# Introduction aux Mathématiques
(Draft)

Master 1 Humanités Numériques
Université PSL
Semestre 1, 2025 - 2026

# Table des matières

# 1 Fondements essentiels

Cette section pose les bases du langage mathématique que nous utiliserons tout au long du cours. Notons que ce language logique est très proche de celui utilisé en algorithmique.

## 1.1 Logique élémentaire

Pour commencer par le commencement : les mathématiques consiste à construire des raisonnements à partir de propositions.

**Définition 1**

On appelle proposition (ou assertion) toute phrase dont on peut dire si elle est vraie ou fausse.

De propositions, on peut construire des théorèmes, étudier des propriétés d'objets, etc. : l'idée étant de partir d'une proposition que l'on estime vrai, et d'étudier ce que l'on en déduit pour obtenir d'autres propositions vraies à leur tour, etc.

**Exemple 1**

Quelques exemples de propositions (vraies ou fausse) :

- A : "Il existe un nombre réel dont le carré vaut 2" (Vrai)
- B : "Émile Zola a écrit Les Misérables" (Faux)
- C : "La popularité des œuvres littéraires suit une distribution en loi de puissance" (Vrai)

*N.B. :* lorsque l'on applique des mathématiques dans des domaines comme la SHS, prouver la véracité de propositions vont beaucoup passer par des raisonnement empiriques ou par des études tendancielles, là où les mathématiques plus fondamentales vont chercher à retrouver des propriétés par raisonnement basé sur des axiomes vrai à la base.

Pour formuler des propositions mathématiques, il est souvent utilisé deux expression : "pour tout", et "il existe". Si bien que ces deux expressions ont un symbole dans ce language formel et sont appelés des quantificateurs.

**Définition 2**

On introduit les symboles $\forall$ et $\exists$ tel que

- $\forall$ se lit "*pour tout*" ou "*quel que soit*"
- $\exists$ se lit "*il existe*" (sous entendu "*il existe au moins un*")
- $\exists!$ se lit "*il existe un unique*"

**Exemple 2**

- $\exists x \in \mathbb{R}, x^2 = 2$ : "il existe un nombre réel dont le carré vaut 2" (se note $\sqrt{2}$)
- Dans un corpus : $\forall$ mot $m$, la fréquence de $m$ est $\geq 0$

*Remarque* 1.1. Une inconnue (comme $x$ par exemple) est un élément non identifié au sein d'un ensemble. Si l'on veut être formel, toute introduction d'inconnu doit être suivi de l'information de l'ensemble dans lequel il se situe : $x \in \mathbb{R}$.

*Remarque* 1.2. L'ordre des quantificateurs est important !

Pour utiliser ensemble plusieurs propositions mathématiques, on utilise des connecteurs ou des opérateurs logiques (le terme n'a pas forcément d'importance) pour construire de nouvelles propositions : par exemple, si on a deux propositions A et B, on peut construire les propositions "A ou B", "A et B", mais aussi "A implique B" et "A si et seulement si B"

**Définition 3**

- On note $A \Rightarrow B$ la proposition "$A$ *implique* $B$", que l'on peut aussi traduire par "si A alors B".
- On note $A \Leftrightarrow B$ : "$A$ *équivaut à* $B$", "A si et seulement si B"
- On note $\neg A$ (appelé "non A") la négation de A, correspondant à la proposition contraire.

**Définition 4**

- Si $A \Rightarrow B$, on dit que $A$ est une **condition suffisante** pour que $B$ soit vraie.
- Si $B \Rightarrow A$, on dit que $A$ est une **condition nécessaire** pour que $B$ soit vraie.
- Si $A \Leftrightarrow B$, on dit que $A$ est une **condition nécessaire et suffisante** pour que $B$ soit vraie.

La valeur de ces nouvelles propositions construites peuvent être résumé dans une table de vérité. Les tables de vérité permettent de comprendre la logique de base. Une proposition est soit vraie (V) soit fausse (F).

| $A$ | $B$ | $A$ et $B$ | $A$ ou $B$ | $A \Rightarrow B$ | $A \Leftrightarrow B$ |
|---|---|---|---|---|---|
| V | V | V | V | V | V |
| V | F | F | V | F | F |
| F | V | F | V | V | F |
| F | F | F | F | V | V |

*Remarque* 1.3. Le cas surprenant : si $A$ est fausse, alors $A \Rightarrow B$ est vraie quel que soit $B$. Exemple : "Si Émile Zola a écrit les misérables, alors 2+2=5" est une phrase vraie !

*Remarque* 1.4 (Lien entre implication et équivalence). Une équivalence $A \Leftrightarrow B$ signifie que les deux implications sont vraies dans les deux sens :

$$A \Leftrightarrow B \text{ équivaut à } (A \Rightarrow B) \text{ ET } (B \Rightarrow A)$$

Prouvons cela en construisant la table de vérité de $(A \Rightarrow B)$ ET $(B \Rightarrow A)$ :

| $A$ | $B$ | $A \Rightarrow B$ | $B \Rightarrow A$ | $(A \Rightarrow B)$ ET $(B \Rightarrow A)$ | $A \Leftrightarrow B$ |
|---|---|---|---|---|---|
| V | V | V | V | V | V |
| V | F | F | V | F | F |
| F | V | V | F | F | F |
| F | F | V | V | V | V |

Les deux dernières colonnes sont identiques : cela prouve que $A \Leftrightarrow B$ est vraie exactement quand $(A \Rightarrow B)$ ET $(B \Rightarrow A)$ est vraie. Autrement dit :

— "A si et seulement si B" = "si A alors B" ET "si B alors A"

— Pour prouver une équivalence, il faut prouver les deux implications séparément

— Si on a seulement une implication dans un sens, on n'a PAS l'équivalence

**Exemple 3**

On définit un sonnet italien comme un poème de 14 vers, composé de 2 quatrains aux rimes embrassées, suivis de 2 tercets dont les 2 premières rimes sont identiques tandis que les 4 dernières sont embrassées (ABBA ABBA CCD EED).
Soit $A$ : "un texte est un sonnet italien" et $B$ : "un texte a 14 vers".

— **Condition nécessaire** : $A \Rightarrow B$
"Si un texte est un sonnet, alors il a 14 vers"
$\Rightarrow$ Avoir 14 vers est *nécessaire* pour être un sonnet (sans ça, impossible d'être un sonnet)

— **Condition suffisante** : $B \Rightarrow A$ est faux ici.
"Si un texte a 14 vers, alors c'est un sonnet" est **FAUX**
$\Rightarrow$ Avoir 14 vers n'est *pas suffisant* pour être un sonnet (il faut aussi la structure des rimes, le mètre, etc.)

— **Condition nécessaire et suffisante** : $A \Leftrightarrow B$ serait donc faux aussi.

**Pour avoir une équivalence**, il faudrait une définition complète :
Soit $C$ : "un texte a 14 vers, des rimes en ABBA ABBA CCD EDE, et est en alexandrins"
Alors $A \Leftrightarrow C$ : "Un texte est un sonnet si et seulement si il satisfait toutes ces conditions"
Ici, $C$ est à la fois nécessaire ET suffisant pour $A$.

*Exercice* 1.1.1. Montrons que les propositions "$A \Rightarrow B$" et "$\neg B \Rightarrow \neg A$" sont équivalentes.

| $A$ | $B$ | $\neg A$ | $\neg B$ | $A \Rightarrow B$ | $\neg B \Rightarrow \neg A$ |
|---|---|---|---|---|---|
| V | V | F | F | V | V |
| V | F | F | V | F | F |
| F | V | V | F | V | V |
| F | F | V | V | V | V |

Les deux dernières colonnes sont identiques : les propositions sont bien équivalentes !

**Application - Raisonnement par contraposée :**

Proposition originale : "Si un texte est un sonnet, alors il a 14 vers"

Contraposée : "Si un texte n'a pas 14 vers, alors ce n'est pas un sonnet"

Les deux affirmations sont logiquement équivalentes. La contraposée est parfois plus facile à vérifier : pour prouver qu'un texte n'est pas un sonnet, il suffit **peut-être** de compter ses vers (mais ça n'est pas sûr).

## 1.2 Ensembles et notations de base

### 1.2.1 Ensembles usuels

**Définition 5**

Les mathématiques travaillent avec différents types de nombres, organisés en ensembles :

- $\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$ : les **entiers naturels** (nombres pour compter)
- $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ : les **entiers relatifs** (avec les négatifs)
- $\mathbb{Q}$ : les **nombres rationnels** (fractions comme $\frac{3}{4}$, $\frac{-7}{2}$)
- $\mathbb{R}$ : les **nombres réels** (tous les nombres...réels, y compris $\sqrt{2}$, $\pi$)

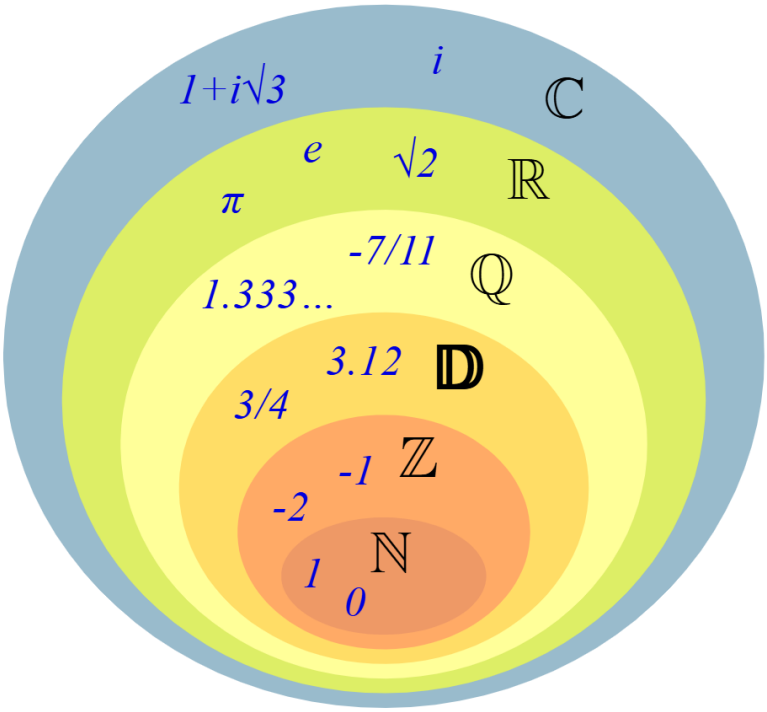


FIGURE 1 – Illustation des différents ensembles de nombre et de leur inclusion.

**Exemple 4**

Dans l'analyse d'un corpus textuel :

- Le nombre de mots dans un texte : $\mathbb{N}$ (on ne peut pas avoir $-5$ mots)
- L'évolution d'un indicateur linguistique : $\mathbb{Z}$ (peut augmenter ou diminuer)
- Une fraction / un pourcentage : $\mathbb{Q}$ (comme 67% ou $\frac{2}{3}$)
- Une mesure de similarité : $\mathbb{R}$ (souvent entre 0 et 1, mais peut être irrationnelle)

**Définition 6**

Le produit cartésien de $E$ et $F$, noté $E \times F$ est l'ensemble de tous les couples dont la première composante appartient à $E$ et la seconde à $F$. $E \times E$ est noté $E^2$.

**Exemple 5**

- $\mathbb{R}^2$ est l'ensemble de couples de réels : $\{(x, y), x \in \mathbb{R}, y \in \mathbb{R}\}$
- Si A est l'ensemble {A, R, D, V, 10, 9, 8, 7, 6, 5, 4, 3, 2} et B l'ensemble {pique, cœur, carreau, trèfle}, alors le produit cartésien de ces deux ensembles est l'ensemble à 52 éléments suivant : (A, pique), (R, pique), ... (2, pique), (A, cœur), ... (3, trèfle), (2, trèfle).

### 1.2.2 Appartenance et inclusion

**Définition 7**

- $x \in E$ : "*$x$ appartient à l'ensemble $E$*"
- $x \notin E$ : "*$x$ n'appartient pas à l'ensemble $E$*"
- $E \subset F$ : "*$E$ est inclus dans $F$*" (tous les éléments de E sont aussi dans F)

Autrement dit : $E \subset F$ signifie que $\forall x \in E$, on a $x \in F$.

*Remarque* 1.5. On a une inclusion naturelle : $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$. Chaque ensemble contient le précédent.

**Exemple 6**

- $5 \in \mathbb{N}$ mais $-3 \notin \mathbb{N}$
- $\frac{1}{2} \in \mathbb{Q}$ et $\frac{1}{2} \in \mathbb{R}$
- L'ensemble des mots d'un poème $\subset$ l'ensemble des mots de la langue française

### 1.2.3 Opérations sur les ensembles

**Définition 8**

Soient E et F deux ensembles :

- $E \cup F$ (**union**) : ensemble des éléments qui sont dans E *ou* dans B (ou les deux)
- $E \cap F$ (**intersection**) : ensemble des éléments qui sont dans E *et* dans B
- $E \setminus F$ (**différence**) : ensemble des éléments qui sont dans E mais *pas* dans F
- $\bar{E}$ (**complémentaire**) : ensemble des éléments qui ne sont pas dans E.

*Remarque* 1.6. Ne pas hésiter à faire des dessins !

**Exemple 7**

Imaginons deux corpus de textes :

- $E$ = ensemble des mots utilisés par Victor Hugo
- $F$ = ensemble des mots utilisés par Emile Zola

Alors :

- $E \cup F$ = vocabulaire total deux auteurs
- $E \cap F$ = mots partagés par les deux auteurs
- $E \setminus F$ = mots spécifiques à Victor Hugo

Il est clair (même si cela mériterait des preuves !) que l'on a les propositions suivantes :

1. $E \cap F = F \cap E$, et $E \cup F = F \cup E$
2. $E \cap F \subset E$
3. $E = F \Leftrightarrow E \subset F$ ET $F \subset E$

**Démontrons le dernier point pour avoir un exemple de rédaction de démonstration.** Pour démontrer une équivalence $A \Leftrightarrow B$, il faut prouver les deux implications :

- Sens direct : $A \Rightarrow B$
- Sens réciproque : $B \Rightarrow A$

**Sens direct :** $E = F \Rightarrow E \subset F$ ET $F \subset E$

*Hypothèse :* $E = F$ (les deux ensembles sont égaux)
*À démontrer :* $E \subset F$ ET $F \subset E$
*Démonstration :* Si $E = F$, alors par définition de l'égalité des ensembles, $E$ et $F$ contiennent exactement les mêmes éléments. Donc tout élément de $E$ est dans $F$, ce qui signifie $E \subset F$. De même, tout élément de $F$ est dans $E$, ce qui signifie $F \subset E$. Conclusion : $E \subset F$ ET $F \subset E$.

**Sens réciproque :** $E \subset F$ ET $F \subset E \Rightarrow E = F$

*Hypothèses :* $E \subset F$ ET $F \subset E$
*À démontrer :* $E = F$
*Démonstration :* Pour montrer que deux ensembles sont égaux, il faut montrer qu'ils contiennent les mêmes éléments, c'est-à-dire $x \in E \Leftrightarrow x \in F$.

— $\forall x \in E$, $E \subset F$, donc par définition $x \in F$

— $\forall x \in F$, $F \subset E$, donc par définition $x \in E$

Donc $E$ et $F$ contiennent exactement les mêmes éléments, ce qui signifie $E = F$.
**Conclusion :** Les deux implications étant démontrées, nous avons bien :

$$E = F \Leftrightarrow E \subset F \text{ ET } F \subset E$$

## 1.3 Fonctions : définition et représentation

### 1.3.1 Définition

**Définition 9**

Une **fonction** $f$ d'un ensemble $E$ vers un ensemble $F$ est une règle qui, à chaque élément $x$ de $E$, associe *un unique* élément $f(x)$ de $F$.
On note : $f : E \to F$ ou $x \mapsto f(x)$.

**Exemple 8**

— $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$ (la fonction "carré")

— len : mots $\to \mathbb{N}$, mot $\mapsto$ longueur du mot

— $h$ : textes $\to \mathbb{R}$, texte $\mapsto$ score de lisibilité

### 1.3.2 Domaine, image

**Définition 10**

Pour une fonction $f : E \to F$ :

— $E$ est **l'ensemble de départ** (ou domaine)

— $F$ est **l'ensemble d'arrivée** (ou codomaine)

— Pour $x \in E$, l'**image** de $x$ par la fonction $f$ est l'élément $f(x)$. L'image de $f$ peut aussi correspondre à l'ensemble $\{f(x) : x \in E\}$ (les valeurs effectivement atteintes)

— Pour $y \in F$, l'**antécédent** de $y$ par la fonction $f$ est **un** $x \in E$ tel que $f(x) = y$

**Attention** : il peut y avoir plusieurs antécédents, par exemple 2 et -2 sont deux antécédent de 4 par la fonction carrée.

**Exemple 9**

Pour $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$ :

- Ensemble de départ : $\mathbb{R}$ (tous les réels)
- Ensemble d'arrivée : $\mathbb{R}$ (tous les réels)
- Image : $\mathbb{R}^+ = [0, +\infty[$ (seulement les réels positifs)
- Les antécédents de 4 sont $-2$ et 2
- $-1$ n'a pas d'antécédent (aucun réel au carré ne donne $-1$)

### 1.3.3 Représentation graphique

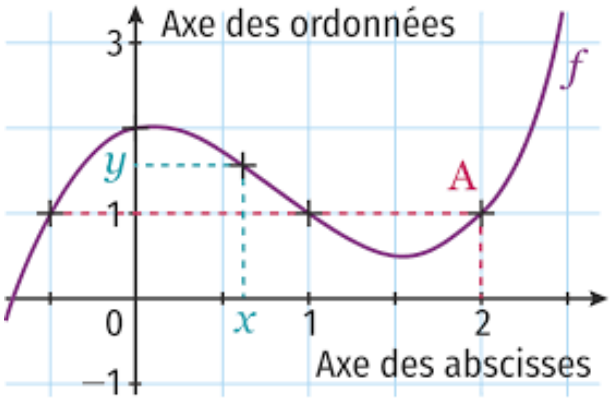


FIGURE 2 – Premier exemple d'un graphique.

*Remarque* 1.7. Un graphique de fonction se lit ainsi :

- L'axe horizontal (abscisses) représente l'ensemble de départ
- L'axe vertical (ordonnées) représente l'ensemble d'arrivée
- Chaque point $(x, y)$ signifie que $f(x) = y$

Les graphiques sont essentiels en humanités numériques :

- *Évolution temporelle* : années → fréquence d'un mot
- *Distribution* : longueur des phrases → nombre de phrases de cette longueur
- *Corrélation* : complexité syntaxique → époque du texte

Savoir lire et interpréter ces graphiques est crucial pour l'analyse de données.

# 2 Concept essentiels d'analyse

Cette section combine l'étude des fonctions usuelles avec leurs propriétés essentielles. **Pour simplifier, nous nous restreignons ici aux fonctions dont l'ensemble d'arrivée est $\mathbb{R}$, et l'ensemble de départ un sous-ensemble de $\mathbb{R}$.** En d'autres termes, les fonctions de type $f : \mathcal{D} \to \mathbb{R}$, avec $\mathcal{D} \subset \mathbb{R}$. La notations des ensembles pourra donc être omise ici lors de la définition des fonctions. Nous introduirons les concepts d'analyse

au fur et à mesure de nos besoins. **Pour chaque concept : un dessin est toujours le bienvenue (voir essentiel) pour visualiser la définition !**

## 2.1 Propriétés de base des fonctions

### 2.1.1 Opération sur les fonctions

Soit $f$ et $g$ deux fonctions défini sur $\mathcal{D}$ à valeur dans $\mathbb{R}$. De façon similaire aux nombres réelles, on peut définir les fonctions suivantes :

- La fonction somme de $f$ et $g$ : $f+g : x \to f(x)+g(x)$.
- La fonction produit de $f$ et $g$ : $fg : x \to f(x)g(x)$.
- La fonction quotient de $f$ et $g$ : $f+g : x \to \frac{f(x)}{g(x)}$. **Attention :** défini seulement $\forall x, g(x) \neq 0$.

**Définition 11**

On appelle composée de $f$ par $g$ et on note $g \circ f$ la fonction définie par $g \circ f : x \to g(f(x))$.

**Attention** : à nouveau, bien vérifier les ensemble de départ et d'arrivée des deux fonctions avant de composer.

### 2.1.2 Parité

**Définition 12**

Soit $f$ une fonction définie sur un domaine symétrique par rapport à 0 :

- $f$ est **paire** si : $\forall x, f(-x) = f(x)$ (symétrie par rapport à l'axe des ordonnées)
- $f$ est **impaire** si : $\forall x, f(-x) = -f(x)$ (symétrie par rapport à l'origine)

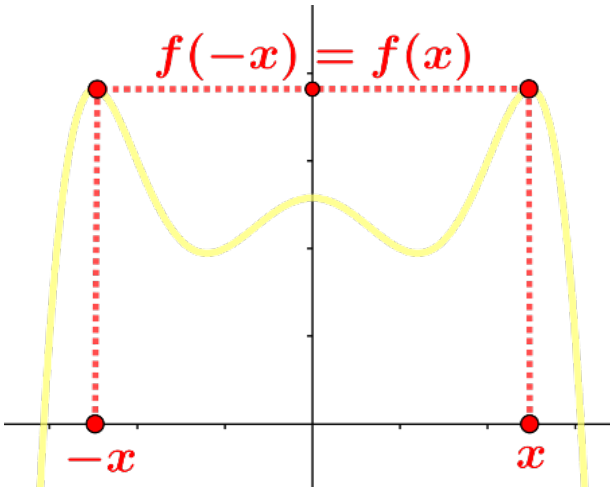


FIGURE 3 – Exemple d'une fonction paire (source : `https://jaicompris.com/lycee/math/fonction/parite/fonction-paire-impaire.php`)

### 2.1.3 Croissance et monotonie

**Définition 13**

Soit $f$ une fonction définie sur un intervalle $I$ :

- $f$ est **croissante** sur $I$ si : $\forall x_1, x_2 \in I, x_1 < x_2 \Rightarrow f(x_1) \leq f(x_2)$
- $f$ est **décroissante** sur $I$ si : $\forall x_1, x_2 \in I, x_1 < x_2 \Rightarrow f(x_1) \geq f(x_2)$
- $f$ est **strictement croissante/décroissante** si l'inégalité est stricte : $f(x_1) < f(x_2)$ et $f(x_1) > f(x_2)$
- Une fonction est dite monotone sur $I$ si elle est croissante ou décroissante sur $I$.

*Remarque* 2.1. Graphiquement :

- Fonction croissante : le graphique "monte" quand on va de gauche à droite
- Fonction décroissante : le graphique "descend" quand on va de gauche à droite

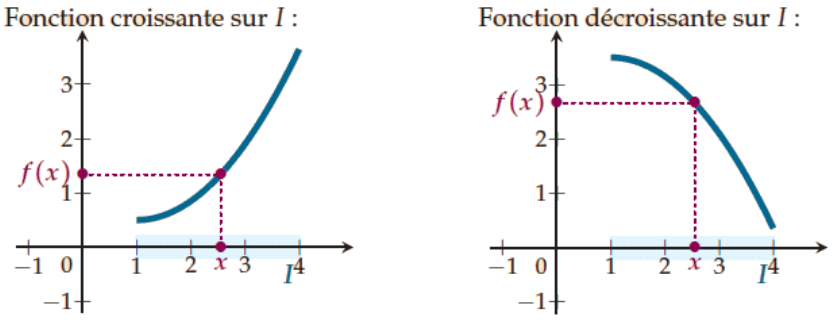


FIGURE 4 – Exemple d'une fonction croissante et d'une fonction décroissante (source : `https://mathovore.fr/variations-de-fonctions-et-extremums` )

### 2.1.4 Notion intuitive de limite

**Définition 14**

- $\lim_{x \to +\infty} f(x) = L$ : quand $x$ devient très grand, $f(x)$ se rapproche de $L$
- $\lim_{x \to +\infty} f(x) = +\infty$ : quand $x$ devient très grand, $f(x)$ devient aussi très grand
- $\lim_{x \to a} f(x) = L$ : quand $x$ se rapproche de $a$, $f(x)$ se rapproche de $L$

**Exemple 10**

On s'intéresse à la popularité d'un terme en France sur internet, sur une période donnée. Pour ce faire, Google Trends permet de visualiser l'évolution de la popularité d'une recherche donnée, et sur l'une de leur plateforme.
Par exemple, on peut s'intéresser à l'évolution de l'intérêt pour les vlogs sur youtube entre 2008 et aujourd'hui. On note $p_{\text{vlog}}$ la fonction qui, pour une date donnée $x$, associe $p_{\text{vlog}}(x)$, la **fraction** du nombre de fois où *vlog* a été recherché, divisé par le nombre de fois maximal où il a été le plus recherché (ici, en août 2025) Ici, $x$ est un mois, que l'on peut renuméroter entre 1 (Janvier 2008) et 212 (Août 2025). On a donc $p_{\text{vlog}}(212) = 1$, et si $p_{\text{vlog}}(x) = 0.1$, alors c'est que *vlog* a été 10 fois moins utilisé au mois $x$ qu'en août 2025.
Quelle propriété a cette fonction ? Difficile à dire !

- **Ensemble de définition :** Elle est définie sur $\mathbb{N}$, plus précisément sur les valeurs entière en 1 et 212.
- **Image :** Elle est à valeur dans $[0, 1]$, elle est donc positive par définition
- **Variation :** D'un point de vue strict, elle n'est ni croissante, ni décroissante, même sur un petit segment, à cause du bruit autour des observations. Une version "lissée" pourrait nous permettre de faire l'hypothèse que la fonction est croissante entre avril 2013 et décembre 2015, puis décroissante, puis croissante à nouveau.
- **Pour aller plus loin :** on observe un pic d'intérêt à chaque mois d'août $\Longrightarrow$ caractère périodique, c'est à dire qui se reproduit à intervalle régulier.

## 2.2 Fonctions de base

L'étude des fonctions permet de modéliser la réalité. Elles représentent des relations entre des quantités qui varient. Le monde qui nous entoure est rempli de phénomènes dont la valeur dépend d'un autre paramètre. Les fonctions que l'on présente ici sont les pierres de base sur lesquelles construire les relations plus complexes (ou non) qui existent entre les phénomènes.

**Définition 15**

Une fonction constante est définie $\forall x \in \mathbb{R}$ par $f(x) = c$ où $c \in \mathbb{R}$ est une constante.

**Propriétés :**

- Graphique : droite horizontale
- Á la fois croissante et décroissante (constante)
- Domaine : $\mathbb{R}$, Image : $\{c\}$
- $\lim_{x \to \pm\infty} f(x) = c$

*Exemple :* On peut supposer que $p_{\text{vlog}}$ entre 2008 et 2013 est une fonction constante, égale à 0.

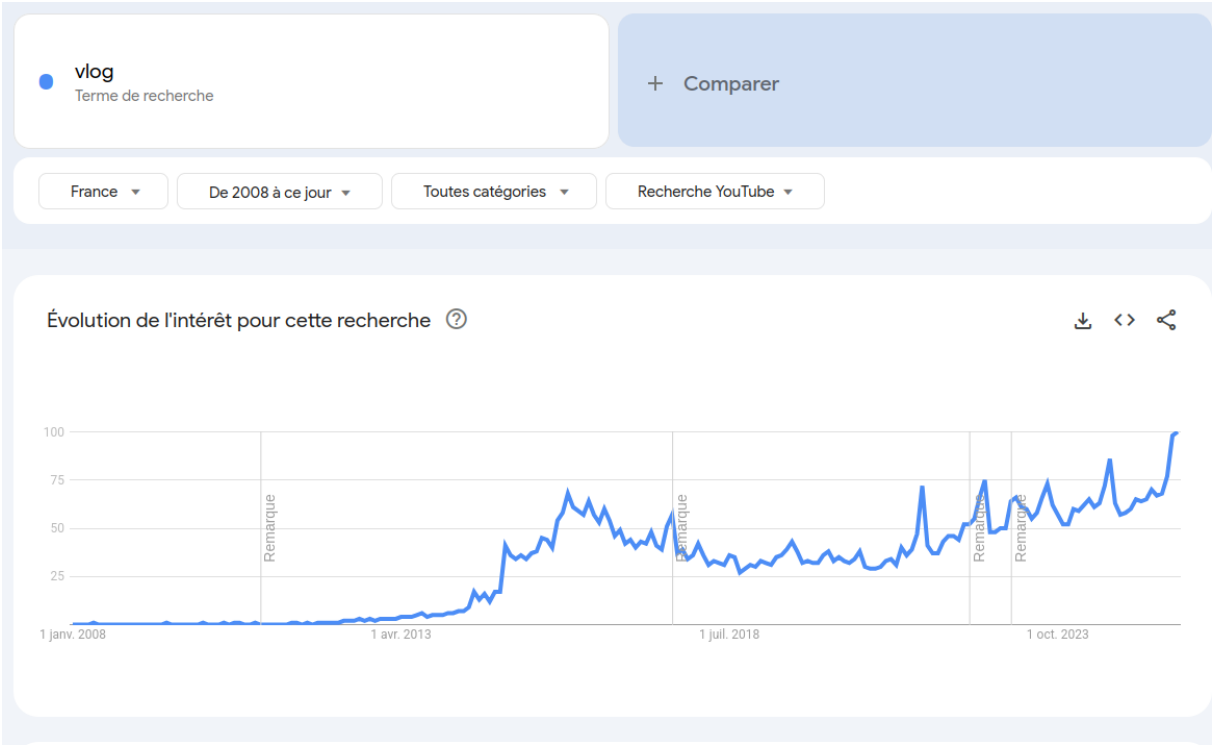


FIGURE 5 – Illustation de la fonction $p_{\text{vlog}}(x)$. Ici, le résultat est multipliée par 100 pour obtenir un pourcentage

**Définition 16**

Une fonction $f$ est dite linéaire s'il existe $a \in \mathbb{R}$ tel que $\forall x \in \mathbb{R}$

$$f(x) = ax$$

Une fonction $f$ est dit affine s'il existe $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$ tel que $\forall x \in \mathbb{R}$

$$f(x) = ax + b$$

**Propriétés :**

- Graphique : droite passant par l'origine
- Domaine : $\mathbb{R}$, Image : $\{c\}$
- Si $a > 0$ : strictement croissante
- Si $a < 0$ : strictement décroissante
- Graphique : droite de pente $a$ et d'ordonnée à l'origine $b$
- $\lim_{x \to +\infty} f(x) = +\infty$ si $a > 0$, $-\infty$ sinon
- $\lim_{x \to -\infty} f(x) = -\infty$ si $a > 0$, $+\infty$ sinon

**Définition 17**

La fonction inverse est définie $\forall x \in \mathbb{R}\backslash\{0\}$ par

$$f(x) = \frac{1}{x}$$

**Propriétés :**

— Graphique : hyperbole

— Domaine : $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ (tous les réels sauf 0)

— Image : $\mathbb{R}^*$

— Strictement décroissante sur $]0, +\infty[$ et sur $]-\infty, 0[$

— Fonction impaire : $f(-x) = \frac{1}{-x} = -\frac{1}{x} = -f(x)$

— $\lim_{x\to 0^+} f(x) = +\infty$ et $\lim_{x\to 0^-} f(x) = -\infty$

— $\lim_{x\to+\infty} f(x) = 0$ et $\lim_{x\to-\infty} f(x) = 0$

**Définition 18**

Pour tout nombre réel $x > 0$, il existe un unique nombre réel positif, noté $\sqrt{x}$ dont le carré vaut $x$. Ce nombre est appelée la racine carrée de x.

**Propriétés :**

— Graphique : courbe concave

— Domaine : $\mathbb{R}^+ = [0, +\infty[$ (réels positifs ou nuls)

— Image : $\mathbb{R}^+ = [0, +\infty[$

— Strictement croissante sur $]0, +\infty[$

— $f(0) = 0$ et $f(1) = 1$

— $\lim_{x\to+\infty} f(x) = +\infty$ (croissance "lente")

— Pour $x > 1$ : $\sqrt{x} < x$ (la racine "ralentit" la croissance)

De ces fonctions, on peut en générer une infinité d'autres en les sommant, les multipliant, les composants, etc. (voir exercices)

**Définition 19**

La fonction valeur absolue est définie $\forall x \in \mathbb{R}$ par

$$f(x) = |x| = \begin{cases} x & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$$

**Propriétés :**

— Graphique : "V" avec sommet à l'origine

— Domaine : $\mathbb{R}$, Image : $\mathbb{R}^+ = [0, +\infty[$

— Décroissante sur $]-\infty, 0]$ et croissante sur $[0, +\infty[$

— Fonction paire : $f(-x) = |-x| = |x| = f(x)$
— $f(x) = 0 \Leftrightarrow x = 0$
— $\lim_{x\to\pm\infty} f(x) = +\infty$

**Définition 20**

La fonction partie entière est définie $\forall x \in \mathbb{R}$ par

$$f(x) = \lfloor x \rfloor = \text{le plus grand entier } \leq x$$

**Propriétés :**

— Graphique : "escalier" avec des sauts aux entiers
— Domaine : $\mathbb{R}$, Image : $\mathbb{Z}$
— Constante par morceaux, croissante mais pas strictement
— Discontinue en tout point entier
— $\forall x \in \mathbb{R}, \lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$
— Exemples : $\lfloor 2.7 \rfloor = 2$, $\lfloor -1.3 \rfloor = -2$, $\lfloor 5 \rfloor = 5$

### 2.2.1 Fonctions trigonométriques

**Définition 21**

La fonction sinus (noté sin) est une fonction périodique définie par l'ordonnée des points du cercle trigonométrique en fonction de la mesure des angles (en radians) du cercle.

**Propriétés :**

— Graphique : courbe sinusoïdale oscillant entre $-1$ et $1$
— Domaine : $\mathbb{R}$, Image : $[-1, 1]$
— Fonction périodique de période $2\pi$ : $\sin(x + 2\pi) = \sin(x)$
— Fonction impaire : $\sin(-x) = -\sin(x)$
— Valeurs remarquables : $\sin(0) = 0$, $\sin\left(\frac{\pi}{6}\right) = \frac{1}{2}$, $\sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}$, $\sin\left(\frac{\pi}{3}\right) = \frac{\sqrt{3}}{2}$, $\sin\left(\frac{\pi}{2}\right) = 1$
— Croissante sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ et décroissante sur $\left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$ (puis périodique)
— La fonction oscille indéfiniment, donc les limites en $\pm\infty$ n'existent pas

**Définition 22**

La fonction cosinus (noté cos) est une fonction périodique définie par l'abscisse des points du cercle trigonométrique en fonction de la mesure des angles (en radians) du cercle.

**Propriétés :**

— Graphique : courbe cosinusoïdale oscillant entre $-1$ et $1$

— Domaine : $\mathbb{R}$, Image : $[-1, 1]$

— Fonction périodique de période $2\pi$ : $\cos(x + 2\pi) = \cos(x)$

— Fonction paire : $\cos(-x) = \cos(x)$

— Valeurs remarquables : $\cos(0) = 1$, $\cos\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2}$, $\cos\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}$, $\cos\left(\frac{\pi}{3}\right) = \frac{1}{2}$, $\cos\left(\frac{\pi}{2}\right) = 0$

— Décroissante sur $[0, \pi]$ et croissante sur $[\pi, 2\pi]$ (puis périodique)

— Relation avec sinus : $\cos(x) = \sin\left(x + \frac{\pi}{2}\right)$

— La fonction oscille indéfiniment, donc les limites en $\pm\infty$ n'existent pas

## 2.3 Introduction au concept de régression

A quoi ça sert en pratique ? Entre autres, à caractériser un phénomène ou une observation par une de ces fonctions, par exemple faire l'hypothèse que la courbe de popularité des vlogs suit une fonction polynomiale. Cet hypothèse induit une **modélisation**, qui ne collera pas parfaitement à la réalité, mais dont on peut se satisfaire si elle est relativement proche. Parmi un ensemble de candidat, par exemple l'ensemble des fonction polynomiales ou l'ensemble des fonctions affines (des droites), le concept de **régression** consiste à trouver celui qui collera le mieux aux observations. Lorsque ces candidats sont les fonctions affines, on parle de relation linéaire entre l'observation et la variable explicative, et donc de régression linéaire.

**Exemple 11**

Vous êtes un chercheur en humanités numériques et que vous voulez savoir s'il existe une relation entre le nombre de chapitres d'un livre et son prix de vente. Vous rassemblez des données sur 10 livres différents, en notant pour chaque livre son nombre de chapitres et son prix :

| Chapitres | 5 | 8 | 10 | 12 | 12 | 15 | 22 | 30 | 45 | 48 |
|---|---|---|---|---|---|---|---|---|---|---|
| Prix | 7 | 9 | 9 | 11 | 20 | 16 | 18 | 21 | 28 | 39 |

Pour explorer cette relation, la première chose à faire est de visualiser les données sur un graphique. On place le nombre de chapitres sur l'axe des abscisses (x) et le prix sur l'axe des ordonnées (y). Cela crée un nuage de points.
En regardant le nuage de points, on observe que les points ne sont pas parfaitement alignés, mais ils semblent suivre une tendance générale : plus le nombre de chapitres augmente, plus le prix augmente.
L'objectif de la régression linéaire est de trouver la ligne droite qui passe le plus près possible de tous les points du nuage. Cette ligne s'appelle la droite de régression. Elle a une équation de la forme $y = ax + b$, c'est à dire une fonction affine :

- $y = f(x)$ : le prix que l'on souhaite déduire du nombre de chapitres.
- $x$ : le nombre de chapitres
- $a$ : C'est la pente de la droite. Elle nous dit de combien le prix augmente en moyenne pour chaque chapitre supplémentaire.
- $b$ : C'est l'ordonnée à l'origine. C'est le prix de départ d'un livre (le prix qu'il aurait s'il avait 0 chapitre, ce qui n'a pas de sens physique, mais est important pour l'équation).

La droite de régression est entièrement caractérisé par $a$ et $b$, qui sont vos *paramètres*. Si on les connaît, on connaît toute la droite, et alors on peut faire des prédictions. Par exemple, si un nouveau livre a 28 chapitres, on peut utiliser l'équation de la droite pour estimer son prix. La régression linéaire nous permet de transformer une simple observation (la tendance du nuage de points) en un modèle de *prédiction*. Si vous y parvenez, félicitations, vous ferez officiellement de *l'intelligence artificiel*. Vous aurez *appris* des données un comportement général et donc fait de l'apprentissage. Comment calculer $a$ et $b$ ? Too soon pour le moment...

## 2.4 Continuité, Dérivation, Limites

### 2.4.1 Continuité

La notion de continuité est fondamentale en analyse. Intuitivement, une fonction est continue si on peut la tracer sans lever le crayon.

**Définition 23**

Une fonction $f$ est continue en un point $x_0$ de son domaine de définition si la limite de $f(x)$ lorsque $x$ tend vers $x_0$ est égale à $f(x_0)$ :

$$\lim_{x \to x_0} f(x) = f(x_0)$$

Une fonction est continue sur un intervalle si elle est continue en tout point de cet intervalle.

*Remarque* 2.2. Les fonctions "usuelles" (polynômes, exponentielles, logarithmes, etc.) que nous allons étudier sont continues sur leur domaine de définition, sauf mention contraire. La fonction "partie entière" est un contre-exemple de fonction continue.

La continuité est une hypothèse forte que l'on fait souvent en modélisation car elle simplifie beaucoup la modélisation. C'est d'ailleurs l'hypothèse que l'on a fait dans nos exemples des vlogs et des prix de livres, alors que théoriquement, ces observations sont dites **discrètes**, c'est à dire qu'elles ne peuvent prendre qu'un nombre fini de valeurs possibles, et donc non continue. En informatique : rien n'est théoriquement continue, car l'ensemble des nombres doivent être stockés, et donc nécessairement n'ont qu'un nombre fini de chiffres après la virgule maximale.

### 2.4.2 Concept de dérivée

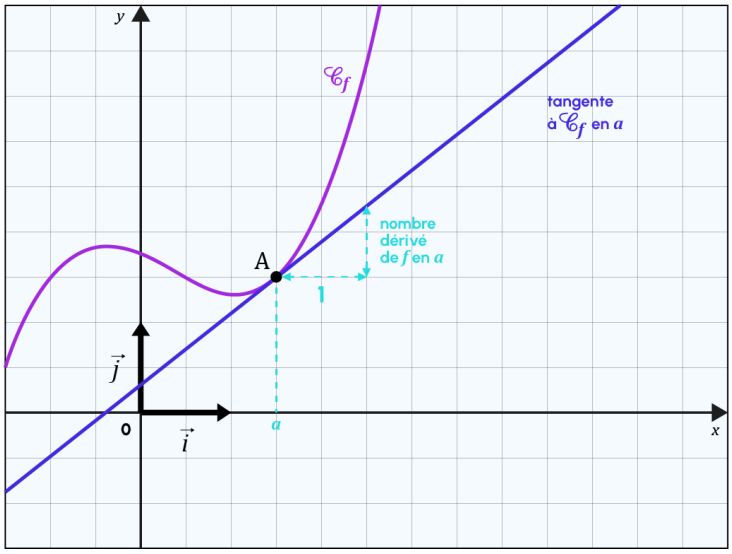


FIGURE 6 – Illustration de la dérivée en un point (source : `https://datascientest.com/coup-de-pouce-math-quest-ce-quune-derivee` )

Une fonction est croissante, mais à quel vitesse ? Quel est sa pente ? Le concept de dérivée est crucial pour comprendre la vitesse de variation d'une fonction. En d'autres termes, elle nous donne des informations sur la pente de la courbe à un instant donné.

**Définition 24**

Soit $f$ une fonction réelle. On appelle nombre dérivé de $f$ en $x_0$ la limite, si elle existe et est finie, du taux de variation de f entre $x_0$ et $x$, lorsque $x$ tend vers $x_0$. On le note $f'(x_0)$ :

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Si cette limite existe, on dit que f est dérivable en $x_0$.

Interprétation graphique : Le nombre dérivé $f'(x_0)$ est la pente de la tangente à la courbe de $f$ au point $(x_0, f(x_0))$. Plus cette pente est grande (en valeur absolue), plus la fonction varie rapidement.

Lien avec la monotonie : La dérivée est un outil puissant pour étudier la croissance d'une fonction.

— Si $f'(x) \geq 0$ sur un intervalle, alors f est croissante sur cet intervalle.

— Si $f'(x) \leq 0$ sur un intervalle, alors f est décroissante sur cet intervalle.

— Si $f'(x) = 0$ sur un intervalle, alors f est constante sur cet intervalle.

*Remarque* 2.3. Toute fonction dérivable est continue, mais toute fonction continue n'est pas dérivable.

La notion de dérivée est cruciale parce qu'elle aussi la base pour l'optimisation : Pour trouver le minimum ou le maximum d'une fonction, on cherche les points où sa dérivée s'annule (où la pente est horizontale).

### 2.4.3 Dérivées usuelles

Pour des fonctions simples, il existe des formules de dérivation qu'il est utile de connaître :

| **Fonction** | **Dérivée** | **Domaine de définition** |
|---|---|---|
| $f(x) = c$ (constante) | $f'(x) = 0$ | $\mathbb{R}$ |
| $f(x) = x$ | $f'(x) = 1$ | $\mathbb{R}$ |
| $f(x) = x^n$ | $f'(x) = nx^{n-1}$ | $\mathbb{R}$ si $n \geq 1$ |
| $f(x) = \frac{1}{x} = x^{-1}$ | $f'(x) = -\frac{1}{x^2}$ | $\mathbb{R}^*$ |
| $f(x) = \sqrt{x} = x^{1/2}$ | $f'(x) = \frac{1}{2\sqrt{x}}$ | $\mathbb{R}_+^*$ |
| $f(x) = e^x$ | $f'(x) = e^x$ | $\mathbb{R}$ |
| $f(x) = a^x$ | $f'(x) = a^x \ln(a)$ | $\mathbb{R}$ (avec $a > 0, a \neq 1$) |
| $f(x) = \ln(x)$ | $f'(x) = \frac{1}{x}$ | $\mathbb{R}_+^*$ |
| $f(x) = \log_a(x)$ | $f'(x) = \frac{1}{x \ln(a)}$ | $\mathbb{R}_+^*$ (avec $a > 0, a \neq 1$) |

TABLE 1 – Dérivées usuelles

| **Règle** | **Formule** |
|---|---|
| Linéarité $(a \in \mathbb{R}, b \in \mathbb{R})$ | $(au + bv)' = au' + bv'$ |
| Produit | $(uv)' = u'v + uv'$ |
| Quotient | $\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$ |
| Composition | $(u \circ v)' = (u' \circ v) \times v'$ |
| Fonction réciproque | $(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$ |

TABLE 2 – Règles de dérivations

Avec ces dérivées usuelles s'accompagnent les règles de dérivations : si l'on connait la dérivée de deux fonctions $u$ et $v$, alors on peut en déduire les dérivées suivantes :

## 2.5 Concept d'intégrale

L'intégration est l'opération inverse de la dérivation. Elle permet de calculer des aires et des sommes continues.

**Définition 25**

Soit $f$ une fonction. On appelle primitive de $f$ sur $\mathbb{R}$ toute fonction $F$ dérivable telle que $F'(x) = f(x) \quad \forall x \in \mathbb{R}$.

Exemple : Une primitive de la fonction $f(x) = 2x$ est la fonction $F(x) = x^2$.

**Définition 26**

L'intégrale d'une fonction $f$ entre deux points $a$ et $b$ est notée $\int_b^a f(x)dx$. Elle correspond à l'aire sous la courbe de f entre a et b. Elle se calcule à l'aide d'une primitive F de f par la formule : $\int_a^b f(x)dx = [F(x)]_a^b = F(b) - F(a)$.

L'opération intégrale peut être vue comme version continue de la somme discrète $\sum$ que l'on verra par la suite : on effectue une somme de rectangle de hauteur $f(x)$ et de largeur infinitésimal $dx$. Cette façon de définir l'intégrale s'appelle l'intégrale de Riemann, et il en existe d'autres !

*Remarque* 2.4. Une fonction $f$ a une infinité de primitives possibles qui ont $f$ pour dérivée, il suffit de considérer toutes les fonctions $\int_a^x f(x)dx$ avec n'importe quel $a$ défini sur l'ensemble de la fonction.

*Remarque* 2.5. Le lien qui relie intégrale et dérivée (où l'on dit intuitivement que ces fonctions sont réciproques) s'appelle le **théorème fondamentale de l'analyse**, c'est dire son importance.

L'application essentielle de cette notion est dans le domaine dans probabilités que nous verront plus tard. En statistiques, la probabilité d'un événement peut être calculée comme l'aire sous une courbe de densité de probabilité. L'intégrale est donc indispensable

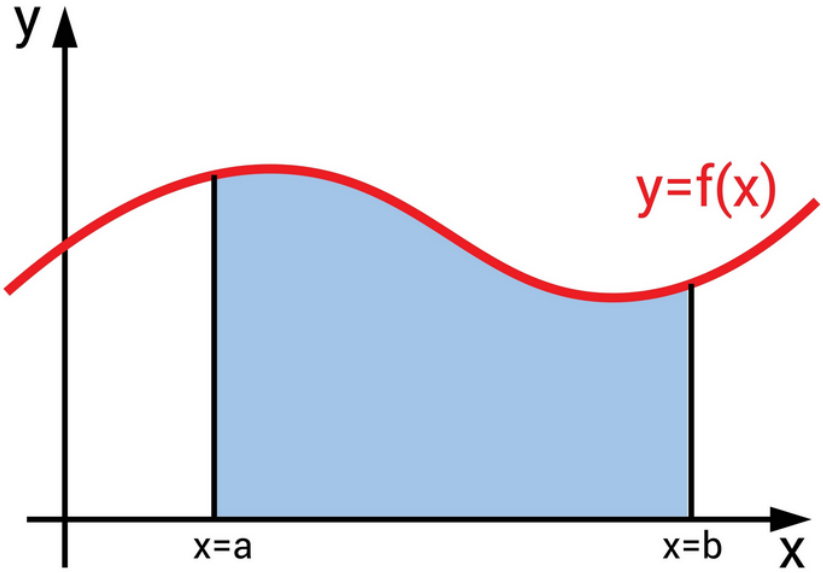


FIGURE 7 – Illustration de ce qu'est $\int_a^b f(x)dx$.

pour de nombreux calculs en statistiques descriptives et inférentielles.

Pour calculer une intégrale, on peut utiliser le tableau des dérivées usuelles (Table 1) et des règles de dérivations (Table 2) dans l'autre sens : en faisant apparaitre une forme dérivée pour en déduire la fonction qui aurait produit une telle forme. Par exemple : la fonction $f(x) = x^2$ est une puissance que l'on peut réécrire comme $f(x) = \frac{1}{3} \times 3x^2$, pour faire apparaître la forme $nx^{n-1}$ avec $n = 3$, et en déduire qu'une primitive de $x^2$ est $\frac{1}{3}x^3$.

## 2.6 Fonctions exponentielle et logarithme

Ces deux fonctions sont essentielles en mathématiques : elles servent à modéliser des phénomènes de croissance (ou de décroissance) très rapide ou, à l'inverse, de plus en plus lente.

### 2.6.1 Fonction exponentielle

**Définition 27**

La fonction exponentielle, notée $\exp(x)$ ou $e^x$, est l'unique fonction dont la dérivée est elle-même et qui vaut 1 en 0.

*Remarque* 2.6. $e$ est en fait un nombre, une constante qui vaut environ 2.71. Il est possible d'étendre la 'famille' des fonctions exponentielles avec les fonctions exponentielles en base $a$, pour tout réel $a$ : par exemple $f(x) = 2^x, 5^x$, etc. Ces fonctions auront les mêmes propriétés si ce n'est une constante qui apparaîtra si on l'a dérive.

**Propriétés :**

— Domaine : $\mathbb{R}$

— Image : $\mathbb{R}^+ = [0, +\infty[$
— Strictement croissante sur $\mathbb{R}$
— $\lim_{x\to-\infty} e^x = 0$, $\lim_{x\to+\infty} e^x = +\infty$
— $\forall x \in \mathbb{R}, \quad , f'(x) = e^x = f(x)$
— $e^{x+y} = e^x e^y$, $e^{x-y} = \frac{e^x}{e^y}$, $(e^x)^y = e^{xy}$

La fonction exponentielle aurait pu alternativement être défini de la manière suivante : c'est l'unique fonction qui transforme les sommes en produit et qui vaut la constante $e$ en 1.

**Exemple 12**

La loi de Moore est une observation empirique faite en 1965 par Gordon Moore, co-fondateur d'Intel, et qui ont trait à l'évolution de la puissance de calcul des ordinateurs et de la complexité du matériel informatique. Il a remarqué que le nombre de transistors sur une puce de microprocesseur doublait environ tous les deux ans. Cette prédiction, bien qu'elle ne soit pas une loi physique stricte, a guidé l'industrie des semi-conducteurs pendant plus d'un demi-siècle.
La loi de Moore peut être modélisée par une fonction exponentielle. Si on note $N(t)$ le nombre de transistors à l'année $t$, on peut écrire :

$$N(t) = N_0 \times 2^{t/2}$$

avec :

— $N_0$ le nombre initial de transistors à $t = 0$
— $t$ le nombre d'années écoulées (t/2 indique que le doublement des transistors intervient tous les 2 ans)

Le graphique d'une fonction exponentielle est une courbe qui démarre lentement avant de s'accélérer de manière spectaculaire : Typiquement, en imaginant que le nombre de transistors était de 1000 pour une puce en 1970 (autrement dit, $N_0 = 1000$), on a avec ce modèle $N(20) = 1000 \times 2^{10} = 1024000$ soit environ 1 million en 1990, et $N(54) = 134217728000$ soit 134 milliard en 2024 !

La puissance de calcul a permis de stocker et d'analyser des quantités massives de données. Le traitement de corpus de plusieurs millions de textes, l'analyse de réseaux sociaux ou la numérisation de vastes bibliothèques sont des tâches devenues possibles grâce à cette croissance exponentielle de la puissance des ordinateurs. Sans cette accélération, de nombreuses disciplines des humanités numériques n'existeraient tout simplement pas sous leur forme actuelle. La loi de Moore a permis de résoudre des problèmes de plus en plus complexes. Cependant, elle a également mis en lumière le concept de complexité algorithmique. Même avec une puissance de calcul exponentielle, un algorithme dont la complexité est également exponentielle reste insoluble pour de grandes entrées.

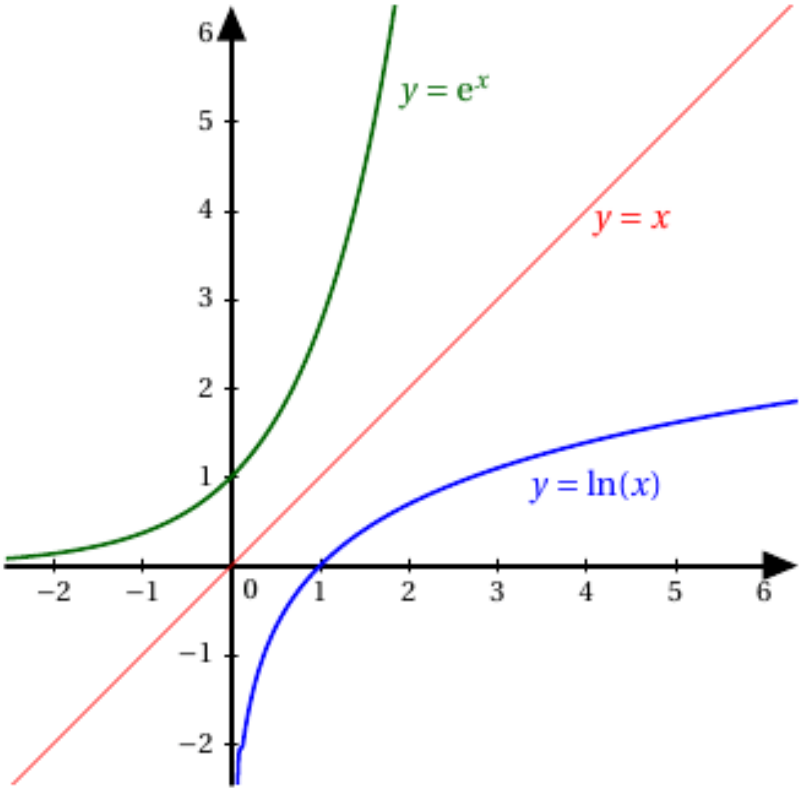


FIGURE 8 – Représentation des fonctions logarithmes et exponentielles.

### 2.6.2 Fonction logarithme

**Définition 28**

La fonction logarithme, notée $\log(x)$ ou $ln(x)$, est la réciproque de la fonction exponentielle, qui transforme les produits en somme.

La fonction logarithme aurait pu alternativement être défini de la manière suivante : c'est la primitive de la fonction inverse ($f(x) = 1/x$) qui vaut 0 en 1.

**Propriétés :**

- Domaine : $\mathbb{R}^{+*} = [0, +\infty[$
- Image : $\mathbb{R}$
- Strictement croissante sur $]0, +\infty[$
- $\lim_{x\to 0} \ln(x) = -\infty$, $\lim_{x\to+\infty} \ln(x) = +\infty$
- $\forall x \in \mathbb{R}, \quad f'(x) = 1/x$
- $\ln(xy) = \ln(x) + \ln(y)$, $\ln(x/y) = \ln(x) - \ln(y)$
- $\ln(e^x) = x$

*Remarque* 2.7. Comme la fonction exponentielle, la fonction logarithme existe en plusieurs bases ! **Il peut y avoir des confusions de notations sur le logarithme.** En France,

la notation courante est ln pour le logarithme népérien, c'est à dire celui en base $e$, dont la dérivée vaut $x \to 1/x$, et log correspond au logarithme en base 10 :

$$\log(x) = \frac{\ln(x)}{\ln(10)}$$

Ici, on obtient $\log(10) = 1$, $\log(100) = 2$, etc. Mais il est possible que parfois la notation log soit désigné pour le logarithme népérien, donc il est nécessaire de s'assure à quoi log correspond quand vous le voyez écrit.

Une utilisation très fréquente de cette fonction est l'**échelle logarithmique**.

Le logarithme a une propriété fondamentale qui le rend indispensable en science des données : il transforme une croissance exponentielle en une croissance linéaire. Pour une série de nombres en progression géométrique, par exemple 1, 10, 100, 1000, 10000... (multiplication par 10 à chaque étape), leurs logarithmes sont en progression arithmétique : $\ln(1) = 0$, $\ln(10) = 2.3$, $\ln(100) = 4.6$, $\ln(1000) = 6.9$, etc. En d'autres termes, les intervalles entre les valeurs deviennent réguliers. Ceci est particulièrement utile pour visualiser des données qui varient sur plusieurs ordres de grandeur. Si vous avez un graphique avec un axe linéaire qui va de 0 à 1 000 000, les valeurs comme 100 ou 1000 seront presque invisibles, tassées près du zéro. En revanche, sur une échelle logarithmique, ces valeurs sont réparties de manière plus équitable, ce qui permet de visualiser le comportement de l'ensemble des données, comme sur la loi de Moore.

Sur un graphique à échelle logarithmique, la courbe de croissance exponentielle se transforme en une ligne droite. Cette transformation permet de mieux observer la tendance et de confirmer que la croissance est bel et bien de nature exponentielle.

## 2.7 Comparaison des croissances

Une question cruciale en analyse et en informatique est de savoir à quelle vitesse une fonction augmente (ou diminue). Il est important de voir que deux fonctions qui tendent vers $+\infty$ en $+\infty$ par exemple ne le feront peut être pas à la même vitesse. C'est le concept de croissance comparée, qui introduit la notion de **complexité** en algorithmique.

On s'intéresse au comportement des fonctions qui tendent vers l'infini lorsque x tend vers l'infini. Les fonctions peuvent être classées par "vitesse de croissance". En général, pour $x \to +\infty$ :

— Logarithme : croissance la plus lente

— Puissances : croissance intermédiaire

— Exponentielle : croissance la plus rapide

Plus formellement, on peut le montrer avec les limites :

— $\lim\limits_{x\to+\infty} \dfrac{\log(x)}{x^n} = 0$ (le logarithme est plus "lent" que toute puissance)

— $\lim\limits_{x\to+\infty} \dfrac{x^n}{e^x} = 0$ (toute puissance est plus "lente" que l'exponentielle)

Comprendre la complexité des algorithmes permet de savoir si un algorithme sera performant pour de grandes quantités de données :

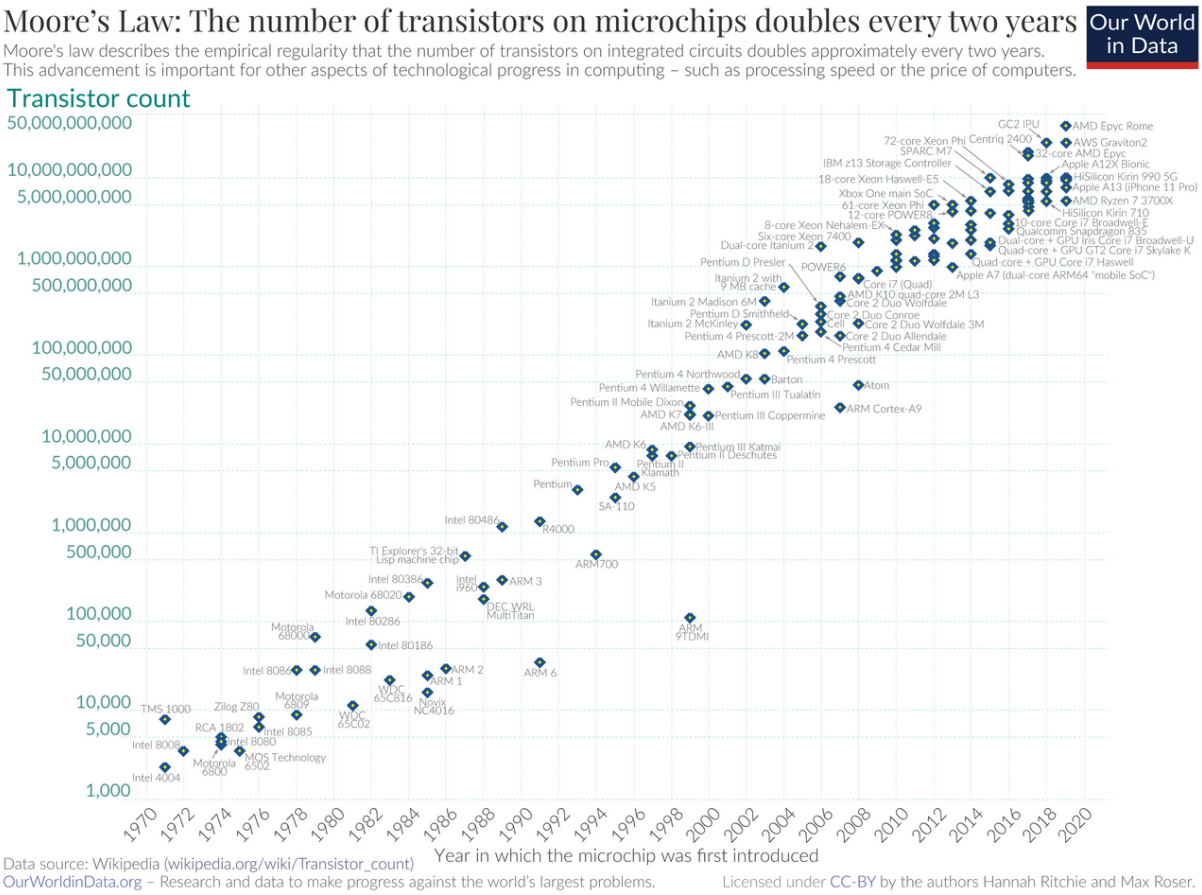


FIGURE 9 – Un graphique semi-logarithmique du nombre de transistors pour les microprocesseurs par rapport aux dates d'introduction, doublant presque tous les deux ans. (source : Wikipedia)

— Si le temps de calcul d'un algorithme est une fonction logarithmique $\log(x)$, il sera très rapide, même avec un grand nombre d'entrées. C'est le cas pour la recherche binaire.

— Si le temps de calcul est linéaire $ax$, il est très acceptable.

— S'il est polynomial ($x^2, x^3$, etc.), il est gérable pour des tailles modérées.

— S'il est exponentiel ($e^x$, $2^x$), il sera impraticable dès que le nombre de données devient un peu grand.

En humanités numériques, travailler avec de grands corpus peut mener à des problèmes de complexité : savoir si une approche (par exemple, un algorithme de classification) est *scalable* (capable de gérer une augmentation des données) dépend de sa complexité, et donc de la fonction qui modélise son temps de calcul.

# 3 Algèbre - bases

L'analyse nous a permis d'étudier les fonctions, qui décrivent des relations entre des variables uniques. Cependant, les données en humanités numériques se présentent rarement sous la forme d'une simple courbe. On travaille plus souvent avec des tableaux, des réseaux, des collections de textes, etc. L'algèbre linéaire est le langage mathématique qui permet de manipuler ces objets complexes. Elle nous donne les outils pour travailler avec des données organisées en listes (vecteurs) et en grilles (matrices).

## 3.1 Vecteurs et opérations de base

### 3.1.1 Vecteurs

Commençons par l'objet le plus fondamental de l'algèbre linéaire.

**Définition 29**

Un **vecteur** est une liste ordonnée de nombres. On le représente généralement comme une suite de nombres entre parenthèses (ou crochets dans un contexte plus 'informatique'). Si un vecteur contient n nombres, on dit qu'il est de **dimension** $n$. L'ensemble de tous les vecteurs de dimension $n$ à composantes réelles est noté $\mathbb{R}^n$.

Un vecteur $v$ de dimension 2, $v = (x, y) \in \mathbb{R}^2$, peut être vu comme une flèche partant de l'origine $(0, 0)$ et pointant vers $(x, y)$. Cette vision géométrique est utile, bien qu'en pratique les vecteurs ont souvent des dimensions bien plus grandes. Pour la notation : pour éviter de devoir trouver une lettre différente par coordonnées lorsque la dimension est trop grande, on utilise souvent une inconnue avec un indice que l'on peut faire varier : par exemple, $(x_1, x_2)$ au lieu de $(x, y)$, où $(x_1, \ldots, x_{500}) = (x_i)_{1 \leq i \leq 500}$.

En maths : un vecteur est un point sont deux objets différents, un vecteur permet de représenter une sorte de 'déplacement', avec une direction et un sens, là où un point est un 'emplacement' plus qu'un déplacement. En informatique : on parle de vecteur souvent pour désigner la structure, l'objet, même quand il s'agit de représenter un point. En fonction des languages, l'objet peut différer avec une liste ou un $n$-uplet, mais tous peuvent servir à représenter la même chose.

**Exemple 13**

Le point de départ du NLP est de considérer le texte comme une structure que l'on peut manipuler mathématiquement. Pour faire cela, l'une des représentations les plus simple d'un texte en traitement du langage naturel est le modèle **bag-of-words** ("sac de mots").
L'idée est de compter la fréquence des mots dans un documents, sans tenir compte de leur ordre ou de leur structure grammaticale. Pour l'exemple, Considérons deux phrases simples qui constitue notre corpus :

1. Document 1 : "Tu aimes les pommes et les bananes."
2. Document 2 : "Tu aimes aussi les bananes."

Du corpus, on commence par créer un vocabulaire de tous les mots uniques des deux documents, après avoir mis les mots en minuscule et retiré la ponctuation : ['tu', 'aimes', 'les', 'pommes', 'et', 'bananes', 'aussi']. Ensuite, chaque document est représenté par un vecteur où chaque position correspond à un mot du vocabulaire :

1. Document 1 : [1, 1, 2, 1, 1, 1, 0]
2. Document 2 : [1, 1, 1, 0, 0, 1, 1]

Le résultat est une représentation numérique des documents qui peut être utilisée par des algorithmes d'apprentissage automatique, par exemple pour de la classification de texte, la détection de spam, de topics, etc. En pratique : la dimension de ces vecteurs peut être de plusieurs dizaines de milliers dans les applications réelles.

Pour manipuler les composantes de vecteur de taille quelconque, des nouvelles notations sont indispensables.

### 3.1.2 Symbole de somme et de produit

**Définition 30**

Le symbole **somme**, noté $\sum_{k=i}^{j}$, permet de noter de façon compacte la somme d'une série de termes allant d'un indice $i$ à un indice $j$ donnée :

$$\sum_{k=i}^{j} x_k = x_i + x_{i+1} + \ldots + x_{j-1} + x_j$$

Ici, k est l'indice de sommation, i est la valeur de départ et j la valeur de fin.

Notons que l'indice $k$ est dit 'muet', car il n'a aucune signification en dehors de la somme.

Quelques propriétés :

— $\sum_{k=i}^{j} x_k$ contient $(j - i + 1)$ termes (**attention** : $i \leq j$, sinon aucuns terme par

convention)

— $\sum_{k=i}^{j} x_k + \sum_{k=j+1}^{l} x_k = \sum_{k=i}^{l} x_k$

— $\sum_{k=i}^{j} (x_k + y_k) = \sum_{k=i}^{j} x_k + \sum_{k=i}^{j} y_k$

— $\sum_{k=i}^{j} a x_k = a \sum_{k=i}^{j} x_k$

**Définition 31**

De manière similaire, le symbole **produit**, noté $\prod_{k=i}^{j}$, permet de noter la multiplication d'une série de termes de $i$ à $j$ :

$$\prod_{k=i}^{j} x_k = x_i \times x_{i+1} \times \ldots \times x_{j-1} \times x_j$$

On retrouve des propriétés similaires à la somme :

— $\prod_{k=i}^{j} x_k$ contient $(j - i + 1)$ termes (**attention** : $i \leq j$, sinon aucuns terme par convention)

— $\prod_{k=i}^{j} x_k \times \prod_{k=j+1}^{l} x_k = \prod_{k=i}^{l} x_k$

— $\prod_{k=i}^{j} x_k y_k = \prod_{k=i}^{j} x_k \times \prod_{k=i}^{j} y_k$

— $\prod_{k=i}^{j} a x_k = a^{j-i+1} \prod_{k=i}^{j} x_k$.

**Exemple 14**

La factorielle d'un nombre $n$, notée $n!$, est le produit de tous les entiers de 1 à $n$ :

$$n! = \prod_{k=1}^{n} k = 1 \times 2 \times \ldots \times n$$

La valeur de la factorielle augmente très rapidement : par exemple, $4! = 24$ et $10! = 3628800$.
Cette fonction est fondamentale en combinatoire (et donc utilisé en probabilité/statistiques). Par exemple, la factorielle de 10 exprime le nombre de combinaisons possibles de placement des 10 convives autour d'une table (on parle de permutation des convives). Le premier convive s'installe sur l'une des 10 places à sa disposition. Chacun de ses 10 placements ouvre 9 nouvelles possibilités pour le deuxième convive, celles-ci 8 pour le troisième, et ainsi de suite.

### 3.1.3 Opération sur les vecteurs

Soient $u = (u_1, u_2, \ldots, u_n)$ et $v = (v_1, v_2, \ldots, v_n)$ deux vecteurs de même dimension $n \in \mathbb{N}$, et c un nombre réel ($c \in \mathbb{R}$). Nous pouvons déjà définir deux opérations de base :

— **Addition/soustraction de vecteurs** : $u + v = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n)$

— **Multiplication par un scalaire** : $c.u = (cu_1, \ldots, cu_n)$

Comment comparer deux vecteurs ? Le produit scalaire est un premier outil.

**Définition 32**

Le **produit scalaire** de deux vecteurs $u$ et $v$ de même dimension $n$, noté $< u, v >$, est le nombre obtenu en sommant les produits de leurs composantes correspondantes :

$$< u, v >= \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

*Remarque* 3.1. Le résultat du produit scalaire n'est pas un vecteur, mais un nombre unique (un scalaire).

Le produit scalaire a une interprétation géométrique très forte : il est lié à l'angle entre les vecteurs. Si $\theta$ est l'angle entre $u$ et $v$, alors $< u, v >= ||u||||v|| \cos(\theta)$ (où $||u||$ est la "longueur" du vecteur $u$, que nous définirons juste après).

**Interprétation :**

— Si $< u, v >> 0$, les vecteurs pointent globalement dans la même direction.

— Si $< u, v >< 0$, ils pointent dans des directions opposées.

— Si $< u, v >= 0$, ils sont dits **orthogonaux** (perpendiculaires). En analyse de données, cela est souvent interprété comme une absence de corrélation ou de similarité.

En résumé, le produit scalaire mesure à quel point deux vecteurs pointent dans la même direction.

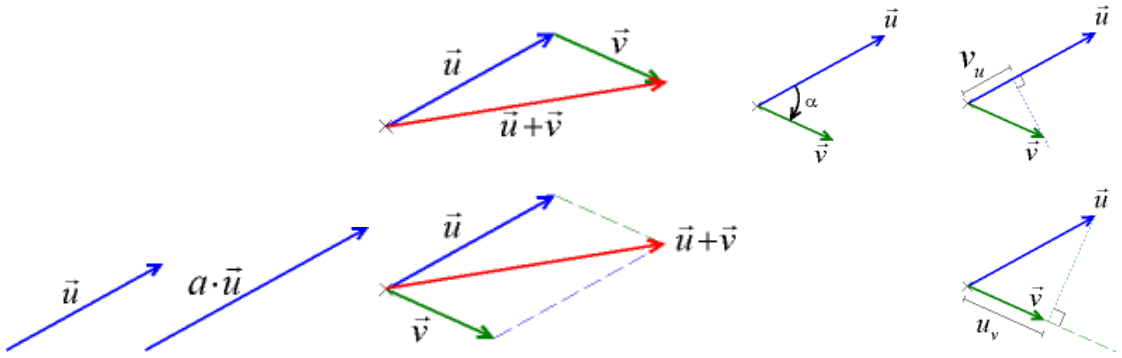


FIGURE 10 – Représentation du produit d'un vecteur par un nombre réel, de la somme et du produit scalaire de deux vecteurs (source : Wikipedia).

### 3.1.4 Norme et distance

**Définition 33**

La **norme** d'un vecteur $v \in \mathbb{R}^n$, notée $||v||$, correspond à sa longueur. Il existe une infinité de norme possibles, en réalité toute application qui satisfait trois conditions qu'on ne détaillera pas ici [a] peut être défini comme une norme. La norme la plus courante et la plus utilisée appelé norme Euclidienne ou norme 2, est définie par

$$||v||_2 = \sqrt{\sum_{i=1}^{n} v_i^2} = \sqrt{v_1^2 + \ldots + v_n^2}$$

*a.* Séparation, homogénéité en valeur absolue et inégalité triangulaire **exercice ?**

*Remarque* 3.2. La norme euclidienne est une généralisation du théorème de Pythagore à n dimensions.

Il existe d'autres normes couramment utilsées :

— La norme 1 :

$$||v||_1 = \sum_{i=1}^{n} |v_i| = |v_1| + \ldots + |v_n|$$

— Plus généralement, la norme $p$ avec $p \geq 1$ qui généralise la norme 1 et la norme 2 :

$$||v||_p = \left(\sum_{i=1}^{n} v_i^p\right)^{1/p} = (v_1^p + \ldots + v_n^p)^{1/p}$$

— La norme infini :

$$||v||_\infty = \max\{v_1, \ldots, v_n\}$$

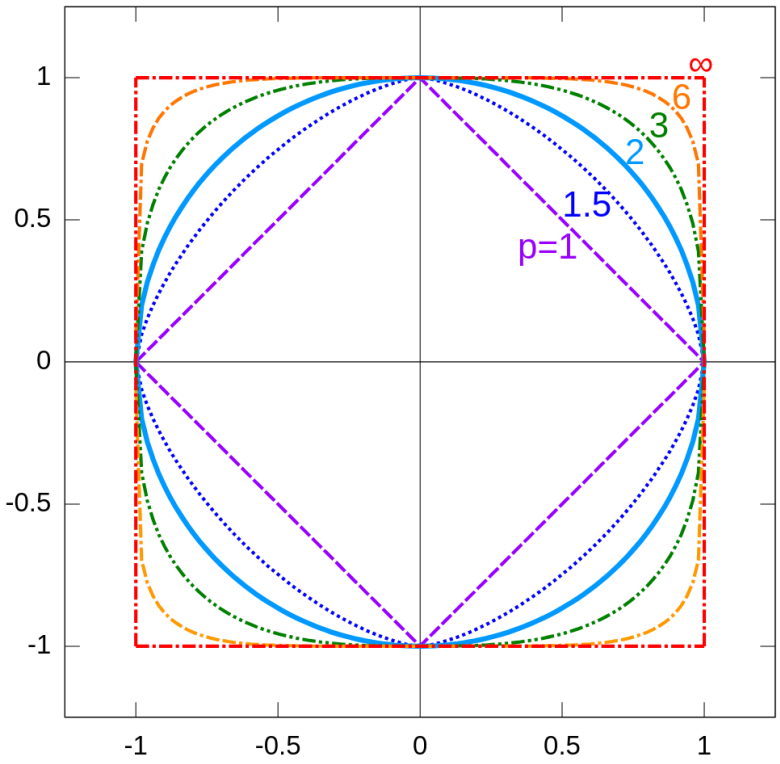


FIGURE 11 – Ensemble des vecteurs de norme $p$ dans $\mathbb{R}^2$ égale à 1 pour différentes valeurs de $p$ (source : Wikipedia)

**Définition 34**

On définit la distance $d(u, v)$ entre deux vecteurs $u$ et $v$ pour une norme donnée comme étant la norme de leur différence :

$$d(u, v) = ||v - u||$$

Dans le cas de la norme 2, la distance euclidienne s'écrit donc

$$||v - u||_2 = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}$$

**Exemple 15**

Reprenons l'exemple de nos deux documents représentés par des vecteurs bag-of-words. La distance euclidienne entre les deux vecteurs que l'on va noter $d_1$ et $d_2$ s'écrit :

$$||d_1 - d_2||_2 = \sqrt{(1-1)^2 + (1-1)^2 + \ldots + (0-1)^2} = \sqrt{4} = 2$$

Ce nombre représente à quel point les "profils de mots" des deux documents sont différents. En visualisant les documents comme des points dans un espace, la distance est simplement la distance qui les sépare. C'est le concept fondamental derrière les algorithmes de classification et de clustering (regroupement), qui cherchent à rassembler les points (données) proches les uns des autres.

Il est important de distinguer le rôle du produit scalaire et celui de la distance :

— **Produit scalaire :** pour mesurer la similarité thématique, la corrélation, l'alignement conceptuel

— **Distance :** pour mesurer la différence quantitative, créer des clusters, faire de la classification

## 3.2 Matrices (introduction)

### 3.2.1 Définition et notation

**Définition 35**

Une **matrice** est un tableau rectangulaire de nombres, organisés en lignes et en colonnes. Une matrice ayant $m$ lignes et $n$ colonnes est dite de taille $m \times n$. On note $A = (a_{i,j})$ où $a_{i,j}$ est l'élément situé à la $i$-ème ligne et à la $j$-ème colonne.

**Exemple 16**

La matrice $A = \begin{pmatrix} 2 & 2 & 1 \\ 4 & 8 & 0 \end{pmatrix}$ est une matrice de taille 2×3. L'élément $a_{2,1} = 4$.

Une matrice peut être vue comme une collection de vecteurs (vecteurs lignes ou vecteurs colonnes). Dans notre exemple "sac de mots", si nous avions 100 documents, nous pourrions les rassembler dans une matrice de taille 7×100, où chaque colonne est le vecteur d'un document. C'est ce qu'on appelle une matrice termes-documents.

### 3.2.2 Opérations de base

Tout comme les vecteurs, les matrices peuvent être additionnées et multipliées par des scalaires, à condition qu'elles aient la même taille. Les opérations se font simplement élément par élément.

**Exemple 17**

Si $A = \begin{pmatrix} 2 & 2 & 1 \\ 4 & 8 & 0 \end{pmatrix}$ et $B = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 4 & 3 \end{pmatrix}$,
alors $A + B = \begin{pmatrix} 3 & 1 & 2 \\ 6 & 12 & 3 \end{pmatrix}$ et $-2A = \begin{pmatrix} -4 & -4 & 1 \\ -8 & -16 & 0 \end{pmatrix}$

### 3.2.3 Transposée d'une Matrice

Transposer une matrice, c'est la "faire pivoter" le long de sa diagonale principale (celle qui part d'en haut à gauche). Les lignes deviennent des colonnes et les colonnes deviennent des lignes.

**Définition 36**

Soit une matrice $A$ de taille $m \times n$. Sa transposée, notée $A^T$, est la matrice de taille $n \times m$ obtenue en échangeant les lignes et les colonnes de $A$. L'élément à la position $(i, j)$ de $A^T$ est l'élément qui était à la position $(j, i)$ de $A$.

**Exemple 18**

Soit la matrice $A$ de taille $2 \times 3$ :

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

Sa transposée $A^T$ est une matrice de taille $3 \times 2$ :

$$A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$

La première ligne de $A$ (1, 2, 3) est devenue la première colonne de $A^T$.

La transposée est un outil de "reformatage" essentiel. Souvent, les données que vous collectez ne sont pas dans le bon format pour l'algorithme que vous souhaitez utiliser. Par exemple, un logiciel peut s'attendre à avoir les individus en lignes et les variables en colonnes, mais votre tableau de données est organisé de manière inverse. Une simple transposition de votre matrice de données résout le problème.

### 3.2.4 Produit matricielle

La multiplication de deux matrices est plus complexe mais aussi beaucoup plus puissante. Sa définition n'est pas intuitive au premier abord, mais elle découle d'une idée fondamentale : une matrice peut représenter une **transformation**.

**Définition 37**

Le produit de deux matrices A (de taille m×n) et B (de taille n×p) est une matrice C=AB (de taille m×p). L'élément $c_{i,j}$ de la matrice C est obtenu en faisant le **produit scalaire** de la i-ème ligne de A avec la j-ème colonne de B :

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$$

**De façon très intuitive, pourquoi cette définition ?**

Imaginez un vecteur comme un point dans un espace. Multiplier ce vecteur par une matrice "déplace" ce point vers un nouvel emplacement. La matrice agit comme une fonction, une "machine à transformer" les vecteurs. Par exemple, la multiplication $Av = w$ transforme le vecteur $v$ en un nouveau vecteur $w$. L'idée clé du produit matricielle est la composition de transformations. Multiplier une matrice A par une matrice B pour obtenir $C = AB$ revient à créer une nouvelle transformation C qui applique d'abord la transformation B, puis la transformation A.

*Remarque* 3.3 (Condition de compatibilité). Pour que le produit AB soit possible, le nombre de colonnes de A doit être égal au nombre de lignes de B.

*Remarque* 3.4. La multiplication matricielle n'est **pas commutative** : en général, $AB \neq BA$.

**Exemple 19**

Si $A = \begin{pmatrix} 2 & 2 & 1 \\ 4 & 8 & 0 \end{pmatrix}$ et $B = \begin{pmatrix} 1 & -1 \\ 1 & 2 \\ 4 & 3 \end{pmatrix}$,
alors le produit $C = AB$ sera de taille 2×2.

- $c_{1,1}$ = (ligne 1 de A) × (colonne 1 de B) $= 2 \times 1 + 2 \times 1 + 1 \times -1 = 8$.
- $c_{1,2}$ = (ligne 1 de A) × (colonne 2 de B) $= 2 \times -1 + 2 \times 2 + 1 \times 3 = 5$.
- $c_{2,1}$ = (ligne 2 de A) × (colonne 1 de B) $= 4 \times 1 + 8 \times 1 + 0 \times 4 = 12$.
- $c_{2,2}$ = (ligne 2 de A) × (colonne 2 de B) $= 4 \times -1 + 8 \times 2 + 0 \times 3 = 12$.

Donc $C = AB = \begin{pmatrix} 8 & 5 \\ 12 & 12 \end{pmatrix}$.

### 3.2.5 Applications

Les matrices permettent de modéliser des relations au sein de grands ensembles de données.

- **La matrice termes-documents** : C'est l'exemple que nous avons déjà esquissé. Une grande matrice où les lignes représentent les mots d'un vocabulaire et les colonnes des documents. L'élément $a_{i,j}$ est le nombre de fois que le mot $i$ apparaît dans le document $j$. Cette matrice est le point de départ de nombreuses méthodes d'analyse de texte, comme le "*topic modeling*" (détection de thèmes).

— **La matrice d'adjacence pour les réseaux** : Pour analyser un réseau social (par exemple, les relations entre personnages dans un roman), on peut construire une matrice carrée où les lignes et les colonnes représentent les personnages. On met un 1 dans la case $(i, j)$ si le personnage $i$ est lié au personnage $j$, et 0 sinon. L'analyse de cette matrice (par exemple en la multipliant par elle-même) permet de révéler des structures cachées dans le réseau : qui sont les personnages centraux, quelles sont les communautés, etc.

— **Une image comme une matrice** : Une image numérique en noir et blanc n'est rien d'autre qu'une matrice où chaque élément représente l'intensité d'un pixel (par exemple, de 0 pour noir à 255 pour blanc). Appliquer des opérations matricielles à cette image permet de la transformer : la rendre plus floue, augmenter le contraste, détecter les contours, etc. Une image couleur peut être vue comme la superposition de trois matrices (une pour le rouge, une pour le vert, une pour le bleu).

En maîtrisant les vecteurs et les matrices, vous ne voyez plus seulement des textes, des images ou des relations sociales, mais des structures de données sur lesquelles vous pouvez appliquer des calculs puissants pour extraire du sens.

**Exemple 20**

Considérons un petit réseau de 4 personnes : Alice (A), Bob (B), Charles (C) et Diane (D). Les relations sont des "suivis" sur un réseau social (une relation orientée : A peut suivre B sans que B suive A). Les liens directs (chemins de longueur 1) sont :

- Alice suit Bob et Charles.
- Bob suit Charles.
- Charles suit Alice.
- Diane suit Alice et Charles.

Construisons la matrice d'adjacence M, où $m_{i,j} = 1$ si i suit j. En ordonnant les lignes/colonnes A, B, C, D :

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Question : Qui peut atteindre qui en passant par exactement un intermédiaire ? (c'est-à-dire, en 2 étapes, ou par un "chemin de longueur 2"). Par exemple, Alice suit Charles ($A \to C$) et Charles suit Alice ($C \to A$), donc Alice peut atteindre Alice en 2 étapes.
Calculons $M^2 = M \times M$.

$$M^2 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Interprétons le résultat. Prenons l'élément (1,1) de $M^2$ (ligne Alice, colonne Alice). Le seul terme non nul dans la multiplication est $m_{1,3} \times m_{3,1} = 1$, ce qui signifie "Alice suit Charles ET Charles suit Alice". C'est bien un chemin de longueur 2 de A vers A.
Prenons l'élément (4,2) (ligne Diane, colonne Bob). Sa valeur est 1. Le calcul est la ligne 4 de M fois la colonne 2 de M, et le terme non nul $m_{4,1} \times m_{1,2}$ signifie "Diane suit Alice ET Alice suit Bob". C'est un chemin de longueur 2 de Diane vers Bob ($D \to A \to B$).
Conclusion : La matrice $M^2$ nous donne le **nombre de chemins de longueur 2** entre chaque paire de nœuds. De même, $M^3$ donnerait le nombre de chemins de longueur 3, et ainsi de suite. Le produit matriciel, qui semblait abstrait, devient un outil d'analyse puissant pour comprendre la connectivité et la propagation de l'information dans un réseau.

### 3.2.6 Inverse d'une Matrice

Pour les nombres, si vous avez l'équation $5x = 10$, vous "divisez" par 5 pour trouver $x = 2$. En réalité, vous multipliez par l'inverse de 5, qui est 1/5. L'inverse d'une matrice,

c'est l'équivalent de la division. C'est la matrice qui "annule" ou "défait" l'action de la matrice originale.

**Définition 38**

Soit $A$ une matrice carrée. Son inverse, si elle existe, est une matrice notée $A^{-1}$ telle que :

$$A \times A^{-1} = A^{-1} \times A = I$$

où $I$ est la **matrice identité** (l'équivalent du nombre 1 pour les matrices : une matrice avec des 1 sur la diagonale et des 0 partout ailleurs).

*Remarque* 3.5. Toutes les matrices n'ont pas d'inverse ! De la même manière que le nombre 0 n'a pas d'inverse pour la multiplication, certaines matrices (dites "singulières") ne peuvent pas être inversées. Le calcul de l'inverse pour de grandes matrices est complexe et est toujours confié à des logiciels.

Inverser une matrice est la clé pour **résoudre des systèmes d'équations linéaires**. De nombreux problèmes de modélisation (en économie, en sociologie, en linguistique computationnelle...) se résument à une équation de la forme $Ax = b$, où :

— $A$ est une matrice qui représente le fonctionnement de votre système (connu).

— $b$ est le résultat que vous observez (connu).

— $x$ est le vecteur des causes ou des paramètres que vous cherchez (inconnu).

Pour trouver $x$, il suffit de multiplier par l'inverse de $A$ :

$$x = A^{-1}b$$

C'est la méthode utilisée par les logiciels de statistiques pour trouver les coefficients d'un modèle de régression, c'est-à-dire pour trouver la "meilleure" relation entre plusieurs variables explicatives et une variable à expliquer.

## 3.3 Introduction à la Décomposition en Valeurs Singulières

La Décomposition en Valeurs Singulières (SVD) est l'une des idées les plus puissantes de l'algèbre. Elle prend une matrice complexe $A$ et la réécrit comme le produit de trois matrices aux rôles très clairs : $A = U\Sigma V^T$.

— $A$ : la matrice de départ, par exemple une matrice Mots × Documents. Elle est grande, pleine de zéros, et "bruyante".

— $U$ : une matrice dites orthogonale, c'est à dire que $U^{-1} = U^T$. Peut être vu comme une matrice qui synthétise toutes les informations de corrélations des mots entre eux.

— $V$ : une matrice également orthogonale, où $V^T = V^{-1}$. Peut être vu comme une matrice qui synthétise toutes les informations de corrélations des documents entre eux.

— $\Sigma$ (Sigma) : une matrice diagonale qui contient les **valeurs singulières**. Chaque valeur est un nombre qui représente la force ou l'importance d'un sujet. La première valeur est la plus grande et correspond au sujet le plus important du corpus, etc.

**Exemple 21**

L'application la plus célèbre de la SVD en humanités numériques est l'Analyse Sémantique Latente (LSA). En analysant une matrice mots-documents avec la SVD, on ne se contente plus de chercher des mots-clés. On cherche des concepts.

- Réduction du bruit : En ne gardant que les quelques sujets les plus importants (les plus grandes valeurs singulières de $\Sigma$), on peut reconstruire une version "nettoyée" de la matrice originale, qui ignore les usages rares ou anecdotiques des mots.
- Découverte de synonymes et de relations : La SVD peut regrouper "voiture" et "automobile" dans le même sujet, car ils apparaissent dans des documents similaires (des contextes similaires).
- Recherche sémantique : Si vous cherchez le mot "bateau", un moteur de recherche basé sur la LSA pourra vous retourner un document qui ne parle que de "voilier" et de "navire", car la SVD aura compris que ces mots appartiennent au même "sujet maritime".

En somme, la SVD est un outil magique qui permet de passer d'une vision de surface (la présence/absence de mots) à une compréhension profonde des structures sémantiques cachées (latentes) dans un grand corpus de textes.

# 4 Probabilités

Jusqu'à présent, nous avons manipulé des objets bien définis : des nombres, des fonctions, des vecteurs, des matrices. Cependant, le monde des données est rarement aussi certain. Un manuscrit a-t-il été écrit par tel auteur ? Quel mot est le plus susceptible de suivre une séquence donnée ? Pour répondre à ces questions, nous devons quantifier l'incertitude. C'est le rôle des probabilités, qui sont le langage mathématique du hasard et le fondement de la statistique et de l'apprentissage automatique.

## 4.1 Concepts fondamentaux

**Définition 39**

Une **expérience aléatoire** est une expérience dont le résultat ne peut pas être prédit avec certitude avant qu'elle ne soit réalisée, même si on peut décrire l'ensemble des résultats possibles.

- L'**espace des possibles** (ou univers), noté $\Omega$, est l'ensemble de tous les résultats possibles d'une expérience aléatoire.
- Un **événement** est un sous-ensemble de l'espace des possibles. C'est un ensemble de résultats auxquels on s'intéresse.

**Exemple 22**

Lancer un dé à six faces est une expérience aléatoire. On sait que le résultat sera dans $\{1, 2, 3, 4, 5, 6\}$, mais on ne sait pas lequel. Ici :

- $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- L'événement "obtenir un nombre pair", notons-le $A$, est le sous-ensemble $A = \{2, 4, 6\}$.
- L'événement "obtenir un 3", notons-le $B$, est le sous-ensemble $B = \{3\}$.

**Définition 40**

Une probabilité est une fonction, notée $P$, qui associe à chaque événement $A$ un nombre entre 0 et 1, représentant la "chance" que cet événement se produise. $P(A) = 0$ signifie que l'événement est impossible, et $P(A) = 1$ signifie qu'il est certain.

### 4.1.1 Opérations sur les événements

Puisque les événements sont des ensembles, nous pouvons leur appliquer les mêmes opérations que celles vues dans la section sur les fondements (union, intersection, etc.). Ces opérations correspondent aux connecteurs logiques "OU", "ET" et "NON". N'hésitez pas à visualiser ces opérations avec des schémas.

Soient A et B deux événements :

- L'événement **union** $A \cup B$ ("A ou B") se réalise si *au moins un* des deux événements se réalise.
- L'événement **intersection** $A \cap B$ ("A et B") se réalise si les *deux* événements se réalisent simultanément.
- L'événement **complémentaire** $\bar{A}$ ("non A") se réalise si l'événement A *ne se réalise pas.*

**Exemple 23**

Prenons l'expérience "tirer un mot au hasard d'un corpus".

- Soit $A$ l'événement "le mot est un verbe".
- Soit $B$ l'événement "le mot a plus de 5 lettres".

Alors :

- $A \cup B$ est l'événement "le mot est un verbe OU a plus de 5 lettres".
- $A \cap B$ est l'événement "le mot est un verbe ET a plus de 5 lettres" (par exemple, "mangeaient").
- $\bar{A}$ est l'événement "le mot n'est pas un verbe".

Pour être valide, une fonction de probabilité doit respecter trois règles simples (dites "axiomes de Kolmogorov") :

1. **Positivité :** Une probabilité est toujours positive ou nulle. $P(A) \geq 0$.

2. **Certitude :** La probabilité de l'espace des possibles $\Omega$ (c'est-à-dire "quelque chose va se passer") est 1 : $P(\Omega) = 1$.
3. **Additivité :** Si deux événements A et B sont *incompatibles* (ils ne peuvent pas se produire en même temps, $A \cap B = \emptyset$), la probabilité que l'un OU l'autre se produise est la somme de leurs probabilités : $P(A \cup B) = P(A) + P(B)$.

De ces règles, on déduit que la somme des probabilités de tous les résultats élémentaires doit être égale à 1.

Comment assigner ces probabilités ?

— En **théorie des probabilités**, on définit un modèle abstrait (ex : "pour un dé parfait, P(6)=1/6").

— En **statistiques** (et donc en humanités numériques), on fait l'inverse. On observe des données réelles et on mesure une **fréquence** observée (l'approche fréquentiste).

La **Loi des grands nombres** est le théorème qui relie les deux : elle nous dit que si on observe un très grand nombre de données, la fréquence observée d'un événement va converger vers sa "vraie" probabilité théorique. En pratique, nous utiliserons donc toujours les fréquences calculées sur nos corpus comme estimations de leurs probabilités.

### 4.1.2 Définir la Probabilité par l'Observation : Le Modèle Unigramme

Construisons un modèle probabiliste pour une langue très simple à partir d'un mini-corpus. Ce modèle nous permettra de "prédire" ou de "générer" des phrases plausibles. En particulier, le modèle "unigramme" est le plus simple dans la famille des modèles dits N-grams.

Considérons le corpus suivant, composé de trois phrases :

1. Le chat mange la souris.
2. La souris aime le fromage.
3. Le chat aime le fromage.

L'expérience aléatoire est "tirer un mot au hasard dans le corpus".

— Espace des possibles ($\Omega$) : L'ensemble des mots uniques {Le, chat, mange, la, souris, aime, le, fromage}.

— Calcul des probabilités : il y a 15 mots au total dans le corpus.
  — Probabilité de tirer le mot "Le" : il apparaît 4 fois. $P(\text{"Le"}) = \frac{4}{15} \approx 0.27$.
  — Probabilité de tirer le mot "fromage" : il apparaît 2 fois. $P(\text{"fromage"}) = \frac{2}{15} \approx 0.13$.
  — Probabilité de tirer le mot "mange" : il apparaît 1 fois. $P(\text{"mange"}) = \frac{2}{15} \approx 0.13$.

**Limite de ce modèle :** Ce modèle n'a aucune notion de l'ordre des mots. Pour lui, "Le fromage le souris aime" est une phrase aussi probable que "Le chat aime le fromage", car elle utilise des mots fréquents.

### 4.1.3 Probabilités conditionnelles

**Définition 41**

La **probabilité conditionnelle** de $A$ sachant $B$, notée $P(A|B)$, est la probabilité que l'événement $A$ se produise, sachant que l'événement $B$ s'est déjà produit. Elle est définie par :

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

où $P(A \cap B)$ est la probabilité que $A$
textitet $B$ se produisent en même temps.

**Exemple 24**

Imaginons un jeu de cartes standard de 52 cartes (13 cartes de chaque couleur : coeur, Carreau, Pique, Trèfle). Les Cœurs et Carreaux sont Rouges (26 cartes), les Piques et Trèfles sont Noirs (26 cartes).

- Soit A l'événement "la carte tirée est un coeur". La probabilité de cet événement est $P(A) = \frac{13}{52} = \frac{1}{4} = 25\%$.
- Soit B l'événement "la carte tirée est Rouge". (Il y a 26 cartes rouges : 13 Cœurs et 13 Carreaux). $P(B) = \frac{26}{52} = \frac{1}{2} = 50\%$.

Quelle est la nouvelle probabilité $P(A \mid B)$ que la carte soit un coeur, *sachant* qu'elle est Rouge ? Intuitivement, notre espace des possibles $\Omega$ s'est réduit. Nous ne regardons plus les 52 cartes, mais seulement les 26 cartes rouges (l'événement B). Parmi ces 26 cartes rouges, 13 sont des coeurs, la probabilité est donc $P(A \mid B) = \frac{13}{26} = \frac{1}{2} = 50\%$.
L'information "Rouge" a fait doubler notre estimation de la probabilité de "coeur" (de 25% à 50%).

### 4.1.4 Application à notre modèle de langue : le modèle bigramme

Maintenant que nous avons défini la probabilité conditionnelle, nous pouvons reprendre notre exemple de corpus et construire un modèle beaucoup plus puissant et contextuel. Au lieu de regarder les mots isolément, regardons les paires de mots qui se suivent (les **bigrammes**).

L'expérience aléatoire est maintenant : "Étant donné un mot, quel est le mot suivant le plus probable ?". Nous allons calculer des probabilités conditionnelles à partir de notre corpus.

- **Quelle est la probabilité que le mot "chat" suive le mot "le" ?** C'est la probabilité conditionnelle $P$(le mot suivant est "chat"|le mot actuel est "le").
  1. On cherche toutes les occurrences du mot "le". Il y en a 4.
  2. Parmi celles-ci, combien de fois sont-elles suivies par "chat" ? 2 fois.
  3. La probabilité estimée est donc :

$$P(\text{"chat"}|\text{"le"}) = \frac{\text{Nombre de fois où ("Le chat") apparaît}}{\text{Nombre de fois où ("Le") apparaît}} = \frac{2}{4} = 0.5$$

— **Quelle est la probabilité que le mot "aime" suive le mot "souris" ?** C'est la probabilité conditionnelle $P(\text{"aime"}|\text{"souris"})$.

1. Le mot "souris" apparaît 2 fois dans le corpus.
2. Une fois il est suivi de "aime", l'autre fois il est en fin de phrase. Il est donc suivi par "aime" 1 fois sur 2.
3. La probabilité estimée est donc :

$$P(\text{"aime"}|\text{"souris"}) = \frac{\text{Nombre de fois où ("souris aime") apparaît}}{\text{Nombre de fois où ("souris") apparaît}} = \frac{1}{2} = 0.5$$

Ce modèle bigramme a "appris" les règles de grammaire et de sens de notre mini-corpus. On peut l'utiliser pour générer une nouvelle phrase de manière bien plus réaliste. Nous avons transformé de simples fréquences en un modèle prédictif qui comprend le contexte. Ce mécanisme, étendu à des contextes plus longs (trigrammes, N-grammes) et entraîné sur des milliards de mots, est le principe de base de nombreuses applications en traitement du langage naturel.

### 4.1.5 Théorème de Bayes

**L'intuition :** Le théorème de Bayes est une formule pour "retourner" une probabilité conditionnelle. Souvent, on connaît $P(B|A)$ mais ce qui nous intéresse, c'est $P(A|B)$. Par exemple, on sait la probabilité d'avoir de la fièvre si on a la grippe, mais on veut savoir la probabilité d'avoir la grippe si on a de la fièvre. Le théorème de Bayes nous permet de faire cette inversion.

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

Il se lit de la manière suivante :

— $P(A|B)$ (**probabilité a posteriori**) : Ce que l'on croit de A *après* avoir vu la preuve B.

— $P(A)$ (**probabilité a priori**) : Ce que l'on croyait de A *avant* de voir la preuve.

— $P(B|A)$ (**vraisemblance**) : La probabilité d'observer la preuve B si notre hypothèse A est vraie.

**Exemple 25**

**Paternité d'un texte** Un historien découvre un texte anonyme et hésite entre deux auteurs possibles : Corneille ou Molière. Il veut calculer la probabilité que l'auteur soit Molière. Il analyse l'ensemble des pièces connues des deux auteurs et observe que :

— sur 30 pièces de **Molière**, 12 d'entre elles contiennent l'expression "par ma foi !".

— sur 30 pièces de **Corneille**, seule 1 d'entre elles contient cette expression.

Le texte anonyme contient l'expression. Nous introduisons les évènements suivants :

— $A$ : "Le texte est de Molière".

— $B$ : "Le texte contient au moins une fois l'expression 'par ma foi '".

On cherche à calculer $P(A|B)$, la probabilité que le texte soit de Molière sachant qu'il contient "par ma foi ". Le théorème de Bayes nous dit :

$$P(\text{Molière}|\text{"par ma foi !"}) = \frac{P(\text{"par ma foi !"}|\text{Molière}) \times P(\text{Molière})}{P(\text{"par ma foi !"})}$$

.

1. **Probabilité a priori, $P$(Molière) :** Sans autre information, l'historien estime qu'il y a autant de chances que ce soit l'un ou l'autre. $P(\text{Molière}) = 0.5$.
2. **Vraisemblance, $P(B|A)$ :** C'est la probabilité d'observer "par ma foi !" *si* l'auteur est Molière. Notre vraisemblances est donc $P(B|A) = 12/30 = 0.4$
3. **Le dénominateur :** sur les 60 pièces des deux auteurs, "par ma foi !" apparaît 13 fois, donc $P(B) = 13/60$

Nous avons maintenant tous les ingrédients pour appliquer Bayes :

$$P(A|B) = \frac{12/30 \times 1/2}{13/60} = \frac{12}{13} \approx 0.92$$

**Conclusion :** Après avoir observé la présence de l'expression "par ma foi !" dans le manuscrit anonyme, la conviction de l'historien que le texte est de Molière passe de 50% à environ 92%. La nouvelle preuve (la présence de ce "marqueur" stylistique) a fortement mis à jour sa croyance initiale. C'est le moteur de nombreuses méthodes de classification automatique.

## 4.2 Variables aléatoires et distributions usuelles

Souvent, les résultats d'une expérience ne sont pas des nombres (ex : "Pile", "Face", "chat", "souris"). Pour pouvoir leur appliquer des outils mathématiques comme la moyenne ou la variance, nous devons les transformer en valeurs numériques. C'est le rôle de la variable aléatoire.

**Définition 42**

Une **variable aléatoire** est une fonction, souvent notée par une lettre majuscule comme $X$, qui associe un nombre à chaque résultat possible d'une expérience aléatoire.

**Exemple 26**

**Expérience :** Lancer une pièce.

- Résultats possibles : {Pile, Face}.
- On définit une variable aléatoire $X$ qui vaut 1 si le résultat est "Pile" et 0 si c'est "Face".

Le concept de variable aléatoire permet de passer de l'étude d'événements qualitatifs à l'analyse quantitative de leurs propriétés.

Une variable aléatoire est caractérisée par sa **distribution** (ou **loi de probabilité**). C'est un objet mathématique qui décrit complètement son comportement : il nous dit quelles valeurs la variable peut prendre et avec quelle probabilité.

### 4.2.1 Résumer une distribution : Espérance et Variance

Avant d'explorer les distributions usuelles, nous devons comprendre comment les caractériser avec deux nombres essentiels : une valeur "centrale" et une mesure de "dispersion".

**Définition 43**

L'**espérance** d'une variable aléatoire $X$, notée $E[X]$, est la moyenne de toutes les valeurs possibles de $X$, pondérée par leur probabilité d'apparition. C'est la valeur moyenne que l'on s'attend à obtenir si l'on répète l'expérience un très grand nombre de fois.

**Exemple 27**

**Pour un dé parfait :** Chaque face (1, 2, 3, 4, 5, 6) a une probabilité de $\frac{1}{6}$.

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

L'espérance est 3.5. Notez que ce n'est pas une valeur que le dé peut prendre, mais c'est la moyenne sur le long terme.

**La variance et l'écart-type (la "dispersion")** L'espérance seule ne suffit pas à caractériser une distribution. Deux variables peuvent avoir la même moyenne mais des comportements très différents.

**Définition 44**

La **variance**, notée $V(X)$ ou $\sigma^2(X)$, mesure la dispersion des valeurs autour de l'espérance. Mathématiquement : $V(X) = E[(X - E[X])^2]$.
L'**écart-type**, noté $\sigma(X)$, est la racine carrée de la variance : $\sigma(X) = \sqrt{V(X)}$. Il a l'avantage d'être dans la même unité que les données.

Interprétation de la variance :

— Faible variance = résultats concentrés autour de la moyenne (phénomène prévisible)

— Forte variance = résultats très étalés (phénomène imprévisible)

### 4.2.2 Visualiser les distributions

Pour comprendre intuitivement une distribution, rien ne vaut une représentation graphique :

— **Courbe de densité :** fonction qui décrit la probabilité d'un événement dans un espace donné.

— **Fonction de répartition :** Courbe croissante montrant la probabilité cumulée $P(X \leq x)$.

— **Histogramme :** permet de représenter la répartition empirique d'une variable aléatoire ou d'une série statistique en la représentant avec des colonnes correspondant chacune à une classe et dont l'aire est proportionnelle à l'effectif de la classe.

Ces visualisations permettent de voir immédiatement où se concentrent les valeurs et quelle est la "forme" de la distribution.

### 4.2.3 Lois de probabilité usuelles

**Lois discrètes** Elles décrivent des variables qui prennent un nombre fini ou dénombrable de valeurs (souvent des entiers).

**Définition 45**

**Loi de Bernoulli $\mathcal{B}(p)$**
Modélise une unique expérience à deux issues : succès (1) ou échec (0).
**Paramètre :** $p$ = probabilité de succès
**Caractéristiques :**

— Espérance : $E[X] = p$

— Variance : $V(X) = p(1-p)$

— Écart-type : $\sigma(X) = \sqrt{p(1-p)}$

**Forme visuelle :** Deux barres, une en 0 (hauteur $1-p$) et une en 1 (hauteur $p$).

**Exemple 28**

Un document est-il écrit par une femme (1) ou un homme (0) ? Si $p = 0.3$, alors $E[X] = 0.3$ (en moyenne, 30% des documents sont écrits par des femmes).

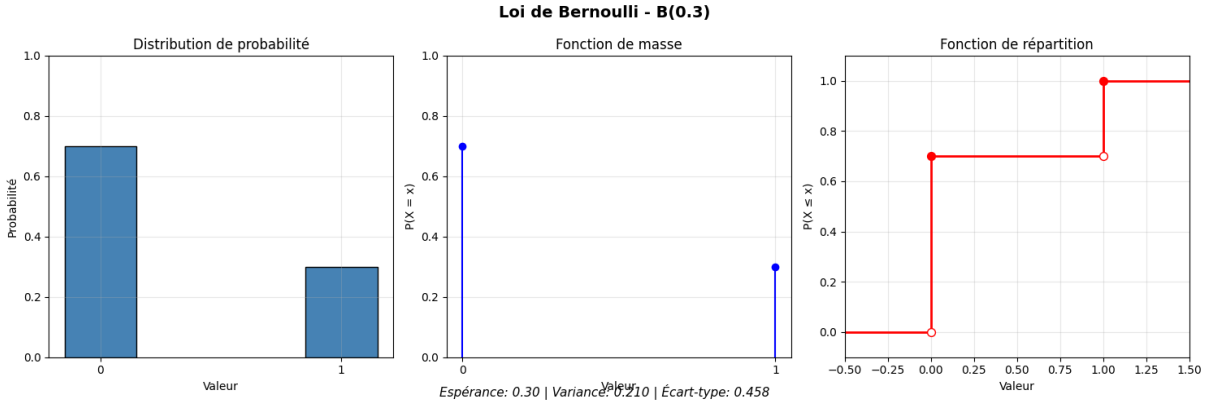


FIGURE 12 – Illustration d'une distribution de Bernoulli

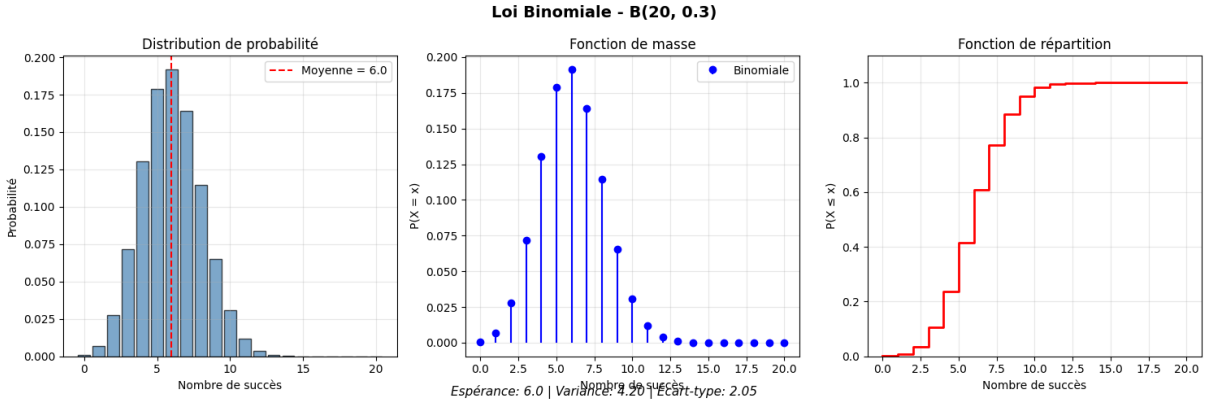


FIGURE 13 – Illustration d'une distribution binomiale

## Définition 46

**Loi Binomiale $\mathcal{B}(n, p)$**
Compte le nombre de succès lors de $n$ répétitions indépendantes d'une expérience de Bernoulli.

**Paramètres :**

- $n$ = nombre d'essais
- $p$ = probabilité de succès à chaque essai

**Caractéristiques :**

- Espérance : $E[X] = n \times p$
- Variance : $V(X) = n \times p \times (1 - p)$
- Écart-type : $\sigma(X) = \sqrt{n \times p \times (1 - p)}$

**Forme visuelle :** Histogramme en forme de cloche (symétrique si $p = 0.5$, asymétrique sinon). Plus $n$ est grand, plus la forme se rapproche d'une courbe en cloche.

**Exemple 29**

Dans un corpus de 50 textes ($n = 50$), si chaque texte a une probabilité $p = 0.3$ d'être écrit par une femme, alors :

- On s'attend à trouver $E[X] = 50 \times 0.3 = 15$ textes écrits par des femmes
- Avec un écart-type $\sigma(X) = \sqrt{50 \times 0.3 \times 0.7} \approx 3.24$
- La plupart des échantillons contiendront entre 12 et 18 textes écrits par des femmes ($\pm$ 1 écart-type)

**Lois continues** Elles décrivent des variables pouvant prendre n'importe quelle valeur dans un intervalle.

*Remarque* 4.1. Pour les lois continues, $P(X = \text{valeur exacte}) = 0$. On calcule toujours des probabilités d'intervalles : $P(a < X < b)$, qui correspond à l'aire sous la courbe entre $a$ et $b$.

**Définition 47**

**Loi Uniforme $\mathcal{U}(a, b)$**
Toutes les valeurs entre $a$ et $b$ sont équiprobables.
**Paramètres :** $a$ (minimum) et $b$ (maximum)
**Caractéristiques :**

- Espérance : $E[X] = \frac{a+b}{2}$
- Variance : $V(X) = \frac{(b-a)^2}{12}$
- Écart-type : $\sigma(X) = \frac{b-a}{\sqrt{12}}$

**Forme visuelle :** Rectangle horizontal entre $a$ et $b$ (densité constante).

**Exemple 30**

Pour simuler des dates de naissance aléatoires entre 1750 et 1850 : $X \sim \mathcal{U}(1750, 1850)$. L'année moyenne est $E[X] = 1800$ avec un écart-type $\sigma = \frac{100}{\sqrt{12}} \approx 28.87$ ans.

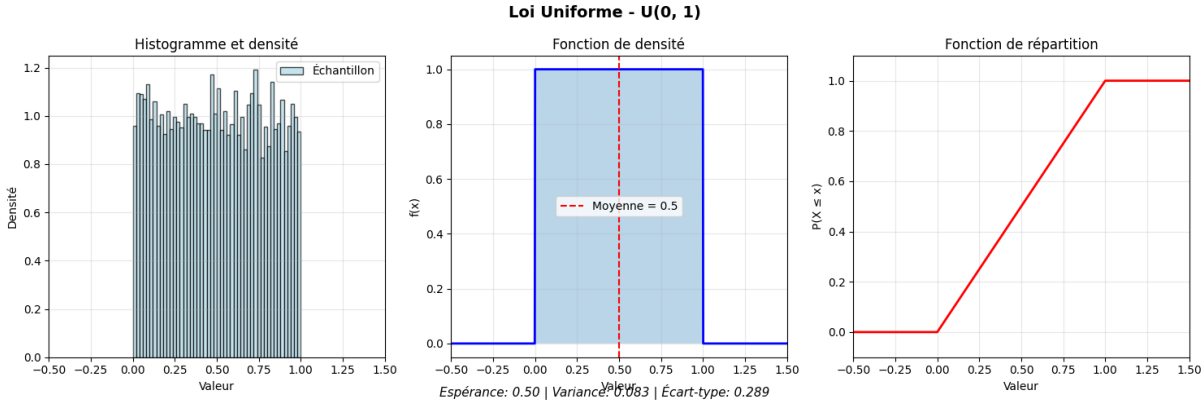


FIGURE 14 – Illustration d'une distribution uniforme

## Définition 48

**Loi Normale $\mathcal{N}(\mu, \sigma^2)$**
La plus importante des lois, reconnaissable à sa "courbe en cloche" symétrique.
**Paramètres :**

- $\mu$ = moyenne (centre de la cloche)
- $\sigma$ = écart-type (largeur de la cloche)

**Caractéristiques :**

- Espérance : $E[X] = \mu$
- Variance : $V(X) = \sigma^2$
- Écart-type : $\sigma(X) = \sigma$

**Règle empirique (68-95-99.7) :**

- 68% des valeurs sont à $\pm$ 1 écart-type de la moyenne
- 95% des valeurs sont à $\pm$ 2 écarts-types
- 99.7% des valeurs sont à $\pm$ 3 écarts-types

**Forme visuelle :** Courbe en cloche symétrique centrée en $\mu$. Plus $\sigma$ est grand, plus la cloche est étalée.

## Exemple 31

Si la longueur des phrases d'un auteur suit $\mathcal{N}(25, 16)$ (moyenne 25 mots, variance 16 donc $\sigma = 4$) :

- 68% des phrases font entre 21 et 29 mots ($25 \pm 4$)
- 95% des phrases font entre 17 et 33 mots ($25 \pm 8$)
- Les phrases de moins de 13 mots ou plus de 37 mots sont exceptionnelles (0.3%)

*Remarque* 4.2. **Théorème Central Limite :** La somme d'un grand nombre de variables aléatoires indépendantes tend vers une loi Normale, quelle que soit leur distribution d'origine. C'est pourquoi la loi Normale apparaît partout dans la nature et les sciences sociales.

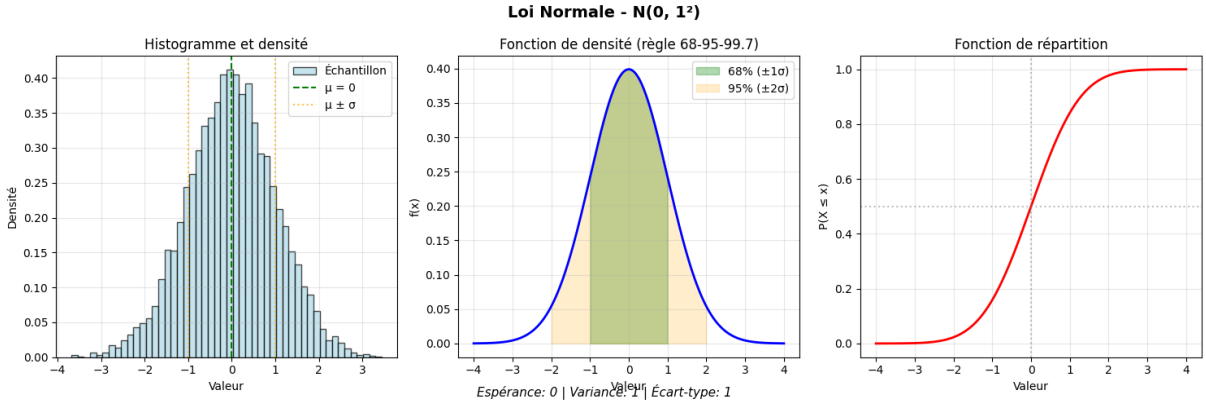


FIGURE 15 – Illustration d'une distribution normale

**Définition 49**

**Loi de Puissance** (Pareto/Zipf)
Décrit les phénomènes "winner-take-all" : quelques valeurs dominent, suivies d'une longue traîne.
**Caractéristique principale :** $P(X > x) \propto x^{-\alpha}$ où $\alpha$ est le paramètre de forme.
**Règle empirique :** Souvent résumée par la "règle 80/20" (20% des éléments concentrent 80% des occurrences).
**Forme visuelle :** Courbe très asymétrique avec un pic abrupt suivi d'une décroissance lente ("longue traîne").

**Exemple 32**

**Loi de Zipf en linguistique :** Dans tout corpus de texte :

- Le mot le plus fréquent ("le", "the") apparaît environ deux fois plus que le 2e
- Le 2e mot apparaît environ deux fois plus que le 4e
- L'immense majorité du vocabulaire n'apparaît qu'une ou deux fois
- Fréquence du $n$-ième mot $\propto \frac{1}{n}$

Cette loi se retrouve aussi dans : nombre de followers sur les réseaux sociaux, richesse (Pareto), trafic web, citations académiques, taille des villes...

*Remarque* 4.3. **Comment choisir la bonne loi ?**

- **Bernoulli/Binomiale :** Je compte des succès/échecs
- **Uniforme :** Je simule du "pur hasard" dans un intervalle
- **Normale :** Je mesure un phénomène naturel/biologique, ou j'ai une somme de nombreux effets
- **Puissance :** Je compte des fréquences (mots, liens, popularité) dans un système social/culturel

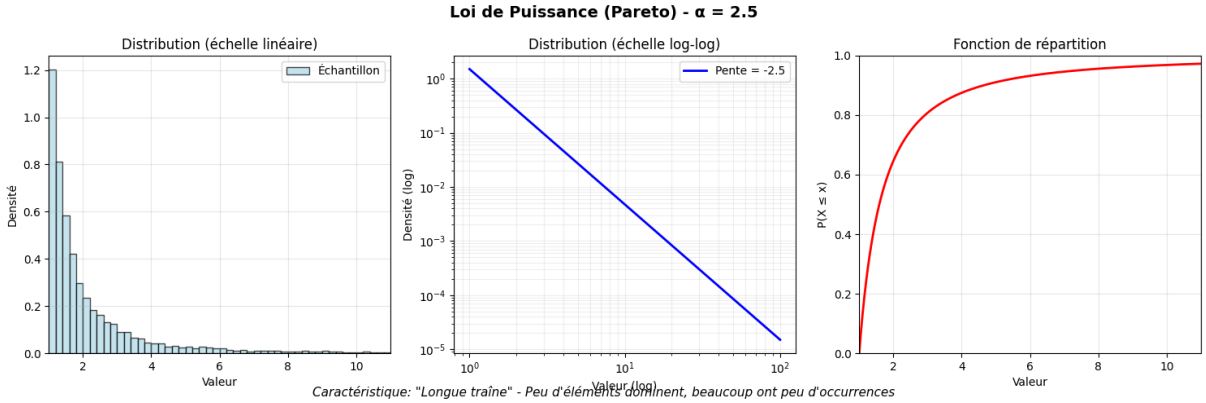


FIGURE 16 – Illustration d'une distribution Pareto

# A Exercices

## A.1 Chapitre 1

*Exercice* A.1.1. Montrer la proposition suivante :

$$\text{Si } E \subset F, \text{ alors } E \cap F = E. \text{ Autrement dit, } E \subset F \Rightarrow E \cap F = E.$$

La réciproque est-elle vraie ? Conclure.

*Exercice* A.1.2. Montrer les propositions suivantes :

1. $E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$
2. $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$
3. $E \bar{\cup} F = \bar{E} \cap \bar{F}$
4. $E \bar{\cap} F = \bar{E} \cup \bar{F}$

*Exercice* A.1.3. **Négation d'une proposition.**

1. Si A et B, sont deux propositions, montrer que $\neg(A \Rightarrow B) \Leftrightarrow A$ et $\neg B$
2. Expliquer par la logique pourquoi $\neg(\forall x, A) \Leftrightarrow \exists x, \neg A$
3. Soit A la proposition : $\forall (x, y) \in \mathbb{R}^2, (x + y = 0 \Rightarrow x = 0 \text{ et } y = 0)$. Exprimer par une phrase en français la proposition énoncée.
4. Expliciter la négation de A.
5. Laquelle de A et $\neg A$ est vraie ?
6. (Bonus) Refaire les questions avec A : $\forall (x, y) \in \mathbb{R}^2, (x^2 + y^2 = 0 \Rightarrow x = 0 \text{ et } y = 0)$

*Exercice* A.1.4. **Ordre des quantificateurs.** Dans les deux cas suivants, exprimer par une phrase en français la proposition énoncée. Expliciter sa négation, et dire laquelle des deux propositions est vraie :

1. $\exists M \in \mathbb{R}, \forall x \in \mathbb{R}, x \leq M$
2. $\forall x \in \mathbb{R}, \exists M \in \mathbb{R}, x \leq M$

## A.2 Chapitre 2

*Exercice* A.2.1. **Parité/Imparité.** Pour chaque fonction donné dans le cours, dire si celle-ci est paire, impaire, ou ni l'un ni l'autre.

*Exercice* A.2.2. **Premières études de fonctions.** Pour chaque fonction, donner l'ensemble de définition, la limite en $+\infty$, la valeur en zéro, et tracer (schématiquement) leur courbe.

1. $f(x) = \dfrac{1}{x^2 - 1}$
2. $g(x) = \sqrt{x - 4}$
3. $h(x) = |-x + 2|$
4. $i(x) = x\ln(x)$

*Exercice* A.2.3. **Calcul de dérivées.** Calculer la dérivée et l'intégrale de 1 à 2 des fonctions suivantes :

1. $f(x) = 5x^3 - 2x + 1$
2. $g(x) = \frac{1}{x} + \sqrt{x - 4}$
3. $h(x) = \frac{e^x}{x+1}$

*Exercice* A.2.4. **Étude de maximum.** On modélise l'intensité lumineuse d'une source à l'aide de la fonction $I(t) = -t^2 + 10t + 5$, pour $t \in [0, 10]$.

1. Déterminer l'instant t où l'intensité lumineuse est maximale.
2. Quelle est cette intensité maximale ?

*Exercice* A.2.5. **Loi de Zipf.** La loi de Zipf est une observation empirique concernant la fréquence des mots dans un texte. Zipf avait entrepris d'analyser une œuvre de James Joyce, Ulysse, d'en compter les mots distincts et de les présenter par ordre décroissant du nombre d'occurrences. La légende dit que :

— le mot le plus courant revenait 8 000 fois ;

— le dixième mot 800 fois ;

— le centième, 80 fois ;

— et le millième, 8 fois. [1]

La loi de Zipf stipule ainsi que dans un grand corpus de textes, la fréquence d'un mot est inversement proportionnelle à son rang. Plus précisément, on peut modéliser la fréquence $f_r$ du mot de rang $r$ par la fonction $f(r) = \frac{C}{r^\alpha}$, où $C$ et $\alpha$ sont des constantes. Pour de nombreux corpus, $\alpha$ est proche de 1. Un chercheur souhaite vérifier cette loi sur un corpus et propose d'utiliser une régression linéaire.

1. En utilisant les propriétés du logarithme, montrer que si on a une relation de la forme $f(r) = \frac{C}{r^\alpha}$, alors une relation linéaire existe entre $\ln(f_r)$ et $\ln(r)$.
2. Pour un corpus, le chercheur a calculé les paires de valeurs $(\ln(f_r), \ln(r))$ pour les 100 mots les plus fréquents. Voici un échantillon des valeurs obtenus :

1. voir la page wikipedia pour plus de détails : `https://fr.wikipedia.org/wiki/Loi_de_Zipf`

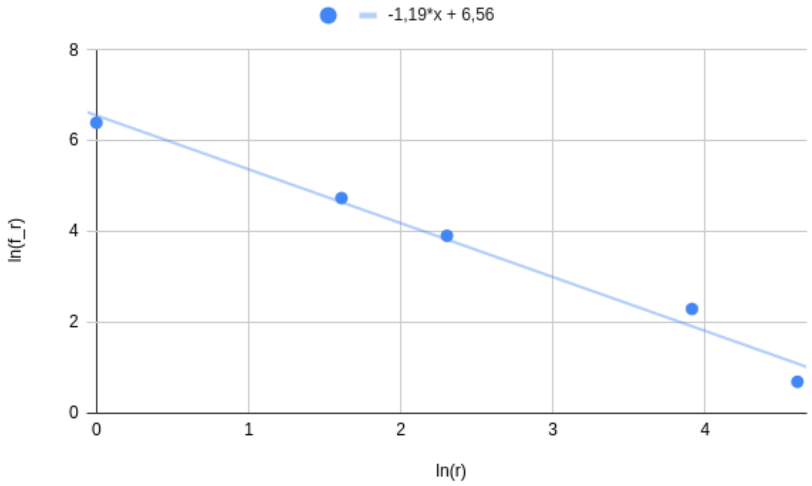


FIGURE 17 – Régression linéaire sur des données de fréquences de mots, en échelle logarithmique.

| Rang ($r$) | $\ln(r)$ | Fréquence ($f_r$) | $\ln(f_r)$ |
|---|---|---|---|
| 1 | 0 | 600 | 6,396 |
| 5 | 1,609 | 115 | 4,744 |
| 10 | 2,302 | 50 | 3,912 |
| 50 | 3,912 | 10 | 2,302 |
| 100 | 4,605 | 2 | 0,693 |

Une régression linéaire est effectué sur les paires de valeurs $(\ln(f_r), \ln(r))$, et le résultat est donné par l'équation $Y = -1.19X + 6.56$ (voir Figure 2).

Déduire de cette équation les valeurs de $\alpha$ et $C$.

3. Utiliser ce modèle pour estimer la fréquence du mot de rang $r = 200$.

## A.3 Chapitre 3

*Exercice* A.3.1. **Opérations sur les vecteurs.** Soient $u$ et $v$ les deux vecteurs de $\mathbb{R}^3$ suivants : $u = (1, 2, 3)$ et $v = (4, -1, 0)$.

1. Calculez le vecteur $w = 2u + v$.
2. Calculez le produit scalaire $< u, v >$.
3. Les vecteurs $u$ et $v$ sont-ils orthogonaux ?

*Exercice* A.3.2. **Produit matriciel et interprétation** Soit la matrice $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ et la matrice $B = \begin{pmatrix} 5 & 6 \\ 0 & -1 \end{pmatrix}$.

1. Calculez le produit matriciel $C = AB$.
2. Calculez le produit matriciel $D = BA$.

3. Que pouvez-vous conclure en comparant $C$ et $D$ ?

*Exercice* A.3.3. **Application en NLP : similarité de documents** En traitement du langage naturel (NLP), on souhaite comparer trois courts documents pour déterminer lesquels sont les plus proches thématiquement. Après avoir défini un vocabulaire de 3 mots { "amour", "guerre", "complice" }, on obtient les vecteurs "sac de mots" pour chaque document :

— $d_1 = (3, 1, 0)$

— $d_2 = (0, 4, 5)$

— $d_3 = (4, 1, 1)$

1. Calculez la distance euclidienne $||d_1 - d_2||_2$, $||d_1 - d_3||_2$, et $||d_2 - d_3||_2$.
2. En vous basant sur ces distances, quels sont les deux documents les plus similaires ? Quels sont les deux plus différents ?
3. Le produit scalaire est une autre mesure de similarité. Calculez $< d_1, d_3 >$. Qu'est-ce que ce score positif suggère sur le contenu des documents D1 et D3 ?

*Exercice* A.3.4. **Multiplication d'une matrice non carrée.** On modélise une transformation d'un espace de dimension 3 vers un espace de dimension 2 à l'aide de la matrice $A$ suivante :

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 3 & 1 \end{pmatrix}$$

Soit le vecteur $v = \begin{pmatrix} 4 \\ 2 \\ -1 \end{pmatrix}$ un point de l'espace de départ.

Calculez le vecteur $w = Av$, qui est la position du point après transformation.

*Exercice* A.3.5. **Application en histoire de l'art : analyse d'un réseau de mécénat.** Un historien de l'art s'intéresse aux relations de commande et d'influence au sein d'un petit groupe d'acteurs de la Renaissance italienne. Il modélise les relations directes par un réseau orienté.

Les acteurs sont :

— Léonard de Vinci (L)

— La famille Médicis (M)

— Michel-Ange (A)

— Le Pape Jules II (P)

Les relations directes (chemins de longueur 1) identifiées sont :

— Les Médicis commandent des œuvres à Léonard.

— Les Médicis commandent des œuvres à Michel-Ange.

— Le Pape commande des œuvres à Michel-Ange.

— Léonard agit comme conseiller auprès du Pape.

1. Construisez la matrice d'adjacence $M$ de ce réseau. Utilisez l'ordre (L, M, A, P) pour les lignes et les colonnes.
2. Calculez la matrice $M^2 = M \times M$.
3. En utilisant $M^2$, répondez aux questions suivantes :
   (a) De combien de manières les Médicis peuvent-ils influencer le Pape via **exactement un** intermédiaire ? Quel est ce chemin d'influence ?
   (b) Existe-t-il un "retour d'influence" sur un acteur en deux étapes ? (C'est-à-dire, une influence partant d'un acteur et lui revenant après être passée par un autre acteur). Justifiez votre réponse en observant la matrice $M^2$.

## A.4 Chapitre 4

*Exercice* A.4.1 (Analyse de corpus et événements). Un chercheur analyse un corpus de 100 documents historiques. Il étudie leur provenance et leur langue. Il obtient les résultats suivants :

— 60 documents proviennent d'Italie (événement $I$).

— 50 documents sont écrits en Latin (événement $L$).

— 40 documents proviennent d'Italie **ET** sont écrits en Latin ($I \cap L$).

Un document est tiré au hasard parmi les 100.

1. Calculez la probabilité $P(I \cup L)$ qu'il vienne d'Italie **OU** qu'il soit en Latin.
2. Calculez la probabilité conditionnelle $P(L|I)$ : sachant que le document vient d'Italie, quelle est la probabilité qu'il soit en Latin ?

*Exercice* A.4.2 (Théorème de Bayes : Le filtre anti-spam). Vous créez un filtre anti-spam rudimentaire. Vous vous concentrez sur la présence d'un seul mot : "gratuit". En analysant votre boîte mail, vous faites les estimations suivantes :

1. 30% de tous les emails que vous recevez sont des spams.
2. Parmi tous vos spams, 80% contiennent le mot "gratuit".
3. Parmi tous vos emails légitimes (non-spams), seuls 5% contiennent le mot "gratuit".

Calculez la probabilité qu'un email contenant le mot "gratuit" soit réellement un spam. Indice : On utilise la formule des probabilités totales :

$$P(B) = [P(B|A) \times P(A)] + [P(B|\bar{A}) \times P(\bar{A})]$$

*Exercice* A.4.3 (Identifier les lois simples). Un chercheur en "cultural analytics" analyse 20 films d'un festival. Il sait que 10% (soit $p = 0.1$) des films de ce festival obtiennent généralement un prix.

1. Il prend un film au hasard. Il définit la variable aléatoire $X$ qui vaut 1 si le film a un prix et 0 sinon. Quelle loi suit la variable $X$ ? Donner la moyenne et la variance de $X$.
2. Il définit la variable aléatoire $Y$ qui compte le nombre total de films primés parmi les 20. Quelle loi suit la variable $Y$ ? Donner la moyenne et la variance de $Y$.
3. Sans faire de calcul, est-il plus probable que $Y = 1$ (un seul film primé) ou que $Y = 20$ (tous les films primés) ? Pourquoi ?

*Exercice* A.4.4 (La loi Normale en stylométrie). Un spécialiste de la stylométrie (l'étude statistique du style) étudie l'œuvre d'un auteur. Il établit que la longueur de ses phrases suit une **loi Normale** avec les paramètres suivants :

— Moyenne $\mu = 25$ mots.

— Écart-type $\sigma = 4$ mots.

1. Que signifie concrètement le fait que la loi soit "Normale" avec une moyenne de 25 ?
2. Un historien découvre un manuscrit anonyme et se demande s'il est de cet auteur. Il analyse le texte et trouve que la longueur moyenne des phrases est de 45 mots. Que peut en conclure le spécialiste en stylométrie ?

*Exercice* A.4.5 (Loi Normale vs. Loi de Puissance). Un chercheur en humanités numériques analyse deux ensembles de données :

— **Donnée A :** La fréquence d'apparition de **tous les mots** (le vocabulaire) dans l'œuvre complète de Victor Hugo.

— **Donnée B :** La **longueur moyenne des mots** (en nombre de lettres) utilisés dans chaque chapitre de l'œuvre de Victor Hugo.

L'une de ces distributions suit une loi Normale, l'autre une loi de Puissance.

1. Laquelle suit la loi de Puissance ? Pourquoi ?
2. Laquelle suit la loi Normale ? Pourquoi ?

# B Solutions

**Solution de A.1.3 :**

1. $\forall (x, y) \in \mathbb{R}^2, (x + y = 0 \Rightarrow x = 0 \text{ et } y = 0)$
   **En français :** "Pour tous réels $x$ et $y$, si $x + y = 0$, alors $x = 0$ et $y = 0$"
   **Négation :** $\exists (x, y) \in \mathbb{R}^2, (x + y = 0 \text{ et } (x \neq 0 \text{ ou } y \neq 0))$
   En français : "Il existe des réels $x$ et $y$ tels que $x + y = 0$ mais $x \neq 0$ ou $y \neq 0$"
   **Quelle proposition est vraie ?** La *négation* est vraie.
   *Contre-exemple :* $x = 1$ et $y = -1$. On a bien $1 + (-1) = 0$ mais $x \neq 0$.
2. $\forall (x, y) \in \mathbb{R}^2, (x^2 + y^2 = 0 \Rightarrow x = 0 \text{ et } y = 0)$
   **En français :** "Pour tous réels $x$ et $y$, si $x^2 + y^2 = 0$, alors $x = 0$ et $y = 0$"
   **Négation :** $\exists (x, y) \in \mathbb{R}^2, (x^2 + y^2 = 0 \text{ et } (x \neq 0 \text{ ou } y \neq 0))$
   En français : "Il existe des réels $x$ et $y$ tels que $x^2 + y^2 = 0$ mais $x \neq 0$ ou $y \neq 0$"

**Solution de A.1.4 :**

1. $\exists M \in \mathbb{R}, \forall x \in \mathbb{R}, x \leq M$
   **En français :** "Il existe un réel $M$ tel que pour tout réel $x$, on ait $x \leq M$"
   Autrement dit : "Il existe un majorant de tous les réels"
   **Négation :** $\forall M \in \mathbb{R}, \exists x \in \mathbb{R}, x > M$
   En français : "Pour tout réel $M$, il existe un réel $x$ tel que $x > M$"
   **Quelle proposition est vraie ?** La *négation* est vraie.
   *Justification :* $\mathbb{R}$ n'est pas majoré. Pour tout $M$ réel, on peut prendre $x = M + 1 > M$.
2. $\forall x \in \mathbb{R}, \exists M \in \mathbb{R}, x \leq M$
   **En français :** "Pour tout réel $x$, il existe un réel $M$ tel que $x \leq M$"
   Autrement dit : "Tout réel admet un majorant"
   **Négation :** $\exists x \in \mathbb{R}, \forall M \in \mathbb{R}, x > M$
   En français : "Il existe un réel $x$ tel que pour tout réel $M$, on ait $x > M$"
   **Quelle proposition est vraie ?** La *proposition originale* est vraie.
   *Justification :* Pour tout $x$ réel, on peut choisir $M = x+1$ (par exemple), et on aura bien $x \leq M$.

**Solution de A.3.1 :**

1. $w = (2, 4, 6) + (4, -1, 0) = (2 + 4, 4 - 1, 6 + 0) = (6, 3, 6)$.
2. On calcule la somme des produits des composantes : $< u, v >= (1 \times 4) + (2 \times -1) + (3 \times 0) = 4 - 2 + 0 = 2$.
3. Deux vecteurs sont orthogonaux si leur produit scalaire est nul, ce qui n'est pas le cas.

**Solution de A.3.2 :**

1. On applique la règle du produit "ligne par colonne" :
   $C = \begin{pmatrix} (1 \cdot 5 + 2 \cdot 0) & (1 \cdot 6 + 2 \cdot -1) \\ (3 \cdot 5 + 4 \cdot 0) & (3 \cdot 6 + 4 \cdot -1) \end{pmatrix} = \begin{pmatrix} 5 & 4 \\ 15 & 14 \end{pmatrix}$.

2. On calcule l'autre produit :
   $D = \begin{pmatrix} (5 \cdot 1 + 6 \cdot 3) & (5 \cdot 2 + 6 \cdot 4) \\ (0 \cdot 1 + -1 \cdot 3) & (0 \cdot 2 + -1 \cdot 4) \end{pmatrix} = \begin{pmatrix} 23 & 34 \\ -3 & -4 \end{pmatrix}$.
3. On observe que $C \neq D$. Cela illustre une propriété fondamentale de la multiplication matricielle : elle n'est pas commutative. L'ordre des opérations est crucial.

**Solution de A.3.3 :**

1. Calcul des distances :
   — $||d_1 - d_2||_2 = \sqrt{3^2 + (-3)^2 + (-5)^2} = \sqrt{43} \approx 6.56$
   — $||d_1 - d_3||_2 = \sqrt{(-1)^2 + 0^2 + (-1)^2} = \sqrt{2} \approx 1.41$
   — $||d_2 - d_3||_2 = \sqrt{(-4)^2 + 3^2 + 4^2} = \sqrt{41} \approx 6.40$
2. Interprétation :
   — La distance la plus faible est $||d_1 - d_3||_2 = \sqrt{2}$. Les documents D1 et D3 sont donc les plus similaires. Cela s'explique par le fait que leurs profils de mots sont très proches.
   — La distance la plus grande est $||d_1 - d_3||_2 = \sqrt{43}$. Les documents D1 et D2 sont les plus différents.
3. Produit scalaire :
   $< d_1, d_3 >= (3 \times 4) + (1 \times 1) + (0 \times 1) = 12 + 1 + 0 = 13$.
   Le produit scalaire est un nombre largement positif. Cela suggère que les mots qui sont fréquents dans un document le sont aussi dans l'autre. Un score élevé indique une forte co-occurrence et donc une similarité thématique, ce qui confirme notre conclusion tirée de la distance euclidienne.

**Solution de A.3.4 :** Le produit est possible car la matrice $A$ est de taille $2 \times 3$ et le vecteur $v$ est de taille $3 \times 1$. Le nombre de colonnes de $A$ (3) est bien égal au nombre de lignes de $v$ (3). Le vecteur résultant $w$ sera de taille $2 \times 1$.

On applique la règle du produit "ligne par colonne" :

$$w = \begin{pmatrix} 4 + 0 + 1 \\ 8 + 6 - 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 13 \end{pmatrix}$$

Le vecteur $v = (4, 2, -1)$ de l'espace de dimension 3 est transformé en vecteur $w = (5, 13)$ dans l'espace de dimension 2.

**Solution de A.3.5 :**

1. Construction de la matrice d'adjacence $M$ :
   On place un 1 dans la case $(i, j)$ s'il y a un lien de $i$ vers $j$.
   — Ligne L (Léonard) : il conseille le Pape. Lien $L \to P$. Ligne : $(0, 0, 0, 1)$.
   — Ligne M (Médicis) : ils commandent à Léonard et Michel-Ange. Liens $M \to L$ et $M \to A$. Ligne : $(1, 0, 1, 0)$.
   — Ligne A (Michel-Ange) : il ne commande personne dans ce modèle. Ligne : $(0, 0, 0, 0)$.
   — Ligne P (Pape) : il commande à Michel-Ange. Lien $P \to A$. Ligne : $(0, 0, 1, 0)$.
   La matrice $M$ est donc :

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

2. **Calcul de $M^2$ :**
   La matrice $M^2$ représente les chemins de longueur 2 dans le réseau.

$$M^2 = M \times M = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

3. **Interprétation de $M^2$ :**
   (a) Pour trouver l'influence des Médicis (ligne 2) sur le Pape (colonne 4) en deux étapes, on regarde l'élément $(2, 4)$ de $M^2$. Cet élément vaut 1. **Il existe donc une seule manière** pour les Médicis d'influencer le Pape via un intermédiaire. En examinant les liens, on trouve facilement le chemin : **Médicis** $\rightarrow$ **Léonard** $\rightarrow$ **Pape**.
   (b) Un retour d'influence en deux étapes sur un acteur $i$ correspondrait à un chemin $i \rightarrow k \rightarrow i$. Dans la matrice $M^2$, de tels chemins seraient comptabilisés sur la diagonale (éléments $(1, 1), (2, 2), (3, 3), (4, 4)$). En observant la matrice $M^2$, on voit que **tous les éléments de sa diagonale sont nuls**. On peut donc conclure qu'il n'existe **aucun retour d'influence** en deux étapes dans ce réseau.

**Solution de A.4.1 :** L'espace des possibles $\Omega$ contient 100 documents.

1. $P(I) = \frac{\text{Nombre de documents italiens}}{\text{Total}} = \frac{60}{100} = 0.6$
2. $P(L) = \frac{\text{Nombre de documents en latin}}{\text{Total}} = \frac{50}{100} = 0.5$
3. $P(I \cap L) = \frac{\text{Nombre de documents italiens ET en latin}}{\text{Total}} = \frac{40}{100} = 0.4$
4. $P(I \cup L) = P(I) + P(L) - P(I \cap L) = 0.6 + 0.5 - 0.4 = 0.7$
   *(Intuitivement :* $60 + 50 = 110$*. On a compté les 40 documents communs deux fois, donc* $110 - 40 = 70$ *documents uniques.* $70/100 = 0.7$*).*
5. $P(L|I) = \frac{P(L \cap I)}{P(I)} = \frac{0.4}{0.6} = \frac{4}{6} \approx 0.67$
   *(Intuitivement : On sait que le document vient d'Italie. Notre univers s'est réduit aux 60 documents italiens. Parmi eux, 40 sont en latin. La probabilité est donc* $\frac{40}{60} \approx 0.67$*).*

**Solution de A.4.2 :** Nous appliquons le Théorème de Bayes :

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

1. **Numérateur (partie facile) :** Nous avons $P(B|A) = 0.8$ et $P(A) = 0.3$. Le numérateur est $0.8 \times 0.3 = 0.24$.
2. **Dénominateur (Probabilité de la preuve $P(B)$) :**
   Quelle est la probabilité totale qu'un email (spam ou non) contienne le mot "gratuit" ? On utilise la formule des probabilités totales :

$$P(B) = [P(B|A) \times P(A)] + [P(B|\bar{A}) \times P(\bar{A})]$$

Nous avons $P(B|A) = 0.8$ et $P(A) = 0.3$. Nous avons $P(B|\bar{A}) = 0.05$. Si 30% des emails sont des spams ($P(A) = 0.3$), alors 70% sont légitimes ($P(\bar{A}) = 1-0.3 = 0.7$).

$$P(B) = (0.8 \times 0.3) + (0.05 \times 0.7)$$

$$P(B) = 0.24 + 0.035 = 0.275$$

3. **Calcul final :**

$$P(A|B) = \frac{0.24}{0.275} \approx 0.873$$

**Conclusion :** Avant de voir le mot, la probabilité qu'un email soit un spam n'était que de 30%. Après avoir vu le mot "gratuit", la probabilité qu'il s'agisse d'un spam grimpe à 87.3%. Le filtre peut donc être raisonnablement confiant en le marquant comme spam.

**Solution**

1. La variable $X$ décrit une unique expérience à deux issues (succès/échec). C'est une **loi de Bernoulli** de paramètre $p = 0.1$.
2. La variable $Y$ compte le nombre de succès (films primés) lors de la répétition de $n = 20$ expériences de Bernoulli indépendantes. C'est une **loi Binomiale** de paramètres $n = 20$ et $p = 0.1$.
3. Il est **beaucoup plus probable que** $Y = 1$. La probabilité de succès (avoir un prix) est faible ($p = 0.1$). La probabilité d'avoir 20 succès d'affilée ($Y = 20$) est extraordinairement faible ($0.1^{20}$), alors qu'il est tout à fait plausible qu'un seul des 20 films obtienne un prix.

**Solution**

1. Cela signifie que la plupart des phrases de l'auteur ont une longueur "autour" de 25 mots. La forme en "cloche" de la loi Normale implique que les phrases très courtes (ex : 5 mots) et les phrases très longues (ex : 45 mots) sont très rares, mais possibles. La majorité (environ 68%) des phrases de l'auteur devrait se situer entre $\mu - \sigma$ et $\mu + \sigma$, c'est-à-dire entre 21 et 29 mots.
2. Le spécialiste conclura qu'il est **extrêmement improbable** que le manuscrit soit de cet auteur. Une moyenne de 45 mots est à $(45 - 25)/\sigma = 20/4 = 5$ écarts-types de la moyenne de l'auteur. Dans une loi Normale, 99.7% des observations se trouvent à moins de 3 écarts-types. Être à 5 écarts-types est si rare que l'hypothèse (le manuscrit est de cet auteur) peut être rejetée avec une quasi-certitude.

**Solution**

1. **La Donnée A (fréquence des mots) suit une loi de Puissance (loi de Zipf). Pourquoi ?** C'est un phénomène de type "le gagnant prend tout". Quelques mots (comme "de", "le", "la", "et") vont être incroyablement fréquents et dominer le classement. En revanche, l'immense majorité du vocabulaire de Hugo (la "longue traîne") sera composée de mots qui n'apparaissent qu'une seule fois dans toute son œuvre. On n'a pas une fréquence "moyenne" des mots.
2. **La Donnée B (longueur moyenne des mots) suit une loi Normale. Pourquoi ?** C'est un phénomène "moyen". Dans un chapitre donné, Victor Hugo utilise des mots courts et des mots longs, mais la *longueur moyenne* des mots d'un chapitre va peu varier. La plupart des chapitres auront une longueur moyenne de mot

"moyenne" (ex : 4.5 lettres). Les chapitres où la moyenne serait de 2 lettres (uniquement des "le", "la", "de") ou de 10 lettres (uniquement des mots très longs) sont quasi-impossibles. Les valeurs se regroupent donc autour d'une moyenne, formant une courbe en cloche.