QML-Mod1-Mathematical and Satistical Preliminaries

Riccardo Marega

March 2025

Indice

1 Mathematical Preliminaries **2**
1.1 Linear Algebra -7/03/2025 . . . 2
1.2 Optimization-8/03/2025 . . . 6
1.2.1 Introduction to Mathematical Programming (or Mathematical optimization): definitions . . . 8
1.2.2 Properties and structure of convex problems . . . 9
1.2.3 Optimality conditions for Constrained and Unconstrained problems: general case . . . 12
1.2.4 Convergence to stationary points . . . 17
1.3 Some algebric results-14/03/25 . . . 18
1.3.1 Eigenvalues and Eigenvectors . . . 20

2 Statistical Preliminaries **22**
2.1 Introduction -7/03/2025 . . . 22
2.2 Statistical methods for business and economics -14/03/25 . . . 29
2.2.1 Hypothesis Testing . . . 29
2.2.2 Statistical learning and classification . . . 33

3 Introduction to Data Protection Regulation **38**
3.1 Introduction-21/03/2025 . . . 38
3.1.1 Privacy or Data Protection . . . 38
3.1.2 Type of consent . . . 40
3.1.3 Cookies and other tracking tool . . . 42
3.1.4 Accountability vs responsability vs liability . . . 43

1 Mathematical Preliminaries

1.1 Linear Algebra -7/03/2025

An algebra is a generalization of arithmetic in which letters representing numbers are combined according to the rules of arithmetics.
A second definition could be the following: an algebra is any of various systems or branches of mathematics concerned with the properties and relationships of abstract entities (such as numbers, matrices, vectors, groups, rings or fields) manipulated in symbolic form under operations often analogous to those of arithmetic.
Our focus now goes to linear Algebra.
To understand what linear means we start by thinking about linear regression, where given a set of points we try to find a linear function able to estimate the behavior of the set. Linearity can be also found in geometrical transformations as also in machine learning when the input to the artificial neuron is calculated by performing a weighted sum of the incoming inputs.

Definition A mathematical space is a set and all of the operations on that set.
A vector is an ordered collection of elements. Each element can be a number, a variable, or a function, depending on the context.

Definition 1 Let V be a set of elements (vector) for which we define the operation $'+'$. In addition, consider the numerical set (field) K, such that among its elements we define the operations of products $'.'$ and sum $\oplus$. Furthermore, between the elements of K and vectors of V the operation of multiplication $'.'$ is defined. Then, the set V(K) is said to be a **Vector Space** over the field K if the following properties hold:

0)	$x+y \in V(K)$,	$\forall x, y \in V(K)$
1)	$(x+y)+z = x+(y+z)$,	$\forall x, y, z \in V(K)$
2)	$\exists w \in V(K) \ : \quad x+w = x$,	$\forall x \in V(K)$
3)	$\forall x \in V(K),\ \exists \bar{x} \in V(K) \ : \quad x+\bar{x} = w$,	
4)	$x+y = y+x$,	$\forall x, y \in V(K)$
5)	$\alpha \bullet x \in V(K), \quad \alpha \bullet (\beta \bullet x) = (\alpha \cdot \beta) \bullet x$,	$\forall x \in V(K),\ \forall \alpha, \beta \in K$
6)	$\exists \sigma \in K \ : \quad \sigma \bullet x = x$,	$\forall x \in V(K)$
7)	$\alpha \bullet (x+y) = \alpha \bullet x + \alpha \bullet y$,	$\forall x, y \in V(K),\ \forall \alpha \in K$
8)	$(\alpha \oplus \beta) \bullet x = \alpha \bullet x + \beta \bullet x$,	$\forall x \in V(K),\ \forall \alpha, \beta \in K$

□

• A first example could be the set R(R), where V = R and K = R (real numbers with the standard operations of sum and product among reals), which is usually indicated as R.
• The se $R^n(R)$, where $V = R^n$ and K =R, which is usually indicated as R^n. This set is given by

those vectors like x, which represents the n-tuple.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where $x_1, ..., x_n \in R$, for which we define the product for the scalar $\alpha \in R$ as

$$\alpha x = \alpha \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix}$$

and the sum $x + y$ (with $x, y \in R^n$) as

$$x + y = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}.$$

• The set $M_{m,n}(R)$, where $V = M_{m,n}$ and K = R, which represents the rectangular matricies of dimension m (rows) and n (columns), with elements in R, for which we define the product for a scalar $\alpha \in R$ as

$$\alpha x = \alpha \begin{pmatrix} x_{11} & \dots & x_{1n} \\ & \vdots & \\ x_{n1} & \dots & x_{nn} \end{pmatrix} = \begin{pmatrix} \alpha x_{11} & \dots & \alpha x_{1n} \\ & \vdots & \\ \alpha x_{n1} & \dots & \alpha x_{nn} \end{pmatrix}$$

and the sum $x + y$ as

$$x + y = \begin{pmatrix} x_{11} 0.1cm \dots & x_{1n} \\ \vdots & \\ x_{n1} & \dots & x_{nn} \end{pmatrix} + \begin{pmatrix} y_{11} & \dots & y_{1n} \\ \vdots & \\ y_{n1} & \dots & y_{nn} \end{pmatrix} = \begin{pmatrix} x_{11} + y_{11} & \dots & x_{1n} + y_{1n} \\ \vdots & \\ x_{n1} + y_{n1} & \dots & x_{nn} + y_{nn} \end{pmatrix}.$$

• The set $P_n(R)$ of polynomials of a real number with a real coefficient, of degree no exceeding n, where $V = P_n$ and K = R. We recall that given two polynomials $p(x) : R \to R$ and $q(x) : R \to R$, of respective degree h and k (with real coefficients $p_0, p_1, ..., p_h$ and $q_0, q_1, ..., q_k$).

$$p(x) = p_0 + p_1 x + p_2 x^2 + \cdots + p_h x^h$$

$$q(x) = q_0 + q_1 x + q_2 x^2 + \cdots + q_k x^k,$$

we define the product $\alpha p(x)$, with $\alpha \in \mathbb{R}$, as

$$\alpha p(x) = (\alpha p_0) + (\alpha p_1)x + (\alpha p_2)x^2 + \cdots + (\alpha p_h)x^h.$$

On the other hand, as regards the sum $p(x)+q(x)$ we have the following:

$$\text{if } h \geq k \Rightarrow p(x)+q(x) = (p_0+q_0)+(p_1+q_1)x+(p_2+q_2)x^2+\cdots+ \\ +(p_k+q_k)x^k+p_{k+1}x^{k+1}+\cdots+p_hx^h$$

$$\text{if } h < k \Rightarrow p(x)+q(x) = (p_0+q_0)+(p_1+q_1)x+(p_2+q_2)x^2+\cdots+ \\ +(p_h+q_h)x^h+q_{h+1}x^{h+1}+\cdots+q_kx^k.$$

Definition 2 Let $A \subseteq R$ be a numerical set, and let $v_1, ..., v_n$ be a set of non-zero vectors in the vector space V(A). Then, the vectors $v_1, ..., v_n$ are linearly independent of A if the relation

$$\alpha_1 v_1 + ... + \alpha_m v_m = w \qquad\qquad \alpha_i \in A, i = 1, ...m.$$

is satisfied iff (i.e. if and only if) $\alpha_1 = ... = \alpha_m = 0$, being w the null vector of V(K).
Note that we can equivalently express the latter condition by saying that if the vecotors $v_1, ..., v_n$ are linearly dependent on A, then at least one vector can be expressed as linear combination (with coefficients in A) of the remaining vectors.

Definition 3 Given the vector space V(K), we say that V(K) has dimension n if the largest number of linearly independent vectors in V(K) is exactly n. Then, the vector space V(K) is indicated $V^n(k)$.
Observe from a more general perspective that linearly independent vectors do not yield "redundancy". Thus, the information associated with each vector is essential. Conversely, if vectors are linearly dependent, then some of them can be expressed as a linear combination of the remaining vectors, so that the information associated with them can be neglected. In this regard, the dimension of a vector space represents the minimum number of vectors which are necessary to build all the vector space properties.

Rank of a matrix is the maximum number of its linearly independent columns/rows.
Note that the rank of a matrix cannot exceed the number of its rows/columns.

Definition 4 Given the set of vectors $v_1, ..., v_n$ of R^n, we say that the vector $s \in R^n$ is an affine combination of vectors $v_1, ..., v_n$ on R if

$$s = \alpha_1 v_1 + ... + \alpha_m v_m, \qquad\qquad \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \in R, i = 1, ..., m.$$

Definition 5 Given the set of vectors $v_1, ..., v_n$ of R^n, we say that the vector $z \in R^n$ is cone combination of vectors $v_1, ..., v_n$ on R if

$$z = \alpha_1 v_1 + ... + \alpha_m v_m, \qquad\qquad \alpha_i \geq 0, i = 1, ..., m.$$

Definition 6 Given the set of vectors $v_1, ..., v_m$ of R^n, we say that the vector $y \in R^n$ is a convex combination of vectors $v_1, ..., v_m$ on R, if it is both affine combination and a cone combination of $v_1, ..., v_m$ on R, i.e.

$$y = \alpha_1 v_1 + ... + \alpha_m v_m, \qquad \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \in R, i = 1, ..., m.$$

Definition 7 Given (for simplicity) the vector space $V^n(R)$, we define the inner product or scalar product "<.,.>" as a function from $V^n(R) \times V^n(R)$ to R, satisfying the following properties:

- $< x, x > \geq 0$, $\forall x \in V^n(R)$ with $< x, x >= 0 \iff x = 0$,
- $< x, y + z >=< x, y > + < x, z >$, $\forall x, y, z \in V^n(R)$,
- $< x, y >=< y, x >$, $\forall x, y \in V^n(R)$.

An example follows:

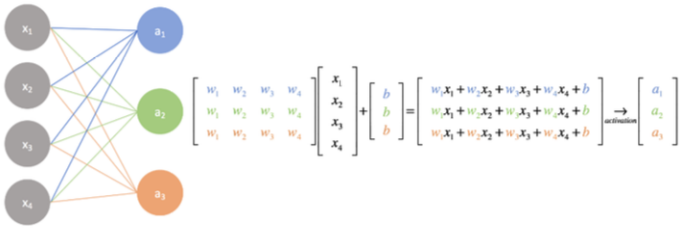


Figura 1: Output calculation of a simple artificial neural network

Definition 8 Given the vectors x,y $\in R^n$, we say that x and y are orthogonal if, considering the standard inner product $x^T y$, we have

$$x^T y = 0.$$

Proposition 1 Given the vectors $v_1, ..., v_m \in R^n/\{0\}$, with $m \leq n$, let $v_1, ..., v_m$ be mutually orthogonal (i.e. $v_i^T v_j = 0$, for any $1 \leq i \neq j \leq m$). Then, the vectors $v_1, ..., v_m$ are linearly independent in R^n.

Definition 9 Let be given (for simplicity) the vector space $V^n(R)$, we introduce the norm of a vector $x \in V^n(R)$, as the function from $V^n(R)$ to $R^+ \cup \{0\}$, indicated as $||.||$, which satisfies the following properties:

- $||x|| \geq 0$, $\forall x \in V^n(R)$ with $||x|| = 0 \iff x = 0$,
- $||\alpha x|| = |\alpha| ||x||$ $\forall x \in V^n(R), \forall \alpha \in R$,

- $||x+y|| \leq ||x|| + ||y|| \ \forall x, y \in V^n(R)$.

From the concept of norm to distance:

$$\text{dist}_p(x,y) = ||x-y||_p, \qquad p \in N/\{0\} \ \text{ and } \ \forall x, y \in R^n$$

Definition 10 Given the function $f : R^n \to R^m$, we say that f(x) is linear in R^n if it satisfies the following two relations:

- $f(x+y) = f(x) + f(y)$, $\forall x, y \in R^n$,
- $f(\alpha x) = \alpha f(x) \ \forall x \in R^m, \forall \alpha \in R$.

Equivalently, these two conditions might be unified in the unique condition:

$$f[\alpha x + \beta y] = \alpha f(x) + \beta f(y), \qquad \forall x, y \in R^n, \forall \alpha, \beta \in R.$$

1.2 Optimization-8/03/2025

We start by getting a general understanding of what optimization is.
Optimization could be thought as the best choice of the inputs values while searching for the optimal solution.
In this section we address and report soma basics of calculus in R^n, which considers real functions with n unknowns. Since we need to find and prove results involving the derivatives of functions, we are committed to report the next definition.

Definition Given the function f(x), with $f : R^n \to R$, we say that f(x) is p times continuously differentiable on the "close set" A, if in the "open set" B, with $A \subset B$, there exist both the partial and mixed derivatives of f(x) up to order p, and they are continuous in B.
Observe that in this section we will use the Gradient $\nabla f(x)$ and the Hessian $\nabla^2 f(x)$ of the function $f : R^n \to R$, which are defined as follows:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 x_1} \cdots \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \vdots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} \cdots \frac{\partial^2 f(x)}{\partial x_n x_n} \end{pmatrix}$$

Definition Given the function $f : R^n \to R$, with domain $A \subseteq R^n$, we say that the poit $\bar{x} \in A$ is of regularity for f(x) if $\nabla f(\bar{x})$ is defined (i.e. it can be computed).
Note that in particular, in case n = 1 (i.e. there is just one unknown), the definition of the gradient $\nabla f(x)$ and the Hessian matrix $\nabla^2 f(x)$ coincide with the very well known definitions of "first derivative" and "second derivative" of the function f(x) at x.

Definition Given the point $\bar{x} \in R^n$ and the norm $||.||$, we define the set $I(\bar{x}, \delta)$ (neighborhood of $\bar{x}$ of radius δ) in the following way:

$$I(\bar{x}, \delta) = \{x \in R^n : ||x - \bar{x}|| \leq \delta\},$$

moreover, we say that the neighborhood $I(\bar{x}, \delta)$ is open if it is defined as:

$$I(\bar{x}, \delta) = \{x \in R^n : ||x - \bar{x}|| < \delta\}.$$

Theorem Given the function f(x), let $f : R^n \to R$ be continuously differentiable at least m times, in the neighborhood $I(\bar{x}, \delta = \{x \in R^n : ||x - \bar{x}|| \leq \delta \ , \delta > 0\})$ of the point $\bar{x}$. Then, we have

$$f(x) = \sum_{h=0}^{m} \frac{D^h f(\bar{x})}{h!}(x - \bar{x})^h + R_{m+1}(\bar{x}), \qquad \lim_{x \to \bar{x}} \frac{R_{m+1}(\bar{x})}{||x - \bar{x}||^m} = 0,$$

where h indicates the order of the partial/mixed derivative (in case of mixed derivative h represents the sum of indices of the derivatives with respect to possibly different unknowns), while $D^h f(\bar{x})$ we shortly indicate the partial/mixed derivative of order h, computed at point $\bar{x}$.
When m= 1 the latter theorem provides the explicit expression

$$f(x) = \frac{1}{0!}f(\bar{x}) + \frac{1}{1!}[\frac{\partial f(\bar{x})}{\partial x_1}(x - x_1) + ... + \frac{\partial f(\bar{x})}{\partial x_n}(x_n - \bar{x_n})] + R_2(\bar{x}) = f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + R_2(\bar{x}),$$

while for m=2 we obtain the expression

$$f(x) = \frac{1}{0!}f(\bar{x}) + \frac{1}{1!}\left[\frac{\partial f(\bar{x})}{\partial x_1}(x_1 - \bar{x}_1) + \cdots + \frac{\partial f(\bar{x})}{\partial x_n}(x_n - \bar{x}_n)\right] +$$

$$+\frac{1}{2!}\left[\sum_{i=1}^{n} \frac{\partial^2 f(\bar{x})}{\partial x_1 \partial x_i}(x_1 - \bar{x}_1)(x_i - \bar{x}_i) + \cdots + \sum_{i=1}^{n} \frac{\partial^2 f(\bar{x})}{\partial x_n \partial x_i}(x_n - \bar{x}_n)(x_i - \bar{x}_i)\right] + R_3(\bar{x})$$

$$= f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}) + R_3(\bar{x}).$$

We strongly highlight the importance of the direction of the gradient $\nabla f(x)$, at the point x, within Mathematical Programming. Indeed, similarly to the trivial case of $n = 1$, where the derivative $f'(x)$ of $f(x)$ at x represents the unit rate of increase of the function at x (being $f'(x)$ the slope of the tangent line to $f(x)$ at the point x), in case $n \geq 2$ we have a similar interpretation which is summarized in the next lemmas.

Lemma Let be given the function f(x), with $f : R^n \to R$, and f(x) continuously differentiable in R^n. Let $\bar{x} \in R^n$, then $d = \nabla f(\bar{x})/||\nabla f(\bar{x})||_2$ represents the direction with unit Euclidean norm which maximizes the Taylor expression $f(\bar{x}) + \nabla f(\bar{x})^T d$ of f(x) at $\bar{x}$. Equivalently, the direction

$$d = \frac{\nabla f(\bar{x})}{||\nabla f(\bar{x})||_2}$$

is a solution to the problem

$$\max_{||d||_2=1} \{f(\bar{x}) + \nabla f(\bar{x})^T d\}$$

Proof
For a given point $\bar{x} \in \mathbb{R}^n$ the quantity $f(\bar{x})$ is constant. Then, changing the direction d, the objective function in (6) outreaches its maximum value when the scalar quantity $\nabla f(\bar{x})^T d$ is maximum. By the definition of inner product, the latter directional derivative is maximum whenever d is parallel and has the same versus of $\nabla f(\bar{x})$. Therefore, since it must be also $\|d\|_2 = 1$, then the objective function in (6) is maximum if $d = \nabla f(\bar{x})/\|\nabla f(\bar{x})\|_2$, which completes the proof. □

Lemma Let be given the function f(x), with $f : R^n \to R$, and f(x) continuously differentiable in R^n. Let $\bar{x} \in R^n$, then $d = -\nabla f(\bar{x})/||\nabla f(\bar{x})||_2$ represents the direction with unit Euclidean norm which minimizes the Taylor expression $f(\bar{x}) + \nabla f(\bar{x})^T d$ of f(x) at $\bar{x}$. Equivalently, the direction

$$d = -\frac{\nabla f(\bar{x})}{||\nabla f(\bar{x})||_2}$$

is a solution to the problem

$$\min_{||d||_2=1} \{f(\bar{x}) + \nabla f(\bar{x})^T d\}$$

Proof
The proof follows guidelines similar to the proof of Lemma 2.1 □

1.2.1 Introduction to Mathematical Programming (or Mathematical optimization): definitions

In this section we give some basic results and definitions which will be widely used to tackle Mathematical Programming problems.

Definition Let be given the set $C \subseteq R^n$ and the function $f : C \to R$, the point $x* \in C$ is a local minimum of f(x) on C, if there exists an open neighborhood $I(x*, \rho) = \{x \in R : ||x - x * || < \rho\}$ with center x* and radius $\rho > 0$, such that:

$$f(x*) \leq f(x) \qquad \forall x \in I(x*, \rho) \cap C.$$

If the latter inequity is satisfied as a "strict" inequity, for any $x \in C$ and $x \neq x*$, then we say that the point x* is a strict local minimum of f(x) on C.

Definition *Let be given the set $C \subseteq \mathbb{R}^n$ and the function $f : C \to \mathbb{R}$, the point $x^* \in C$ is a **local maximum** of $f(x)$ on C, if there exists an open neighborhood $I(x^*, \rho) = \{x \in \mathbb{R}^n : \|x - x^*\| < \rho\}$ with center x^* and radius $\rho > 0$, such that*

$$f(x^*) \geq f(x), \quad \forall x \in I(x^*, \rho) \cap C.$$

If the latter inequality is satisfied as a 'strict' inequality, for any $x \in C$ and $x \neq x^*$, then we say that the point x^* is a **strict local maximum** of $f(x)$ on C.

Definition *Let be given the set $C \subseteq \mathbb{R}^n$ and the function $f : C \to \mathbb{R}$, the point $x^* \in C$ is a **global minimum** of $f(x)$ on C, if*

$$f(x^*) \leq f(x), \quad \forall x \in C.$$

If the latter inequality is satisfied as a 'strict' inequality, for any $x \in C$, $x \neq x^*$, then we say that the point x^* is a **unique global minimum** of $f(x)$ on C.

Definition *Let be given the set $C \subseteq \mathbb{R}^n$ and the function $f : C \to \mathbb{R}$, the point $x^* \in C$ is a **global maximum** of $f(x)$ on C, if*

$$f(x^*) \geq f(x), \quad \forall x \in C.$$

If the latter inequality is satisfied as a 'strict' inequality, for any $x \in C$, $x \neq x^*$, then we say that the point x^* is a **unique global maximum** of $f(x)$ on C.

1.2.2 Properties and structure of convex problems

In this section we specifically want to introduce the definition of convex set and convex function.

Definition Given the nonempty set $C \subseteq R^n$, we say that C is a convex set if

$$\alpha x + (1 - \alpha)y \in C, \qquad \forall x, y \in C \forall \alpha \in [0, 1].$$

Note that equivalently, the nonempty set $C \subseteq \mathbb{R}^n$ is convex if the convex combination of any pair of points in the set C yet belongs to C. Using a geometric viewpoint observe that the point $\alpha x + (1 - \alpha)y$, when $\alpha \in [0, 1]$, represents any point in the closed segment joining x and y (the latter segment is often shortly indicated as $[x, y]$).

Proposition Given the convex sets $C_1, ..., C_m$, with $m \geq 1$, then the intersection of the m sets, indicated as

$$C = C_1 \cap \cdots \cap C_m,$$

is a convex set.

Definition Given the function $f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$, and given the convex nonempty set $C \subseteq \mathbb{R}^n$, we say that $f(x)$ is **convex on the set** C, if for any pair of points $x, y \in C$, the following property holds

$$f[\alpha x + (1-\alpha)y] \leq \alpha f(x) + (1-\alpha)f(y), \quad \forall \alpha \in [0,1].$$

The function $f(x)$ is said to be **strictly convex on** C if the above inequality is a strict inequality (i.e. $<$ in place of $\leq$) for any $x \neq y$ and $\alpha \in (0,1)$.

Definition Given the function $f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$, given the nonempty convex set $C \subseteq \mathbb{R}^n$, we say that the function $f(x)$ is **concave on the set** C, if for any pair of points $x, y \in C$, the following property holds

$$f[\alpha x + (1-\alpha)y] \geq \alpha f(x) + (1-\alpha)f(y), \quad \forall \alpha \in [0,1].$$

The function $f(x)$ is said to be **strictly concave on** C if the above inequality is a strict inequality (i.e. $>$ in place of $\geq$) for any $x \neq y$ and $\alpha \in (0,1)$. Any linear function is both a concave and a convex function.

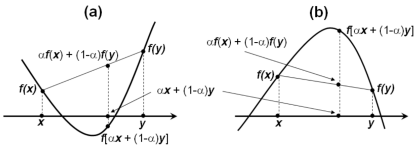


Figure 15: Geometric meaning of the definition of *strict convexity* **(a)** and *strict concavity* **(b)** for the continuous function $f(x)$.

Proposition Given the affine function g(x), with $g : R^n \to R$, then g(x) is at once convex and concave on R^n.

Proposition Given the function f(x) with $f : R^n \to R$, let f(x) be convex on R^n. Then, the level set (possibly empty) L_γ defined as

$$L_\gamma \doteq \{x \in R^n : f(x) \leq \gamma\}$$

is convex for any $\gamma \in R$.
This proposition is important because if we are solving a constraint optimization problem we are typically minimizing a function subject to some constraints typically represented by an inequality. Any inequality can be reversed giving an inequality similar to the one of the definition. A set of inequalities on a convex function gives a convex set.

Definition Given the function f(x) with $f : R^n \to R$, and the real parameter γ, we define the level curve $c_\gamma(x)$ of the function f(x), as the set (possibly empty)

$$c_\gamma(x) \doteq \{x \in R^n : f(x) = \gamma\}$$

being $\gamma \in R$.

Important observation 1 Considering, without loss of generality, the constraint $v_i \leq a_i$, and, recalling the previous proposition, the values of $x \in R^n$ which satisfy the i-th constraint are exactly the point of the level set

$$L_{a_i} = \{x \in R^n : v_i(x) \leq a_i\}$$

Theorem Given the convex set $C \subseteq R^n$, let $f_i(x)$, i = 1,...,m. be convex functions on C. Let be given the coefficients $\lambda_i \geq 0$, i = 1,...,; then, the functions

$$g(x) = \sum_{i=1}^{m} \lambda_i f_i(x),$$

$$f(x) = \max_{1 \leq i \leq m} \{\lambda_i f_i(x)\}$$

are convex on C.
What that means is that the linear combination of convex functions using positive coefficients gives a convex function. That explains the first result of the theorem which among the two is the most important.
What follows is a very strong result.

Proposition Given the function f(x), with $f : R^n \to R$, and the convex set $C \subseteq R^n$, let f(x) be convex on C. Then, any local minimum $x* \in C$, for the mathematical programming problem

$$\min_{x \in C} f(x)$$

is also a global minimum for the same problem.

Proposition Given the function f(x), with $f : R^n \to R$, and the convex set $C \subseteq R^n$, let f(x) be convex on C. Then, the set of the solutions for the mathematical programming problem

$$\min_{x \in C} f(x)$$

is convex.
What that proposition means is that if one, solving an optimization problem, finds two possible solutions (and all the previous conditions are satisfied), then all the points joining these two solutions are still a solution of the problem.

1.2.3 Optimality conditions for Constrained and Unconstrained problems: general case

In this section we deal with the solution of the following general Nonolinear Programming problem:

$$\min_{x \in X} f(x) \quad [*]$$

where the feasible set $X \subseteq R^n$ is exprecially defined throughout a finite number of equalities and/or inequalities i.e.

$$X = \begin{cases} h_j(x) = 0, & j = 1, \ldots, p, \\ g_i(x) \leq 0, & i = 1, \ldots, m. \end{cases}$$

Observe that the first equation [*] has a very general structure. Indeed, in particular we do not assume any convexity/concavity for f(x) or for any constraints functions.
We also remark that the solution of latter equation requires in principle the solution of two distinct problems, namely:

- the minimization of f(x),

- the feasibility of the final solution (if any) found.

Introduction to Fitz-john and Karush-Kuhn necessary optimality conditions First we observe that the optimization methods proposed in literature to solve the problem previously introduced, typically can guarantee to detect only stationary points (which are possibly not also minima), i.e. points which satisfy the so called first order optimality conditions for the problem. Thus, it is first relevant to focus on the latter conditions, in order to clarify some specif aspects. On this purpose we have to report the next definitions.

Definition Given the previous problem and the point $x* \in X$, we define the set I(x*) of the active constraints at x*, as the set of indices of the constraints which satisfy (at the optimal solution x^*)

$$I(x*) = \{j : h_j(x*) = 0, j = 1, ...p\} \cup \{i : g_i(x*) = 0, i = 1, ..., m\}.$$

Observation As a further simple but relevant observation on I(x*), note that if from the constraints in the problem we remove those constraints which are not active at the solution x*, then the solution will play a relevant role in the optimally conditions for the problem.

Definition Let us consider once again the previous problem [*]. Introducing the additional variables $\lambda_0 \in R$, $\lambda \in R^p$ and $\mu \in R^m$, we define the lagrangian function $L(x, \lambda_0, \lambda, \mu)$ as

$$L(x, \lambda_0, \lambda, \mu) = \lambda_0 f(x) + h(x)^T \lambda + g(x)^T \mu$$

where $\lambda = (\lambda_1, ..., \lambda_p)^T$ and $\mu = (\mu_1, ..., \mu_m)^T$.
No specific conditions where given to g(x) and h(x).
Using the latter definitions we can consider the next I order (necessary) optimality conditions for the problem, which are known as the Fritz-Joh Optimality Conditions.
If a point x* is a solution of the problem, then the following conditions (Fritz-John) hold

Proposition: Fritz-John (necessary) Optimality Conditions Given the problem [*], let $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^p$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, with $f(x)$, $h(x)$ and $g(x)$ continuously differentiable in an open set containing X. Let x^* be a local minimum of $f(x)$ on the set X. Then, there exists a vector $(\lambda_0^*, \lambda^*, \mu^*) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^m$ such that

$$\lambda_0^* \nabla f(x^*) + \sum_{j=1}^{p} \lambda_j^* \underbrace{\nabla h_j(x^*)}_{\substack{\text{gradient equality} \\ \text{constraints}}} + \sum_{i=1}^{m} \mu_i^* \underbrace{\nabla g_i(x^*)}_{\substack{\text{gradient inequality} \\ \text{constraints}}} = 0, \quad (36)$$

$$h_j(x^*) = 0, \quad j = 1, \ldots, p, \quad (37)$$

$$g_i(x^*) \leq 0, \quad i = 1, \ldots, m, \quad (38)$$

$$\mu_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m, \quad (39)$$

$$(\lambda_0^*, \lambda^*, \mu^*) \neq 0, \quad (40)$$

$$(\lambda_0^*, \mu^*) \geq 0. \quad (41)$$

The vector $(\lambda_0^*, \lambda^*, \mu^*)$ is the vector of **generalized Lagrange multipliers**, while $(x^*, \lambda_0^*, \lambda^*, \mu^*)$ is a **Fritz-John point**.
Fritz-Jhon optimality condition generalize the unconstrained optimality condition (that is asking for $\nabla f(x) = 0$) when considering $X \equiv R^n$. That is because, being h(x) and g(x) the constraints of the problem, if there are no constraints (a.k.a. $X \equiv R^n$) then h(x)=g(x)=0.
The proof of the latter Proposition can be found in [?, ?]. Using the notation

$$\nabla h(x^*) = (\nabla h_1(x^*) \cdots \nabla h_p(x^*)) \in \mathbb{R}^{n \times p},$$
$$\nabla g(x^*) = (\nabla g_1(x^*) \cdots \nabla g_m(x^*)) \in \mathbb{R}^{n \times m},$$

i.e., $\nabla h(x^*)$ and $\nabla g(x^*)$ respectively represent the gradient matrices associated with the equality and inequality constraints. Recalling that each multiplier $\mu_i^*, i = 1, \dots, m$, is nonnegative, we can rewrite conditions (36)-(41) in the more compact form:

$$\begin{aligned} \lambda_0^* \nabla f(x^*) + \nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^* &= 0, \\ h(x^*) = 0, \quad g(x^*) &\leq 0, \\ g(x^*)^T \mu^* &= 0, \\ (\lambda_0^*, \lambda^*, \mu^*) \neq 0, \quad (\lambda_0^*, \mu^*) &\geq 0. \end{aligned}$$

Moreover, observe that the latter Proposition simply yields a **necessary condition** (which is possibly not also sufficient) for the point x^* to be a local minimum of [*]. Also note that in case $X \equiv \mathbb{R}^n$, the conditions (36)-(41) reduce to the standard necessary conditions for the unconstrained case:

$$\nabla f(x^*) = 0.$$

(Indeed, it suffices to observe that (40) yields $\lambda_0 \neq 0$, and no constraints are present).

As a further consideration, we strongly remark that in case the problem [*] were a maximization problem, then it would easily be **first reformulated as a minimization problem** (i.e. changing the sign of the objective function), before writing the relative Fritz-John necessary optimality conditions. In other words, the results of **the previous proposition specifically refer to a minimization problem, and cannot be immediately extended in case of a maximization**.

We highlight that in the Fritz-John conditions (36)-(41) the coefficient λ_0^* might possibly be zero. In the latter case the optimality conditions (36)-(41) do not depend on the function $f(x)$, which is surely an **unusual fact**. Indeed, since the optimality conditions are conceived to detect local minima for the function $f(x)$ in the constrained problem (35), it is definitely an **anomaly** if such points are independent of $f(x)$, inasmuch as any function would yield exactly the same minima points.

Moreover, observe that conditions (36)-(41) include a set of $n+p+m$ equalities (which are given by the equations (36),(37) and (39)) in $n+p+m+1$ unknowns (n given by x^*, 1 given by λ_0, p given by λ^* and m given by μ^*) (i.e. the $1+p+m$ multipliers $\lambda_0^*, \lambda^*, \mu^*$, plus the n original variables x^*). In other words, there is an additional unknown with respect to the number of equality constraints. To solve this problem, due to the fact that we cannot modify the problem itself, we have to try to see weather the constraints fullfill some properties.

To avoid the latter two drawbacks we can introduce some additional requirements on the constraints (the so called **constraint qualification conditions**), in such a way that the condition

$$\lambda_0 * \neq 0$$

can be fullfilled. As a consequence, in case $\lambda_0 * \neq 0$, we can divide the n equations by λ_0, and re-denominating

$$\lambda^* \leftarrow \lambda^*/\lambda_0^*, \quad \mu^* \leftarrow \mu^*/\lambda_0^*,$$

the conditions (36)-(41) simply become (we recall that now $\lambda_0^* \neq 0$)

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* + \nabla g(x^*)\mu^* = 0,$$

$$h(x^*) = 0, \quad g(x^*) \leq 0,$$

$$g(x^*)^T \mu^* = 0,$$

$$\mu^* \geq 0.$$

Now, we report the next three (among many others) well known **Constraint Qualification Conditions**, each of which ensuring that $\lambda_0^* \neq 0$.

- Linear independency constraint qualification (LICQ): this is one of the most used constraint qualification conditions in the literature, and simply states that in case the gradients of the vectors of the active constraints are linearly independent, the $\lambda_0 * \neq 0$ if the vectors

  $$\{\nabla h_j(x*), \nabla g_i(x*) \qquad s.t. \qquad j, i \in I(x*)\}$$

  are linearly independent.
  To prove the latter result, let by contradiction the vectors in (42) be linearly independent, with $\lambda_0 * = 0$. Then, relation (36) reduces to

  $$\sum_{j=1}^{p} \lambda_j * \nabla h_j(x*) + \sum_{i=1, i \in I(x*)} \mu_i * \nabla g_i(x*) = 0.$$

  As a consequence, the linear independecy of the gradients in the latter equations yields $\lambda * = 0$ and $\mu * = 0$, which gives a contradiction with the assumption $\lambda_0 * = 0$ and condition (40)

- Mangasarian-Fromowitz Constraint Qualification (MFCQ): which is also much used in the literature and represents a weaker condition with respect to LICQ. In partic- ular, we have that MFCQ holds in case the following two conditions are fulfilled:

  - the gradients of the equality contraints at the solution point x* are linearly independet, i.e. the vectors $\{\nabla h_j(x*)\}$ are linearly independent
  - there exist a nonzero vector $d \in R^n$ such that

    $$\nabla g_i(x*)^T d < 0 \qquad \forall i \in I(x*)$$
    $$\nabla h_j(x*)^T d = 0 \qquad \forall j = 1, ..., p$$

- Linearity (or affinity) of Equality constraints and Concavity of Inequality constraints (LECI): which is obviously equivalent to say that for any $\alpha \in [0,1]$ and for any $x, y \in X$

$$h_j(x) = c_j^T x, \qquad c_j \in R^n \qquad \forall j = 1, ..., p$$
$$g_i[\alpha x + (1-\alpha)y] \geq \alpha g_i(x) + (1-\alpha) g_i(y) \qquad \forall i \in I(x*).$$

Proposition (KKT - Optimality Conditions) Given the problem [*], let $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^p$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, with $f(x)$, $h(x)$ and $g(x)$ continuously differentiable in an open set containing X. Let x^* be a local minimum of $f(x)$ on X. If any of the above Constraint Qualification Conditions is satisfied at x^*, then there exists a vector $(\lambda^*, \mu^*) \in \mathbb{R}^p \times \mathbb{R}^m$ such that

$$\nabla f(x^*) + \sum_{j=1}^{p} \lambda_j^* \nabla h_j(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) = 0,$$

$$h_j(x^*) = 0, \quad j = 1, \ldots, p,$$

$$g_i(x^*) \leq 0, \quad i = 1, \ldots, m,$$

$$\mu_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m,$$

$$\mu^* \geq 0.$$

In particular, if LICQ is satisfied at x^*, then there exists a unique vector of generalized Lagrange multipliers $(\lambda^*, \mu^*) \in \mathbb{R}^p \times \mathbb{R}^m$ satisfying the above conditions. The vector (x^*, λ^*, μ^*) is a so called **Karush-Kuhn-Tucker point**.
What we see is that KKT conditions are excactly Fritz-John optimality conditions while one of the constraint qualification condition holds.

Proposition Given the problem (35), let $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^p$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, with $f(x)$, $h(x)$ and $g(x)$ continuously differentiable in an open set containing X. Let $f(x)$ be convex on X, $g_i(x)$ be convex on X, $i = 1, \ldots, m$, and let $h_j(x)$ be linear (affine) $j = 1, \ldots, p$. If there exist vectors $\lambda^* \in \mathbb{R}^p$ and $\mu^* \in \mathbb{R}^m$ such that the next conditions at x^* hold

$$\begin{aligned}
\nabla f(x^*) + \sum_{j=1}^{p} \lambda_j^* \nabla h_j(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) = 0, & \\
h_j(x^*) = 0, & \qquad j = 1, \ldots, p, \\
g_i(x^*) \leq 0, & \qquad i = 1, \ldots, m, \\
\mu_i^* g_i(x^*) = 0, & \qquad i = 1, \ldots, m, \\
\mu^* \geq 0, &
\end{aligned}$$

then x^* is a local (and global) minimum of f [*]. In particular, if $f(x)$ is also strictly convex on X, then x^* is the unique local (global) minimum of [*].
We remark that the hypotheses on h(x) and g(x) in Proposition (6.3) ensure that the set X is convex, which justifies the coincidence of local and global minima, according to proposition 4.6.
To summarize we have that if the objective function is a convex function and the feasible set is a convex set, the KKT are also necessary and sufficient optimality conditions and all the solutions are not just local minima but also global minima.

1.2.4 Convergence to stationary points

We want to study iterative optimization methods, which generate a sequence $\{x_k\}$ of points in R^n. Moreover, we also study when $\{x_k\}$ converges to a stationary point for the problem in hand, or it admits limit points satisfying to some extent the Karush-Kuhn-Tucker conditions. In particular, we want to study convergence regardless of the initial point x_0 of the sequence. On this purpose, we introduce the concept of global convergence, in place of simply speaking about convergence, by means of the next definition.

Definition The sequence $\{x_k\}$, with $x_k \in R^n$, k=0,1,2,... , is globally convergent to the point $\bar{x}$, if for any $x_0 \in R^n$ we have that at least one subsequence of $\{x_k\}$ converges to $\bar{x}$.
Depending on the features of any iterative method we distinguish among four different kinds of global convergence to stationary points:

- finite convergence of $\{x_k\}$ i.e. there exist a finite index $k^* < \infty$ such that $\nabla f(x_{k^*}) = 0$
- convergence of $\{x_k\}$ i.e. we simply have that $\lim_{k\to\infty} x_k = x^*$, with $\nabla f(x^*) = 0$
- convergence of $\{\nabla f(x_k)\}$ i.e. we have $\lim_{k\to\infty} ||\nabla f(x_k)||_2 = 0$
- the sequence $\{\nabla f(x_k)\}$ admits limit points: i.e. we have that the condition $\liminf_{k\to\infty} ||\nabla f(x_k)||_2 = 0$ if fullfilled.

Gradient Methods This subsection reports basic results on the so called Gradient methods, for the solution of (46); further information on the latter methods can be easily found.
Assuming that $f \in C^1(R^n)$ and starting from the point $x \in R^n$, let us consider the scheme for generating the novel point x_α

$$x_\alpha = x - \alpha \nabla f(x), \qquad \alpha \geq 0,$$

i.e. the point x_α is taken moving from the point $x \in R^n$, along the direction $-\nabla f(x)$, with a steplength $\alpha \geq 0$. By the mean value Theorem at x we obtain

$$f[x - \alpha\nabla f(x)] = f(x) + \nabla f(x)^T[x - \alpha\nabla f(x) - x] + o(||\alpha\nabla f(x)||)$$

i.e.

$$f[x - \alpha\nabla f(x)] = f(x) - \alpha||\nabla f(x)||^2 + o(||\alpha\nabla f(x)||),$$

so that if we take α sufficiently small, i.e. such that $\alpha||\nabla f(x)||^2 > o(||\alpha\nabla f(x)||)$, then we obtain

$$f[x - \alpha\nabla f(x)] < f(x).$$

Observe that as proved in Proposition 5.1, as long as $f \in C^1(\mathbb{R}^n)$, $d \in \mathbb{R}^n$ is a descent direction for $f(x)$ at x if and only if $\nabla f(x)^T d < 0$, which is surely satisfied in case $d = -\nabla f(x)$, being indeed at any non-stationary point x

$$\nabla f(x)^T d = -||\nabla f(x)||^2 < 0.$$

Using the above arguments we can now define the following class of iterative methods, known in the literature as **Gradient Methods**, for $k = 0, 1, 2, \ldots$

$$x_{k+1} = x_k + \alpha_k d_k, \quad \nabla f(x_k)^T d_k < 0,$$

where α_k is chosen so that $f(x_{k+1}) < f(x_k)$, $k = 0, 1, 2, \ldots$. More in general, if $D_k \in \mathbb{R}^{n\times n}$ is a positive definite matrix and $d_k = -D_k\nabla f(x_k)$, then we can consider the Generalized Gradient Methods

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k),$$

for which the direction d_k satisfies (at any non-stationary point x_k)

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T D_k \nabla f(x_k) < 0,$$

i.e. d_k is always a descent direction at x_k.

1.3 Some algebric results-14/03/25

Let us consider the function $f : R^n \to R^m$, and suppose it is linear. Now, for any $x \in R^n$ we can write

$$x = x_1 e_1 + ... + x_n e_n$$

where $x_i \in R$, for any i = 1,...,n and e_i is the i-th unit vector, i.e.

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}.$$

For the sake of simplicity in the sequel we indicate with x any point in R^n, while X indicates the corresponding n-tupla of real entries

$$X = (x_1 x_2 ... x_n)^T.$$

This implies that by linearity of f(x) we can write

$$f(x) = f(x_1 e_1 + ... + x_n e_n) = x_1 f(e_1) + ... + x_n f(e_n),$$

where $f(e_i) \in R^m$ is the vector obtained transforming the vector $e_i \in R^n$ by f. Now, introducing the following notation

$$A = [f(e_1) \vdots ... \vdots f(e_n)] \in R^{m \times n},$$

and recalling that for point $x, y \in R^n$ we can write the corresponding n-tuplas as

$$X = (x_1 x_2 ... x_n)^T \qquad\qquad Y = (y_1 y_2 ... y_n)^T$$

we can write the relation $y = f(x)$ as $Y = AX$.
Note that canonical basis is not the only possible basis.

$$x = x_1 e_1 + ... + x_n e_n = \bar{x_1} u_1 + ... + \bar{x_n} u_n, \qquad\qquad for\ \ some\ \ \bar{x_i} \in R, i = 1, ..., n,$$

and we would like to determine the analytical relation (if any) between the n-tuplas

$$X = (x_2 x_2 ... x_n)^T \qquad\qquad \bar{X} = (\bar{x_1} \bar{x_2} ... \bar{x_n}).$$

On this purpose, writing $x = (e_1 \vdots \cdots \vdots e_n)X = (u_1 \vdots \cdots \vdots u_n)\bar{X}$, and recalling that the vectors $\{u_i\}$ can be expressed as a linear combination of the vectors $\{e_i\}$, being

$$u_i = c_{1i} e_1 + \cdots + c_{ni} e_n, \quad i = 1, \ldots, n,$$

we have

$$\begin{aligned} x &= (e_1 \vdots \cdots \vdots e_n)X = x_1 e_1 + \cdots + x_n e_n = \bar{x}_1 u_1 + \cdots + \bar{x}_n u_n \\ &= \bar{x}_1 (c_{11} e_1 + \cdots + c_{n1} e_n) + \cdots + \bar{x}_n (c_{1n} e_1 + \cdots + c_{nn} e_n) \\ &= (e_1 \vdots \cdots \vdots e_n) \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ldots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix} \\ &= (e_1 \vdots \cdots \vdots e_n) C \bar{X}, \end{aligned}$$

having set

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ldots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}.$$

Determinant Square matrix that satisfies:

- $|.| : R^{n \times n} \to R$
- $|A| = |A|^T$, with $A \in R^{n \times n}$ and where A^T idicates the transpose of A.
- $|c.A| = c^n |A|$, with $A \in R^{n \times n}$ and $c \in R$
- if two rows/columns of a square matrix interchange between themselves, then the determinant of this matrix changes it sign
- if all the elements of a row/column of a square matrix are zero, then the determinant of this matrix is equal to zero

Determinant and linear dependence **Proposotion** Given the n vectors $v_1, ..., v_n \in R^n$, they are linearly independent (on R) iff the determinant of the matrix

$$(v_1 ... v - n) \in R^{n \times n}$$

is nonzero.

1.3.1 Eigenvalues and Eigenvectors

An eigenvalue of a square matrix is a scalar value that indicates how the square matrix "stretches" or "shrinks" a vector when multiplied by that square matrix.
An eigenvector of a square matrix is a non-zero vector that, when multiplied by the square matrix, results in a "scaled" version of that matrix.
Formally, let $A \in R^{n \times n}$ be a square matrix representing a linear transformation, and let λ be a scalar. λ is said to be an eigenvalue of A if there exist a non-zero vector $x \in R^n$ such that:

$$Ax = \lambda x$$

The vector x is called eigenvector of the square matrix A corresponding to the eigenvalue λ.

Definition Let $A \in R^{n \times n}$, we say that

$$det(A - \lambda I) = 0$$

is the secular (or characteristics) equation associated with A. As a consequence, $\lambda \in K$ is an eigenvalue of matrix A if and only of λ is a solution of the **secular equation**.
Note that there is no solution in radicals to general polynomial equations of degree five or higher with arbitrary coefficients.
Therefore, in case the secular equation is of degree five or higher, numerical methods must be used to calculate the associated eigenvalues.

Definition Let the square matrix A be symmetric, and consider its (real) eignevalues:

- if the eigenvalues of A are all positive, the A in a positive definite (A $>$ 0) matrix
- if the eigenvalues of A are all nonnegative, the A is a positive semi-definite ($A \geq 0$) matrix
- if the eigenvalues of A have unspecified sign, then A is an indefinite matrix.

2 Statistical Preliminaries

2.1 Introduction -7/03/2025

Our study of statistics starts with some definitions.

Definition A random variable is a function between its domain Ω (space of events) and a space of possible results (sampling space): $f(\omega) = x$.
This function maps the points of the domain to the points of the outcome space where each point of the latter is characterized by a finite probability. The sum of all the probability corresponds to the unity.
Probability is defined through axioms:

- Each probability $0 \leq p_i \leq 1$. This means that by definition a probability is a number included between 0 and 1 (bounds included)
- The probability of observing at least one element in the sampling space is 1, that is P(S) = 1. This means that at least one event on the sampling space in going to happen.
- If two elements A and B are distinct (not overlapping), then P(A or B) = $p_A + p_B$. If we consider the former example of throwing two dices, with $A = (i, j)$ and $B = (i', j')$, $\forall i, i', j, j' \in 1, ..., 6$, then the probability of observing A or B is the sum of the probability of A and the probability of B.

If A and B are independent events, the probability of P(A and B) = P(A)P(B).
Conditional probability is a relevant concept in statistics. It defines probability in case events are not independent.
Two events are not independent, if the probabilit of the event B is influenced by the outcome of another experiment, let's say A. In this case, the probability of observing A together with B is P(A and B) = P(A)P(B given A) = P(A)P(B|A).
This concept is fundamental and represents the rule that defines the probability of two non independents events. Suppose we set a second stage experiment. The admissible outcomes of the second stage reduce the dimension of the sampling space. The unconditional sampling space is the set of all the admissible results before the two stage experiments.
The probability distribution represents a graphical representation of the probability space.

Discrete Distributions The simplest discrete random variable is called Bernoulli, that assumes values 0 or 1, and for which the probability distribution is defined by the function:

$$P(Y = i) = p^Y (1-p)^{1-Y}, i = 0, 1 \text{ or } Y = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability 1-p} \end{cases}$$

This random variable is characterized by the parameter p that is the probability of success (the probability of observing 1). This random variable is really useful to describe populations that are characterized by variables that allow for just two outcomes.
A generalization of the Bernoulli, is the binomial random variable, that describes the number of successes over N experiments. Its probability function is defined as

$$P(Y = i) = \binom{N}{i} p^i (1-p)^{N-i}, \quad i = 0, 1, 2, \ldots, N.$$

This setup describes a sequence of Bernoulli independent experiments and counts the number of successes over N experiments. Here the parameters of interest are the number N and the probability of success on each experiment p.

Another relevant discrete random variable is the Poisson, for which the variable distribution is given by

$$Y \sim P(\lambda) \Rightarrow P(Y = i) = \frac{e^{-\lambda}\lambda^i}{i!}, i = 0, 1, 2, ..$$

This is a count variable, which can take on non-negative integer values $0, 1, 2, ...$ We are especially interested in cases where Y takes on relatively few values, including 0.
The most important random variable in statistic is the Gaussian or Normal, since it well represents the distributions of a large number of variables in the real world. Furthermore, it is important for inference.
The normal distribution is a **continuous**, symmetric, bell shaped distribution, characterized by its mean μ and its standard deviation σ. The density is defined by the following function

$$Y \sim N(\mu, \sigma) \rightarrow \frac{1}{\sqrt{2\pi}\sigma} exp - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}, y \in R.$$

The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to produce values that fall far from its mean. This makes it useful for understanding the statistical behavior of certain type of ratios od random quantities, in which variation in the denominator is amplified and may produce outlying values when the denominator of the ratio falls close to zero.
It is characterized by three parameters, that are μ, σ and ν. The latter one is called degree of freedom and controls the behavior of the tails of the distribution.
If ν is larger than 20-30, then the t distribution and the Normal one are substantially identical.

Continuous random variables have an infinite continuum of possible values. It does not make sense to assign a probability to each event y_i (they are too many). However, probability is assigned to intervals. The graph of continuous probability distribution is in general a smooth function called density and the probability of an event is the area under the curve for an interval of values.

Random variables: moments The expected value E[Y] or μ_Y is the average value of a random variable computed on the basis of the probabilities associated to the elementary events.

In case of discrete random variables that takes values $y_1, ..., y_n$ with probability respectively $p_1, ..., p_n$, this is the weighted average

$$E[Y] = \sum_{i=1}^{k} y_i p_i.$$

In case of continuous random variables we have a sum over an infinity of numbers, and then we need to recur to integrals

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

The variance V(Y) or σ_Y^2 measures the dispersion of the distribution with respect to the mean.
In case of a discrete random variable it is defined as

$$V(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i.$$

For a continuous random variable it is defined as

$$V(Y) = E[(Y - \mu_Y)^2] = \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y) dy.$$

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. Skewness is measured by third order moments, and in particular

$$\text{skewness} = \frac{E[(Y - \mu)^3]}{\sigma^3}.$$

The skewness for a normal distribution in zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.

Kurtosis is a measure of whether the data are peaked or flat relative to normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. It is defined as

$$\text{kurtosis} = \frac{E[(Y - \mu)^4]}{\sigma^4}.$$

Kurtosis of a random variable is 3.
variance of a random variable can be computed as $V(Y) = E[Y^2] - E^2[Y]$.
Lets consider a vector $z = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ then $E[z] = \begin{bmatrix} E[x_1] \\ E[x_2] \end{bmatrix}$.

Properties Given $E[x] = \mu$, let's now consider $z = a + bx$, then, $E[z] = E[a + bx] = a + E[bx] = a + bE[x]$. In the same way $V(z) = E[(z - E[z])^2] = E[(bx - bE[x])^2] = b^2\sigma^2 = b^2V(x)$.

In $\mathbb{R}^2$

$$\mathrm{Cov}(x) = \mathbb{E}\left[\begin{pmatrix} (x_1 - \mathbb{E}x_1)^2 \equiv V(x_1) & (x_1 - \mathbb{E}x_1)(x_2 - \mathbb{E}x_2) \equiv cov(x_1, x_2) \\ (x_1 - \mathbb{E}x_1)(x_2 - \mathbb{E}x_2) \equiv cov(x_2, x_1) & (x_2 - \mathbb{E}x_2)^2 \equiv V(x_2) \end{pmatrix}\right].$$

Suppose considering $y = \alpha + \beta x$, we want to compute

$$cov(x, y) = cov(\alpha + \beta x, x) = cov(\alpha, x) + cov(\beta x, x) = \beta V(x)$$

that means, given $cov(y, x) = \beta V(x)$, $\beta = \frac{cov(x,y)}{V(x)}$ and that means that we can consider covariance as a measure of linear dependency between the two variables since V(x) is always positive.

2.2.1 Properties

Linear Transformation

$$\mathbb{E}(Ax + b) = A\mathbb{E}(x) + b.$$
$$\mathrm{Cov}(Ax + b) = A\mathrm{Cov}(x)A^T.$$

Proof.

$$\begin{aligned}
\mathrm{Cov}(Ax + b) &\overset{(4)}{=} \mathbb{E}\left[(Ax + b) - \mathbb{E}(Ax + b))((Ax + b) - \mathbb{E}(Ax + b))^T\right] \\
&= \mathbb{E}\left[(Ax + b - A\mathbb{E}(x) - b)(Ax + b - A\mathbb{E}(x) - b)^T\right] \\
&= \mathbb{E}\left[A(x - \mathbb{E}(x))(x - \mathbb{E}(x))^T A^T\right] \\
&= A\mathbb{E}\left[(x - \mathbb{E}(x))(x - \mathbb{E}(x))^T\right] A^T \\
&= A\mathrm{Cov}(x)A^T.
\end{aligned}$$

Conditional probability We try to predict the variable y given x, to do so we'll use conditional probability. Suppose having a discrete random variable y such that

	y_1	y_2	y_3
x_1	p_{11}	p_{12}	p_{13}
x_2	p_{21}	p_{22}	p_{23}

where $p_{ij} = P(x = x_i, y = y_j)$.
Suppose studying the variable y knowing that $x = x_1$, we want to use this info to predict the outcome. We consider the conditional probability $P(y|x = x_i) = \sum_{y=1}^{3} P(y_i|x)$.
The probability defined in that way as however the problem that it doesn't sum to one. To get a

probability that sums up to one we have to divide by the probability of getting $x = x_1$. Given that we can conclude that

$$p(y|x = x_1) = \frac{p(y, x_1)}{p(x_1)}.$$

This formula is called **Bayes Rule**.

Suppose having a random variable y, what we want to do now is to sample the statistical distribution of this random variable; for the following example we assume wanting to extrapolate the value of μ_y (centrality parameter). In order to evaluate that parameter what we use is $\frac{1}{n}\sum_{i=1}^{n} y_i$.

Given the strong correlation between the value of μ_y and the number of people composing the sample (n), we concern now on how to build a sample.

Suppose having a population of 4 people (A,B,C,D), we call Y_1 a random outcome given by this set of people. This random outcome has an equal probability of $\frac{1}{4}$ of coming from any of the members of the sample (A,B,C,D) and that means that the statistical distribution of the first value matches the exact probability distribution of the whole population. The same goes for $Y_2, ..., Y_n$.

Two variables Y_1 and Y_2 are independent if $P(Y_1|Y_2) = P(Y_2)$ and in the same way they are dependent if $Cov(Y_1, Y_2) = 0$.

Let's consider now a set of independent variables $Y_1, ..., Y_n$ described by the same probability distribution. All of them are going to have the same variance and same expected value.

The average is going to be

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

where y_i is the characteristic associated with Y_i. This formula represents all the averages that I can obtain from all the possible samples of size n. In statistics the quantity $\bar{y}$ is called estimate. What we want is $\bar{y}$ to be as close as possible to the value of μ_Y. Indeed,

$$E[\bar{Y}] = E[\frac{1}{n}\sum Y_1] = E[\frac{1}{n}(Y_i + ... + Y_n)] = \frac{1}{n}(\underbrace{E[Y_1]}_{\mu_1} + ... + \underbrace{E[Y_n]}_{\mu_n}) = \frac{1}{n}(\mu_1 + ... + \mu_n) = \mu_Y.$$

Recap

Population and Random Variables

- Consider a population of size N, where each individual i has an associated characteristic (e.g., GPA score) denoted by y_i.
- Before sampling, the characteristic of the i-th selected individual is represented by a **random variable** Y_i.
- Since we select individuals randomly, Y_i is unknown before the selection.

Random Sampling and Observations

- We extract a **random sample** of size n from the population.

- Each selected subject has a corresponding realization of the random variable:

$$Y_1 \to y_1$$
$$Y_2 \to y_2$$
$$\vdots$$
$$Y_n \to y_n$$

- Once extracted, the sample is a sequence of observed values $y_1, y_2, \ldots, y_n$.

Estimator and Estimate

- Since we do not know the population mean $\mu_Y = E[Y]$, we use the sample data to **estimate** it.
- A function of the sample values that provides an estimate of a population parameter is called an **estimator**.
- The **sample mean** is a natural **estimator** of the population mean:

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

- This estimator is a **random variable** because different samples yield different values.
- The **estimate** $\bar{y}$ is the specific numerical value obtained from a given sample:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2}$$

Sampling Distribution of the Estimator

- Since the sample mean $\bar{Y}$ is computed from random variables Y_i, it has its own probability distribution, called the **sampling distribution** of the estimator.
- If we take many different samples, we obtain different estimates $\bar{y}$, which together form the sampling distribution of $\bar{Y}$.

Key Terms Recap

- **Estimator**: A function of the sample used to approximate a population parameter (e.g., $\bar{Y}$ estimates μ_Y).
- **Estimate**: The specific value obtained from a sample (e.g., $\bar{y}$, the computed sample mean).
- **Sampling Distribution**: The probability distribution of an estimator over all possible samples of size n.

Thus, the sample mean $\bar{Y}$ is an **unbiased estimator** of the population mean μ_Y, meaning that its expected value equals the true population mean:

$$E[\bar{Y}] = \mu_Y \tag{3}$$

This justifies the use of the sample mean as a good strategy for estimating the population mean.

Averaging over a wide number of samples, what we obtain is that the formula is going to work.

The problem is that in general we don't have a big number of samples, so we can't conclude if the result will be valuable.
The variance will be

$$V(\bar{Y}) = E[(\bar{Y} - \mu_y)^2].$$

$V(\bar{Y}) = V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}V(\sum_{i=1}^{n} Y_i)$.
Knowing that $V(Y_1+Y_2) = V(Y_1)+V(Y_2)+2cov(Y_1,Y_2)$, we get that $V(\bar{Y}) = \frac{1}{n^2}V(Y_1+...+Y_n) = \frac{1}{n^2}(V(Y_1)+...+V(Y_n)) = \frac{1}{n^2}\sigma_Y^2$.
We have just seen how on average $E[\bar{Y}] = \mu_Y$. If I compute $V(\bar{Y}) = \frac{1}{n}\sigma_Y^2$ then if n is large enough $V(\bar{Y}) \to 0$ and we get a strategy that is going to converge with probability 1 to the true parameter. In principle this result is one of the possible versions of the **law of the large numbers**.
Up to now we have derived the mean and the standard deviation of $\bar{Y}$. However, it is possible to describe the whole distribution of $\bar{Y}$. In fact it can be proved that the sampling average is well approximated by a Gaussian random variable. The approximation is accurate when the sample size is large. Furthermore, the approximation does not depend on the population's distribution.

Central Limit theorem For random sampling with a large sample size n, the distribution of the sample mean $\bar{Y}$ is approximately a normal distribution i.e.,

$$\sqrt{n}(\bar{Y} - \mu_Y) \sim N(0, \sigma_Y^2).$$

To estimate the expected value of a given population the sampling average $\bar{Y}$ is an outstanding tool:

- it is unbiased
- it is efficient
- we know that the sampling distribution is normal.

However, in practical applications, a single number dosn't provide a sufficient set of information for inferential purposes. In fact, it is always useful to define a margin of error tha summarizes the degree of uncertainty associated to the estimate.
An interval estimate or confidence interval is an interval of numbers around the point estimate, within which the parameter value is expected to fall.
From what we got before

$$\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \sim N(0, \sigma_Y^2),$$

let's try now to define a confidence interval in which our parameter is going to fall with high probability (confidence level).
To get a better understanding we start by computing

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

what we get is

$$P(z_{-\frac{\alpha}{2}}\sigma_{\bar{Y}} + \bar{Y} \leq \mu_Y \leq z_{\frac{\alpha}{2}}\sigma_{\bar{Y}} + \bar{Y}) = 1 - \alpha.$$

We want now to use this formula to compute an interval in which our variable is going to fall with that given probability. What we get is then:

$$[\bar{Y} - z_{\frac{\alpha}{2}}\sigma_{\bar{Y}}, \bar{Y} + z_{\frac{\alpha}{2}}\sigma_{\bar{Y}}].$$

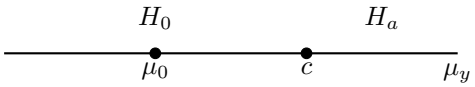
2.2 Statistical methods for business and economics -14/03/25

2.2.1 Hypothesis Testing

In statistics an hypotesis is a statement. The goal is to decide whether there is some empirical evidence against H_0 or not. The data are summarized through a test statistic. The decisions based on the observations could be:

- H_0 is correct $\rightarrow$ Reject H_0 (type 1 error) or Do not reject H_0 (correct decision)
- H_0 is false $\rightarrow$ Reject H_0 (correct decision) or Do not reject H_0 (type 2 error)

The null hypothesis, H_0, is a statement that a parameter takes a particular value whereas the alternative hypotesis H_a states that a parameter falls in an alternative range of values.
We need to find some rules to make a possibly correct decision or, at least, keep under control the probability of making a mistake.
First we consider as test statistic the sampling average $\bar{Y}$. The intuition is to divide the set of all possible μ_Y in two parts. One part collects the values of $\bar{Y}$ consistent with the alternative hypothesis whereas the other collects the values of $\bar{Y}$ that are more likely in case the null hypothesis is right.



The intuition is that, if we observe a really high $\bar{y}$, then the emprical evidence suggests there is a stronger evidence on H_a than on H_0. On the other side, a small value of $\bar{y}$ indicates that H_a is not likely and then there is not enough evidence to reject the null hypothesis.
Wanting to detect μ_0 we compute $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and we expect it to converge to E[y] for the rule of large numbers. If I see that the result is much larger than what expected what explained before will follow.
We want to understand now how to decide the threshold (c).
We start by considering the t ration as

$$t = \frac{\bar{y} - \mu_0}{\sqrt{V(\bar{y})}} = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma_y}$$

If n is large enough that tends to N(0,1). The cental limit theorem is going to tell that

$$\frac{\sqrt{n}(\bar{y}-\mu_y)}{\sigma_y} \to N(0,1)$$

Considering this first case of study, we consider our hypothesis being

$$H_0 : \mu_y = \mu_0,$$

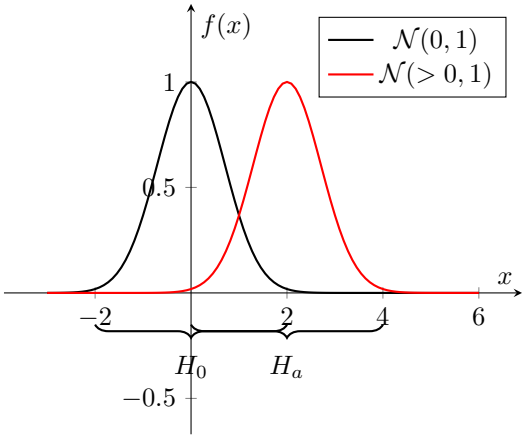
then,

$$t_{H_0} = \frac{\sqrt{n}(\bar{y}-\mu_y)}{\sigma_y} \to N(0,1).$$

If we now compute the t-ration under the alternative hypothesis then the previous result is not gonna be valid anymore

$$t_{H_a} = \frac{\sqrt{n}(\bar{y}-\mu_y)}{\sigma_y} \to N(>0,1)$$

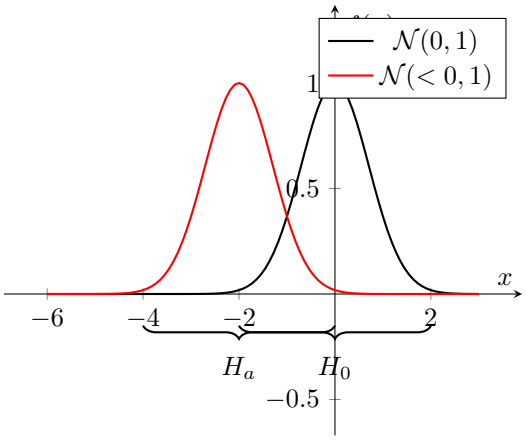
that is because $\mu_y \geq \mu_0$, and indeed, if $\mu_y \geq \mu_0$ then $\bar{y} - \mu_0 > 0$ and $E[\mu_y] = 0$.



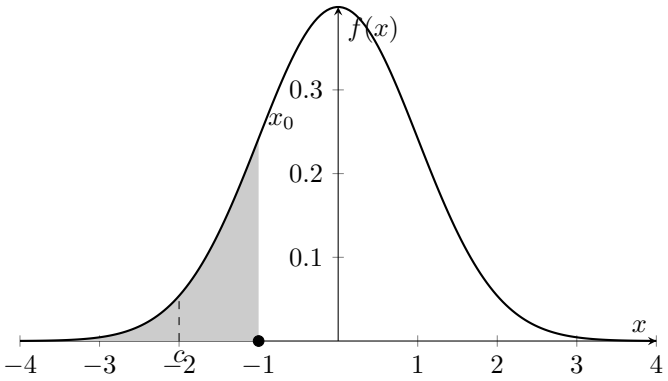
Considering where the empirical evidence is going to fall is going to determine which of the two hypothesis is to be considered. But what if the empirical falls in part in common of the two graphs? In this case we have to impose a threshold such that I can take a definitive decision. The threshold is to be defined such that $P(t > c) = 0.05$.
That goes for $H_0 : \mu_y = \mu_0 H_a : \mu_y > \mu_0$.
In case $H_0 : \mu_y = \mu_0; H_a : \mu_y < \mu_0$, then

$$\bar{y} - \mu_0 < 0$$

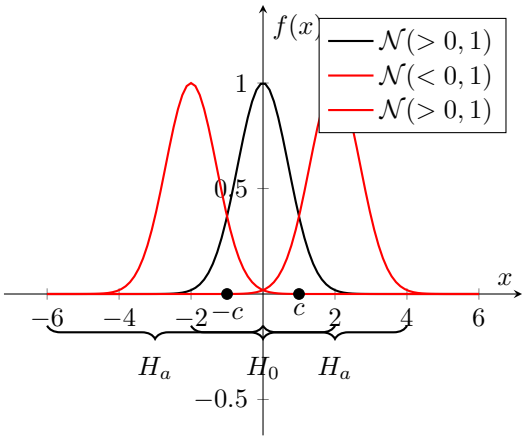
under H_0.



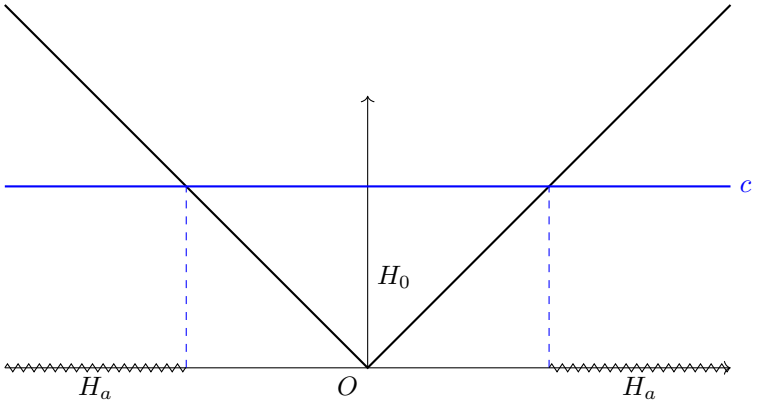
The problem is the same as before; we have to find a threshold with the same characteristics as before.
Given an empirical evidence, we want to compute the probability of finding events more extreme compared to the one witnessed (that is how the empirical evidence is far away from the alternative hypothesis).



The area under the curve that goes from the event to the extreme limit of the distribution is the p-value. For events very far from the mean the p-value gets very small values. In the first example, the p-value is computed on the right while for the second one on the left, and in general that depends on the test we are considering.
As for the last test we consider a mix of the two previous cases $H_0 : \mu_y = \mu_0 \;\; H_a : \mu_y \neq \mu_0$.



Since we can make mistakes in two directions we have to find two different thresholds. Given α as the overall probability, if for example $\alpha = 5\%$, then $-c = -1,96$ and $c = 1,96$.
The last thing regarding testing is how to compute the p-value in the case one has to reject both extremes. The idea is that instead of computing the t-statistics and looking at the properties of the t-statistic, lets compute $|t|$.



We see how for absolute values of t larger than c corresponds very large values of t or very small values of t, both of them corresponding to the alternative hypothesis domain.
Let's consider $P(|t| \geq |t^{obs}|) = P(t \leq -|t^{obs}| \text{ or } t \geq |t^{obs}|)$ and that is equal to

$$P(t \leq -|t^{obs}|) + P(t \geq |t^{obs}|)$$

and because of the symmetry this two probability are going to be the same

$$= 2P(t \leq -|t^{obs}|) = 2P(t \geq |t^{obs}|).$$

If the p-value is larger than α than the hypothesis is not rejected, while if it is smaller the hypothesis is rejected.

2.2.2 Statistical learning and classification

The Bayesian Logic We start by studying a specific case in which P(H) = 0.99 and P(D) = 0.01. If the subject is H the probability of the test P(test|H) = $\begin{cases} 90.4\% \\ 9.6\% \end{cases}$ and if the subject is D we got P(test|D) = $\begin{cases} 80\% \\ 20\% \end{cases}$ and given

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

the possible outcomes of the test P(H|test) are

–	positive	negative
disease	0.8%	0.2%
healty	9.5%	89.5%

P(D|positive) = $\frac{\text{favorable cases}}{\text{total cases}}$ = 7,76. %.
This outcome is called posterior probability.

Definitions In a Bayesian setup there are two different probabilities to summarize uncertainty:

- A first amount of uncertainty describe just the proposition, independently by the empirical evidence provided by E. It is the prior probability (in our case P(H) and P(D)).
- The second level of uncertainty refers to the reliability of the empirical evidence E, and it is called likelihood. In a nutshell, the likelihood summarize how the empirical evidence is consistent with the hypothesis E (in our case P(test|D) and P(test|H)).

Combining prior and likelihood, we get the posterior probability through the Bayes' formula

$$P(H_i|E_i) = \frac{P(H_i, E_i)}{P(E_i)} = \frac{P(E_j|H_i)P(H_i)}{\sum_k P(H_k)P(E_j|H_k)} \qquad \forall i, j.$$

This probability describes the uncertainty on H_i and how this uncertainty is updated once observed the event E_j.
Suppose now we get two random variables X and Y and we aim at checking how information on X can help predicting the behavior of Y. For instance, we might look at the probability Y takes some values.
If X in not observed, the best we can do is to evaluate P(Y=y), maybe using some naive estimator. However, knowing some extra information, say X=x, we might obtain improved estimates on the

probability to observe Y=y, in the sense the knowledge if X can help on removing uncertainty regarding Y. Therefore, we might be interested to compute, for whatever possible value of X and Y

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(P(Y))}{\int P(y)P(X|y)dy}.$$

A related concept is to understand how expected behavior of given variable change in correspondence of changes of some predictors X.
Without knowing X, the average behavior of Y is summarized by E[Y], that can be estimated through a sampling average $\bar{Y}$.
Let's consider the following case:

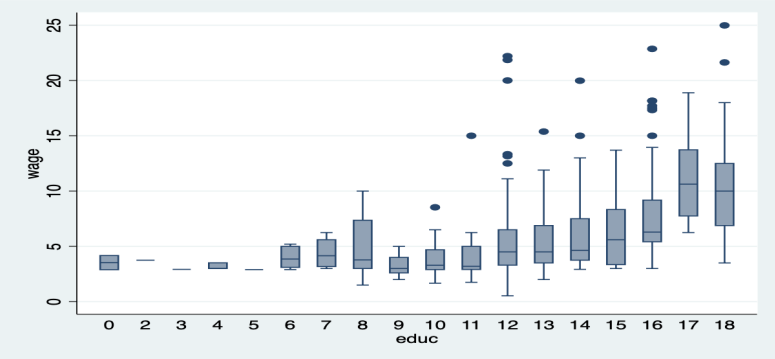


Figure 1: Distributions of salaries over (conditional) the number of years of education. How expected salaries changes as a function of X?

From the following data we get that the average salary is of 5.89 dollars per hour. It does not seem a good centrality indicator for too many subjects in the population. For instance, workers with higher level of education are not well represented by the average of 5.89$.
In might be convenient then to update our knowledge on the expected Y once we get information about X. We aim at computing E[Y|X], for whatever different level of X.
The conditional expectation is the average of Y computed from the conditional probability P(Y|X). Often this conditional expectation is approximated by the analytical function

$$E(Y|X) \sim f(X)$$

The function f(.) is sometimes called **statistical learning**.

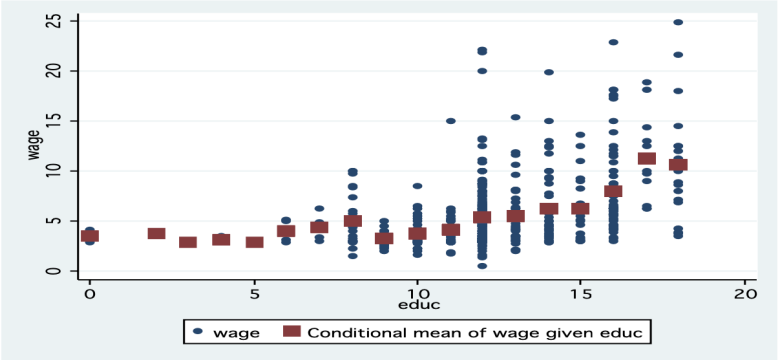


Figure 3: Conditional averages (red squares $\approx E[Y|X]$).

No matter what assumptions we make, E[Y|X] is still the best predictor of Y since it minimizes the mean squared error $E(Y - f(X))^2$, for function f(.).
Note that, Y-f(X) = Y- E[Y|X]+E[Y|X] - f(X). Therefore, the mean squared error (MSE) is

$$E[(Y - f(x))^2] = ... \geq E[(Y - E[Y|X])^2]$$

So, $f(X) = E[Y|X]$ is the optimal predictor under the MSE rule.
In principle f(x) can be whatever: $Y = f(X) + \epsilon$ where ϵ represents how Y is deviating from the conditional expectation and its formally a general random error.
Here f is some fixed but unknown function of $X_1, ..., X_p$ and ϵ is a random error term, which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about Y.
When the variable Y is quantitative, its prediction through f is referred as regression, whereas if is qualitative or categorical prediction referes to classification.
There are two main reasons that we may wish to estimate f: prediction and inference.
In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In the setting, since the error term averages zero, we can predict Y using

$$\hat{Y} = \hat{f(X)}$$

where $\hat{f(X)}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.
The prediction error is:

$$E[(Y - \hat{Y})^2|X] = E[f(X) - \hat{f(X)}]^2 + Var(\epsilon|X)$$

where the first term is reducible and the second is irreducible.
Let f(x) be $f(x) = E[Y|X] = \alpha + \beta x$ on the population. We want to find a way to estimate α and β. In principle we can use $\hat{\alpha}$ and $\hat{\beta}$ and for n large enough we know that those two converge to α and β. What we can prove, indeed, is that $(\alpha - \hat{\alpha}) \to 0$ and the same goes for β.
Parametric methods involve a two-step model-based approach. First, we make an assumption about the functional form, or shape, of f:

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

With parametric models, even nonlinearities are possible:

$$f(X) = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}.$$

Full knowledge of f is obtained once we find estimates for the parameter $\beta_0, ..., \beta_p$.
Non-parametric methods do not explicitly assume a parametric form for f(X), and thereby provide an alternative and more flexible approach for performing regression.

To provide an example, the N-Nearest Neighbors (KNN) works as follows: given a value for K and a prediction point x_0, the KNN regression first indicates the K training observations that are closest to x_0, represented by N_0. It then estimates $f(x_0)$ using the average of all the training responses in N_0, or

$$\hat{f(x)} = \frac{1}{K} \sum_{x_i \in N_0} y_i.$$

Suppose that I got a dataset and I need to predict y for a given $x = x_0$. We start by considering a neighborhood of x_0, then I m gonna average with respect of all these numbers. The neighborhood is to be decided in the way that an exact number of k observation are going to be included in the interval.
Another popular way to fit some data is through spline functions, that in a nutshell are regressions with predictors $b_j(x_j)$ such as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + ... + \beta_k b_k(x_i) + \epsilon_i.$$

They involve dividing the range of X into K distinct regions. Within each region, a polynomial function is fit to the data. However, these polynomials are constrained so that they join smoothly at the region boundaries, or knots. Provided that the interval is divided into enough regions, this can produce an extremely flexible fit.
In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data. In the regression setting, the most commonly used measure is the mean squared error (MSE), given by

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f(x_i)})^2$$

The MSE is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE. But in general, we do not really care how well the method works training on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

3 Introduction to Data Protection Regulation

3.1 Introduction-21/03/2025

Managing accounts means having a collection of personal data.
In the last 20 years we've assisted to a massive digitalization which imposed a so called datafication (i.e. automated collection of data: ai, machine learning, etc.).

Personal Data this is composed of any type of knowledge based on personal attributes. Personal Data relates only to natural person (human being), not to animals, places, legal subjects (e.g. companies, foundations, etc.).
Data should be considered as source of information. Information is what is conveyed or represented by data.

3.1.1 Privacy or Data Protection

The concept of privacy is related to a more personal sphere, it was first introduced in the 20s and was related to the articles in the newspaper regarding personal stuff of the people (aka gossip) while the concept of data protection was first introduced about 40 years ago and was related to the usage of personal data.

Charter of Fundamental Rights of the EU

- **article 7:** Everyone has the right to respect for his or her private and family life, home and communications.
- **article 8:**
  1. Everyone has the right to the protection of personal data concerning him or her.
  2. Such data must be processed for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it recified
  3. Compliance whit these rules shall be subject to control by independent authority

Data Regulation (in Europe) at first there was only General Data Protection and Regulation (GDPR). **Art 1 GDPR:** this regulation lays down rules relating to the protection...
The "ambiguity" of personal data: any information relating to an identified or identifiable natural person (i.e. having control over your own data is an expression of a fundamental right recognized to individuals and also personal data are collected and used in order to perform operations for different purposes).

Personal data is referred to data subject (aka natural person). From the point of view of the nature of the information, the concept of personal data includes any sort of statements about a person. It covers "objective" information. It also include "subjective" information. Moreover, data information it is not necessary be or to be proved.
From the point of view of the content of the information, the concept of personal data includes data providing sort of information.
Considering the format or the medium on which that information is contained, the concept of personal data includes information available in whatever form.

Identified or Identifiable while talking about identified we mean "distinguished" from other members of the group (e.g. passport copy, mail, ecc) as while talking about identifiable data we mean data which at first glance cannot identify directly a person (e.g. birth data).

Anonymous data To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.
The pro is that the GDPR does not apply to anonymous/anonymized data.
Randomization is a family of techniques that alters the veracity of the data in order to remove the strong link between the data and the individual.
Generalization consists of generalizing, or diluting, the attributes of data subjects by modifying the respective scale or order of magnitude.

Pseudonymised data Despite of "anonymization" for "pseudonymised" is meant that the personal data can no longer be attributed to a specific data subject without the use of additional information. Additional information is kept separately and is subject to technical and organizational measures.

Data processing is any operation or set of operations which is performed on personal data or on set of personal data.

Data controller natural or legal person, public authority, which, alone or jointly, determines the purposes and means of processing personal data.

Data processor the activities entrusted to a processor may be limited to a specific task or context or may be quite general and comprehensive. The processor shall not engage another processor without prior specific or general written authorization of the controller.
Recipients and third paries Recipient is a natural or legal person, public authority, or agency to which the personal data are disclosed, whether a third party or not. The third party is a natural

or legal person, public authority, or agency who, under the authority of the controller or processor, are authorized to process personal data.

Material scope of application GDPR does not apply to data processing.

3.1.2 Type of consent

Informed Informed consent will usually comprise a precise and easily understandable description of the subject matter requiring consent. The data subject should be aware at least of the identity of the controller, the purposes of the processing for which the personal data, what type of data will be collected and used and the existence of the right to withdraw consent.

Specific Purpose specification as a safeguard against "function creep" → if a controller processes data based on consent and wishes to process the data for another purpose, than the controller needs to seel additional consent.
Granularity in consent requests for each purpose for each purpose and for each processing.
Clear separation of information related to the request for consent for data processing from information about other matters.
Consent is likely to degrade over time, but how long it lasts will depend on the context.

Unambiguous indication of the data subject's wishes The data subject must have taken a deliberate action or physical motions to consent.
Consent could be obtained through a written statement, filling an electronic form or recorded oral statement.
The usage of pre-ticked or opt-out boxes are considered invalid.

Example of invalid consent

- in case of doubts
- in case of not clear records
- no genuine free choice
- etc

Another important consideration is that the consent in an alternative legal basis for data processing (not the main one). The request shall be presented in a manner which is clearly distinguishable form the other matters, in an intelligible and easily accessible form, using clear and plain language.

Legitimate interest assessment (LIA)

- data disclosed must pursue a legitimate interest
- the processing of personal data must be necessary for the purposes of the legitimate interests pursued
- the fundamental rights and freedom of the data subject must not take precedence over the controller's or third parties' legitimate interest.

Special categories of personal data are regulated in art 9 of GDPR. In principle, processing of "sensitive data" is forbidden. However, several conditions may render the processing to be lawful if:

- explicit content
- employment
- vital interests
- not-for-profit bodies with political,philosophical,religious or trade union purposes
- data manifestly made public by the data subject
- legal claims or judical acts
- etc.

Principles relating to processing of personal data

- Lawfulness, fairness and transparency.
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity/security
- Accountability

Privacy notice completely free and without any cost. Information provided in a concise, transparent, intelligible and easily accessible form. Using clear and plain language.

Where personal data have been obtained from the data subject

- the identity and the contact details of the controller
- the period for which the data will be strored
- etc.

3.1.3 Cookies and other tracking tool

Directive 2002/EC on the processing of the personal data and the protection of privacy in the electronic communications sector and also GDPR (in particular, personal data and consent).
Information encoded in cookies may include personal data, such as an IP address, a username, a unique identifier or an email address.
There are different type of "cookies":

- first party/third party
- session/persistent
- techincal
- analytical
- profiling

The problem with cookies is that they are not regulated by GDPR. Profiling cookies do usually contain personal data. Consent can only be considered to have been validly given if the result of an affirmative action by the user and if that action can be appropriately identified and demonstrated so that the consent can be in line with all the requirements set up by the GDPR.

Purpose of processing

- any processing of personal data must be done for a specific and well-defined purpose
- the processing of personal data without a certain purpose, just based on the consideration they may be useful sometime in the future, is also not lawful
- the controller should strictly limit data collection
- legitimate processing is limited to its initially specified purpose and any new purpose of processing will require a separate new legal basis

Data subject rights

Right of access

Right of erasure Everyone has the right to erasure in case of inaccurate, false or unlawfully processed data. The GDPR does make a clear list of all the specific cases. Data erasure should be obtained at the request and the controller shall have the obligation to erase personal data without undue delay. When data have been transferred, the controller shall take reasonable steps to inform controllers which are processing the personal data.

ECJ case: Google Spain (2014) The right to be forgotten may be invoked where information relating to an individual is inaccurate, inadequate, irrelevant or excessive for the data processing purposes. [...]

ECJ case: Manni (2017) Third parties should thus have access and be able to examine the basic documents of a company and other information concerning the company, especially particulars of the persons who govern the company. The court further noted that even after the passage of time, and even after a company is dissolved, rights and legal obligations related to the company often continue to exist. Due to the legitimate aim of the disclosure and the difficulties in establishing a period [...]

Right to data portability

Right to object The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on public interest or legitimate interest. The controller shall stop or prevent for the processing of personal data.

3.1.4 Accountability vs responsability vs liability

Art 24 (responsability),5 (accountability) and 82 (liability) of the GDPR.

Risk-based approach

- nature
- scope
- context
- purpose of processing
- cyber-sicurity

- costs

The controller and the processor shall implement appropriate technical measures.

Data protection by design vs data protection by default

Data breach notification In case of a personal data breach the controller shall inform the data protection authorities.

Data protection impact assessment It helps to identify and minimize the data protection risks of any new data processing likely to result in high risk to individuals' interests.

Data protection officer (DPO) The controller and the processor shall designate a data protection officer accounted to deal with specific situations regarding data protection activities.