MASTER IMALIS - ENS PSL

Training in Mathematics and Statistics

September 2020

Benoît Perez-Lamarque - benoit.perez@ens.psl.eu

Antoine Sicard - antoine.sicard@ens.psl.eu

Planning

Lecture 1: Few revisions

1.1 Sets

1.1.1 Common sets

By convention, the following symbols are reserved for the most common sets of numbers:

$\varnothing$ – empty set;

$\mathbb{N}$ – natural numbers, $\mathbb{N} = \{0, 1, 2, ...\}$;

$\mathbb{Z}$ – integers, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$;

$\mathbb{Q}$ – rational numbers (from quotient), $\mathbb{Q} = \left\{\frac{p}{q}, p \in \mathbb{Z}, q \in \mathbb{N}^*\right\}$;

$\mathbb{R}$ – real numbers, $\mathbb{R} = \{a_1 a_2 \dots a_p.a_{p+1}..., \forall i \in \mathbb{N}, a_i \in \mathbb{N}, p \in \mathbb{N}^*\}$;

$\mathbb{C}$ – complex numbers, $\mathbb{C} = \{\alpha + i\beta, (\alpha, \beta) \in \mathbb{R}^2\}$. α (resp. β) is referred to as the real part (resp. the imaginary part), and the imaginary unit i is defined by its property $i^2 = -1$.

1.1.2 Product of sets

Let E and F be two sets:

— $E \times F = \{(x, y), x \in E, y \in F\}$;

— $E \times E = E^2$ is the set of all couples of E;

— $E \times \dots \times E = E^n$ is the set of n-tuple of E.

1.2 Functional analysis

1.2.1 Asymptotic notation

Let f and g be two functions in the neighbourhood of a, such as g is not equal to 0 in the neighbourhood of a.

The function f is **negligible** with respect to g in the neighbourhood of a, if $\lim\limits_{x \to a} \frac{f(x)}{g(x)} = 0$, and f is denoted: $f = \circ(g)$ (called *little-o*).

In other words, $f(x)/g(x)$ tends to zero as x tends to a and the limit of f/g at a is zero.

1.2.2 Continuity

A function $f : E \to \mathbb{R}$ is **continuous** at $x_0 \in E$ if $\lim\limits_{x \to x_0} f(x) = f(x_0)$.

To go further, f is continuous at x_0 if, for $\epsilon \to 0$, $f(x_0 + \epsilon) = f(x_0) + \circ(1)$.

1.2.3 Derivability

A function f is **differentiable** at $x_0 \in E$ if $\dfrac{f(x) - f(x_0)}{x - x_0}$ has a limit when $x \to x_0$. This limit is referred to as the **derivative** of f at x_0, denoted $f'(x_0)$.

Other notation: $f' = \dfrac{df}{dx}$.

If $f(x, y)$ is a function of several variables (x and y), the **partial derivatives** of f are the derivatives of f with respect to one of its variables (either x or y), denoted:

$$\frac{\partial f(x,y)}{\partial x} \text{ or } \frac{\partial f(x,y)}{\partial y}$$

Common derivative:

Let $c \in \mathbb{R}$ be a constant , $\forall x \in \mathbb{R}$:

$f(x) = c$ has for derivative $f'(x) = 0$;

$f(x) = cx$ has for derivative $f'(x) = c$;

$\forall x \in \mathbb{R}$, $\forall n \in \mathbb{N}$, $f(x) = cx^n$ has for derivative $f'(x) = cnx^{n-1}$;

$\forall x \in \mathbb{R}^*$, $\forall \alpha \in \mathbb{Z}$, $f(x) = cx^\alpha$ has for derivative $f'(x) = c\alpha x^{\alpha-1}$ (and so $f(x) = x^{-1} = \dfrac{1}{x}$ has for derivative $\dfrac{-1}{x^2}$);

$\forall x \in \mathbb{R}_+^*$, $\forall \alpha \in \mathbb{R}$, $f(x) = cx^\alpha$ has for derivative $f'(x) = c\alpha x^{\alpha-1}$ (and so $f(x) = x^{1/2} = \sqrt{x}$ has for derivative $\dfrac{1}{2\sqrt{x}}$);

$f(x) = e^{cx}$ has for derivative $f'(x) = ce^{cx}$;

$\forall x \in \mathbb{R}_+^*$, $f(x) = \ln(x)$ has for derivative $f'(x) = \dfrac{1}{x}$.

$\forall\ a$ a constant $\in \mathbb{R}_+^*, \forall x \in \mathbb{R}$, $f(x) = a^x$ has for derivative $f'(x) = a^x \ln(a)$.

Operations on derivative: Let $c \in \mathbb{R}$ be a constant and f and g two functions :

— scalar multiplication: $(cf)' = cf'$;

— sum of two functions: $(f+g)' = f' + g'$;

— product of two functions: $(fg)' = f'g + fg'$;

— function composition: $(f \circ g)' = g'\ f' \circ g$;

— inverse function: $\left(\frac{1}{f}\right)' = -\left(\frac{-f'}{f^2}\right)$

— quotient of two functions: $\left(\frac{f}{g}\right)' = \left(\frac{f'g - fg'}{g^2}\right)$.

1.2.4 Bijectivity

A function $f : E \to F$ is **injective** for all a and b in E, if and only if, $f(a) = f(b)$ implies $a = b$.

A function $f : E \to F$ is **surjective**, if and only if, for every element $y \in F$, there is at least one element $x \in E$ such that $f(x) = y$.

A function $f : E \to F$ is **bijective** (or one-to-one correspondence), if and only if, f is injective and surjective at the same time, *i.e.* every $y \in F$ has a unique counterimage with f:

$$\forall y \in F, \exists! x \in E, f(x) = y$$

If f is bijective, one can define a function g that associates to every $y \in F$ its counterimage with f. It verifies $g \circ f = IdE$ and $f \circ g = IdF$, where IdE and IdF represent the identity function: $\forall x \in E,\ g \circ f(x) = x$ and $\forall y \in F,\ f \circ g(y) = y$).

g is called **inverse function** of f, $g = f^{-1}$.

1.2.5 Differential equation

A **differential equation** is an equation involving an unknown function f and at least one of its derivatives (f', f'',...). If the unknown function f only involves derivatives with respect to one variable, then the differential equation is called an **ordinary differential equation** (ODE).

For example, $\forall (a, b) \in \mathbb{R}$, the differential equation of first order $f' + af = b$ has for set of solutions the functions defined by:

$$\forall \lambda \in \mathbb{R},\ \forall x \in \mathbb{R},\ f(x) = \lambda e^{-ax} + \frac{b}{a}$$

The value of the arbitrary constant λ can be found by assuming particular conditions (e.g. initial conditions).

If the unknown function involves derivatives with respect to two or more variables (x, y, ...), then the differential equation is called a **partial differential equation** (PDE).

1.3 Matrix

1.3.1 Definitions

— A **matrix** is any rectangular array of numbers. If the array has n rows and m columns, then it is an $n \times m$ matrix, denoted $A_{n,m}$. One dimensional matrices are called row vectors for a $1 \times m$ matrix or column vectors for a $n \times 1$ matrix. One uses the notation $(a_{i,j})$ to refer to the number in the $i-th$ row and $j-th$ column. If $n = m$, $A_{n,m} = A_{n,n} = A_n$ is called a **square matrix**.

— The zero matrix or null matrix is a matrix with all its elements equal to zero, denoted $0_{n,m}$.

— The **identity matrix** is a square matrix with ones on the main diagonal and zeros elsewhere, called I_n. The identity matrix is neutral with regard to products: $\forall A_n$, $A \times I_n = I_n \times A = A$.

— The **trace**, called tr(A), of a square matrix A is the sum of its diagonal elements.

1.3.2 Matrix operation

— The **transpose** of a matrix flips a matrix $A = [a_{i,j}]$ over its diagonal: it switches the row and column indices of the matrix and gives another matrix denoted as ${}^tA$ (also called A', A^{tr}, or A^T): ${}^tA = [a_{j,i}]$.

— The matrix addition is the operation of adding two matrices of the same dimensions, $A_{n,m}$ and $B_{n,m}$, by adding the corresponding elements together.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}$$

— The multiplication by a scalar λ: $\lambda(a_{i,j}) = (\lambda a_{i,j})$.

$$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}$$

— The matrix product : we can only multiply two matrices together if the number of columns of the first matrix equals the number of rows of the second matrix.
Let $A_{n,m}$ and $B_{m,p}$ be two matrices: $A_{n,m}B_{m,p}$ exists but $B_{m,p}A_{n,m}$ does not exist if $n \neq p$.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

Some properties on the matrix product:

Let A, B, and C be three matrices (such that their products exist), and μ and λ two scalars :

i) $AB \neq BA$ in general: the matrix product is not commutative;

ii) $\lambda(AB) = (\lambda A)B = A(\lambda B)$: the matrix product is associative;

iii) ${}^t(AB) = {}^tB\,{}^tA$

iv) $A(B+C) = AB + AC$ and $(A+B)C = AC + BC$.

v) $AB = 0$ does not imply $A = 0$ or $B = 0$. Moreover, $AC = BC$ does not imply $A = B$.

1.3.3 Determinant of a square matrix

The **determinant** is a value that can be computed from the elements of a square matrix A_n, denoted $\det(A) = |A|$.

For $n = 2$, if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $\det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$.

If $n > 2$, the determinant is defined recursively using the Laplace formula with regard to a row or a column and using cofactors. For example, if $n = 3$:

$$\det(A) = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} \oslash & \oslash & \oslash \\ \oslash & e & f \\ \oslash & h & i \end{vmatrix} - b \begin{vmatrix} \oslash & \oslash & \oslash \\ d & \oslash & f \\ g & \oslash & i \end{vmatrix} + c \begin{vmatrix} \oslash & \oslash & \oslash \\ d & e & \oslash \\ g & h & \oslash \end{vmatrix}$$

$$= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = a(ei - hf) - b(di - gf) + c(dh - ge)$$

For a triangular matrix, its determinant is the product of its diagonal elements.

1.4 Counting

The **cardinality** of a set E, called $card(E)$ is the number of elements of the set E.

$\forall n \in \mathbb{N}$, the **number of permutations** of the n elements, denoted $\boldsymbol{n!}$ (and called *n-factorial*), is defined as:

$$n! = \begin{cases} 1 \times 2 \times \ldots \times (n-1) \times n & \text{if } n > 0 \\ 1 & \text{if } n = 0. \end{cases}$$

An **arrangement** is an ordered subset of k elements among n. The **number of arrangement** A_n^k of k elements among n is defined as:

$$A_n^k = \frac{n!}{(n-k)!}$$

A **combination** is a (unordered) subset of k elements among n. The **number of combination** C_n^k is defined as:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

1.5 Discrete probability

1.5.1 Probability space

Let's assume a randomized experiment (when the outcome is not deterministic, but the probability of each event is known) defined by a **probability space $(\boldsymbol{\Omega, P})$**:

— Ω is the set of all possible outcomes, called **sample space**.

— P is the **probability distribution** associated to the outcomes of the experiment. P verifies:

$$\begin{cases} \forall x \in \Omega, P(x) \in [0,1] \\ \\ \sum\limits_{x \in \Omega} P(x) = 1 \end{cases}$$

An **event** E is a subset of Ω and verifies: $P(E) = \sum\limits_{x \in E} P(x)$

If all events of Ω are elementary events (i.e. all events are equiprobable), then $\forall E \in \Omega$:

$$P(E) = \frac{card(E)}{card(\Omega)}$$

Let (Ω, P) be a probability space and A and B two events from this space:

(i) $P(A) \in [0,1]$;

(ii) $P(\varnothing) = 0$ and $P(\Omega) = 1$;

(iii) The **complementary event** of A, denoted $\overline{A}$ or A^c, verifies: $P(\overline{A}) = 1 - P(A)$;

(iv) The probability of having A and B is denoted $P(A \cap B)$;

(v) The probability of having A or B is: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

(vi) The events A and B are **incompatible** if and only if $A \cap B = \varnothing$. Then, $P(A \cup B) = P(A) + P(B)$.

1.5.2 Conditional probability and independence

A. Conditional probability

Given a probability space (Ω, P) and two events A and B with $P(B) \neq 0$. The conditional probability of A given B, denoted $P(A|B)$ or $P_B(A)$, is defined by:

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

Consequently, $P(A \cap B) = P(A|B)P(B)$

One can deduce:

(i) the **Bayes' theorem**:

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

(ii) the **law of total probability**:

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) = P(A|B)P(B) + P(A \cap \overline{B})P(\overline{B})$$

B. Independence

Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

Similarly, if $P(B) \neq 0$, A and B are independent if and only if $P(A|B) = P(A)$.

1.6 Taylor series

The Taylor series of a function is a series expansion of the function in the neighbourhood of a point. For example, the Taylor series of a function $f(x)$ around a certain value a is

$$f(x) = f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + ... + \frac{f^n(a)(x-a)^n}{n!}$$

The Taylor series is very useful to approximate a complex function around a certain point and is often used in the analysis of non-linear biological system.

1.7 Other revisions

— $\forall\ (a,b) \in \mathbb{R}^2$, $(a+b)^2 = a^2 + 2ab + b^2$, and $a^2 - b^2 = (a-b)(a+b)$.

— $\forall\ (a_1, \ldots, a_n) \in \mathbb{R}^n$, $(a_1 + \ldots + a_n)^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} a_i a_j$

— Two vectors $v_1 = (x, y)$ and $v_2 = (x', y')$ are collinear if $\exists\ a \in \mathbb{R}, v_1 = a v_2$ that is to say, $xy' + yx' = 0$;

— $\forall\ \theta \in \mathbb{R}$, $\cos(\theta) + i \sin(\theta) = e^{i\theta}$.

— $f : \mathbb{R} \to \mathbb{R}$ is an even function if and only if $\forall\ x \in \mathbb{R},\ f(-x) = f(x)$.

— $f : \mathbb{R} \to \mathbb{R}$ is an odd function if and only if $\forall\ x \in \mathbb{R},\ f(-x) = -f(x)$.

Lecture 2: Elementary linear algebra

2.1 Linear map and matrix

2.1.1 Linear map

A **linear map** $f : E \to F$ is a mapping that preserves the operations of addition and scalar multiplication:

$$\forall\ (\mathbf{x}, \mathbf{y}) \in E,\ \forall\ \lambda \in \mathbb{R} \begin{cases} f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) \\ \\ f(\lambda\mathbf{x}) = \lambda f(\mathbf{x}) \end{cases};$$

Any vector $\mathbf{x}$ in E can be represented as $c_0\ \mathbf{b_0}+c_1\ \mathbf{b_1}+c_2\ \mathbf{b_2}+.....+c_n\ \mathbf{b_n}$ where $c_0, c_1, ..., c_n$ are the coefficients and $\mathcal{B} = (\mathbf{b_0}, \mathbf{b_1}, ..., \mathbf{b_n})$ is a basis for E. Therefore, $dim(E) = n = dim(\mathbb{R}^n)$.

The **canonical basis** (or standard basis), denoted $\mathcal{C}$, is the set of unit vectors pointing in the direction of the axes of a Cartesian coordinate system. For $n = 2$, the canonical basis is $\mathbf{e_1}$=(1,0) and $\mathbf{e_2} = (0, 1)$. For $n = 3$, the canonical basis is $\mathbf{e_1} = (1, 0, 0)$, $\mathbf{e_2} = (0, 1, 0)$, and $\mathbf{e_3} = (0, 0, 1)$.

2.1.2 Matrix representation of a linear map

A linear map from E to F can always be represented by a matrix. Reciprocally, one can associate a unique linear map to any matrix. If A is a real $m \times n$ matrix, then $f(\mathbf{x}) = A\ \mathbf{x}$ describes a linear map $\mathbb{R}^n \to \mathbb{R}^m$.

For instance, a linear map

$$g : \begin{cases} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \longmapsto & \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix} \end{cases} \text{ can be represented by the matrix } A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

A is called **canonical matrix** denoted $Mat(g)$. Generally, the canonical matrix of a $\mathbb{R}^n$ linear map is a unique $n \times n$ array.

For example, $h : \begin{cases} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \longmapsto & \begin{pmatrix} x_1 + 3x_2 \\ -2x_2 \end{pmatrix} \end{cases}$ can be represented by the matrix $A = \begin{pmatrix} 1 & 3 \\ 0 & -2 \end{pmatrix}$.

The matrix A turns the vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1\mathbf{e_1} + x_2\mathbf{e_2}$ into the vector $h(\mathbf{x}) = A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 3x_2 \\ -2x_2 \end{pmatrix}$, where (x_1, x_2) are the coordinates of the vector x in the canonical basis defined by the vectors of the canonical basis $\mathbf{e_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{e_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

2.1.3 Operations on linear maps

Let's consider $n \in \mathbb{N}$, two linear maps f and g in $\mathbb{R}^n \to \mathbb{R}^n$, a vector $\mathbf{x} \in \mathbb{R}^n$, and a basis $\mathcal{B}$ of $\mathbb{R}^n$. Based on the properties of matrices, one can deduce the following properties of linear maps:

— $Mat_{\mathcal{B}}(f+g) = Mat_{\mathcal{B}}(f) + Mat_{\mathcal{B}}(g)$;

— $Mat_{\mathcal{B}}(f \circ g) = Mat_{\mathcal{B}}(f)Mat_{\mathcal{B}}(g)$;

— $Mat_{\mathcal{B}}(f(x)) = Mat_{\mathcal{B}}(f)Mat_{\mathcal{B}}(x)$.

2.2 Invertible matrix

A square matrix A_n is **invertible** if it exists a matrix B_n such as $AB = BA = I_n$. Then, B is the inverse of A, denoted A^{-1}.

Let's consider the linear maps f and g corresponding to the matrices A and B, one can deduce that $f \circ g = Id$, *i.e.* f is bijective and $g = f^{-1}$. In other words, a matrix is invertible if and only if its associated linear map is bijective.

For $n = 2$, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible, if and only if $ad - bc \neq 0$, *i.e.* if and only if $\det(A) \neq 0$, and its inverse is:

$$A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{\det(A)}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

In general ($n \geq 2$), A_n is invertible, if and only if $\det(A) \neq 0$.

For higher dimension matrices, 2 methods can be used to find the inverse:

Let's consider a new matrix A which is now a 3×3 matrix,

— *Method n°1:* Set a 3×3 matrix $B = \begin{pmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{pmatrix}$ and solve the system $A \times B = I_3$. So, B is the inverse of A.

— *Method n°2:* Take two column vectors $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$. Consider the system $A \times X = Y$ and express all the x_1, x_2, x_3 as linear combinations of y_1, y_2, y_3. The inverse is the matrix formed by taking the coefficients of the previous linear combinations in the right order.

Systems of linear equations can be written with matrices:

For example, let's consider the system (E) of linear equations:

$$(E): \begin{cases} x_0 + x_1 & = 0 \\ x_0 + 2x_1 + x_2 & = 1 \\ x_0 + x_1 + 2x_2 & = 1 \end{cases}$$

(E) can be written using matrices:

$$A\,\mathbf{x} = \mathbf{c} \text{ with } A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}, \text{ and } \mathbf{c} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

2.3 Eigenvectors and eigenvalues

An **eigenvector** of a matrix A is a non-zero vector, $\mathbf{v} \in \mathbb{R}^n \setminus \{0\}$, such that it exists a scalar $\lambda \in \mathbb{R}$ that satisfies $A\mathbf{v} = \lambda\mathbf{v}$. λ is called the **eigenvalue** associated with $\mathbf{v}$.

$$\begin{aligned} \lambda \text{ is a eigenvalue of A} &\iff \exists \mathbf{v} \neq 0, A\mathbf{v} = \lambda\mathbf{v} \\ &\iff \exists \mathbf{v} \neq 0, A\mathbf{v} - \lambda I_n \mathbf{v} = 0 \\ &\iff \exists \mathbf{v} \neq 0, (A - \lambda I_n)\mathbf{v} = 0 \\ &\iff 0 \text{ is an eigenvalue of } (A - \lambda I_n) \\ &\iff (A - \lambda I_n) \text{ is not invertible} \\ &\iff \det(A - \lambda I_n) = 0. \end{aligned}$$

To get the eigenvalues of A and their associated eigenvectors, one has to:

— Find the eigenvalues λ by solving $\det(A - \lambda I_n) = 0$.

— For each eigenvalue λ, find the associated eigenvector by finding $\begin{pmatrix} x_1 \\ x_2 \\ \dots \end{pmatrix}$ such as:

$$A \begin{pmatrix} x_1 \\ x_2 \\ \dots \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ \dots \end{pmatrix}$$

Geometrically, an eigenvector associated with a real non-zero eigenvalue points in a direction of the vector space that is stretched by the transformation A and its eigenvalue corresponds

to the factor by which it is stretched. If the eigenvalue is negative, the direction is reversed. Conversely, an eigenvalue of zero indicates that the transformation is collapsing at least one dimension (some non-zero vectors are transformed into the zero vector).

2.4 Change of basis

A vector space can have many bases and sometimes working with one basis is easier than with another. Therefore, a change of basis may be necessary. For instance, one can perform a change of basis from the canonical basis $\mathcal{C}$ to a new basis $\mathcal{B}$.

Given two bases $\mathcal{B}_1 = (\mathbf{u}_1, \mathbf{u}_2)$ and $\mathcal{B}_2 = (\mathbf{v}_1, \mathbf{v}_2)$, the matrix of change of basis from $\mathcal{B}_1$ to $\mathcal{B}_2$, denoted $P_{\mathcal{B}_1,\mathcal{B}_2}$, is the matrix composed by the coordinates of the vectors $\mathcal{B}_2$ in the basis $\mathcal{B}_1$:

$$P_{\mathcal{B}_1,\mathcal{B}_2} = \begin{array}{c} \\ \mathbf{u}_1 \\ \mathbf{u}_2 \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{v}_1 & \mathbf{v}_2 \end{array} \\ \begin{pmatrix} \diamond & \diamond \\ \diamond & \diamond \end{pmatrix} \end{array}$$

The matrix of change of basis has the following properties:

$$Mat_{\mathcal{B}2}(f) = P^{-1}_{\mathcal{B}_1,\mathcal{B}_2}\ Mat_{\mathcal{B}1}(f)\ P_{\mathcal{B}_1,\mathcal{B}_2} \tag{1}$$

$$Mat_{\mathcal{B}2}(x) = P^{-1}_{\mathcal{B}_1,\mathcal{B}_2}\ Mat_{\mathcal{B}1}(x). \tag{2}$$

2.5 Diagonalizable matrix

Eigenvectors of a linear map f (and its associated canonical matrix A) directly indicate the directions of the stretching operated by the linear transformation. Thus, working in the basis $\mathcal{B}$ formed by the eigenvectors of A is much easier, as the linear map f can be directly represented by a diagonal matrix in the basis $\mathcal{B}$. Thus, a change of basis from the canonical basis $\mathcal{C}$ to basis $\mathcal{B}$ of eigenvectors is essential.This is the general idea of the **diagonalization**.

A square matrix A is **diagonalizable** if it exists an invertible matrix P that verifies $A = PDP^{-1}$, where D is a diagonal matrix. The diagonal entries of the matrix D are the eigenvalues of A, and the column vectors of P are the right eigenvectors of A.

$$A = \begin{pmatrix} a_{00} & a_{01} & ... & a_{0n} \\ a_{10} & a_{11} & ... & a_{1n} \\ ... & & & \\ a_{n0} & a_{n1} & ... & a_{nn} \end{pmatrix} = PDP^{-1} = P \begin{pmatrix} \lambda_0 & 0 & ... & 0 \\ 0 & \lambda_1 & ... & 0 \\ & ..... & & \\ 0 & 0 & ... & \lambda_n \end{pmatrix} P^{-1}$$

where $P = \begin{pmatrix} v_{00} & v_{01} & ... & v_{0n} \\ v_{10} & v_{11} & ... & v_{1n} \\ & ..... & & \\ v_{n0} & v_{n1} & ... & v_{nn} \end{pmatrix}$, and $\lambda_0, \lambda_1, ..., \lambda_n$ are the eigenvalues that respectively correspond to the eigenvectors $\begin{pmatrix} v_{00} \\ v_{10} \\ ... \\ v_{n0} \end{pmatrix}, \begin{pmatrix} v_{01} \\ v_{11} \\ ... \\ v_{1n} \end{pmatrix}, ..., \begin{pmatrix} v_{0n} \\ v_{1n} \\ ... \\ v_{nn} \end{pmatrix}$

Matrix operations are far easier on diagonal matrices, for example:

$$A^2 = PDP^{-1}PDP^{-1} \iff A^2 = PD^2P^{-1} \longrightarrow A^n = PD^nP^{-1}$$

thus, one would rather work with the diagonalized matrices instead of the original ones. This is very useful when analyzing biological systems.

2.6 Other properties

— A symmetrical matrix is always diagonalizable ;

— Given a diagonalizable matrix A, $\det(A) = \prod \lambda_i$ and $\mathrm{tr}(A) = \sum \lambda_i$;

— The characteristic polynomial associated with the matrix A_n is defined as:

$$P_A : \begin{cases} \mathbb{R} & \longmapsto & \mathbb{R} \\ \lambda & \longrightarrow & \det(A - \lambda I_n) \end{cases} ;$$

P_A has the eigenvalues of A as roots. A is diagonalizable if and only if P_A has n roots or $p < n$ roots with a total roots' degree equals to n.

— A is not invertible if and only if $det(A) = 0$, *i.e.* 0 is an eigenvalue of A.

Lecture 3: Dynamical systems

3.1 Mathematical modeling of biological systems

Biological systems are dynamical systems changing through time. Thus, they are often modelled by mathematical equations that describe the evolutions of a given variable through time (e.g. the ratio of an activated enzyme $X(t)$, the membrane potential $V(t)$, or the number of individual of a species in a population $N(t)$). The analysis of the variable, X, V, or N as a function of time belongs to the study of dynamical systems. Systems that relate the variables to their time derivatives constitute systems of differential equations.

3.1.1 Example in one dimension

Let's consider a bacterial population characterized by a constant division time of 1 hour between each generation and $N(t)$ is the number of bacteria at time t. One can deduce:

$$N(t_1) = 2N(t_0),\ N(t_2) = 2N(t_1) = 2^2N(t_0)\ldots\ \ \forall n \in \mathbb{N},\ N(t_n) = 2^nN(t_0),$$

$$\text{that can be generalized in continuous time: } \forall t \in \mathbb{R},\ N(t) = 2^tN(t_0)$$

In a short time period, between t and $t + \mathrm{d}t$ with $\mathrm{d}t \longrightarrow 0$, one can quantify the variation of number of bacteria, *i.e.* by definition of the derivative:

$$\frac{N(t+\mathrm{d}t) - N(t)}{\mathrm{d}t} \longrightarrow N'(t) = \ln(2)2^tN(t_0) = \ln(2)N(t)$$

then,

$$N(t+\mathrm{d}t) - N(t) = \lambda N(t)\mathrm{d}t + o(\mathrm{d}t)$$

with $\lambda = \ln(2)$: at the first order $\mathrm{d}t$, the variation of number of bacteria between t and $t+\mathrm{d}t$ is proportional to $\mathrm{d}t$ and to $N(t)$: biological systems can often be modelled in first order differential equations.

3.1.2 Example in two dimensions

Given a system of two interacting species X and Y, with the variables $x(t)$ the number of individuals from species X and $y(t)$ the number of individuals from species Y at time t, one can propose this model for a small time variation:

$$\begin{cases} x(t+\mathrm{d}t) - x(t) = a\ \mathrm{d}t\ x(t) + b\ \mathrm{d}t\ y(t) + o(\mathrm{d}t) \\ \\ y(t+\mathrm{d}t) - y(t) = c\ \mathrm{d}t\ x(t) + d\ \mathrm{d}t\ y(t) + o(\mathrm{d}t) \end{cases};$$

with the parameters $a, b, c, d \in \mathbb{R}$. By dividing the system by $\mathrm{d}t \to 0$:

$$\begin{cases} x'(t) = ax(t) + by(t) \\ \\ y'(t) = cx(t) + dy(t). \end{cases}$$

Combing these systems of differential equations with matrices can strongly help their resolutions.

3.2 Phase space of a dynamical system

Given a general first order dynamical system:

$$\begin{cases} x' = f(x, y) \\ \\ y' = g(x, y). \end{cases} ;$$

One can represent a **phase space**, *i.e.* a plot (x, y) that has the following objects:

(i) the **solutions**: trajectories followed by the system given a set of initial conditions.

(ii) the **force field**: tangents of the trajectories in different points of the space phase.

(iii) the **isoclines** (or **nullclines**): the curves in which each variable remains constant from one time point to the next (in the x', y' system, it is the curves $x' = 0$ and $y' = 0$).

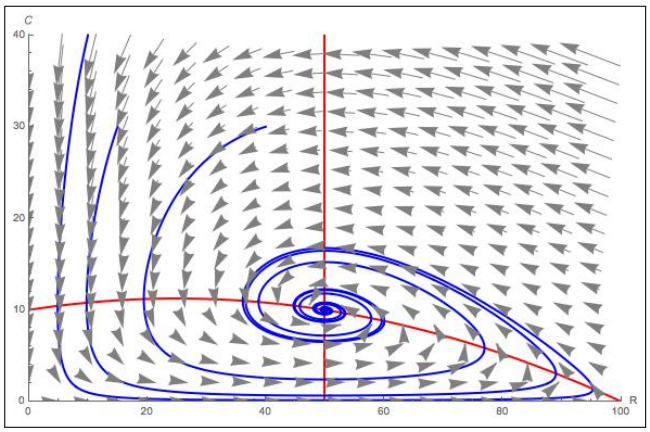

Figure 1: Phase space (x,y): some solutions are plotted in blue, examples of isocline are represented in red, and force fields in grey.

A **fixed point** of a dynamical system is reached when all the temporal derivatives of the system are equal to zero, *i.e.* it corresponds to:

$$f(\hat{x}, \hat{y}) = g(\hat{x}, \hat{y}) = 0.$$

If the system is initially located at the fixed point, it will never move: **fixed points are the intersection of the isoclines** $x' = 0$ **and** $y' = 0$.

The **stability of a fixed point** is defined by the evolution of the system when it starts at an initial point close to this fixed point. The fixed point is **stable (resp. instable)** is the system converging toward the fixed point (resp. diverging away from the fixed point).

Sometimes, differential equations do not present mathematical explicit solutions. However, it is always possible to approximate the trajectories of the system in the phase space for a given initial condition. Several methods exist. For example, the **Euler method** (the simplest method) proposes the approximation $(\tilde{x}, \tilde{y})$ of the trajectory given the initial conditions (x_0, y_0) and a short time step $\mathrm{d}t$:

$$\forall t, \begin{cases} \tilde{x}(t + \mathrm{d}t) = \tilde{x(t)} + \mathrm{d}t \ f(\tilde{x}, \tilde{y}) \\ \\ \tilde{y}(t + \mathrm{d}t) = \tilde{y(t)} + \mathrm{d}t \ g(\tilde{x}, \tilde{y}). \end{cases}$$

3.3 Solving a linear system

Let's consider the dynamical system (E):

$$\mathbf{x}'(t) = M\mathbf{x}, \text{ where } M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Let's assume that M is diagonalizable with real eigenvalues λ_1 and λ_2 and their corresponding eigenvectors $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$.

In the basis $\mathcal{B}$ formed by its eigenvectors, (E) can thus be expressed using the diagonal matrix $D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = P^{-1}MP$, with:

$$P = \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \end{pmatrix} \quad \text{and} \quad P^{-1} = \frac{1}{u_1 \ v_2 - u_2 \ v_1} \begin{pmatrix} v_2 & -v_1 \\ -u_2 & u_1 \end{pmatrix}$$

Thus, the system (E) is equivalent to:

$$\mathbf{x}'(t) = P \ D \ P^{-1} \ \mathbf{x}(t) \iff P^{-1}\mathbf{x}'(t) = D \ P^{-1}\mathbf{x}(t)$$

Let's define $\mathbf{n}(t) = \begin{pmatrix} n_1(t) \\ n_2(t) \end{pmatrix} = P^{-1}\mathbf{x}(t)$, the system (E) can thus be written as (E'):

$$\mathbf{n}'(t) = D\mathbf{n}(t)$$

Instead of working with M, we can now work with the diagonal matrix D, which corresponds to a change of basis (see Lecture 2).

Then, the system (E') can be directly solved (see Lecture 1):

$$\mathbf{n}(t) = N\mathbf{n}(0) \text{ where } N = \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \iff \mathbf{x}(t) = PNP^{-1}\mathbf{x}(0)$$

Thus, given an initial condition $\mathbf{x}(0)$, one can always find the value of $\mathbf{x}(t)$. However, knowing the value of $\mathbf{x}(t)$ is not the only goal when studying a system. Frequently, one would like to how the system behaves given enough time, *e.g.*, whether the system converges into a set of values, diverge toward infinity or cycle. This can be achieved by studying the fixed points of the system and their stability.

3.4 Stability of the fixed points

3.4.1 Linear case

Let's suppose the system (E) :

$$\mathbf{x}'(t) = M\mathbf{x}(t), \text{ where } M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The eigenvalues of M are the solutions of the equations $\det(A - \lambda I_n)$. M is diagonalizable if the equation admits two solutions λ_1 and λ_2. The eigenvalues can be real or complex (in this latter case λ_1 and λ_2 are conjugated).

A. If λ_1 and $\lambda_2 \in \mathbb{R}*$:

As λ_1 and $\lambda_2 \in \mathbb{R}*$, A is invertible, and $ad - cd \neq 0$. Thus, $(0, 0)$ is the only fixed point of (E). The stability is determined by the behavior of the system in the neighborhood of $(0, 0)$ with $\mathbf{x}(0) \neq (0, 0)$. According to the result of section 3.3, we could look at the stability of $\mathbf{n}$ since it has a simpler form: $\mathbf{n}(t) = (e^{\lambda_1 t}, e^{\lambda_2 t})$:

(i) if $\lambda_1 > 0$ and $\lambda_2 > 0$: $\mathbf{n}(t)$ goes to $+\infty$. Then, $(0, 0)$ is an **unstable fixed point**.

(ii) if $\lambda_1 < 0$ and $\lambda_2 < 0$: $\mathbf{n}(t)$ goes to 0. Then, $(0, 0)$ is a **stable fixed point**.

(iii) if $\lambda_1 < 0 < \lambda_2$ or $\lambda_2 < 0 < \lambda_1$: one axis converges and the other diverges. $(0, 0)$ is a **saddle point**.

B. If $\lambda_1 = 0$ and $\lambda_2 \in \mathbb{R}*$:

It exists an infinity of fixed points (the line $n_2 = 0$) and $\begin{pmatrix} n_1(t) \\ n_2(t) \end{pmatrix} = \begin{pmatrix} n_1(0) \\ n_2(0)e^{\lambda_2 t} \end{pmatrix}$

So, if $\lambda_2 < 0$ (resp. $\lambda_2 > 0$) all fixed points are stable (resp. unstable).

C. If λ_1 and λ_2 are not real:

If the eigenvalues of M are complex with the form $\lambda = r \pm i\omega$ with $(r, \omega) \in \mathbb{R} \times \mathbb{R}^*$, we can still look at the behavior of $\mathbf{n}(t)$:

$$\mathbf{n}(t) = N\mathbf{n}(0) = \begin{pmatrix} e^{(r+i\omega)t} & 0 \\ 0 & e^{(r-i\omega)t} \end{pmatrix} \mathbf{n}(0) = \begin{pmatrix} e^{rt} & 0 \\ 0 & e^{rt} \end{pmatrix} \begin{pmatrix} e^{i\omega t} & 0 \\ 0 & e^{-i\omega t} \end{pmatrix} \mathbf{n}(0)$$

Thus, the system (E) has the following solutions:

$$\begin{cases} x_1(t) &= C_1 e^{rt}\cos(\omega t + \phi_1) + C_2 e^{rt}\sin(\omega t + \phi_1) \\ x_2(t) &= C_1 e^{rt}\cos(\omega t + \phi_2) + C_2 e^{rt}\sin(\omega t + \phi_2). \end{cases}$$

The behavior of the system depends on the sign of the real part r of the eigenvalues:

(i) if $r < 0$, the fixed point is stable and the system oscillates with decreasing amplitude.

(ii) if $r > 0$, the fixed point is unstable and the system oscillates with increasing amplitude.

(iii) if $r = 0$, the fixed point is not unstable or stable and the system oscillates on an ellipse (depending on the initial conditions).

D. If M is not diagonalizable:

For example, $M = \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} 1/2 & -1 \\ 1 & -1/2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$ only has one eigenvalue of order 1, $\lambda = -1$.

Thus, it is not possible to find two non-collinear eigenvectors. It is then impossible to find a general formula for the systems. The trajectory of the system can only be approximated by simulations.

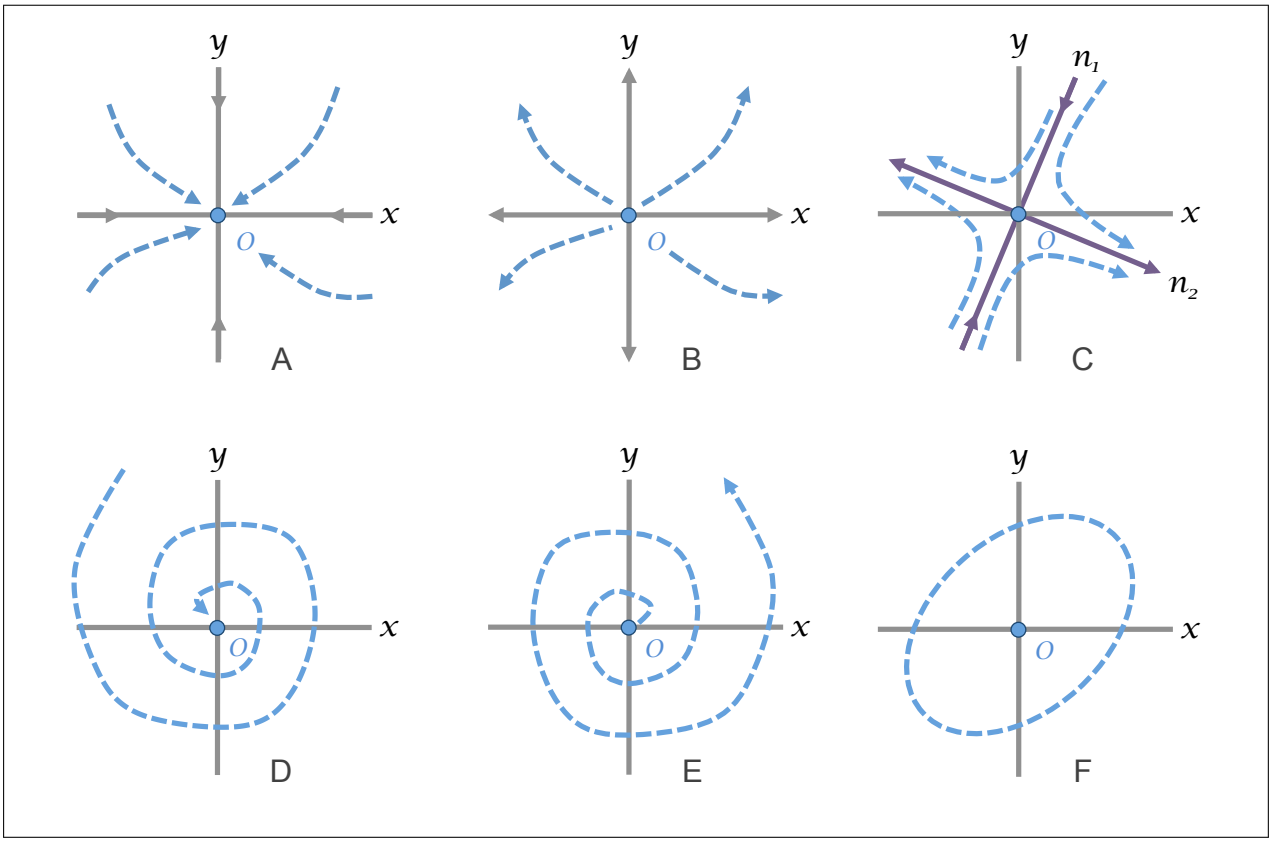

Figure 2: Stability of fixed points in $\mathbb{R}^2$: if the eigenvalues are real: A) negative eigenvalues, B) positive eigenvalues, and, C) one positive, one negative;
if the eigenvalues are complex: D) their real part is negative, E) their real part is negative, and, F) if their real part equals 0.

3.4.2 Non-linear case

Most of the dynamical systems are not linear, hence, often do not have explicit solutions. The study of a non-linear system is limited to the study of its fixed points and their stability. The idea is to find a fixed point of the system and approximate the non-linear system in the neighbourhood of this fixed point by a linear system. Given a simple system:

$$(E) : \begin{cases} x' = f(x, y) \\ y' = g(x, y). \end{cases}$$

Let's suppose that the system (E) has a fixed point $(\hat{x}, \hat{y})$. At the point $(x, y) = (\hat{x}+\epsilon_x, \hat{y}+\epsilon_y)$ that is very near the fixed point, we want to analyse whether the system is converging or is pushed away from the fixed point by looking at the changes induced by these small displacements (ϵ_x, ϵ_y)

$$\begin{cases} \dfrac{d\epsilon_x}{dt} = \dfrac{d(x - \hat{x})}{dt} = \dfrac{dx}{dt} \\ \dfrac{d\epsilon_y}{dt} = \dfrac{d(y - \hat{y})}{dt} = \dfrac{dy}{dt} \end{cases}$$

given that $\hat{x}$, $\hat{y}$ are the equilibrium point, therefore, they do not change with time.

Because we are analysing the point that is very near the fixed point, we have:

$$\begin{cases} \dfrac{d\epsilon_x}{dt} = f(x,y) \\ \\ \dfrac{d\epsilon_y}{dt} = g(x,y) \end{cases}$$

The Taylor series of $f(x,y)$ and $g(x,y)$ around the fixed point $(\hat{x},\hat{y})$ are

$$\begin{cases} f(x,y) = f(\hat{x},\hat{y}) + \dfrac{\partial f(x,y)}{\partial x}(x-\hat{x}) + \dfrac{\partial f(x,y)}{\partial y}(y-\hat{y}) + \text{higher order terms} \\ g(x,y) = g(\hat{x},\hat{y}) + \dfrac{\partial g(x,y)}{\partial x}(x-\hat{x}) + \dfrac{\partial g(x,y)}{\partial y}(y-\hat{y}) + \text{higher order terms} \end{cases}$$

At the fixed point, $f(\hat{x},\hat{y}) = g(\hat{x},\hat{y}) = 0$, and since we consider the point very near the fixed point, the higher order terms are negligible. Therefore, this system

$$\begin{cases} \dfrac{d\epsilon_x}{dt} = \dfrac{\partial f(x,y)}{\partial x}\Big|_{(\hat{x},\hat{y})}\epsilon_x + \dfrac{\partial f(x,y)}{\partial y}\Big|_{(\hat{x},\hat{y})}\epsilon_y \\ \\ \dfrac{d\epsilon_y}{dt} = \dfrac{\partial g(x,y)}{\partial x}\Big|_{(\hat{x},\hat{y})}\epsilon_x + \dfrac{\partial g(x,y)}{\partial y}\Big|_{(\hat{x},\hat{y})}\epsilon_y \end{cases}$$

become a linear system with respect to (ϵ_x, ϵ_y). The Jacobian matrix, denoted $J_{(\hat{x},\hat{y})}$, is then used to characterize the behavior of the system around the fixed point.

$$J_{(\hat{x},\hat{y})} = \begin{pmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \\ \\ \dfrac{\partial g}{\partial x} & \dfrac{\partial g}{\partial y} \end{pmatrix}_{(\hat{x},\hat{y})}$$

Thus, the study of the stability of the fixed point can be done by analysing $J_{(\hat{x},\hat{y})}$ (see previous section).

Example: the Lotka-Volterra system:
Let's consider two species of prey and predator with x the density of preys and y the density of predators and the system:

$$\begin{cases} x' &= \alpha x - \beta xy \\ y' &= -\gamma y + \delta yx. \end{cases} \quad \text{with } \alpha, \beta, \gamma, \text{ and } \delta \in \mathbb{R}^+.$$

Here, we suppose $\alpha = \beta = r$ and $\gamma = \delta = m$:

$$\begin{cases} x' &= rx(1-y) \\ y' &= -my(1-x). \end{cases}$$

The two fixed points are $(0,0)$ and $(1,1)$ with the respective Jacobian matrices:

$$J_{(0,0)} = \begin{pmatrix} r & 0 \\ 0 & -m \end{pmatrix} \quad \text{and} \quad J_{(1,1)} = \begin{pmatrix} 0 & -r \\ m & 0 \end{pmatrix}$$

with the respective eigenvalues: $J_{(0,0)} : \lambda \in \{r, -m\}$ and $J_{(1,1)} : \lambda \in \{i\sqrt{rm}, -i\sqrt{rm}\}$.

The fixed point (0,0) is then a saddle point, and the fixed point (1,1) is not stable or unstable: the system oscillates in an ellipse in the neighbourhood of this point.

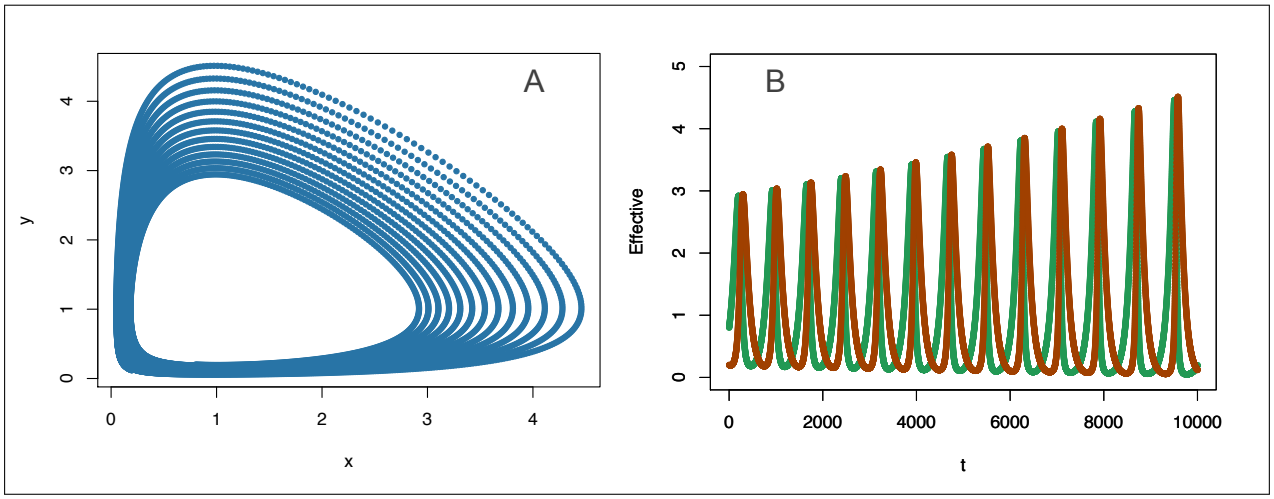

Figure 3: Lotka-Volterra system (example with $m = r = 0.1$ and $(x_0, y_0) = (0.8, 0.2)$):
A) Phase space (x, y) between t=0 and t=10,000
B) Temporal representation of the system: $x(t)$ in green and $y(t)$ in orange.

3.5 Bifurcation

A **bifurcation** in a dynamical system occurs when a small continuous change of a parameter value (the bifurcation parameter) causes a sudden qualitative change of the behavior of the system.

There are different kinds of bifurcations depending on the change of the behavior of the system:

The transcritical bifurcation: a transcritical bifurcation occurs when a fixed point interchanges its stability with another fixed point as the parameter varies. For instance, given the system $\forall \alpha \in \mathbb{R},\ x' = \alpha x - x^2 = x(\alpha - x)$, the two fixed point $x = 0$ and $x = \alpha$ interchanges their stability according to the sign of α, the bifurcation occurs at $\alpha = 0$.

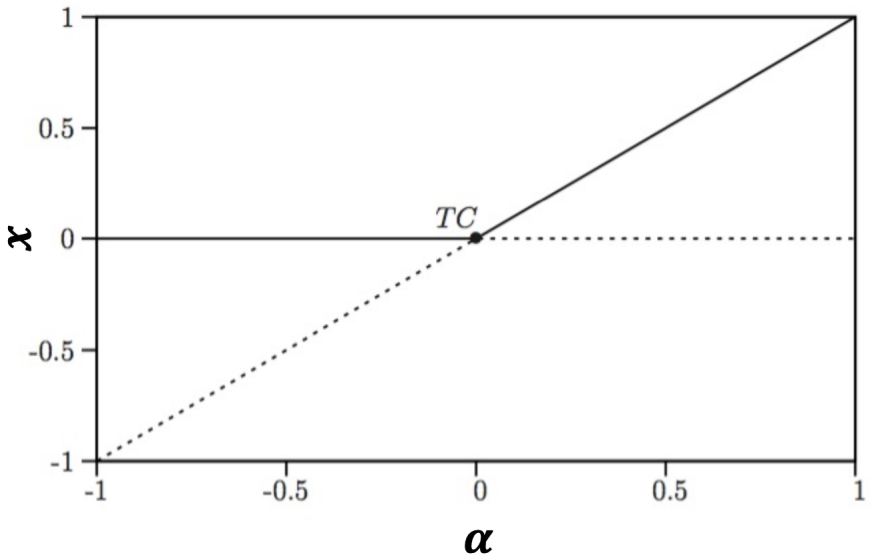

Figure 4: Transcritical bifurcation (TC). Thick lines are stable fixed points, dashed lines are unstable fixed points

The pitchfork bifurcation: a pitchfork bifurcation occurs when the system transitions from one fixed point to three fixed points. For instance, given the system $\forall \alpha \in \mathbb{R},\ x' = \alpha x - x^3 = x(\alpha - x^2)$, for $\alpha < 0$, the system has only one stable fixed point at $x = 0$, but for $\alpha > 0$, there is one unstable fixed point at $x = 0$ and two stable fixed point at $x = \pm\sqrt{\alpha}$.

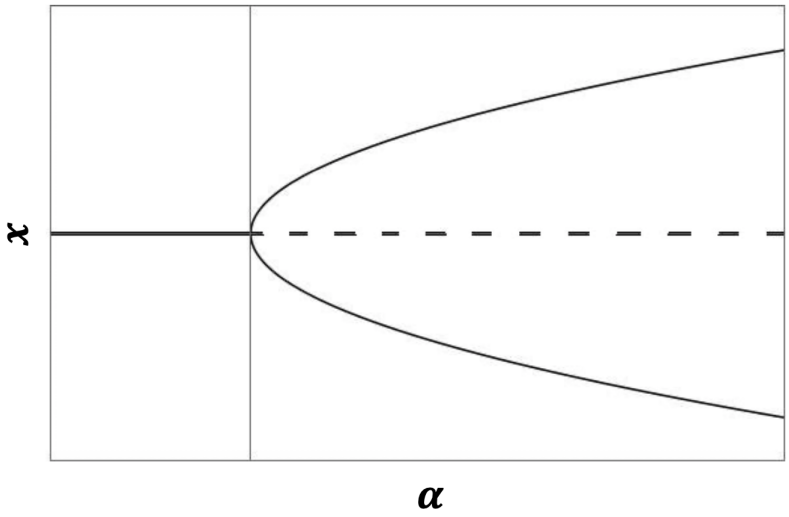

Figure 5: Pitchfork bifurcation. Thick lines are stable fixed points, dashed lines are unstable fixed points

The saddle-node bifurcation: a saddle-node bifurcation (or fold bifurcation) is a bifurcation in which two fixed points collide and annihilate each other. For instance, given the system $\forall \mu \in \mathbb{R}, x' = \mu + x^2$, for $\mu < 0$, the system has two fixed points (a stable fixed point at $-\sqrt{-\mu}$ and an unstable fixed point at $+\sqrt{-\mu}$). For $\mu = 0$, there is a unique stable fixed point called a saddle-node fixed point. For $\mu > 0$, there is no more fixed point.

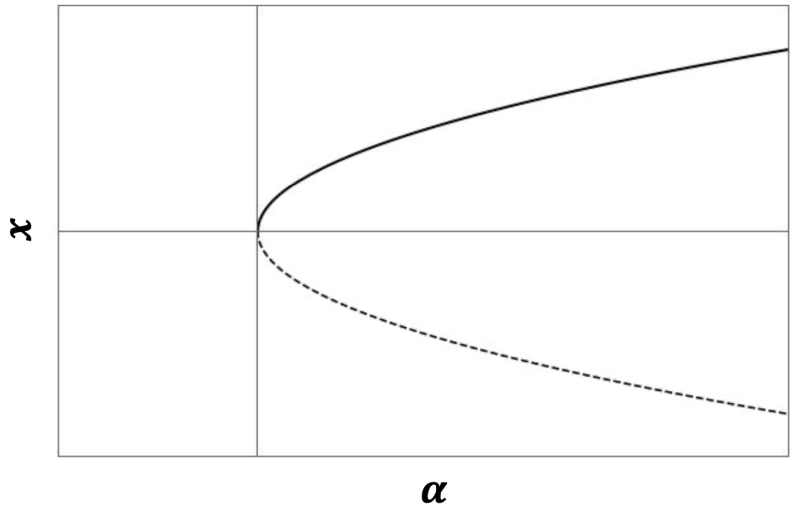

Figure 6: Saddle-node bifurcation. Thick lines are stable fixed points, dashed lines are unstable fixed points

Lecture 4: Probability

4.1 Discrete probability

The goal of this section is to study **randomized experiments**, *i.e.* experiments with non-deterministic outcomes, but for which the probability of each outcome is known.
Discrete probabilities deal with experiments that have a **finite number N of outcomes**.

4.1.1 Random variable

A **real-valued random variable** X is a **function** defined on the set of outcomes of a randomized experiment: $X : \Omega \to \mathbb{R}$. It is any function that gives a real value depending on the outcome of the experiment. $X(\Omega)$ is the codomain of X.

A. Probability distribution

Let's consider a random variable X on (Ω, P), its **distribution** P_X is the function that gives the probability of each value of X.

$$\forall x \in X(\Omega),\ P_X(x) = P(X = x) = \sum_{X(\omega)=x} P(\omega).$$

B. Cumulative distribution function

Let's consider X a random variable, its **cumulative distribution function (CDF)**, F_X, is the function defined as:

$$\forall x \in \mathbb{R},\ F_X(x) = P(X \leq x).$$

The cumulative distribution function is increasing, and has for limits 0 and 1 in $-\infty$ and $+\infty$ respectively. A cumulative distribution function fully describes its corresponding random variable.

C. Expected value and variance

The **expected value** of a random variable X is its mean value on the set of possible outcomes:

$$E(X) = \sum_{\omega \in \Omega} P(\omega)X(\omega) = \sum_{x \in X(\Omega)} xP(X = x).$$

The **variance** of a random variable X is the mean of the squared deviation of a random variable from its mean:

$$V(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$$

The **standard deviation** of X is $\sigma(X) = \sqrt{V(X)}$.

The expected value and the variance have the following properties, $\forall (a, b) \in \mathbb{R}^2$, and X and Y two random variables:

(i) $E(aX + b) = aE(X) + b$ and $E(X + Y) = E(X) + E(Y)$

(ii) $V(aX + b) = a^2V(X)$ and if X and Y are independent $V(X + Y) = V(X) + V(Y)$

D. Independence

Two random variables X and Y are independent if and only if:

$$\forall (i, j) \in X(\Omega) \times Y(\Omega),\ P\left((X = i) \cap (Y = j)\right) = P(X = i)P(Y = j)$$

4.1.2 Common discrete distributions

A. Uniform distribution $\mathcal{U}(n)$

A random variable X follows a uniform distribution with a unique parameter n, $X \sim \mathcal{U}(n)$, if its n outcomes are equally likely. In other words, X takes the values $1, 2, \ldots, n$ with equiprobability:

$$\forall i \in \{1, 2, \ldots, n\},\ P_X(i) = P(X = i) = \frac{1}{n}$$

$$E(X) = \frac{n + 1}{2} \quad \text{and} \quad V(X) = \frac{n^2 - 1}{12}$$

B. Bernoulli distribution $\mathcal{B}(p)$

A random variable X follows a Bernoulli distribution with a unique parameter p, $X \sim \mathcal{B}(p)$, if X takes the value 1 with probability p and the value 0 with probability $q = 1 - p$:

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = p \quad \text{and} \quad V(X) = pq = p(1 - p)$$

A Bernoulli experiment corresponds to an experiment with two outcomes: a success with probability p and failure with probability $1 - p$.

C. Binomial distribution $\mathcal{B}(n, p)$

Let's consider n independent repetitions of a Bernoulli experiment of parameter p, $\mathcal{B}(p)$, and an associated random variable X that counts the total number of success(es) among these n repetitions. Thus, $X(\Omega) = \{0, 1, \ldots, n\}$ and follows the binomial distribution, $X \sim \mathcal{B}(n, p)$:

$$\forall k \in X(\Omega),\ P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$E(X) = np \quad \text{and} \quad V(X) = np(1-p)$$

D. Geometric distribution $\mathcal{G}(p)$

Let's consider a succession of independent Bernoulli experiments of parameter p, the experiment ends once it reaches the first success. X is the random variable that counts the number of Bernoulli experiments until the first success. Thus, $X(\Omega) = \mathbb{N}^*$ and X follows the geometric distribution of parameter p, $X \sim \mathcal{G}(p)$:

$$\forall k \in \mathbb{N}^*,\ P(X = k) = (1-p)^{k-1} p$$

$$E(X) = \frac{1}{p} \quad \text{and} \quad V(X) = \frac{1-p}{p^2}$$

E. Poisson distribution $\mathcal{P}(\lambda)$

The Poisson distribution of parameter λ describes rare probabilistic events (or a very large number of individually unlikely events) happening in a certain time interval: the associated random variable X gives the probability of a number of occurrence of a rare event given the average occurrence λ of this event. Thus, $X(\Omega) \in \mathbb{N}$ and $X \sim \mathcal{P}(\lambda)$:

$$\forall k \in \mathbb{N},\ P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$E(X) = \lambda \quad \text{and} \quad V(X) = \lambda$$

The Poisson distribution is a continuous version of the binomial distribution: if X follows $\mathcal{B}(n, p)$ with $n \geq 30$, $p \leq 0.1$, and $np \leq 15$, X can be approximated by a Poisson distribution of parameter $\lambda = np$.

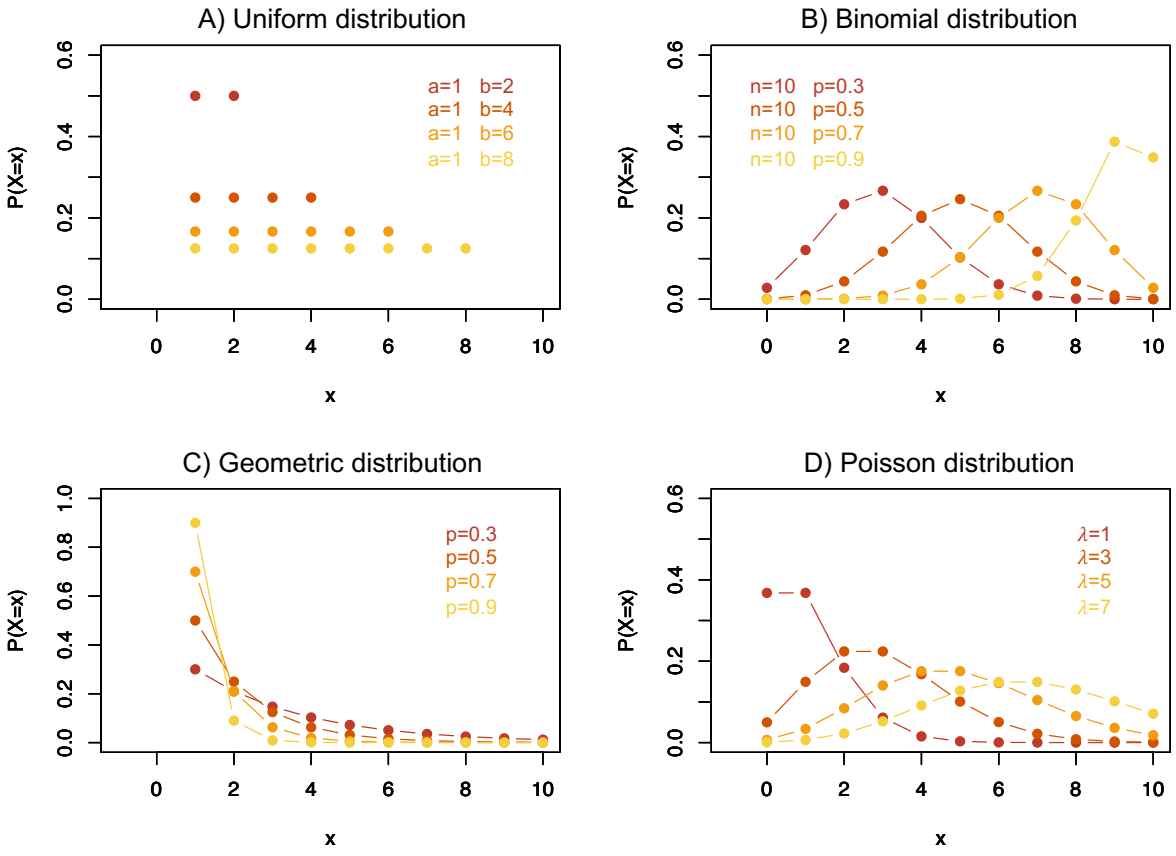

Figure 7: Common discrete distributions

4.2 Continuous probability

4.2.1 Probability density

In an infinite sample space Ω (e.g. $\mathbb{R}, [0,1], \dots$), the probability of an event $\omega \in \Omega$ is $P(\omega) = \dfrac{1}{Card(\Omega)} = \dfrac{1}{+\infty} = 0$.

Thus, the probability of a single event in an infinite sample space always equal 0 and the approach is instead to look at the probability of having an outcome close to $X(\omega) = x$. The **probability density** f_X of a random variable X is a function define as:

$$\forall x \in X(\Omega),\ dx \longrightarrow 0,\ P(X \in [x, x+dx]) = f_X(x)dx$$

Thus, to get the probability of a event $\{X \in [a,b]\}$ with $a < b$, we have to sum this small interval dx between a and b:

$$P(X \in [a,b]) = P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

4.2.2 Cumulative distribution function

A **real-valued continuous random variable** X is characterized by its density f_X. The cumulative distribution function F_X is defined by:

$$\forall x \in X(\Omega),\ F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du.$$

$$\text{By definiton,} \quad \int_{-\infty}^{+\infty} f_X(u)du = 1.$$

4.2.3 Expected value and variance

The expected value of a continuous random variable X is defined by:

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx$$

The variance of a continuous random variable X is defined by:

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x)dx$$

The standard deviation of X is $\sigma(X) = \sqrt{V(X)}$.

4.2.4 Common continuous distributions

A. Uniform distribution $\mathcal{U}(a,b)$

A continuous random variable X follows a uniform distribution between a and b (with $a < b$), $X \sim \mathcal{U}(a,b)$, if X takes a random value $x \in [a,b]$ with equiprobability and f_X is defined by:

$$\forall x \in \mathbb{R},\ f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

$$\forall x \in \mathbb{R},\ F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } x \in [a,b] \\ 1 & \text{if } b < x \end{cases}$$

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad V(X) = \frac{(a-b)^2}{12}$$

B. Exponential distribution $\mathcal{E}(\lambda)$

The exponential distribution is a time-continuous version of the discrete geometric distribution. A continuous random variable X follows an exponential distribution of parameter λ, $X \sim \mathcal{E}(\lambda)$, if X is defined by:

$$\forall x \in \mathbb{R}^+,\ f_X(x) = \lambda e^{-\lambda x}$$

$$\forall x \in \mathbb{R}^+,\ F_X(x) = 1 - e^{-\lambda x}$$

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad V(X) = \frac{1}{\lambda^2}$$

C. Normal distribution $\mathcal{N}(\mu,\sigma^2)$

A continuous random variable X follows a normal distribution of parameter μ and σ^2, $X \sim \mathcal{N}(\mu,\sigma^2)$, if f_X is defined by:

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

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2$$

The cumulative distribution function of $\mathcal{N}(\mu, \sigma^2)$, does not have a simple expression.

$Z \sim \mathcal{N}(0,1)$ is called the **standard normal distribution**. Z can be obtained from any normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ with the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$

Z is an even function, thus $f_Z(-z) = f_Z(z)$. Its cumulative distribution function is denoted Φ and we have: $\Phi(-z) = 1 - \Phi(z)$.

The two following distributions are derived from the standard normal distribution.

D. Chi-squared distribution $\chi^2(k)$

Let's consider the independent continuous random variables $X_1, X_2, \ldots, X_k$ following the standard normal distribution - the random variables X_i are called *independent and identically distributed* random variables. Then, the random variable Y, defined as the sum of the squares of the variables X_i, follows the χ^2 distribution with k degree of freedom:

$$Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$$

$$E(Y) = k \quad \text{and} \quad V(Y) = 2k$$

E. Student distribution $t(k)$

Let's consider Z a random variable following a standard normal distribution $\mathcal{N}(0,1)$ and V a random variable following a χ^2 distribution with k degree of freedom. Then, the random variable $Y = \dfrac{Z}{\sqrt{V/k}}$ follows a Student distribution with k degree of liberty, $t(k)$.

$$\text{For } k \geq 2,\ E(Y) = 0 \text{ et } V(Y) = \frac{k}{k-2}$$

F. Fisher distribution $F(d_1, d_2)$

Let's consider two independent and identically distributed random variables V_1 and V_2 following two χ^2 distributions with the degrees of freedom d_1 and d_2 respectively. Then, the random variable $Y = \dfrac{V_1/d_1}{V_2/d_2}$ follows a Fisher distribution of parameters d_1 and d_2.

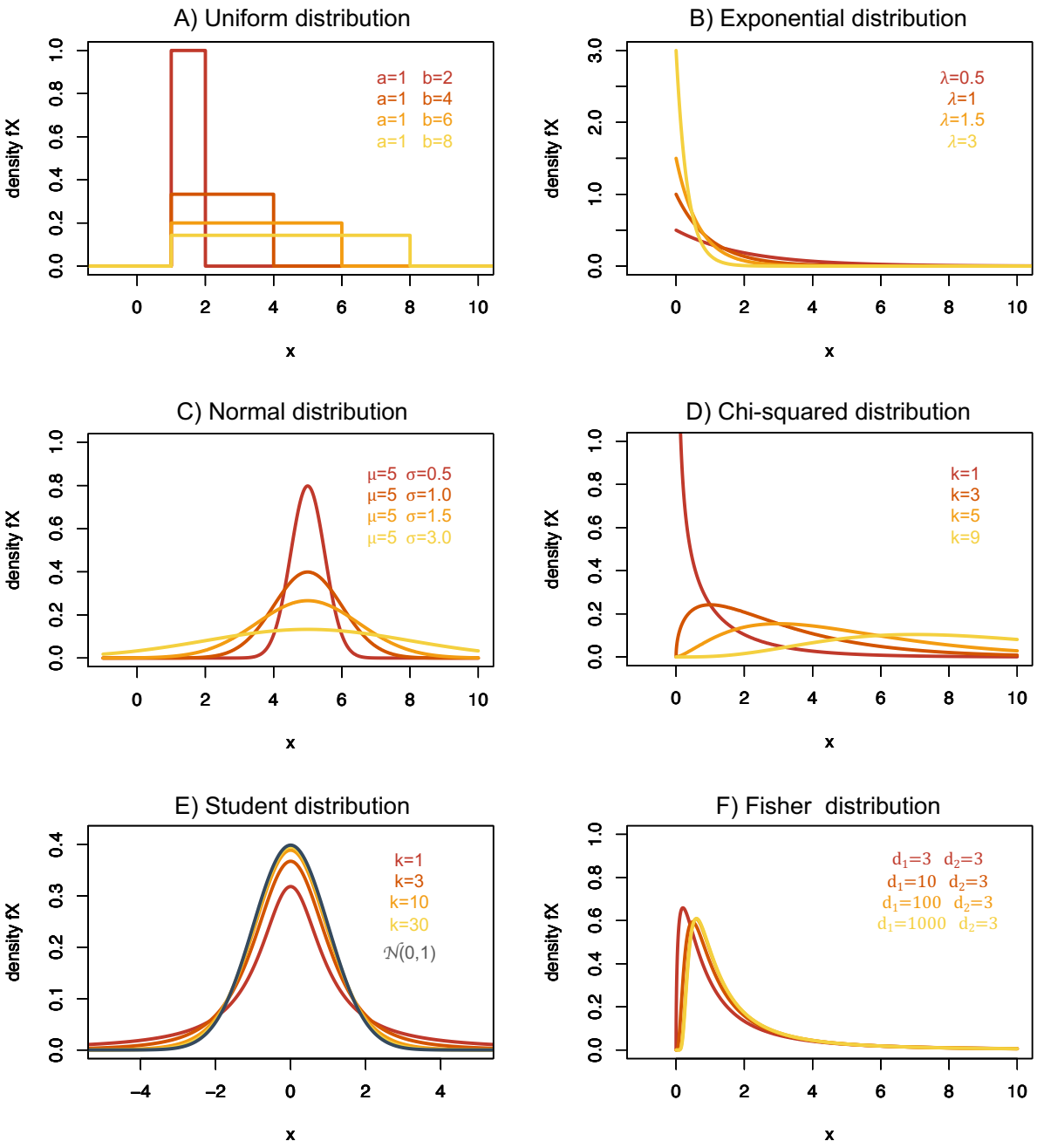

Figure 8: Common continuous distributions

4.2.5 Law of large numbers and Central limit theorem

The law of large numbers (LLN) describes the result of performing the same experiment a large number of times n. Let's consider n independent and identically distributed random variable $X_1, X_2, \ldots, X_n$, with μ the expected value and σ the standard deviation of the common distribution of all X_i. We can define the mean, M_n, of these random variables:

$$M_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

According to the law of large numbers, the mean of the results obtained from a large number of trials, M, should be close to the expected value of the common distribution, μ:

$$n \longrightarrow +\infty, \quad M_n \longrightarrow \mu$$

The central limit theorem (CLT) describes how fast M_n converges toward n:

$$\sqrt{n}(M_n - \mu) \longrightarrow \mathcal{N}(0, \sigma^2)$$

$$\frac{M_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \longrightarrow \mathcal{N}(0, 1)$$

Then, for large values of n,

$$M_n \longrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

For instance, if a random variable X follows a binomial distribution, $X \sim \mathcal{B}(n, p)$, that counts the number of successes of a large enough sample (usually when $np \geq 5$ and $n(1-p) \geq 5$), the CLT applies and the discrete distribution of X can be approximated by a continuous normal distribution, with a mean np and variance $np(1-p)$: $X \sim \mathcal{N}(\mu = np, \sigma^2 = np(1-p))$.

4.3 Introduction to Markov chains

4.3.1 Markov chains in discrete time

Markov chains are a simple class of mathematical models for random events. A given chain describes a sequence of possible moves between states. We consider here the simple case of a finite number of states (the set of possible values of the chain, called the **state space**, denoted S) and discrete time. The probabilities of the transition between states are here considered to be constant over time.

Let X_n be the random variable indicating the state of the chain at time n.

A Markov chain is a **memoryless** property: the future of the chain only depends on the present state, and is independent of the past states (this is called the **Markov property**):

$$\forall (x_0, ..., x_{n+1}) \in S,$$
$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

Thus, the distribution of a Markov chain is only determined by its initial condition and the transition probabilities between states.

4.3.2 Representation

Markov chains can be represented using a **transition matrix** or a **state diagram**: a directed graph with nodes representing the individual states and directed edges indicating the probability of transitions between states).

4.3.3 Properties of a Markov chain

If it exists a path in the Markov chain from state $i \in S$ to state $j \in S$, then state j is **accessible** from state i:

$$\exists\ n,\ P(X_n = j | X_0 = i) > 0$$

Two states $(i, j) \in S$ **communicate** if it exists a path in the Markov chain from i to j and from j to i. The communications between states two by two form partitions of the state space into disjoint **communication classes**.

A Markov chain is **irreducible** if it has only one communication class.

A state $i \in S$ is **recurrent** if from state i there is a probability of one to return to state i after some time.

A state $i \in S$ is **transient** if it is not recurrent, *i.e.* it exists a non-zero probability that the chain will never return to state i.

A state i is **absorbing** if it is impossible to leave this state once reached:

$$P(X_{n+1} = i | X_n = i) = 1$$

A state $i \in S$ has a **period** k if any return to state i must occur in multiples of k time steps from i. A Markov chain is **aperiodic** if all its states are aperiodic.

A state $i \in S$ is **ergodic** if it is aperiodic and recurrent. A Markov chain is ergodic if all its states are ergodic.

Lecture 5: Statistics

5.1 The field of statistics

Statistics in biology aim to explain the observed biological variations: is the variation due to biological factors or is it due to random noise of the biological processes or sampling protocols?

5.1.1 Sampling and estimators

Unlike the field of probabilities that looks at the likelihood of different outcomes given a random variable, the goal of the statistics is to find the probability distribution of the observed random variable given some hypotheses (*i.e.* to find the theoretical distribution from the empirical observations).

In statistics, the **sample** of a study is a subset of individuals from the studied **population**. The observed characteristics of each **individual** $(1, ..., i, ..., n)$, called the **randomized experiment**, are supposed to be some realizations $(x_1, ..., x_i, ..., x_n)$ of the random variables $(X_1, ..., X_i, ..., X_n)$ following a general distribution P_X. The goal of the statistical inference is to estimate the characteristics of the entire population, called **parameters**, thanks to the sample. In other words, the goal is to estimate the expected value μ, the standard deviation σ, and distribution P_X of the observed random variable X.

To get valid estimations, the **sampling** (*i.e.* the selection of sample's individuals) has to be **random** and **non-biased**. For instance, if one selects individuals from the population that share another characteristic Y (*i.e.* a subpopulation), the sampling is biased and estimates $P_{X|Y}$.

Estimators can be deduced from the observations $(x_1, ..., x_i, ..., x_n)$. An estimator is thus a random variable depending on the sample.
For instance, an estimator of the expected value $\mu = E(X)$ of the random variable X is the empirical mean M of the sample $M = (X_1 + \ldots + X_n)/n$ (M is the mean of the observed random variables X_i so it is also a random variable).

A good estimator T of a parameter θ verifies two properties:

1. T is **unbiased**: $E(T) = \theta$;
2. $V(T)$ **converges** toward 0 with n: $\lim_{n \to +\infty} V(T) = 0$

The mean M directly verifies these properties and is a good estimator of the expected value $\mu = E(X)$. However, the empirical variance S^2 is not a good estimator of the variance $\sigma^2 = V(X)$:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 \; - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right) = \frac{X_1^2 + \ldots + X_n^2}{n} - M^2$$

$$E(S^2) = E\left(\frac{X_1^2 + \ldots + X_n^2}{n} - M^2\right) = \frac{1}{n}E(nX^2) - E(M^2)$$

$$\text{Given} \quad E(M^2) = \frac{1}{n^2}E((X_1 + \ldots + X_n)^2) = \frac{1}{n}E(X^2) + \frac{n(n-1)}{n^2}E(X)^2$$

$$E(S^2) = E(X^2) - \frac{1}{n}E(X^2) - \frac{n-1}{n}E(X)^2$$

$$E(S^2) = \frac{n-1}{n}(E(X^2) - E(X)^2) = \frac{n-1}{n}V(X) = \frac{n-1}{n}\sigma^2$$

Thus, the empirical variance S^2 is a biased estimator of $\sigma^2 = V(X)$, but the estimator $\frac{n}{n-1}S^2$ is not biased. The standard estimator of the variance is then:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$$

5.1.2 Example

Let's suppose the experiment of flipping 10 times a coin and getting tail 10 times . We can wonder whether this coin is well equilibrated (*i.e.* if the probability of having a tail equals 0.5)?

The classical approach in frequentist statistical inference is to suppose that the coin C is equilibrated (the null hypothesis, denoted H_0) and to test how likely is the fact of getting tail 10 times.

Thus, we can design the following probabilistic model: we consider n flipping of the coin C and we define the random variables $X_1, \ldots, X_n$ following a Bernoulli distribution $\mathcal{B}(p = 0.5)$ equal to 1 if the result is tail, 0 if the result is face. Then, the random variable $Y_n = X_1 + \ldots + X_n$ follows a binomial distribution $\mathcal{B}(n, p)$. In the example $n = 10$, the distribution of Y_{10} is:

$$\forall k \in \{0; 1; \ldots; 10\}, P(Y_{10} = k) = \binom{10}{k} 0.5^k (0.5)^{10-k}$$

Let's look at the shortest interval of $k \in \{0; 1; \ldots; 10\}$ that have 95% of chance to containing Y_{10}:

$$P(3 \leq Y_{10} \leq 7) = \sum_{k=3}^{7} \binom{10}{k} 0.5^k (0.5)^{10-k} = 0.891$$

$$P(2 \leq Y_{10} \leq 8) = \sum_{k=3}^{7} \binom{10}{k} 0.5^k (0.5)^{10-k} = 0.979$$

$$P(1 \leq Y_{10} \leq 9) = \sum_{k=3}^{7} \binom{10}{k} 0.5^k (0.5)^{10-k} = 0.998$$

The probability of having $Y_{10} \in [2; 8]$ is more than 95%. Thus, $\{0, 1, 9, 10\}$ are in the **region of rejection**: these extreme results have a probability lower than 5% under the null hypothesis H_0. The rejection of H_0 is done with a risk, called the **alpha risk** (here 5%), that is the risk that the null hypothesis is rejected when it is actually true (this type of error is called **false positive**). The risk alpha is chosen **before** the experiment (often in biology, $\alpha = 0.05 = 5\%$).

Then , observing $Y_{10} = 10$ can be interpreted as a rejection of the null hypothesis H_0, *i.e.* the coin is not well equilibrated. We can even calculate the probability $P(Y_{10} = 10) = 1/1024$ under the null hypothesis. $P(Y_{10} = 10) < \alpha$ so the null hypothesis is rejected, and $1/1024$, called the **p-value**, is the probability, when the H_0 is true, that Y_{10} would be the same as or of greater magnitude than the observed results, *i.e.* 10.

If $n = 2$, the same statistical test can not define a region of rejection: flipping only two times a coin is not enough for determining if the coin is well equilibrated or not. This example shows the importance of having a large number of samples n in statistical tests, and illustrates the central limit theorem: the distribution is closer to a normal distribution when n increases.

However, not rejecting the null hypothesis H_0 does not mean that H_0 is true: this type of error is called a **false negative**. The less false negatives a statistical test has, the more false positives it has, and reciprocally.

To conclude, the statistical test is a kind of *proof by contradiction*. To show that something is true (e.g. the coin is biased), we suppose that this thing is wrong (e.g. the coin is well equilibrated): this hypothesis is called null hypothesis H_0. Then, we have to find the probability distribution associated with the null hypothesis: it corresponds to the *probability part* (that can be done by solving equations in the classical tests, or by doing simulations in the more complex models). Finally, the *statistical part* takes care of designing an acceptation interval of likely values and a rejection perimeter.

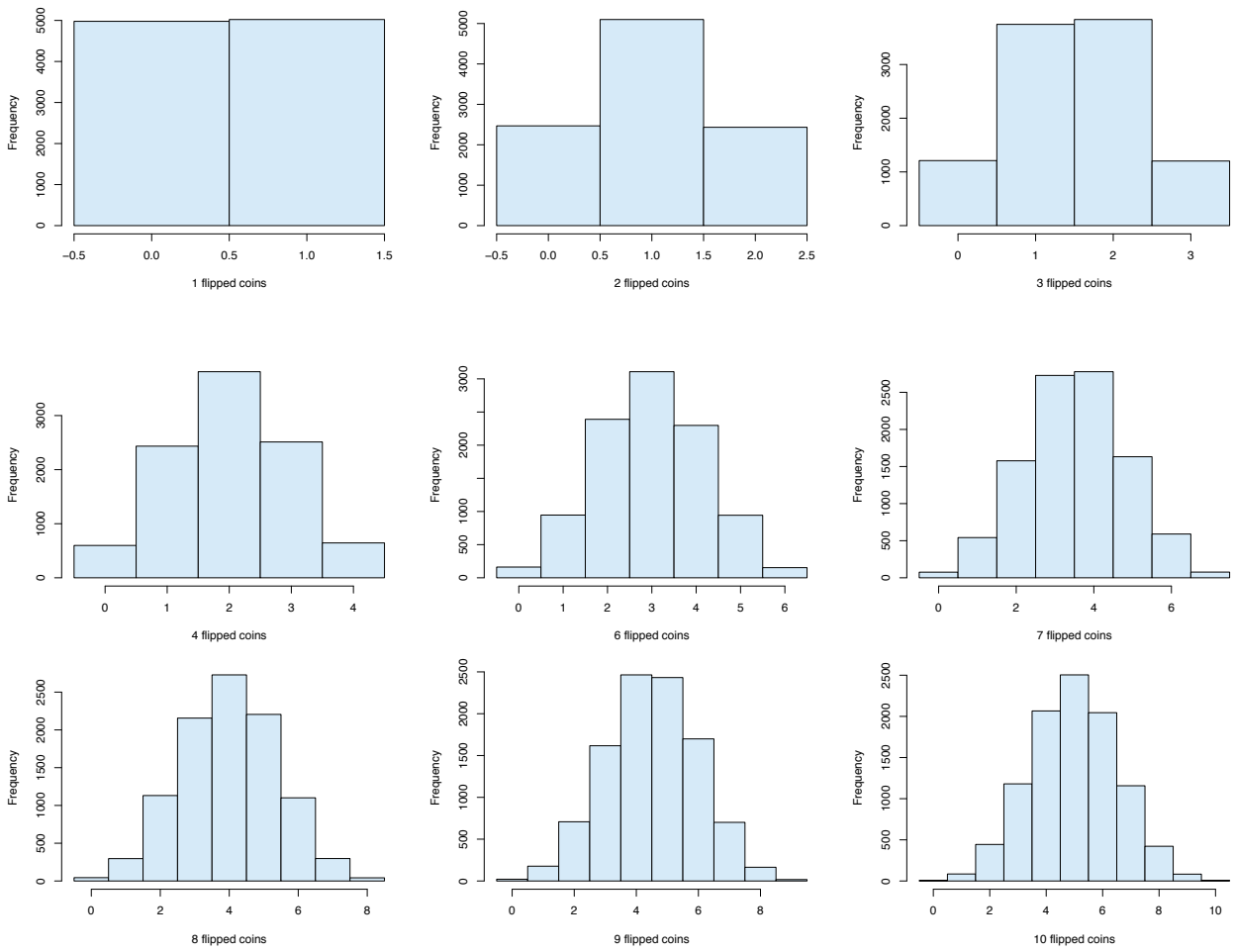

Figure 9: Experimental distributions of the number of tails obtained given a certain number of coin flipping (from 1 to 10, repeated 10,000 times)

5.2 The statistical test

Let's suppose the observations $x_1, x_2, \ldots, x_n$ of the random variables $X_1, X_2, \ldots, X_n$ following an unknown distribution. Let's consider the random variable Y linking X_i together $Y = f(X_1, \ldots, X_n)$ and its observation, $y = f(x_1, \ldots, x_n)$.

5.2.1 Null hypothesis and alternative hypothesis

Let's consider a default hypothesis H_0, called the **null hypothesis**, concerning the distribution of the observed random variable Y, and $H_1 = \overline{H_0}$ is the **alternative hypothesis**. A **hypothesis testing** is a decision rule that, given the observation y, decides whether H_0 **is rejected or not**.

A test is characterized by its region of rejection, W, that describes a subset of $Y(\Omega)$ defined as:
— if $y \in W$, H_0 is rejected and H_1 is accepted;
— if $y \notin W$, H_0 is not rejected.

Generally, the region of rejection is chosen for being as large as possible.

5.2.2 Statistical errors

The **type I error** is the rejection of H_0 when H_0 is true (*i.e.* false positive). The type I error is measured by the alpha risk (α) that is the probability of rejecting H_0 when H_0 is true. Generally, in biology, α is chosen equal to 5%.

$$\alpha = P(Y \in W | H_0)$$

The **type II error** is the non-rejection of H_0 when H_0 is wrong (*i.e.* false negative). The type II error is measured by the beta risk (β) that is the probability of accepting H_0 when H_1 is true.

$$\beta = P(Y \notin W | H_1)$$

The **power** of a statistical test is the probability that the test correctly rejects H_0 when H_1 is true.

$$P(Y \in W | H_1) = 1 - \beta$$

By convention, we tend to minimize the type I error associated with a statistical test before minimizing the type II error. Thus, the choice of H_0 depends on the question: we must choose as the null hypothesis a hypothesis for which false positives are the most problematic.

Frequently, β is not well quantified, thus we cannot be able to conclude that H_0 is true, but only that H_0 is not rejected based on α.

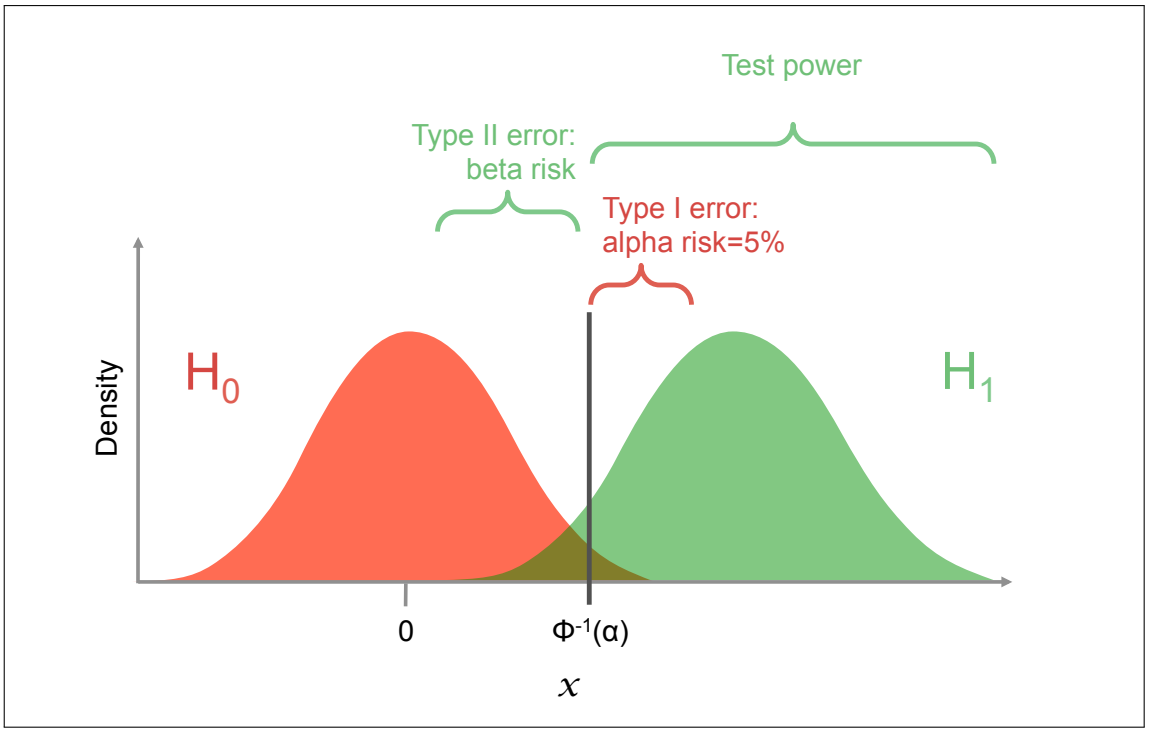

Figure 10: A statistical test with the distributions of its associated H_0 (in red) and H_1 (in green). Representation of type I and II errors, and the statistical power of the test.

5.2.3 Unilateral or bilateral tests

Based on the definition of the region of rejection W, that is as large as possible, it exists different type of test:

Unilateral test: Let's consider Y following a Chi-squared distribution $\chi^2(1)$, the rejection zone is $W = [\gamma, +\infty[$ where γ is defined such as:

$$P_{H_0}(0 \leq Y \leq \gamma) = 1 - \alpha \quad i.e., \quad P_{H_0}(Y \geq \gamma) = \alpha;$$

Bilateral test: Let's consider Y following the standard normal distribution $\mathcal{N}(0, 1)$, the rejection zone is $W =]-\infty; \gamma] \cup [\gamma, +\infty[$ where γ is defined such as:

$$P_{H_0}(-\gamma \leq Y \leq \gamma) = 1 - \alpha \quad i.e., \quad P_{H_0}(Y \geq \gamma) = \frac{\alpha}{2}.$$

5.2.4 P-value

The **p-value** p is the probability, given H_0, that Y has a more extreme value than the observation y.

For a unilateral test, the p-value p is defined by:

$$p = P_{H_0}(Y \geq y) \quad \text{or} \quad p = P_{H_0}(Y \leq y)$$

Otherwise, for a bilateral test, the p-value p is defined by:

$$p = P_{H_0}(Y \leq -|\gamma| \ \cup \ |\gamma| \leq Y)$$

Thus, the rejection of the null hypothesis is given by:
— H_0 is rejected if $p \leq \alpha$;
— H_0 is not rejected if $p > \alpha$.

A p-value gives qualitative information on the degree of confidence in the rejection of H_0. However, it does not indicate anything about the quantitative difference. For instance, a very small p-value indicates a very significant difference but it can correspond to a very small difference (especially if the sample size n is large).

5.2.5 Parametric and nonparametric tests

When we suppose that the sample comes from a parametric distribution (*i.e.* a distribution that depends on a certain number of parameters, such as a normal distribution), the corresponding statistical test is called a **parametric test**. For example, if we suppose that a random variables X follows a normal distribution, the hypothesis of normality of the observations $(x_1, \ldots, x_n)$ has to be verified first in order to apply a parametric test.

Otherwise, we have to use a **nonparametric test**. Nonparametric tests often do not need to verify strong hypotheses, but have generally a weaker statistical power compared to the parametric tests. For instance, for a given α risk (*i.e.* the same chance of having false positives), a parametric test would generally have a smaller β risk than its equivalent nonparametric test (*i.e.* more false negatives).

5.2.6 Quantitative and qualitative/categorical variables

It exists two main types of variables. The **quantitative variables** regroup the discrete and continuous random variables (*i.e.* random variables with values in $\mathbb{N}$ or $\mathbb{R}$). Otherwise, the **categorical variables (or qualitative variables)** are variables that do not represent quantity (*e.g.* male or female, treatment or control, ...). Categorical variables are usually not ordered and do not tolerate sum.

5.2.7 Multiple testing

Let's suppose that we observed 20 realizations $(x_1, \ldots, x_{20})$ of some random variables X_i that follows a standard normal distribution $\mathcal{N}(0, 1)$ under H_0. Given an α risk of 5%, we can compute the p-value for each of the 20 observations $(x_1, \ldots, x_{20})$, and one average we can expect one p-value to be < 0.05 even if $H_0 : X_i \sim \mathcal{N}(0, 1)$ is true.

Thus, observing one p-value < 0.05 does not mean that H_0 has to be rejected, and we can even compute the probability p_n of rejecting H_0 when H_0 is true as a function of the number of observations n:

$$p_n = P_{H_0}(\text{at least one } p < 0.05) = 1 - P_{H_0}(\text{all } \ p > 0.05) = 1 - 0.95^n.$$

p_n is called the **family-wise error rate** (FWER): it is the probability of finding one or more false positives (type I errors) when performing multiple tests.

For $n = 100$, $p_n = 0.99$: it is almost certain to reject H_0 whereas H_0 is true. On average, 5 observations over 100 can be false positive: the global test has a α risk of almost 100%.

Thus, we must correct for multiple testing:

— **Bonferroni correction**: the Bonferroni correction controls the FWER, *i.e.* it lowers the individual level of significance (α) in order to keep a global level of significance equal to 0.05 (*i.e.* $p_n = 0.05$): the corrected α risk is then α/n.

— **FDR-controlling procedure**: the false discovery rate (FDR) is the expected proportion of type I errors. The FDR procedure is less strict than the Bonferroni correction: it has greater power, at the cost of increased numbers of type I errors.

5.2.8 How to design a statistical test

A classical statistical approach is composed of the following steps:

1. Choose a α risk (e.g. 0.05);
2. Realize n experiments that give the observations $x_1, \ldots, x_n$;
3. Make hypotheses about the $X_1, \ldots, X_n$ distributions (e.g. the random variables are independent and identically distributed);
4. Set the hypothesis to test: the null hypothesis H_0 and its alternative H_1;
5. Introduce the random variable Y as a function of $X_1, \ldots, X_n$ (e.g. the mean of $X_1, \ldots, X_n$);
6. Compute the probability distribution of Y;
7. Deduce the region of rejection as a function of α;
8. Look at the position of the observed value y regarding the region of rejection;
9. Reject or do not reject H_0.

Thanks to statistical software like R, the approach is simplified by using common statistical tests that directly compute a p-value through the following steps:

1. Choose a α risk (e.g. 0.05);
2. Realize n experiments that give the observations $x_1, \ldots, x_n$;
3. Make hypotheses about the $X_1, \ldots, X_n$ distributions (e.g. the random variables are independent and identically distributed);
4. Set the hypothesis to test: the null hypothesis H_0 and its alternative H_1;
5. Find a common statistical test given the observations and the probability distributions;
6. Compare the obtained p-value associated with the observation y to the α risk:
   — if $p \leq \alpha$, H_0 is rejected with the risk α;
   — if $p > \alpha$, H_0 is not rejected with the risk α.

5.3 Confidence interval

Let's consider n independent observations $(x_1, \dots, x_n)$ of some random variables X_i that follow a same distribution with mean μ and variance σ^2. Let's Y be the mean of these random variables (so y is the mean of the observations $(x_1, \dots, x_n)$). By applying the Central limit theorem (CLT), if n is large enough, Y follows a normal distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

The **confidence interval**, associated with a given confidence level α, is an interval estimated from the observations $(x_1, \dots, x_n)$, that might contain the true value of the parameter μ.

The confidence interval **is not** the interval that contains the true value of the parameter with a high probability. But the probability of having observed y is very low ($\leq \alpha$) if the parameter μ is not within this interval. In other words, if confidence intervals are constructed using α from an infinite number of samplings, the proportion of those intervals that contain μ will equal α.

Given α and the observed data y, the confidence interval of the parameter μ is defined by:

$$P(\mu - e \leq Y \leq \mu + e) = 1 - \alpha$$

where e has to be determined based on the expected distribution.

By applying the CLT, $Z = \dfrac{Y - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$, *i.e.* :

$$P(\mu - e \leq Y \leq \mu + e) = 1 - \alpha$$

$$P(-e/\sqrt{\frac{\sigma^2}{n}} \leq (Y - \mu)/\sqrt{\frac{\sigma^2}{n}} \leq e/\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha$$

$$P(-e/\sqrt{\frac{\sigma^2}{n}} \leq Z \leq e/\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha$$

$$\text{Thus,} \quad e/\sqrt{\frac{\sigma^2}{n}} = \Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$$

Therefore, the confidence interval corresponds to $\left[y - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \; ; \; y + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right]$.

If the variance σ^2 is unknown, we have to estimate it using the unbiased estimator S^2 of the variance, and then $Z = \dfrac{Y - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n - 1)$.

5.4 Common statistical tests

5.4.1 One sample t-test

A **one sample t-test** (or Student's t-test) is a statistical test to determine whether the mean of the sampled observations (given by the random variables $X_1, \dots, X_n$ independent

and identically distributed) could come from a population with a given expected value μ.

The null hypothesis of the t-test corresponds to $H_0 : E(X) = \mu$.

At least one of the two following hypotheses has to be verified:

1. the random variables X_i follow a normal distribution;
2. n is large (generally $n \geq 30$, *i.e.* the CLT applies).

Let's consider the estimator M of the expected value μ and the unbiased estimator S^2 of the variance σ^2:

$$M = \frac{X_1 + \ldots + X_n}{n} \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_i (X_i - M)^2$$

The test statistic of a one sample t-test corresponds to:

$$Y = \sqrt{n}\,\frac{M - \mu}{S} \sim t(n-1)$$

The result of the t-test relies on the empirical mean M, standard deviation S and the sample size n.

Example: The temperature of a healthy human body is generally equal to $37°C$. To verify it, we sampled the temperature of 50 healthy adults, represented by the independent and identically distributed random variables $X_1, \ldots, X_n$. The question is to check whether $E(X) \neq 37$. Because $n > 30$, it fits the hypothesis of the Student's t-test with the null hypothesis $H_0 : E(X) = 37$. We fix the risk $\alpha = 0.05$. Let's consider M and S^2 the empirical mean and estimator of the variance, and Y defined as:

$$Y = \sqrt{50}\,\frac{M - 37}{S} \sim t(49)$$

H_0 is rejected if the observed y is greater than $F^{-1}_{t(49)}(1 - \alpha/2)$ or lower than $F^{-1}_{t(49)}(\alpha/2)$.

Function on R: `t.test(x=x,mu=mu)` where `x` is the vector of observations $(x_1, \ldots, x_n)$ and `mu` the expected value to test.

5.4.2 Paired sample t-test

The **paired sample t-test** (or the dependent sample t-test) is a statistical test to determine whether the mean difference between two sets of observations equals zero. Each individual is measured twice, resulting in pairs of observations (X_i, Y_i). The idea is to define a third variable Z such as $Z_i = X_i - Y_i$. If X and Y follow the same distribution, $E(Z) = 0$. The question of testing if it exists a significant difference between X and Y is equivalent to test if $E(Z) = 0$: it corresponds to a t-test with a null hypothesis $H_0 : E(Z) = 0$.

Similarly, one of the following hypotheses has to be verified: $Z = X - Y$ follows a normal distribution, or the sample size n is large.

Example: We want to verify the effect of a new drug for reducing fever. 100 adults with fever are randomly selected for taking part in the experiment. We measured for each individual i their temperature X_i before taking the drug and their temperature Y_i 30 minutes after ingestion. Let's consider $Z_i = X_i - Y_i$ and M and S^2 the empirical estimator of Z. The experimental design fits with a paired sample t-test:

$$Z = \sqrt{100}\,\frac{M}{S} \sim t(99)$$

Function on R: `t.test(x=x,y=y,paired=TRUE)` or `t.test(x=x-y,mu=0)` where `x` (resp. `y`) is the vector of observations $(x_1, \ldots, x_n)$ (resp. $(y_1, \ldots, y_n)$).

5.4.3 Unpaired sample t-test

The **unpaired sample t-test** is a statistical test to determine whether the mean difference between two sets of unpaired observations equals zero, *i.e.* whether the two sets of sampled observations could come from the same population. In other words, given the unpaired independent and identically distributed random variable $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, it tests whether X and Y follow the same distribution with an expected value μ.

The null hypothesis of the unpaired sample t-test corresponds to $H_0 : E(X) = E(Y)$.

At least one of the two following hypotheses has to be verified:

1. the random variables X_i and Y_i follow a normal distribution;
2. n is large (generally $n \geq 30$).

Let's define $M_X = \dfrac{X_1 + \ldots + X_n}{n}$ and $S_X^2 = \dfrac{1}{n-1}\sum_i (X_i - M_X)^2$ (and reciprocally M_Y and S_Y^2). The test statistic corresponds to:

$$Z = \sqrt{n}\,\frac{M_X - M_Y}{\sqrt{S_X^2 + S_Y^2}} \sim t(2(n-1))$$

The result of the t-test relies on the empirical means, the standard deviations, and the sample size n. An unpaired sample t-test can also be done in the case of unequal sample sizes n_X and n_Y:

$$Z = \frac{M_X - M_Y}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim t(n_X + n_Y - 2)$$

Example: We want to verify the effect of a new drug for reducing fever. 100 adults infected by the flu and 100 adults infected by the dengue virus are randomly selected for taking part

in the experiment. We measured the temperature X_i of the adults infected by the flu and Y_i of the adults infected by the dengue virus. Such an experimental design fits with an unpaired sample t-test.

Function on R: t.test(x=x,y=y) where x (resp. y) is the vector of observations $(x_1, \ldots, x_n)$ (resp. $(y_1, \ldots, y_n)$).

5.4.4 One-way ANOVA

The **one-way ANOVA** (or one-factor ANOVA) is a statistical test to determine whether the means of different sampled observations (generated through $p > 2$ different treatments) are different, *i.e.* whether individuals from different treatments could come at least two different populations.

For each treatment $i \in (1, ..., p)$, given n_i the number of observations for this treatment, we consider the n_i independent and identically distributed random variables $X_1^i, \ldots, X_{n_i}^i$, following the distribution $\mathcal{N}(\mu_i, \sigma^2)$.

The null hypothesis of a one-way ANOVA corresponds to $H_0 : \mu_1 = \ldots = \mu_p$.

The hypotheses of a one-way ANOVA are:

1. the independence of all random variables;
2. all random variables follow a normal distribution;
3. the variance of the different experimental treatments $(1, \ldots, p)$ are equal (called homoscedasticity).

Let's denote:

$$n = \sum_{i=1}^{p} n_i \ , \qquad M_i = \frac{X_1^i + \ldots + X_{n_i}^i}{n_i}, \quad \text{and} \quad M = \frac{\sum\limits_{i,j} X_j^i}{n}$$

We can define the factor and the residual sums of square of the observations, SS_{factor} and $SS_{residual}$, such as:

$$SS_{factor} = \sum_i n_i (M_i - M)^2, \quad \text{and} \quad SS_{residual} = \sum_i \sum_j (X_j^i - M_i)^2$$

The test statistic of the one-factor ANOVA *(ANalysis Of VAriance)* corresponds to:

$$F = \frac{SS_{factor}/(p-1)}{SS_{residual}/(n-p)} \sim F(p-1, n-p)$$

Example: We want to verify whether the human body temperature depends on the age class. 100 children (C), 100 teenagers (T), 100 adults (A), and 100 seniors (S) are randomly

selected for taking part in the experiment. We measured their temperature C_i, T_i, A_i, and S_i for all $i \in [1, 100]$. Such an experimental design fits with a one-factor ANOVA with the null hypothesis $H_0 : \mu_C = \mu_T = \mu_A = \mu_S$.

Function on R: `aov(y~f, data=df)` where `df` is a data frame with the columns `y` (with all the observation X_j^i) and `f` (with the corresponding factors).

Two-way ANOVA: The two-way ANOVA is similar to the one-way ANOVA but allows the interactions between factors, e.g. that it exists an effect of the variable A, an effect of the variable B, and an interaction between A and B.

5.4.5 Nonparametric tests for quantitative variables

The previous tests assume (1) that all random variables follow normal distributions, or (2) that the sample size n is greater than 30. If none of these hypotheses is verified, we must use the equivalent nonparametric test. These tests have generally a lower statistical power. They are based on the rank of the observations (instead of the observed values), and thus, do not rely on a hypothetical distribution.

Quantitative variable(s)	Parametric test	Nonparametric test
One sample	One-sample t-test	Wilcoxon signed rank test
Two paired samples	Paired-sample t-test	Wilcoxon signed rank test
Two unpaired samples	Unpaired-sample t-test	Mann-Whitney U test
$n \geq 3$ samples	One-factor ANOVA	Kruskal-Wallis test

5.4.6 Chi-squared test

A. Chi-squared goodness-of-fit test

The **Chi-squared goodness-of-fit test** is used for comparing the observed distribution of a categorical random variable with p categories to a theoretical distribution that present the theoretical frequencies E_i with $i \in [1, p]$. Let's consider n observations attributed to one of the p categories. We can then calculate their observed frequencies O_i with $i \in [1, p]$.

The null hypothesis of a Chi-squared goodness-of-fit test corresponds to H_0: O and E have the same distribution.

When n is large, the test statistic of a Chi-squared goodness-of-fit test corresponds to:

$$\sum_{i \in [1,p]} \frac{(nO_i - nE_i)^2}{nE_i} \sim \chi^2((p-1))$$

Example: We want to verify if the risk of having fever depends on the class age. 1,000 persons with fever are randomly selected for taking part to the experiment and for each of them we recorded their age class: children (C), teenagers (T), adults (A), and seniors (S), which gives the frequencies O_C, O_T, O_A, and O_S. Moreover, we know the frequencies of the different age classes in the total population, denoted E_C, E_T, E_A, and E_S. Such an experimental design fits with a Chi-squared test with the null hypothesis: the categorical

random variables O and E (with the categories C, T, A, and S) follow the same distribution.

Function on R: `chisq.test(x=x, p=p)` where `x` is the vector of observed frequencies and `p` is the vector of theoretical frequencies (same length as `x`).

B. Chi-squared test of independence

The **Chi-squared test of independence** can be used for testing the independence of two categorical random variables X and Y, with respectively n_X and n_Y categories. These two categorical variables can be represented within a two-way table, called a $n_X \times n_Y$ contingency table, displaying the multivariate frequency distribution.

The null hypothesis of a Chi-squared test of independence corresponds to H_0: X and Y are independent.

Let O be the observed contingency table from empirical observation (i.e. $O_{x,y}$ is the frequency of jointly observing $X = x$ and $Y = y$). One can also define E, the expected contingency table (i.e. the contingency table in a case of independence between X and Y) by computing:

$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$

When n is large, the test statistic is of the Chi-squared test of independence corresponds to:

$$\sum_{x \in [1,n_X]} \sum_{y \in [1,n_Y]} \frac{(O_{x,y} - E_{x,y})^2}{E_{x,y}} \sim \chi^2((n_X - 1)(n_Y - 1))$$

Example: We want to verify the correlation between having fever (F) and having high blood pressure (P). 10,000 adults are randomly selected for taking part in the experiment and for each of them we recorded if they experienced a fever in the previous six months (yes or no) and if they have high blood pressure (yes or no): this experiment is resumed in a contingency table. Such an experimental design fits with a Chi-squared test with the null hypothesis: having fever and having high blood pressure are independent.

Function on R: `chisq.test(x=x)` where `x` is the matrix of the table of contingency. The table of expected frequencies E is automatically computed.

French-English translation

Codomain (or target set): ensemble d'arrivée (d'une fonction)

Complementary event: événement contraire

Cumulative distribution function: fonction de répartition

Eigenvalue: valeur propre

Eigenvector: vecteur propre

Even function (contrary odd function): fonction paire (fonction impaire)

Expected value: espérance

Inverse function: fonction bijective

Gaussian elimination: pivot de Gauss

Probability distribution: loi de probabilité

Proof by contradiction: raisonnement par l'absurde

Saddle point: point-selle

Sample space: univers (de probabilités)

Standard deviation: écart-type