Data analysis : Lesson 1

Mounir Atiq

atiq.mounir@gmail.com

September 2, 2024

What does data analysis stand for ?

To analyse data means to extract information from these data.
It alludes to various methods and tools to prepare, transform, examine data for extracting useful information, as well as the different ways to conclude a make decisions based on obtained information.

- **Prepare** : cleaning, pre-processing data, providing context about data gathering.
- **Transform** : applying operations on data to facilitate interpretation, reducing data dimension while keeping the main part of contained information, building models to summarize data based on assumptions.
- **Examine** : employing tools to help data visualisation (like projecting in 2d subspaces), pointing out patterns and trends, computing metrics about data variables and how they are related, testing hypothesis, detecting clusters.
- **Conclude** : answering questions that motivated data analysis and sometime deciding actions to undertake based on these conclusions.

What does data analysis stand for ?

So data analysis is also driven by the kind of targeted conclusions that can be :

- **Descriptive** : summarize data, presents an overview of their particularities, determine reasonable assumptions.
- **Explanatory** : explain how or why certain observations can be made on data, try to find underlying phenomenon producing data.
- **Predictive** : use extracted information to guess hidden variables (like noise removal) or to forecast future outcomes.
- **Recommendations** : use previous types of analysis to recommend concrete actions (like production launch), especially important in industry as it is often the main desired output.

Course outline and evaluation

1. Multivariate statistics
   - Multivariate Gaussian distributions
   - Parametric estimation
   - Hypothesis testing
2. Factorial analysis methods
   - Principal component analysis
   - Factorial correspondence analysis
   - Multiple correspondence analysis
3. Clustering
   - Hierarchical clustering
   - K-means

Evaluation = (Group project + Practical Sessions)*60% + Exam*40%

Probability space

Definition

We call $(\Omega, \mathbb{P}, \mathscr{F})$ a **probability space**

- Ω : sample space (all possible outcomes)
- $\mathscr{F} \subset \mathscr{P}(\Omega)$ and $\omega \in \mathscr{F}$ is called an event
- $\mathbb{P}$ an application from $\mathscr{F}$ to $[0,1]$

Such that :

- $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$
- σ-**additivity** : For any finite or countable series $(A_i)_{i\in I} \in \mathscr{F}$ of **disjoint** events,

$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$$

$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B \cup A_i)$

Considering an increasing sequence $(A_n)_{n\in\mathbb{N}}$:

$\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \lim_{n\to+\infty} \mathbb{P}(A_n)$

Conditional probability and independence of events

Conditional probability

Let A, B be two **events**, the probability of "A knowing B" is :
$\mathbb{P}(A/B) \overset{\text{def}}{=} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

We can easily see that, with a fixed B, the new application $\mathbb{P}_B : A \mapsto \mathbb{P}(A/B)$ verifies probability measure properties.
It is called **conditional probability relatively to** B and is denoted.

Independence

A, B are said to be **independent** iif : $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
$\iff \mathbb{P}(A/B) = \mathbb{P}(A)$

Random variables

Definition

A real (resp. complex) **random variable** X is a measurable application
$(\Omega, \mathscr{F}) :\longrightarrow (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ (resp. $(\mathbb{C}, \mathscr{B}(\mathbb{C}))$).
Then, $X^{-1}(v)$ is an event and $\mathbb{P}(X(\omega) \in v) = \mathbb{P}(X^{-1}(v))$

It can be discrete or continuous or mixed.

Definition

A random variable X is a "variable with density" if there exists a positive measurable application f_X that satisfies:

$$\forall A \in \mathscr{B}(\mathbb{R}), \mathbb{P}(A) = \mathbb{P}(X \in A) = \int_A f_X(x)\,dx$$

Change of variable

X a random variable such that : $\int_{B_X} f_X(u)\,du = 1$ and $Z = h(X)$ with h a diffeomorphism of $\mathbb{R}^d$ from B_X to B_z.

Then,

$$z \in B_z, \quad f_Z(z) = \frac{1}{|h'(h^{-1}(z))|} f_X\left(h^{-1}(z)\right)$$

In particular, if $Z = aX + b$ (with $a \neq 0$), $\quad f_Z(z) = \frac{1}{|a|} f_X\left(\frac{z-b}{a}\right)$

Conditional distribution and independence of random variables

Conditional probability

Let X, Y be two real random variables, the probability of "$X \in A$ knowing $Y \in B$" is :

- $\mathbb{P}(X \in A / Y \in B) \overset{\text{def}}{=} \frac{\mathbb{P}(\{X \in A\} \cap \{Y \in B\})}{\mathbb{P}(Y \in B)}$
- With $f_{X,Y}$ the joint density : $f_{X/Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ and $\forall A, B, \mathbb{P}(X \in A / Y \in B) = \int\limits_{(x,y) \in A \times B} f_{X/Y}(x,y)\,dxdy$

We can easily see that, with a fixed B, the new application $\mathbb{P}_B : A \longmapsto \mathbb{P}(A/B)$ verifies probability measure properties.
It is called **conditional probability relatively to** B and is denoted.

Independence

X, Y are said to be **independent** iif : $f_{X,Y} \overset{a.s.}{=} f_X f_Y$.
$\Longleftrightarrow f_{X/Y} = f_X$

Expected value

Definition

Expected value :

- For discrete variables : $\mathbb{E}[X] = \sum_{i \in I} X(A_i)\mathbb{P}(A_i) = \sum_{i \in I} X_i P_i$
- For density variables : $\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,dx$

Not all random variables have an expected value.
$\mathbb{P}(A) = \mathbb{E}[1_A(X)]$

Definition

Conditional expected value : for a fixed $B \subset \mathbb{R}$,

- $f : x \mapsto \int_B f_{X/Y}(x, y)\,dy$ is a probability measure
- $\mathbb{E}[X/Y \in B] = \int_{\mathbb{R}} f(x)\,dx$

Characteristic function

Definition

X a real random variable, the associated **characteristic function** is defined by :

$$\begin{array}{rccc} \phi_X & : & \mathbb{R} & \to & \mathbb{C} \\ & & t & \mapsto & \mathbb{E}\left[e^{itX}\right] \end{array}$$

It always exists and characterizes completely the distribution of X.

Proposition

- ϕ is continuous
- $|\phi(t)| \leq 1$
- $\phi(0) = 1$
- $\phi(-t) = \overline{\phi(t)}$

Variance, covariance and correlation

Definition

- **Variance** : $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
- **Moments** : $M^{(k)}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^k\right]$
- **Covariance** : $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$ X and Y are "uncorrelated" $\overset{\text{def}}{\iff} \mathrm{Cov}(X,Y) = 0$

As for the expected value, these quantities do not necessarily exist.

Exercise 1

Show that independence implies non-correlation.
What is the general expression of $\mathrm{Var}(X+Y)$?
So, what if X and Y are independent ?

Exercise 2

Consider X, Y such that $\mathrm{Cov}(X,Y) \neq 0$,
Create a new variable Z using X, Y such that X and Z are uncorrelated.

Link with geometry

As a functional space, random variables constitute a vector space. We would like to provide it with an inner product and a norm.
Intuitively, if X is the zero vector $\mathbb{P}(X=0)=1$. But there are numerous variables verifying this statement.
So the equality requires to be defined "almost surely".

Random variable vector space

$X \equiv Y \iff X \overset{a.s.}{=} Y \iff \mathbb{P}(X=Y)=1$
Then, **equivalence classes** are a vector space E and $0_E = \{X \text{ r.v. } / \mathbb{P}(X=0)=1\}$.

Orthogonality on random variable vector space

$\langle X,Y\rangle \overset{\text{def}}{=} \mathbb{E}[XY]$ and associated **norm** : $||X||_2 = \sqrt{\mathbb{E}[X^2]}$ $\quad (\ ||X||_k = \mathbb{E}[|X|^k]^{1/k}$ and $\forall k \in \mathbb{N}, ||X||_{2k}^{2k} = M^{(2k)}(X))$

- $\langle X,X\rangle = 0 \implies X \overset{a.s.}{=} 0$
- Symmetry is direct with definition
- Linearity of expected value : $\langle aX+Z,Y\rangle = a\mathbb{E}[XY]+\mathbb{E}[ZY] = a\langle X,Y\rangle + \langle Z,Y\rangle$

Then,

- $\mathrm{Var(X)} = ||\mathrm{X}-\mathbb{E}[\mathrm{X}]||_2^2$ and $\mathrm{Cov(X,Y)} = \langle \mathrm{X}-\mathbb{E}[\mathrm{X}], \mathrm{Y}-\mathbb{E}[\mathrm{Y}]\rangle$, so $\boxed{X, Y \textbf{ uncorrelated} \iff (X-\mathbb{E}[X]) \perp (Y-\mathbb{E}[Y])}$

Inequalities

Expected value inequalities

- Cauchy-Schwartz : $\langle X, Y\rangle^2 = \mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2] = ||X||_2^2 ||Y||_2^2$
- Markov : $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[X]}{a}$ with X a.s. positive.
- Tchebychev : $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[X^\beta]}{a^\beta}$ with $\beta > 0$.
- Jensen : $\phi\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[\phi(X)\big]$ with ϕ a **convex** function.

1-dimensional Gaussian distribution

Definition

Density function:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cumulative function:

$$\mathbb{P}(X \geq x) = F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\left(\frac{u-\mu}{\sigma}\right)^2} du$$

Linearity stability

$X \sim \mathcal{N}(\mu, \sigma^2)$ a 1-dimensional Gaussian,

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2) \text{ (with } a \text{ and } b \text{ deterministic) .}$$

Every Gaussian $Y \sim \mathcal{N}(\mu, \sigma^2)$ can be seen as $Y = \sigma X + \mu$ with $X \sim \mathcal{N}(0, 1)$.

Multivariate distributions

Definition

A d-dimensional real random variable $X = (X_1, \dots, X_d)$ of density $f_X : \mathbb{R}^d \mapsto \mathbb{R}$:

$$\forall A \in \mathscr{B}(\mathbb{R}), \mathbb{P}(A) = \mathbb{P}(X \in A) = \int_A f_X(x)dx$$

Then,

$$\mathbb{P}(X_1 < x_1, \dots, X_d < x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(x_1, \dots, x_d)dx_1 \dots dx_d$$

- Independence : $f_{X_1,\dots,X_d}(x_1, \dots, x_d) = f_{X_1}(x_1) \dots f_{X_d}(x_d)$
- Expected value : $\mathbb{E}[X] = \big(\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]\big)$ and $\mathbb{E}[X_i] = \int_{\mathbb{R}^d} x_i f_X(x_1, \dots, x_d)dx_1 \dots dx_d$
- Characteristic function : $\forall u \in \mathbb{R}^d, \phi_X(u) = \mathbb{E}[e^{i(u_1X_1 + \dots + u_dX_d)}] = \mathbb{E}[e^{i\langle u, X\rangle}]$ (Note that $\phi_X(u) = \phi_{\langle X,u\rangle}(1)$)

Multivariate distributions

Independence

$X = (X_1, \dots, X_d)$ a multivariate random variable,

1. If $(X_1, \dots, X_d)$ are independent, $\phi_{X_1+\dots+X_d}(t) = \prod_{i=1}^{d} \phi_{X_i}(t)$
2. $(X_1, \dots, X_d)$ are independent iif $\phi_{X_1,\dots,X_d}(u) = \prod_{i=1}^{d} \phi_{X_i}(u_i)$

Multivariate distributions

To generalize notions of **variance/covariance** we need a value for each couple (X_i, X_j) of variables among the vector.

Covariance matrix

$X = (X_1, \dots, X_d)$ a multivariate random variable,

$$\forall (i,j) \in \{1,\dots,d\}, \mathrm{Cov(X)_{ij}} = \mathrm{Cov(X_i, X_j)}$$

This matrix is symmetric with separate variance of each univariate variable in the diagonal.

Exercise 3

Show that $\forall u \in \mathbb{R}^d, \mathbb{E}[\langle u, X\rangle] = \langle u, \mathbb{E}[X]\rangle$ and $\mathrm{Var(\langle u, X\rangle)} = \langle \mathrm{u, Cov(X)u}\rangle = \mathrm{u^T Cov(X)u}$
Deduce that $\forall A \in \mathcal{M}_{n,d}(\mathbb{R}), \mathrm{Cov(AX)} = \mathrm{ACov(X)A^T} \in \mathcal{M}_n(\mathbb{R})$.

Change of variable

Proposition

X a random variable such that : $\int_{B_X} f_X(u)du = 1$

- h a diffeomorphism of $\mathbb{R}^d$ from B_X to B_z
- J Jacobian matrix of h

$$z \in B_z, \quad f_Z(z) = \frac{1}{|\det(J)^{-1}|} f_X\left(h^{-1}(z)\right)$$

Exercise 4

$X = (X_1, \ldots, X_d)$ a random variable of dimension d of density f_X.
$A \in \mathcal{M}_d(\mathbb{R})$ invertible and $Z = AX + B$
Show that density of Z is :

$$z \in \mathcal{R}^d, \quad f_Z(z) = \frac{1}{|\det(A)|} f_X\left(A^{-1}(z - B)\right)$$

Exercise 5

X, Y two **independent** 1-dimensional random variables,
Show that : $f_{X+Y} = f_X * f_Y$

Multivariate Gaussian variable

Definition

A random variable $X = (X_1, \dots, X_d)$ with values in $\mathbb{R}^d$ is defined to be a **Gaussian vector** iif all linear combination of X_i is a univariate Gaussian.

- $\iff \forall c \in \mathbb{R}^d, \langle c, X \rangle$ is Gaussian.
- In particular each X_i is Gaussian.

Proposition

If $X = (X_1, \dots, X_d)$ is a **Gaussian vector**, then X_i are independent $\iff X_i$ are uncorrelated.

Proof.

Assume $Z = (X_i, X_j)$ are uncorrelated,

$$\phi_Z \quad : \quad t = (t_1, t_2) \quad \mapsto \quad \mathbb{E}\left[e^{i\langle t, Z\rangle}\right] = \phi_{\langle t, Z\rangle}(1)$$

By definition $\forall t, \exists (\mu_t, \sigma_t) : \langle t, X\rangle \sim \mathcal{N}(\mu_t, \sigma_t^2)$ then, $\phi_{\langle t, X\rangle}(1) = e^{i\mu_t} e^{-\frac{\sigma_t^2}{2}}$
And $\mu_t = t_1 \mu_i + t_2 \mu_j, {\sigma_t}^2 = t_1^2 {\sigma_1}^2 + t_2^2 {\sigma_2}^2$ (because non-correlated)
Thus, $\phi_{X_i, X_j}(t_1, t_2) = \phi_{X_i}(t_1)\phi_{X_j}(t_2)$

Multivariate Gaussian variable

Warning

Gaussian vector assumption, which allows to write $\langle t, X\rangle \sim \mathcal{N}(\mu_t, \sigma_t^2)$, is **crucial** to ensure equivalence between dependence and correlation.

Example

Consider $X \sim \mathcal{N}(0,1)$ and an independent ϵ Bernouilli of parameter 0.5 with values in $\{-1,1\}$.

- $Y = \epsilon X$ so X and Y are not independent.
- $Y \sim \mathcal{N}(0,1)$
- $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] = 0.5\mathbb{E}[\epsilon X^2/\epsilon = 1] + 0.5\mathbb{E}[\epsilon X^2/\epsilon = -1] = 0.5(\mathbb{E}[X^2] + \mathbb{E}[-X^2]) = 0$

$\longrightarrow$ X and Y are uncorrelated normal variables but are dependent. So (X,Y) is **not** a 2-d Gaussian vector.

Proposition

$X = (X_1, \ldots, X_d)$ is a Gaussian vector iif :
$\exists \mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{M}_d(\mathbb{R})$:

$$\forall u \in \mathbb{R}^d, \phi_X(u) = \exp\left(i\langle u, \mu\rangle - \frac{\langle u, \Sigma u\rangle}{2}\right)$$

Operations on multivariate Gaussian

Linearity stability

$X \sim \mathcal{N}(\mu, \Sigma)$ a d-dimensional Gaussian, $S \in \mathcal{M}_{n,d}(\mathbb{R})$ and $T \in \mathbb{R}^n$ (S and T deterministic).

$$SX + T \sim \mathcal{N}(S\mu + T, S\Sigma S^T) \text{ is a } n\text{-dimensional Gaussian.}$$

Proof.

Denote $Y = SX + T$,

$\forall u \in \mathbb{R}^n, \phi_Y(u) = e^{i\langle u, T\rangle}\mathbb{E}[e^{i\langle u, SX\rangle}] = e^{i\langle u, T\rangle}\mathbb{E}[e^{i\langle S^T u, X\rangle}] = e^{i\langle u, T\rangle}\phi_{\langle S^T u, X\rangle}(1)$

By definition, $\langle S^T u, X\rangle$ is univariate Gaussian with :

- $\mathbb{E}[\langle S^T u, X\rangle] = \langle S^T u, \mathbb{E}[X]\rangle$
- $\mathrm{Var}(\langle \mathrm{S^T u, X}\rangle) = \langle \mathrm{S^T u, Cov(X) S^T u}\rangle = \langle \mathrm{u, S Cov(X) S^T u}\rangle$

So $\phi_{\langle S^T u, X\rangle}(1) = \exp\left(i\langle S^T u, \mathbb{E}[X]\rangle - \frac{\langle u, S\mathrm{Cov(X)S^T u}\rangle}{2}\right)$

Putting everything together :

$$\phi_Y(u) = \exp\left(i\langle u, T + S\mu_X\rangle - \frac{\langle u, S\mathrm{Cov(X)S^T u}\rangle}{2}\right)$$

□

Multivariate Gaussian distribution

Definition

$X = (X_1, \dots, X_d)$ a d-dimensional Gaussian vector,
Density function:

$$f_X(x) = \frac{1}{|\det(\Sigma)|^{1/2}(2\pi)^{\mathrm{d}/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Cumulative density function:

$$\mathbb{P}\Big(X \in]-\infty, x_1] \times \cdots \times]-\infty, x_d]\Big) = \frac{1}{|\det(\Sigma)|^{1/2}(2\pi)^{\mathrm{d}/2}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} e^{-\frac{1}{2}(u-\mu)^T \Sigma^{-1}(u-\mu)} du_1 \dots du_d$$

Operation on multivariate Gaussian

Proof.

Let X be a multivariate Gaussian random variable with values in $\mathbb{R}^d$,
Let be $T = \mathbb{E}[X]$ its expected value and $\Sigma = \mathrm{Cov(X)}$ its covariance matrix :

So $\forall c \in \mathbb{R}^d, \exists \mu_c, \sigma_c : \sum_{i=1}^{d} c_i X_i \sim \mathcal{N}(\mu_c, \sigma_c^2)$, in particular each $\forall k, X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$

$\exists \widetilde{X}$ a Gaussian vector such that :

- $X = S\widetilde{X} + T = h(X)$
- $\mathbb{E}[\widetilde{X}] = 0_{\mathbb{R}^d}$ and $\mathrm{Cov}[\widetilde{\mathrm{X}}] = \mathrm{I_d}$
- $\Sigma = P\Lambda P^T$ with P orthogonal (as symmetric real matrix), $\Lambda = \mathrm{Diag}\left\{\sigma_1^2, \ldots, \sigma_{\mathrm{d}}^2\right\}$
- $\Sigma = SS^T$ with $S = P\Lambda^{1/2}$, so $\det(\Sigma) = \det(\Lambda)^2$ and $\det(\mathrm{S}^{-1}) = \det(\Sigma)^{-1/2}$

This means marginals $\widetilde{X_k} \sim \mathcal{N}(0,1), \forall k$ and $\widetilde{X_1}, \ldots, \widetilde{X_d}$ are independent as uncorrelated Gaussians.
Thus,

$$f_X(u) = \frac{1}{\left|\det(\mathrm{S})\right|} f_{\widetilde{X}}(S^{-1}(u-T)) = \frac{1}{\left|\det(\Lambda)^{1/2}\right|} \prod_{k=1}^{d} f_{\widetilde{X_k}}(h^{-1}(u_k)) = \frac{1}{(2\pi)^{d/2}\left|\det(\Sigma)^{1/2}\right|} \prod_{k=1}^{d} e^{-\frac{1}{2}\left(h^{-1}(u_k)\right)^2} = \frac{1}{(2\pi)^{d/2}\left|\det(\Sigma)^{1/2}\right|} e^{-\frac{1}{2}\sum_{k=1}^{d}\left(h^{-1}(u_k)\right)^2}$$

□

Operation on multivariate Gaussian

Proof.

$$f_X(u) = \frac{1}{|\det(S)^{-1}|} f_{\widetilde{X}}(h^{-1}(u)) = \frac{1}{|\det(S)^{-1}|} \prod_{k=1}^{d} f_{\widetilde{X}_k}(h^{-1}(u_k)) = \frac{1}{(2\pi)^{d/2}|\det(\Sigma)^{1/2}|} \prod_{k=1}^{d} e^{-\frac{1}{2}\left(h^{-1}(u_k)\right)^2} = \frac{1}{(2\pi)^{d/2}|\det(\Sigma)^{1/2}|} e^{-\frac{1}{2}||h^{-1}(u)}$$

$$= \frac{1}{(2\pi)^{d/2}|\det(\Sigma)^{1/2}|} e^{-\frac{1}{2}h^{-1}(u)^T h^{-1}(u)} = \frac{1}{(2\pi)^{d/2}|\det(\Sigma)^{1/2}|} e^{-\frac{1}{2}[S^{-1}(U-T)]^T[S^{-1}(U-T)]}$$

By noticing : $[S^{-1}(u-T)]^T[S^{-1}(u-T)] = (u-T)^T P^T \Lambda^{-1/2}\Lambda^{-1/2} P(u-T) = (u-T)^T \Sigma^{-1}(u-T)$
Then,

$$\boxed{f_X(u) = \frac{1}{(2\pi)^{d/2}|\det(\Sigma)^{1/2}|} e^{-\frac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)}}$$

□

Motivation

In practice we rarely know the exact distribution of data we are confronted with.

- **Parametric estimation :** approximate some distribution parameters from data.
- **Confidence interval :** estimation comes with a quantifiable uncertainty under some assumptions.
- **Hypothesis testing :** try to answer yes/no questions about the distribution with a certain level of uncertainty.

Most common assumptions to work with are that samples come from **independent** random variables following the **same distribution**.

Quality of an estimator

Definition

Data are modelled as $(x_1, \dots, x_n)$ samples coming from $(X_1, \dots, X_n)$ random variables.
We call an **estimator** any function $\hat{\theta} : (x_1, \dots, x_n) \mapsto \hat{\theta}(x_1, \dots, x_n)$.

Quantify the difference between $\hat{\theta}$ and θ :

- Bias : $\mathbb{E}[\hat{\theta}] - \theta$
- Quadratic error : $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2$
- For any loss function ℓ : $\mathrm{err}_{\hat{\theta}}(\theta) = \mathbb{E}[\ell(\hat{\theta}, \theta)]$

Given $\hat{\theta}_1, \hat{\theta}_2$ two estimators, $\forall \theta, \mathrm{err}_{\hat{\theta}_1}(\theta) < \mathrm{err}_{\hat{\theta}_2}(\theta) \iff \hat{\theta}_1$ is **better** than $\hat{\theta}_2$.
But if it depends on θ, we can not conclude.

Estimators designing methods : using moments

Moment estimation method

Consider a parametric family of probability distributions denoted $\mathbb{P}_\theta$ with $\theta = (\theta_1, \dots, \theta_d)$.
Assume a series of random variables $(X_1, \dots, X_n)$ follows the **same** probability distribution among this family.
$x = (x_1, \dots, x_n)$ samples from these variables.
Assume that X_i moments exist until K then,

- Raw moments $\mathbb{E}[X^k]$ can be estimated by $\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$
- Centered moments $\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^k\right]$ can be estimated by $\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{m}_1)^k$
- Each moment is a function of $(\theta_1, \dots, \theta_d)$: $m_k = u_k(\theta_1, \dots, \theta_d)$

Moments method consists in trying to solve this equation system :

$$\begin{cases} u_1(\theta_1, \dots, \theta_d) = \hat{m}_1 \\ \dots \\ u_K(\theta_1, \dots, \theta_d) = \hat{m}_K \end{cases}$$

Estimators designing methods

Example

$X \sim \text{Unif}([a,b])$, so $m_2 = \text{Var}[X] = \frac{(b-a)^2}{12}$ and $m_1 = \mathbb{E}[X] = \frac{(a+b)}{2}$

$$b = 2m_1 - a \qquad b = 2m_1 - a \qquad a = m_1 - \sqrt{3m_2}$$

$$m_2 = \frac{(m_1 - a)^2}{3} \qquad a = \pm\sqrt{3m_2} + m_1 \qquad b = m_1 + \sqrt{3m_2}$$

Then,

- $\widehat{m_1} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $\widehat{m_2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{m_1})^2$
- $\widehat{a} = \widehat{m_1} - \sqrt{3\widehat{m_2}}$
- $\widehat{b} = \widehat{m_1} + \sqrt{3\widehat{m_2}}$

Estimators designing methods : maximum likelihood

Maximum likelihood estimation method

Consider a parametric family of probability distributions denoted $\mathbb{P}_\theta$.
Assume a series of random variables $(X_1, \dots, X_n)$ follows the **same** probability distribution among this family.
$x = (x_1, \dots, x_n)$ samples from these variables.
Likelihood function:

$$\mathscr{L}_n(x, \theta) : \theta \mapsto \mathbb{P}_\theta(x_1, \dots, x_n)$$

Maximum likelihood estimation:

$$\hat{\theta}(x_1, \dots, x_n) = \arg\max_\theta \mathscr{L}_n(x, \theta)$$

In general we consider independence between $(X_1, \dots, X_n)$ and so:

$$\hat{\theta}(x_1, \dots, x_n) = \arg\max_\theta \sum_{i=1}^{n} \log\big(\mathbb{P}_\theta(x_i)\big)$$

Estimators designing methods

Example

$X \sim \mathcal{N}(\mu, \sigma^2)$,

Log-likelihood is : $\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 - \frac{n}{2}\log(2\pi\sigma^2)$

- $\frac{\partial \log\left(\mathcal{L}_n(x,\mu)\right)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$
- $\frac{\partial \log\left(\mathcal{L}_n(x,v)\right)}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2, \quad (\text{with } v=\sigma^2)$

If we want to estimate mean : $\frac{\partial \mathcal{L}_n(x,\mu)}{\partial \mu} = 0 \Longleftrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

If we want to estimate variance : $\frac{\partial \mathcal{L}_n(x,v)}{\partial v} = 0 \Longleftrightarrow v = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

Estimators designing methods

Exercise

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid. and $\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Show that $U_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$ is biased and find an unbiased estimator.

Next lesson

Imagine independent N trials with a coin or a dice. You build an estimator $\widehat{p}$ of the associated random variable parameter.

- How to build **confidence interval** that quantifies uncertainty of this estimator ?
- How to attempt to detect whether the coin/dice is rigged or not ?
- How these two questions behave while N increases ?

Khi-squared distribution

Khi-squared

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid.

- $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{S_n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, $\overline{S_{n-1}}' = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$
- $V_n = \frac{n\overline{S_n}}{\sigma^2}$, $V_{n-1} = \frac{(n-1)\overline{S_{n-1}}'}{\sigma^2}$
- $\boxed{V_n \sim \chi^2(n)}$ and $\boxed{V_{n-1} \sim \chi^2(n-1)}$

Density function: $f_V(x) = \frac{1}{2^{-n/2}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-\frac{x^2}{2}}$, **Characteristic function:** $\phi_V(t) = (1 - i2t)^{-n/2}$

Student distribution

Student

Consider $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(n-1)$ independent,

Then, $\boxed{T_{n-1} = \dfrac{Z}{\sqrt{\frac{V}{n-1}}} \sim \text{Stud(n}-1)}$

Density function:

$$f_T(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

In particular,

- $Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \sim \mathcal{N}(0,1)$
- $\overline{S_{n-1}}' = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 \qquad V_{n-1} = \frac{n-1}{\sigma^2} \overline{S_{n-1}}' \sim \chi^2(n-1)$

Then, $T_{n-1} = \dfrac{Z_n}{\sqrt{\frac{V_{n-1}}{n-1}}} \sim \text{Stud(n}-1)$

Estimation of Gaussian parameters : expected value

Estimate μ while σ is known

$X = (X_1, \dots, X_n)$ independent following same distribution $\mathcal{N}(\mu, \sigma^2)$.

This means $Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \sim \mathcal{N}(0,1)$, so $\mathbb{P}\left(-q_{1-\alpha/2} \leq \overline{X_n} \leq q_{1-\alpha/2}\right) = 1 - \alpha$ with:

- $\mathbb{P}$ probability measure of centered and scaled normal distribution.
- $q_{1-\alpha/2}$: quantile of order $1 - \frac{\alpha}{2}$.

Thus,

$$\mathbb{P}\left(\overline{X_n} - \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} \leq \mu \leq \overline{X_n} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2}\right) = 1 - \alpha$$

Estimation of Gaussian parameters : expected value

Estimate μ while σ is unknown

$X = (X_1, \dots, X_n)$ independent following same distribution $\mathcal{N}(\mu, \sigma^2)$.

$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \sim \mathcal{N}(0,1), \quad T_{n-1} = \frac{Z_n}{\sqrt{\frac{V_{n-1}}{n-1}}} = \frac{(\overline{X_n} - \mu)\sqrt{n}}{\sqrt{S_{n-1}'}} \sim \text{Stud(n} - 1)$

So $\mathbb{P}\left(-t_{1-\alpha/2} \leq T_{n-1} \leq t_{1-\alpha/2}\right) = 1 - \alpha$ with:

- $\mathbb{P}$ probability measure of Student distribution.
- $t_{1-\alpha/2}$: quantile of order $1 - \alpha/2$.

Thus,

$$\mathbb{P}\left(\overline{X_n} - \frac{\sqrt{S_{n-1}'}}{\sqrt{n}} t_{1-\alpha/2} \leq \mu \leq \overline{X_n} + \frac{\sqrt{S_{n-1}'}}{\sqrt{n}} t_{1-\alpha/2}\right) = 1 - \alpha$$

Estimation of Gaussian parameters : variance

Estimate σ while μ is known

$X = (X_1, \dots, X_n)$ independent following same distribution $\mathcal{N}(\mu, \sigma^2)$.

$\overline{S_n} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2, \quad V_n = \frac{n}{\sigma^2} \overline{S_n} \sim \chi^2(\mathrm{n})$

So $\mathbb{P}\left(\chi^2_{\alpha/2}(n) \leq \frac{n}{\sigma^2} \overline{S_n} \leq \chi^2_{1-\alpha/2}(n)\right) = 1 - \alpha$ with:

- $\mathbb{P}$ probability measure of Khi-squared distribution.
- $\chi^2_{\alpha/2}(n)$: quantile of order $\alpha/2$.

Thus,

$$\mathbb{P}\left(\frac{n}{\chi^2_{1-\alpha/2}(n)} \overline{S_n} \leq \sigma^2 \leq \frac{n}{\chi^2_{\alpha/2}(n)} \overline{S_n}\right) = 1 - \alpha$$

Estimation of Gaussian parameters : variance

Estimate σ while μ is unknown

$X = (X_1, \dots, X_n)$ independent following same distribution $\mathcal{N}(\mu, \sigma^2)$.

$\overline{S_{n-1}}' = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2, \quad V_{n-1} = \frac{n-1}{\sigma^2} \overline{S_{n-1}}' \sim \chi^2(n-1)$

So $\mathbb{P}\left(\chi^2_{\alpha/2}(n) \leq \frac{n-1}{\sigma^2} \overline{S_{n-1}}' \leq \chi^2_{1-\alpha/2}(n)\right) = 1 - \alpha$ with:

- $\mathbb{P}$ probability measure of Khi-squared distribution.
- $\chi^2_{\alpha/2}(n-1)$: quantile of order $\alpha/2$.

Thus,

$$\mathbb{P}\left(\frac{n}{\chi^2_{1-\alpha/2}(n)} \overline{S_{n-1}}' \leq \sigma^2 \leq \frac{n}{\chi^2_{\alpha/2}(n)} \overline{S_{n-1}}'\right) = 1 - \alpha$$

Estimation of Gaussian parameters : summary

There are four situations depending on which parameter to estimate and whether the other parameter is known or not.

- $Z_n = \frac{\sqrt{n}(\overline{X_n}-\mu)}{\sigma}$
- $V_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$
- $V_{n-1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2$
- $T_{n-1} = \frac{Z_n}{\sqrt{\frac{V_{n-1}}{n-1}}}$

Est. param. \ Information	μ known	μ unknown	σ known	σ unknown
μ	-	-	$Z_n \sim \mathcal{N}(0,1)$	$T_{n-1} \sim \text{Stud(n}-1)$
σ	$V_n \sim \chi^2(n)$	$V_{n-1} \sim \chi^2(n-1)$	-	-

Table 1: Building confidence intervals for Gaussian parameters

Hypothesis testing

There are numerous important yes/no question we can wonder about data :

- **Parameter** test : Is a distribution parameter consistent with what we assumed ?
- **Matching** test : Are data compatible with the distribution we assumed ?
- **Homogeneity** test : Are several groups of samples coming from the same distribution ?
- **Independence** test : Are these variables related or independent ?

Hypothesis testing can be sorted in two main categories :

- **Parametric** hypothesis testing : more constrained assumptions but more powerful conclusions.
- **Non-parametric** hypothesis testing : weaker assumptions but less powerful conclusions.

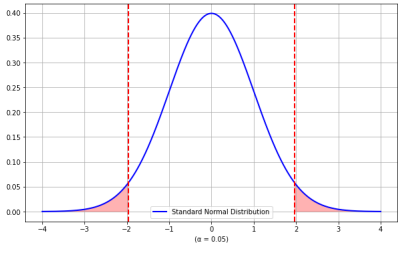
Steps of hypothesis testing :

1. Define working assumption to check, namely "null hypothesis" : $\mathscr{H}_0$ (how to reject it with a fixed level of certainty).
2. Define a statistic ξ_n to check $\mathscr{H}_0$.
3. Find the distribution of ξ_n under assumption $\mathscr{H}_0$.
4. Fix the test level of **significance** α and the associated **critical region**.
5. Compute the value of chosen statistic using observed samples.
6. Conclude about $\mathscr{H}_0$.

Hypothesis testing : Gaussian parameters

Example

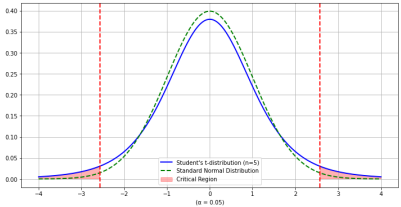
- Working assumption : $\mathscr{H}_0$: " $\mu = \mu_0$ "
- Statistic : $Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{\sigma} \sim \mathcal{N}(0,1)$
- Critical region : $R_\alpha =]-\infty, -q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[$
- $\mathscr{H}_0$ is rejected if observed statistic value $z_n \in R_\alpha$



Hypothesis testing : Gaussian parameters

Example

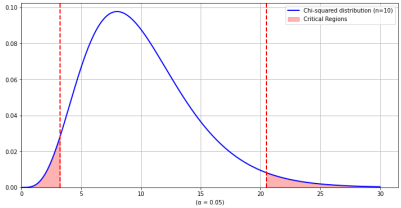
- Working assumption : $\mathscr{H}_0$: " $\mu = \mu_0$ "
- Statistic : $T_{n-1} = \dfrac{Z_n}{\sqrt{\frac{V_{n-1}}{n-1}}} \sim \text{Stud(n} - 1)$
- Critical region : $R_\alpha =]-\infty, -t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, +\infty[$
- $\mathscr{H}_0$ is rejected if observed statistic value $t_n \in R_\alpha$



Hypothesis testing : Gaussian parameters

Example

- Working assumption : $\mathscr{H}_0$: " $\sigma = \sigma_0$ "
- Statistic : $V_n = \frac{n\overline{S_n}}{\sigma_0^2} \sim \chi^2(n)$
- Critical region : $R_\alpha = [0, \chi^2_{\alpha/2}(n)[\cup]\chi^2_{1-\alpha/2}(n), +\infty[$
- $\mathscr{H}_0$ is rejected if observed statistic value $v_n \in R_\alpha$



Confidence interval

Imagine you want to know whether a coin is rigged or not.
In other words it follows a Bernouilli distribution of parameter p and you want to check if $p = 0.5$.

- Estimator : $\hat{p}(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$
- $\forall k \in \{0, \dots, n\}, \mathbb{P}(\hat{p} = \frac{k}{n}) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\mathbb{P}(\hat{p} \in [p-\Delta, p+\Delta]) = \mathbb{P}(p \in [\hat{p}-\Delta, \hat{p}+\Delta]) = \sum_{k=\lceil n(p-\Delta) \rceil}^{\lfloor n(p+\Delta) \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$
- Given a fixed α, there is a minimum Δ_α such that : $\mathbb{P}(p \in [\hat{p}-\Delta_\alpha, \hat{p}+\Delta_\alpha]) > 1-\alpha$

Confidence interval

If we assume $\mathscr{H}_0$ true :

- $\mathbb{P}(|\hat{p}-0.5| \leq \delta]) = \frac{1}{2^n} \sum_{k=k_1}^{k_2} \binom{n}{k} = \frac{1}{2^n}\left[\binom{n+1}{k_2+1} - \binom{n+1}{k_1}\right]$ with $k_1 = \lceil n(0.5-\delta)\rceil$ and $k_2 = \lfloor n(0.5+\delta)\rfloor$
- $\Delta_0 = \arg\min_{\Delta}\left\{\mathbb{P}(|\hat{p}-0.5| \leq \Delta]) > 1-\alpha\right\}$ is computable.
- Check if observed $\hat{p}$ is in critical region $R_\alpha =]-\infty, 0.5-\Delta_\alpha[\cup]0.5+\Delta_\alpha, +\infty[$

Asymptotic approximation

However, sometime the estimator distribution is hard to express exactly or very heavy to compute while n is large. That's why in these cases we often use asymptotic confidence intervals.

Asymptotic behavior of estimators and confidence intervals

Central limit theorem

$(X_n, n \in \mathbb{N})$ series of i.i.d. real random variables, assuming that $\mu = \mathbb{E}[X_i]$ and $\sigma = \mathrm{Var}(\mathrm{X_i})$ exist,

$$\frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Central limit theorem

$(X_n, n \in \mathbb{N})$ series of i.i.d. d-dimensional real random variables,
Assuming that $\mu = \mathbb{E}[X_i] \in \mathbb{R}^d$ and $\Sigma = \mathrm{Cov}(\mathrm{X_i}) \in \mathcal{M}_{\mathrm{d}}(\mathbb{R})$ exist,

$$\sqrt{n}(\overline{X_n} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0_{\mathbb{R}^d}, \Sigma)$$

Asymptotic behavior of estimators and confidence intervals

Asymptotic confidence interval

Under TCL assumptions,

- $\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}$
- $I_n = \left[\overline{X_n} - \frac{a\sigma_n}{\sqrt{n}}, \overline{X_n} + \frac{a\sigma_n}{\sqrt{n}}\right]$

$$\mathbb{P}(\mu \in I_n) \underset{n\to+\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-\frac{x^2}{2}}\, dx$$