the CLT is better than the one obtained by Bienaymé Chebychev

- The interest of Bienaymé Tchebychev is that it is true for all n. This can give information for small sample.

Confidence set

- In general for a n-sample $(X_1, \ldots, X_n)$ of a law $\mathbb{P}_{\theta^*}$ for using Bienaymé Tchebychev we need to control the variance independently of θ^*. Here for $\mathcal{B}(\theta^*)$ we have used

$$Var(\bar{X}_n) = \frac{\theta^*(1-\theta^*)}{n} \leqslant \frac{1}{4n}$$

- For Poisson random variable $\mathcal{P}(\theta^*)$ we have

$$Var(\bar{X}_n) = \frac{\theta^*}{n}$$

  and conditions on θ^* have to be known to construct a confidence interval with B-T (example you know that $\theta^* \leqslant M$ for a known value M.
- For using CLT one can use the same trick by replacing the variance in terms of $\bar{X}_n$ and justify it via Slustsky theorem.

Confidence set

- In general if we are not in such a situation, in order to use the CLT, we have to estimate the variance. To this end we have the following estimator

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- and the corresponding confidence interval is

$$\left[\bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\ ;\ \bar{X}_n + q_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right]$$

Confidence set

- Let us concentrate on this estimator

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\bar{X}_n)^2$$

- As we said it is an estimator of the variance. If you come back to the previous chapter, let us adress the usual question, bias, consistency....

Confidence set

- Let us start with the bias

$$
\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_iX_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}
$$

Confidence set

- Let us start with the bias

$$
\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_i X_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}
$$

Confidence set

- Let us start with the bias

$$
\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_iX_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}
$$

Confidence set

- Let us start with the bias

$$\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_iX_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}$$

Confidence set

- Let us start with the bias

$$
\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_iX_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}
$$

- Let us start with the bias

$$
\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\sum_{i,j}\mathbb{E}[X_i X_j] \\
&= \mathbb{E}(X_1^2) - \frac{1}{n^2}\left(\sum_{i=j}\mathbb{E}[(X_i)^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \mathbb{E}(X_1^2) - \frac{1}{n}\mathbb{E}[X_1^2] - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_1]^2 \\
&= \frac{n-1}{n}\mathbb{E}[X_1^2] - \frac{n-1}{n}\mathbb{E}[X_1]^2 = \frac{n-1}{n}Var(X_1)
\end{aligned}
$$

Confidence set

- Let us start with the bias

$$\mathbb{E}[\sigma_n^2] = \frac{n-1}{n} Var(X_1)$$

- Then considering

$$S_n^2 = \frac{n}{n-1}\sigma_n^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X}_n)^2,$$

we have an unbiased estimator.

- Let assume that $(X_1, \ldots, X_n)^t$ be a Gaussian vector of law $\mathcal{N}(m, \sigma^2)$. We have

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

- Indeed note that $Y = \frac{1}{\sigma}(X_1 - m, \ldots, X_n - m)^t \sim \mathcal{N}_n(0, I_n)$
- Define $F = Vect(1_n)$ where $1_n = (1, \ldots, 1)^t$. We easily have $dim(F) = 1$ and $dim(F^\perp) = n - 1$.
- Now note that $P_F(X) = \left\langle \frac{1_n}{\sqrt{n}}, X \right\rangle \frac{1_n}{\sqrt{n}} = \frac{1}{\sigma}(\bar{X}_n - m, \ldots \bar{X}_n - m)^t$ and then

$$P_{F^\perp}(X) = X - P_F(X) = \frac{1}{\sigma}(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)^t$$

- The Cochran Theorem then says that $||P_{F^\perp}(X)||^2 \sim \chi^2(n-1)$. Now it is easy to see that

$$||P_{F^\perp}(X)||^2 = \frac{n-1}{\sigma^2} S_n^2$$

- Let assume that $(X_1, \dots, X_n)^t$ be a Gaussian vector of law $\mathcal{N}(m, \sigma^2)$. We have
$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$
- Indeed note that $Y = \frac{1}{\sigma}(X_1 - m, \dots, X_n - m)^t \sim \mathcal{N}_n(0, I_n)$
- Define $F = Vect(1_n)$ where $1_n = (1, \dots, 1)^t$. We easily have $dim(F) = 1$ and $dim(F^\perp) = n - 1$.
- Now note that $P_F(X) = \left\langle \frac{1_n}{\sqrt{n}}, X \right\rangle \frac{1_n}{\sqrt{n}} = \frac{1}{\sigma}(\bar{X}_n - m, \dots \bar{X}_n - m)^t$ and then
$$P_{F^\perp}(X) = X - P_F(X) = \frac{1}{\sigma}(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^t$$
- The Cochran Theorem then says that $||P_{F^\perp}(X)||^2 \sim \chi^2(n-1)$. Now it is easy to see that
$$||P_{F^\perp}(X)||^2 = \frac{n-1}{\sigma^2} S_n^2$$

Confidence set

- Let assume that $(X_1, \dots, X_n)^t$ be a Gaussian vector of law $\mathcal{N}(m, \sigma^2)$. We have

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

- Indeed note that $Y = \frac{1}{\sigma}(X_1 - m, \dots, X_n - m)^t \sim \mathcal{N}_n(0, I_n)$
- Define $F = Vect(1_n)$ where $1_n = (1, \dots, 1)^t$. We easily have $dim(F) = 1$ and $dim(F^\perp) = n - 1$.
- Now note that $P_F(X) = \left\langle \frac{1_n}{\sqrt{n}}, X \right\rangle \frac{1_n}{\sqrt{n}} = \frac{1}{\sigma}(\bar{X}_n - m, \dots \bar{X}_n - m)^t$ and then

$$P_{F^\perp}(X) = X - P_F(X) = \frac{1}{\sigma}(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^t$$

- The Cochran Theorem then says that $||P_{F^\perp}(X)||^2 \sim \chi^2(n-1)$. Now it is easy to see that

$$||P_{F^\perp}(X)||^2 = \frac{n-1}{\sigma^2} S_n^2$$

Confidence set

- Let assume that $(X_1, \dots, X_n)^t$ be a Gaussian vector of law $\mathcal{N}(m, \sigma^2)$. We have

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

- Indeed note that $Y = \frac{1}{\sigma}(X_1 - m, \dots, X_n - m)^t \sim \mathcal{N}_n(0, I_n)$
- Define $F = Vect(1_n)$ where $1_n = (1, \dots, 1)^t$. We easily have $dim(F) = 1$ and $dim(F^\perp) = n-1$.
- Now note that $P_F(X) = \left\langle \frac{1_n}{\sqrt{n}}, X \right\rangle \frac{1_n}{\sqrt{n}} = \frac{1}{\sigma}(\bar{X}_n - m, \dots \bar{X}_n - m)^t$ and then

$$P_{F^\perp}(X) = X - P_F(X) = \frac{1}{\sigma}(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)^t$$

- The Cochran Theorem then says that $||P_{F^\perp}(X)||^2 \sim \chi^2(n-1)$. Now it is easy to see that

$$||P_{F^\perp}(X)||^2 = \frac{n-1}{\sigma^2} S_n^2$$

Confidence set

- Let assume that $(X_1, \ldots, X_n)^t$ be a Gaussian vector of law $\mathcal{N}(m, \sigma^2)$. We have
$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$
- Indeed note that $Y = \frac{1}{\sigma}(X_1 - m, \ldots, X_n - m)^t \sim \mathcal{N}_n(0, I_n)$
- Define $F = Vect(1_n)$ where $1_n = (1, \ldots, 1)^t$. We easily have $dim(F) = 1$ and $dim(F^\perp) = n-1$.
- Now note that $P_F(X) = \left\langle \frac{1_n}{\sqrt{n}}, X \right\rangle \frac{1_n}{\sqrt{n}} = \frac{1}{\sigma}(\bar{X}_n - m, \ldots \bar{X}_n - m)^t$ and then
$$P_{F^\perp}(X) = X - P_F(X) = \frac{1}{\sigma}(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)^t$$
- The Cochran Theorem then says that $\|P_{F^\perp}(X)\|^2 \sim \chi^2(n-1)$. Now it is easy to see that
$$\|P_{F^\perp}(X)\|^2 = \frac{n-1}{\sigma^2} S_n^2$$

Confidence set

- This allows to construct confidence interval for the variance of a Gaussian law. Let denote $\chi^k_{1-\alpha}$ the quantile of the $\chi^2(k)$ law that is if $T \sim \chi^2(k)$ then

$$\mathbb{P}[\chi^k_{\alpha/2} \leqslant T \leqslant \chi^k_{1-\alpha/2}] = 1 - \alpha$$

- Then we have

$$\mathbb{P}\left[\chi^k_{\alpha/2} \leqslant \frac{n-1}{\sigma^2} S_n^2 \leqslant \chi^{n-1}_{1-\alpha/2}\right] = 1 - \alpha$$

- This implies

$$\mathbb{P}\left[\frac{n-1}{\chi^{n-1}_{1-\alpha/2}} S_n^2 \leqslant \sigma^2 \leqslant \frac{n-1}{\chi^{n-1}_{\alpha/2}} S_n^2\right] = 1 - \alpha$$

and then the interval

$$\left[\frac{n-1}{\chi^{n-1}_{1-\alpha/2}} S_n^2, \frac{n-1}{\chi^{n-1}_{\alpha/2}} S_n^2\right]$$

is a confidence interval of level α for the variance σ^2 of X_1.

Confidence set

- Other possible interesting result when $X_1, \ldots, X_n$ are Gaussian $\mathcal{N}(m, \sigma^2)$

$$\sqrt{n}\left(\frac{\bar{X}_n - m}{\sigma}\right) \sim \mathcal{N}(0, 1)$$

  then if σ^2 is known this allows to construct a confidence interval for μ
- If σ^2 is not known replace σ by S_n and we have

$$\sqrt{n}\left(\frac{\bar{X}_n - m}{S_n}\right) \sim \mathcal{T}_{n-1}$$

  where $\mathcal{T}_{n-1}$ is a r.v distributed along a Student law of $n-1$ degree of freedom.

Confidence set

- In the above example the confidence interval are bounded but we can also consider bounds which are infinite (only one of course)

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ xhere $a(X_1,\ldots,X_n)$ and $b(X_1,\ldots,X_n)$ are statistics (independent of θ^*) satisfying

$$\mathbf{P}_\theta(\theta \in [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]) = 1-\alpha.$$

   1. if $a(X_1,\ldots,X_n) > -\infty$ and $b(X_1,\ldots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\ldots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\ldots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\ldots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\ldots,X_n)$ in a measurable way and is independent of θ satisfying

$$\mathbf{P}_\theta(\theta \in R(X_1,\ldots,X_n)) = 1-\alpha.$$

Confidence set

- In the above example the confidence interval are bounded but we can also consider bounds which are infinite (only one of course)

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ xhere $a(X_1,\ldots,X_n)$ and $b(X_1,\ldots,X_n)$ are statistics (independent of θ^*) satisfying

$$\mathbf{P}_\theta(\theta \in [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]) = 1-\alpha.$$

   1. if $a(X_1,\ldots,X_n) > -\infty$ and $b(X_1,\ldots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\ldots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\ldots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\ldots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\ldots,X_n)$ in a measurable way and is independent of θ satisfying

$$\mathbf{P}_\theta(\theta \in R(X_1,\ldots,X_n)) = 1-\alpha.$$

Confidence set

- In the above example the confidence interval are bounded but we can also consider bounds which are infinite (only one of course)

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ xhere $a(X_1,\ldots,X_n)$ and $b(X_1,\ldots,X_n)$ are statistics (independent of θ^*) satisfying

$$\mathbf{P}_\theta(\theta \in [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]) = 1-\alpha.$$

   1. if $a(X_1,\ldots,X_n) > -\infty$ and $b(X_1,\ldots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\ldots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\ldots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\ldots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\ldots,X_n)$ in a measurable way and is independent of θ satisfying

$$\mathbf{P}_\theta(\theta \in R(X_1,\ldots,X_n)) = 1-\alpha.$$

Confidence set

- In the above example the confidence interval are bounded but we can also consider bounds which are infinite (only one of course)

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ xhere $a(X_1,\ldots,X_n)$ and $b(X_1,\ldots,X_n)$ are statistics (independent of θ^*) satisfying

$$\mathbf{P}_\theta\left(\theta \in [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]\right) = 1-\alpha.$$

   1. if $a(X_1,\ldots,X_n) > -\infty$ and $b(X_1,\ldots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\ldots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\ldots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\ldots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\ldots,X_n)$ in a measurable way and is independent of θ satisfying

$$\mathbf{P}_\theta\left(\theta \in R(X_1,\ldots,X_n)\right) = 1-\alpha.$$

Confidence set

- We can relax the previous definition by allowing $\geqslant$ instead of $=$

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\dots,X_n), b(X_1,\dots,X_n)]$ xhere $a(X_1,\dots,X_n)$ and $b(X_1,\dots,X_n)$ are statistics (independent of θ^*) satisfying

$$\mathbf{P}_\theta\left(\theta \in [a(X_1,\dots,X_n), b(X_1,\dots,X_n)]\right) \geqslant 1-\alpha.$$

   1. if $a(X_1,\dots,X_n) > -\infty$ and $b(X_1,\dots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\dots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\dots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\dots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\dots,X_n)$ in a measurable way and is independent of θ satisfying

$$\mathbf{P}_\theta\left(\theta \in R(X_1,\dots,X_n)\right) \geqslant 1-\alpha.$$

Confidence set

- We can also have asymptotic confidence set

Definition

Let $\alpha \in [0,1]$ fixé and let $\theta^* \in \mathbb{R}^k$

1. When $k = 1$, we call confidence interval of level $1-\alpha$ for θ^* all random interval I of the form $[a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]$ xhere $a(X_1,\ldots,X_n)$ and $b(X_1,\ldots,X_n)$ are statistics (independent of θ^*) satisfying

$$\lim_n \mathbf{P}_\theta\left(\theta \in [a(X_1,\ldots,X_n), b(X_1,\ldots,X_n)]\right) = 1-\alpha.$$

   1. if $a(X_1,\ldots,X_n) > -\infty$ and $b(X_1,\ldots,X_n) < \infty$ we speak about bilateral interval
   2. if $a(X_1,\ldots,X_n) = -\infty$ we speak about left unilateral interval
   3. if $b(X_1,\ldots,X_n) = \infty$ we speak about right unilateral interval

2. When $k > 1$ we speak about confidence set of level $1-\alpha$ for θ all random subset $R(X_1,\ldots,X_n)$ of $\mathbf{R}^k$ which depends on $(X_1,\ldots,X_n)$ in a measurable way and is independent of θ satisfying

$$\lim_n \mathbf{P}_\theta\left(\theta \in R(X_1,\ldots,X_n)\right) = 1-\alpha.$$

Confidence set

- One can also use open set for confidence set
- In general there is an infinity of confidence interval. For example with the CLT we can choose

$$\left]-\infty, \bar{X}_n - q_{1-\alpha}\sqrt{\frac{\sigma_n^2}{n}}\right]$$

- Can it be interested to have a interval bound which is infinite? It looks like not sharp!
- Imagine that you known that the unknow quantity is non negative (decibel of a night club, number of student attending the summer school in France); then the part $]-\infty, 0]$ is useless and the interval

$$\left]0, \bar{X}_n - q_{1-\alpha}\sqrt{\frac{\sigma_n^2}{n}}\right] \subset \left]0, \bar{X}_n - q_{1-\alpha/2}\sqrt{\frac{\sigma_n^2}{n}}\right]$$

which makes the interval $\left]0, \bar{X}_n - q_{1-\alpha}\sqrt{\frac{\sigma_n^2}{n}}\right]$ more relevant.

Basic of Regression

Basic of Regression

- First let us start with a simple situation. Let Y be a L^2 r.v. You want to approximate Y by a constant a by minimizing the quadratic error that is you want to find

$$argmin_{a\in\mathbb{R}}\mathbb{E}[(Y-a)^2]$$

- In fact it is easy to check that

$$\min_{a\in\mathbb{R}}\mathbb{E}[(Y-a)^2]$$

is reached for $a=\mathbb{E}[Y]$.

- Indeed one can think in terms of projection of Y onto the subspace of constant function.
- If you do not have the possibility to consider the L^2 norma, one could have thought

$$argmin_{a\in\mathbb{R}}\mathbb{E}[|Y-a|]$$

and you would have founded the median

Basic of Regression

- First let us start with a simple situation. Let Y be a L^2 r.v. You want to approximate Y by a constant a by minimizing the quadratic error that is you want to find

$$argmin_{a \in \mathbb{R}} \mathbb{E}[(Y - a)^2]$$

- In fact it is easy to check that

$$\min_{a \in \mathbb{R}} \mathbb{E}[(Y - a)^2]$$

  is reached for $a = \mathbb{E}[Y]$.
- Indeed one can think in terms of projection of Y onto the subspace of constant function.
- If you do not have the possibility to consider the L^2 norma, one could have thought

$$argmin_{a \in \mathbb{R}} \mathbb{E}[|Y - a|]$$

  and you would have founded the median

Basic of Regression

- First let us start with a simple situation. Let Y be a L^2 r.v. You want to approximate Y by a constant a by minimizing the quadratic error that is you want to find
$$argmin_{a\in\mathbb{R}}\mathbb{E}[(Y-a)^2]$$
- In fact it is easy to check that
$$\min_{a\in\mathbb{R}}\mathbb{E}[(Y-a)^2]$$
is reached for $a=\mathbb{E}[Y]$.
- Indeed one can think in terms of projection of Y onto the subspace of constant function.
- If you do not have the possibility to consider the L^2 norma, one could have thought
$$argmin_{a\in\mathbb{R}}\mathbb{E}[|Y-a|]$$
and you would have founded the median

- Now imagine you have a couple (X, Y) whose you know the joint distribution. Suppose that X and Y are L^2.
- Consider the situation where you only observe a realization of X let say x. You want to estimate Y knowing this realization. Without further information it is not possible since Y knowing x is random.
- An idea is to approximate Y as an affine function of X, i.e $Y = aX + b$ and you to minimise

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2]$$

- Here, you see that, you need to find the orthogonal projection onto the subspace of affine function of X. Computations give

$$a = \frac{Cov(X, Y)}{\sigma^2(X)}, \quad b = \mathbb{E}[Y] - \frac{Cov(X, Y)}{\sigma^2(X)}\mathbb{E}[X]$$

Basic of Regression

- Now imagine you have a couple (X, Y) whose you know the joint distribution. Suppose that X and Y are L^2.
- Consider the situation where you only observe a realization of X let say x. You want to estimate Y knowing this realization. Without further information it is not possible since Y knowing x is random.
- An idea is to approximate Y as an affine function of X, i.e $Y = aX + b$ and you to minimise

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2]$$

- Here, you see that, you need to find the orthogonal projection onto the subspace of affine function of X. Computations give

$$a = \frac{Cov(X, Y)}{\sigma^2(X)}, \quad b = \mathbb{E}[Y] - \frac{Cov(X, Y)}{\sigma^2(X)}\mathbb{E}[X]$$

Basic of Regression

- Now imagine you have a couple (X, Y) whose you know the joint distribution. Suppose that X and Y are L^2.
- Consider the situation where you only observe a realization of X let say x. You want to estimate Y knowing this realization. Without further information it is not possible since Y knowing x is random.
- An idea is to approximate Y as an affine function of X, i.e $Y = aX + b$ and you to minimise

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2]$$

- Here, you see that, you need to find the orthogonal projection onto the subspace of affine function of X. Computations give

$$a = \frac{Cov(X, Y)}{\sigma^2(X)}, \quad b = \mathbb{E}[Y] - \frac{Cov(X, Y)}{\sigma^2(X)}\mathbb{E}[X]$$

- At this stage let us introduce the so called correlation coefficient

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}, \quad |\rho| \leqslant 1$$

- Note that X and Y independent implies $\rho = 0$
- In terms of ρ one can check

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2] = \sigma^2(Y)(1 - \rho^2)$$

- The error is small when $|\rho|$ is close to 1
- When $\rho = 0$ the error is maximum. In this case the best approximation is $\mathbb{E}[Y]$

- At this stage let us introduce the so called correlation coefficient

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}, \quad |\rho| \leqslant 1$$

- Note that X and Y independent implies $\rho = 0$
- In terms of ρ one can check

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2] = \sigma^2(Y)(1 - \rho^2)$$

- The error is small when $|\rho|$ is close to 1
- When $\rho = 0$ the error is maximum. In this case the best approximation is $\mathbb{E}[Y]$

- At this stage let us introduce the so called correlation coefficient

$$\rho = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}, \quad |\rho| \leqslant 1$$

- Note that X and Y independent implies $\rho = 0$
- In terms of ρ one can check

$$\min_{a,b} \mathbb{E}[(Y - aX + b)^2] = \sigma^2(Y)(1 - \rho^2)$$

- The error is small when $|\rho|$ is close to 1
- When $\rho = 0$ the error is maximum. In this case the best approximation is $\mathbb{E}[Y]$

- In statistics, i.e in the true life we do not know the law of the couple (X, Y). We have n realizations $((X_1, Y_1), \dots, (X_n, Y_n))$ and you want to minimize

$$\min_{a,b} \sum_{i=1}^{n} (Y_i - (aX_i + b))^2$$

- In terms of realizations, in concrete terms you want to minimize

$$\min_{a,b} \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

- Concretely, you replace

$$a = \frac{Cov(X, Y)}{\sigma^2(X)}, \quad b = \mathbb{E}[Y] - \frac{Cov(X, Y)}{\sigma^2(X)}\mathbb{E}[X]$$

by their empirical versions (variance, covariance, expectation...)

Basic of Regression

- More generally you can ask to approximate Y as a function $u(X)$ and then minimize
$$\min_u \mathbb{E}[(Y - u(X))^2]$$
- As we already seen this quantity is obtained by using the conditional expectation that is
$$\mathbb{E}[Y|X]$$
- The curve
$$x \to \mathbb{E}[Y|X = x]$$
is called the regression curve (regression function).

- More generally you can ask to approximate Y as a function $u(X)$ and then minimize
$$\min_u \mathbb{E}[(Y - u(X))^2]$$
- As we already seen this quantity is obtained by using the conditional expectation that is
$$\mathbb{E}[Y|X]$$
- The curve
$$x \to \mathbb{E}[Y|X = x]$$
is called the regression curve (regression function).

Basic of Regression

- Example of a couple (X, Y) with density

$$f(x,y) = 2e^{-(x+y)}\mathbf{1}_{0\leqslant x\leqslant y}$$

- The conditional expectation is then $f_{Y|X=x} = f_{x,y}(x,y)/f_X(x)$ where

$$f_X(x) = 2e^{-2x}\mathbf{1}_{0\leqslant x}, \quad (exponential \quad law)$$

- We then have

$$f_{Y|X=x}(y) = e^{x-y}\mathbf{1}_{0\leqslant x\leqslant y}$$

- We can then compute

$$\mathbb{E}[Y|X=x] = \int y f_{Y|X=x}(y)dy = \int_x^{+\infty} e^x y e^{-y} dy = x+1$$

- Come back to the Gaussian case
- Let (X, Y) be a Gaussian vector, one can check

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X])$$

Theorem

In the Gaussian world the regression curve and the regression line are the same!

- $\mathbb{E}[Y|X]$ is supposed to be the orthogonal projection of Y onto

$$L^2(X) = \{f(X), \mathbb{E}[f(X)^2] < \infty\}$$

but here it reduces to the orthogonal projection onto

$$Vect\{1, X\}$$

Basic of Regression

- Come back to the Gaussian case
- Let (X, Y) be a Gaussian vector, one can check

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X])$$

Theorem

In the Gaussian world the regression curve and the regression line are the same!

- $\mathbb{E}[Y|X]$ is supposed to be the orthogonal projection of Y onto

$$L^2(X) = \{f(X), \mathbb{E}[f(X)^2] < \infty\}$$

but here it reduces to the orthogonal projection onto

$$Vect\{1, X\}$$

Basic of Regression

- Come back to the Gaussian case
- Let (X, Y) be a Gaussian vector, one can check

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X,Y)}{Var(X)}(X - \mathbb{E}[X])$$

Theorem

In the Gaussian world the regression curve and the regression line are the same!

- $\mathbb{E}[Y|X]$ is supposed to be the orthogonal projection of Y onto

$$L^2(X) = \{f(X), \mathbb{E}[f(X)^2] < \infty\}$$

but here it reduces to the orthogonal projection onto

$$Vect\{1, X\}$$

- On $Vect\{1, X\}$ one can check that

$$1, \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

is an orthonormal basis.

- One can then check

$$\mathbb{E}[Y|X] = \langle 1, Y\rangle 1 + \left\langle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}, Y \right\rangle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

which is exactly another way of writting

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X])$$

- On $Vect\{1, X\}$ one can check that

$$1, \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

is an orthonormal basis.

- One can then check

$$\mathbb{E}[Y|X] = \langle 1, Y\rangle 1 + \left\langle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}, Y \right\rangle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

which is exactly another way of writting

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X])$$

- On $Vect\{1, X\}$ one can check that

$$1, \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

is an orthonormal basis.

- One can then check

$$\mathbb{E}[Y|X] = \langle 1, Y\rangle 1 + \left\langle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}, Y \right\rangle \frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}}$$

which is exactly another way of writting

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X])$$

Regression Hyperplan

- Let $X = (X_1, \ldots, X_n)$ be a random vector, we aim to approximate Y by a hyperlan which minimizes

$$\min_{a_1,\ldots,a_n,b} \mathbb{E}\left[\left(Y - \left(b + \sum_{i=1}^{n} a_i X_i\right)\right)^2\right]$$

- We suppose that the dispersion matrix

$$\Gamma_X = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^t]$$

- The regression hyperplan is given by

$$\pi_H(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X]),$$

where $\Gamma_{Y,X} = \mathbb{E}[(Y - \mathbb{E}(Y))(X - \mathbb{E}(X))]$ is the covariance line matrix $(Cov(Y, X_1), \ldots, Cov(Y, X_n))$

Regression Hyperplan

- Let $X = (X_1, \ldots, X_n)$ be a random vector, we aim to approximate Y by a hyperlan which minimizes

$$\min_{a_1,\ldots,a_n,b} \mathbb{E}\left[\left(Y - \left(b + \sum_{i=1}^{n} a_i X_i\right)\right)^2\right]$$

- We suppose that the dispersion matrix

$$\Gamma_X = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^t]$$

- The regression hyperplan is given by

$$\pi_H(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X]),$$

where $\Gamma_{Y,X} = \mathbb{E}[(Y - \mathbb{E}(Y))(X - \mathbb{E}(X))]$ is the covariance line matrix $(Cov(Y, X_1), \ldots, Cov(Y, X_n))$

Regression Hyperplan

- Let $X = (X_1, \ldots, X_n)$ be a random vector, we aim to approximate Y by a hyperlan which minimizes

$$\min_{a_1,\ldots,a_n,b} \mathbb{E}\left[\left(Y - \left(b + \sum_{i=1}^{n} a_i X_i\right)^2\right]$$

- We suppose that the dispersion matrix

$$\Gamma_X = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^t]$$

- The regression hyperplan is given by

$$\pi_H(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X]),$$

where $\Gamma_{Y,X} = \mathbb{E}[(Y - \mathbb{E}(Y))(X - \mathbb{E}(X))]$ is the covariance line matrix $(Cov(Y, X_1), \ldots, Cov(Y, X_n))$

Regression Hyperplan

- We can also compute the quadratic error

$$\mathbb{E}[(Y - \pi_H(Y))^2] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

- Gaussian situation

Theorem

In the Gaussian world if $(X_1, \ldots, X_n, Y)$ is a Gaussian vector, we have

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

then the Hyperplan of regression is equal to the conditional expectation.

Regression Hyperplan

- We can also compute the quadratic error

$$\mathbb{E}[(Y - \pi_H(Y))^2] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

- Gaussian situation

Theorem

In the Gaussian world if $(X_1, \ldots, X_n, Y)$ is a Gaussian vector, we have

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

then the Hyperplan of regression is equal to the conditional expectation.

Principal Component Analysis: Overview

PCA

- Will be developed in details in the 3rd week
- Assume you have access to p datas (age, sex, color of hair, rate of alcohol in the blood ...) of n people
- The parameter p can be huge and unless for $p \leqslant 3$ it is not possible to represent these datas on a graph
- We want to determine $q < p$ variables which explains the phenomena, we study, and which can be represented in a graph $(q = 2, 3)$

- The datas are grouped in a matrix X of size $n \times p$

$$X = \left(X^1, \ldots\ldots\ldots\ldots, X^p \right) \tag{15}$$

$$X = \begin{pmatrix} X_{1,1} & \ldots & \ldots & X_{1,p} \\ \vdots & \ldots & \ldots & \vdots \\ X_{i,1} & \ldots & \ldots & X_{i,p} \\ \vdots & \ldots & \ldots & \vdots \\ X_{n,1} & \ldots & \ldots & X_{n,p} \end{pmatrix} = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{pmatrix} \tag{16}$$

- Introduce $\bar{X} = \begin{pmatrix} \bar{X}^1 & \ldots & \ldots & \bar{X}^p \end{pmatrix}$, where $\bar{X}^k$ is the mean of the variable X^k. Denote $s_k^2 = Var(X^k) = \frac{1}{n} \sum_{i=1}^{n} (X_{ik} - \bar{X}^k)^2$ the corresponding variance.
- The number of people belongs to $\mathbb{R}^n$ and the variables to $\mathbb{R}^p$ where the average is made by column

PCA

- The centered version

$$Y = \begin{pmatrix} X_{1,1} - \bar{X}^1 & \dots & \dots & X_{1,p} - \bar{X}^p \\ \vdots & \dots & \dots & \vdots \\ X_{j,1} - \bar{X}^1 & \dots & \dots & X_{j,p} - \bar{X}^p \\ \vdots & \dots & \dots & \vdots \\ X_{n,1} - \bar{X}^1 & \dots & \dots & X_{n,p} - \bar{X}^p \end{pmatrix} \quad (17)$$

- The centered and reduced version

$$Z = \begin{pmatrix} \frac{X_{1,1}-\bar{X}^1}{s_1} & \dots & \dots & \frac{X_{1,p}-\bar{X}^p}{s_p} \\ \vdots & \dots & \dots & \vdots \\ \frac{X_{j,1}-\bar{X}^1}{s_1} & \dots & \dots & \frac{X_{j,p}-\bar{X}^p}{s_p} \\ \vdots & \dots & \dots & \vdots \\ \frac{X_{n,1}-\bar{X}^p}{s_1} & \dots & \dots & \frac{X_{n,p}-\bar{X}^p}{s_p} \end{pmatrix}, \quad Var(Z^j) = 1, j = 1, \dots, p \quad (18)$$

PCA

- Let us speak about the distance between two people. To this end consider a symmetric definite positive matrix M of size $p \times p$ and denote
$$\langle x, y\rangle_M = \langle x, My\rangle = x^t My$$
and $||x||_M = \sqrt{\langle x, x\rangle_M}$ as well as
$$d_M(x, y) = ||x - y||_M$$
- Often we consider matrix M of diagonal form $M = diag(m_i)$ and in this case
$$\langle x, y\rangle_M = \sum_{i=1}^{p} m_i x_i y_i$$
$$d_M^2(x, y) = \sum_{i=1}^{p} m_i (x_i - y_i)^2$$

PCA

- Let us speak about the distance between two people. To this end consider a symmetric definite positive matrix M of size $p \times p$ and denote

$$\langle x, y \rangle_M = \langle x, My \rangle = x^t My$$

and $||x||_M = \sqrt{\langle x, x \rangle_M}$ as well as

$$d_M(x, y) = ||x - y||_M$$

- Often we consider matrix M of diagonal form $M = diag(m_i)$ and in this case

$$\langle x, y \rangle_M = \sum_{i=1}^{p} m_i x_i y_i$$

$$d_M^2(x, y) = \sum_{i=1}^{p} m_i (x_i - y_i)^2$$

PCA

- Let us make the link between the matrix X, Y, Z and the above distance. Let us consider a diagonal matrix $M = diag(m_i)$

$$\|X_i\|_M^2 = \sum_{k=1}^{p} m_k X_{ik}^2, \quad d_M^2(X_i, X_j) = \sum_{k=1}^{p} m_k (X_{i,k} - X_{j,k})^2$$

- In the case where $M = I_p$ we have

$$d_{I_p}^2(X_i, X_j) = \sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2 = d_{I_p}^2(Y_i, Y_j)$$

- In the case where $M = diag(1/s_1^2, \ldots, 1/s_p^2)$ we have

$$d_M^2(X_i, X_j) = d_{I_p}^2(Z_i, Z_j)$$

PCA

- Let us make the link between the matrix X, Y, Z and the above distance. Let us consider a diagonal matrix $M = diag(m_i)$

$$\|X_i\|_M^2 = \sum_{k=1}^{p} m_k X_{ik}^2, \quad d_M^2(X_i, X_j) = \sum_{k=1}^{p} m_k (X_{i,k} - X_{j,k})^2$$

- In the case where $M = I_p$ we have

$$d_{I_p}^2(X_i, X_j) = \sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2 = d_{I_p}^2(Y_i, Y_j)$$

- In the case where $M = diag(1/s_1^2, \ldots, 1/s_p^2)$ we have

$$d_M^2(X_i, X_j) = d_{I_p}^2(Z_i, Z_j)$$

- Let us make the link between the matrix X, Y, Z and the above distance. Let us consider a diagonal matrix $M = diag(m_i)$

$$\|X_i\|_M^2 = \sum_{k=1}^{p} m_k X_{ik}^2, \quad d_M^2(X_i, X_j) = \sum_{k=1}^{p} m_k (X_{i,k} - X_{j,k})^2$$

- In the case where $M = I_p$ we have

$$d_{I_p}^2(X_i, X_j) = \sum_{k=1}^{p} (X_{i,k} - X_{j,k})^2 = d_{I_p}^2(Y_i, Y_j)$$

- In the case where $M = diag(1/s_1^2, \ldots, 1/s_p^2)$ we have

$$d_M^2(X_i, X_j) = d_{I_p}^2(Z_i, Z_j)$$

- Now let us define the notion of inertia. Introducing the diagonal matrix $M = diag(m_i)$ allows to consider weight. We define the inertia as

$$I(X) = \sum_{k=1}^{p} m_i d^2(X_i, \bar{X}) = \sum_{k=1}^{p} m_i s_j^2$$

  It measures the dispersion of the data X_i with respect to the barycenter $\bar{X}$.
- In the case $M = diag(1/s_1^2, \dots, 1/s_p^2)$ we have

$$I(Z) = p$$

- The p column of X represent a so-called scatter graph.
- Regarding the weight introduced before we shall concentrate on $m_j = 1$ in the context of PCA.
- If we analyze Y we shall say we do non-normalized PCA
- If we analyze Z we do normed PCA and we are going to focus on this case

PCA

- In PCA you can have two points of view
  - Either you analyze the n point people and you will choose the metric with $M = I_p$
  - Or you analyze the p datas and you will choose the metric given by $N = \frac{1}{n} I_n$
- We already have seen the effect of $M = I_p$ on the line of the matrix
- The effect of the matrix N is on the column. Note that

$$Var(X^j) = \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}^i)^2 = \|Y^j\|_N^2$$

$$Var(Z^j) = \|Y^j\|_N^2 = 1$$

- In PCA you can have two points of view
  - Either you analyze the n point people and you will choose the metric with $M = I_p$
  - Or you analyze the p datas and you will choose the metric given by $N = \frac{1}{n} I_n$
- We already have seen the effect of $M = I_p$ on the line of the matrix
- The effect of the matrix N is on the column. Note that

$$Var(X^j) = \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}^i)^2 = \|Y^j\|_N^2$$

$$Var(Z^j) = \|Y^j\|_N^2 = 1$$

PCA

- The covariance between X_j and $X_{j'}$ is given by

$$c_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}^j)(X_{i,j'} - \bar{X}^{j'}) = \langle Y^i, Y^j \rangle_N$$

- In particular one can easily see that the covariance matrix

$$C = Y^t N Y$$

- The correlation between X_j and $X_{j'}$ is given by

$$r_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(\frac{X_{i,j} - \bar{X}^j}{s_j})(\frac{X_{i,j'} - \bar{X}^{j'}}{s_{j'}}) = \langle Y^i, Y^j \rangle_N$$

- In particular one can easily see that the correlation matrix

$$R = Z^t N Z$$

PCA

- The covariance between X_j and $X_{j'}$ is given by

$$c_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}^j)(X_{i,j'} - \bar{X}^{j'}) = \langle Y^i, Y^j \rangle_N$$

- In particular one can easily see that the covariance matrix

$$C = Y^t N Y$$

- The correlation between X_j and $X_{j'}$ is given by

$$r_{jj'} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{i,j} - \bar{X}^j}{s_j}\right)\left(\frac{X_{i,j'} - \bar{X}^{j'}}{s_{j'}}\right) = \langle Y^i, Y^j \rangle_N$$

- In particular one can easily see that the correlation matrix

$$R = Z^t N Z$$

PCA

- Let us start by concentrating on the people
- For example an reasonable objective is to find the projection plan such that the distance between the people are the better conserved.
- Let us speak about the projection of a guy. We are in the case $M = I_p$ and we want to project $Z_j \in \mathbb{R}^p$ for example on an axis defined by Δ_α which is directed by a vector v_α of norm 1. The coordinate will be given by

$$f_{j\alpha} = \langle Z_j, v_\alpha \rangle = Z_j^t v_\alpha$$

- Define now

$$f^\alpha := (f_{1\alpha}, \ldots, f_{n\alpha})^t = Zv_\alpha$$

  this the vector of each coordinate of each projection of the Z_j
- We can rewrite

$$f^\alpha = Zv_\alpha = \sum_{j=1}^{p} v_{j\alpha} Z^j$$

PCA

- Method: we are looking for an axis Δ_1 with generator v_1 such that

$$v_1 = argmax_{v_1/||v_1||=1} Var(Zv_1)$$

- We can show that this optimization problem can be written as

$$\max_{v/||v||=1} ||Rv||^2$$

  with $R = \frac{1}{n} Z^t Z$

- Then this maximum is reached for v_1 the eigenvector associated to the maximum eigenvalue of R

PCA

- Then $f_1 = Zv_1$ is the first principal coomponent
- If we want to find a plan we look for v_2 such that

$$v_2 = argmax_{v_2/v_2 \perp v_1 \|v_2\|=1} Var(Zv_2)$$

- v_2 appears as the second eigenvector corresponding to the second higher eigenvalue. The vector $f_2 = Zv_2$ is the second principal component
- and so on
- Note that f_1 and f_2 are orthogonal and then non correlated.
- Conclusion: to find the principal component we need to diagonalize R.

PCA

- If you denote $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r$ the eigenvalues of R (here r corresponds to the rank of Z), we can show easily that

$$Var(f_i) = \lambda_i$$

- An important question is how many component shall we need. This can be quantified by looking at the quantity

$$\frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_r} = \frac{\lambda_1 + \ldots + \lambda_q}{Tr(R)}$$

- You can fix a level $1 - \alpha$ and you stop to the first time (first q) where

$$\frac{\lambda_1 + \ldots + \lambda_q}{Tr(R)} \geqslant 1 - \alpha$$

- In practice to find the first eigenvector v_1 and the first eigenvalue λ_1 you can use the power method. Define

$$w_{n+1} = \frac{Rw_n}{||Rw_n||}$$

- We have

$$||Rw_n|| \to_n \lambda_1$$

and

$$w_n \to v_1$$

- In order to find the second eigenvector and the second eigenvalue you do the same job on the orthogonal $vect v_1{}^{\perp}$

- In practice to find the first eigenvector v_1 and the first eigenvalue λ_1 you can use the power method. Define

$$w_{n+1} = \frac{Rw_n}{||Rw_n||}$$

- We have

$$||Rw_n|| \to_n \lambda_1$$

and

$$w_n \to v_1$$

- In order to find the second eigenvector and the second eigenvalue you do the same job on the orthogonal $vect v_1{}^{\perp}$

- You can also take the problem from the the p variable size by considering Z^t instead of Z and do the same job.

- Moment method for $\mathcal{N}(\mu, \sigma^2)$
- MLE for $\mathcal{U}([0, \theta])$. Consistency? Confidence set ?
- Consider the density

$$f_\theta(x) = \frac{|x - \theta|}{2} e^{-|x-\theta|} ,$$

Moment method ? Two type of confidence interval ?