Chapitre 2

Statistiques et probabilités pour Data Mining

(Rappel)

Plan

1. **Introduction**
2. **Statistique descriptive**
3. **Statistique multivarié**
4. **Variables aléatoires**
5. **Distributions de variables aléatoires**

1. Introduction

- Le *Data Mining* (ou exploration de données) est un champ situé au *carrefour* de plusieurs disciplines, dont la *Statistique* et la *Probabilité* sont les fondements théoriques et méthodologiques principaux.
- En substance, la Statistique et la Probabilité fournissent *l'ossature mathématique* qui permet aux algorithmes de Data Mining de fonctionner, d'extraire des connaissances significatives, et d'évaluer la fiabilité de ces connaissances.

1. Introduction

objet des statistiques :

- c'est l'étude d'un ensemble d'individus sur lesquels on observe des caractéristiques appelées variables.
- Selon le nombre de variables, on distingue :

✓Techniques simples résumant les caractéristiques d'une variable (moyenne, médiane, etc.), permettant de détecter les valeurs atypiques. ⟹ Statistiques univariée

✓Techniques s'appliquant à deux variables ou plus (corrélation, nuage de points). ⟹ Statistique multivariée

2. Statistique descriptive

Population et individus

- **Individu** ou unité statistique
  - ✓ Une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques données.

Enregistrement, tuple , objet,…

2. Statistique descriptive

Population et individus

- Population
  - ✓ Ensemble des individus (ou unités statistiques) pour lequel on considère une ou plusieurs caractéristiques
- Taille de la population
  - ✓ Le nombre d'individus constituant la population.

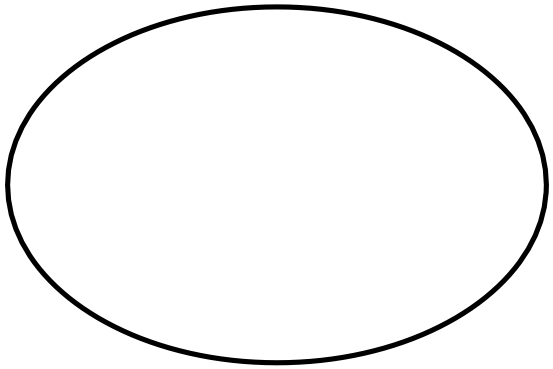
Notation : ***N***

- Exemple:
  - ✓ paramètre étudié : note d'un étudiant dans un groupe de TD. un individu = un étudiant

2. Statistique descriptive

Échantillon

Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la population dans son ensemble *(ce serait alors un recensement). On se restreint à un sous-ensemble, l'***échantillon** pour tirer des conclusions sur la population.



Taille de l'échantillon : le nombre d'observations dans l'échantillon. Notation : ***n***

2. Statistique descriptive

Caractères statistiques

- Caractère ou Variable = propriété observable des individus qui prend différents états appelés modalités.

- Plusieurs catégories de caractères => méthodes statistiques différentes.

- Deux types de variables sont à distinguer :
  - ✓ Variables quantitatives
  - ✓ Variables qualitatives

2. Statistique descriptive

Variables

- Variables quantitatives : les valeurs prises sont numériques.

Ces valeurs peuvent être :

- ✓ Discrètes : c'est à dire appartenant à une liste dénombrable. Exemple : le nombre de pannes d'une machine, le nombre des jours de travail.
- ✓ Continues : les valeurs prises ne peuvent pas être comptées et appartiennent à un intervalle. Exemple : la température, la moyenne annuelle d'un étudiant.

- Variables qualitatives : les valeurs prises sont des labels.

Ces valeurs peuvent être :

- ✓ Nominales : quand elles ne sont pas ordonnables. Exemple : la couleur.
- ✓ Ordinales : quand il est possible de les ordonner selon un sens : petit <moyen < grand ; faible < normal < puissant.

2. Statistique descriptive

Central Tendency

- Moyenne : la moyenne arithmétique est la somme des valeurs d'une variable quantitative, divisée par le nombre d'individus.
  - ✓ **Exp** : la moyenne des âges 3, 12, 18 est 11.
- Médiane : c'est la valeur qui sépare les valeurs d'une série statistique en deux.
  - ✓ **Exp :** la médiane de la série 1 3 5 7 9 est 5.
- Mode : il correspond à la valeur la plus fréquente.
  - ✓ **Exp :** le mode de la série 2, 3, 3, 4, 7, 3, 2, 1, 3 est 3 avec l'effectif 4.

2. Statistique descriptive

Data Dispersion

- Variance : la variance sert à caractériser la dispersion des valeurs de la moyenne.

$$\sigma^2 = \frac{\sum \left(X - \bar{X} \right)^2}{N}$$

  - ✓ variance de zéro: toutes les valeurs sont identiques,
  - ✓ petite variance : les valeurs sont proches les unes des autres,
  - ✓ variance élevée : celles-ci sont très écartées.

- Ecart-type : C'est la racine carrée de la variance. Il permet de mesurer l'écart entre les valeurs de la série avec la même grandeur que celle des valeurs.

2. Statistique descriptive

Représentations graphiques

• Les données statistiques peuvent être représentées graphiquement pour une meilleure *interprétation* et *mémorisation*.
• Ces représentations sont parfois suffisantes à elles-mêmes pour visualiser une population.

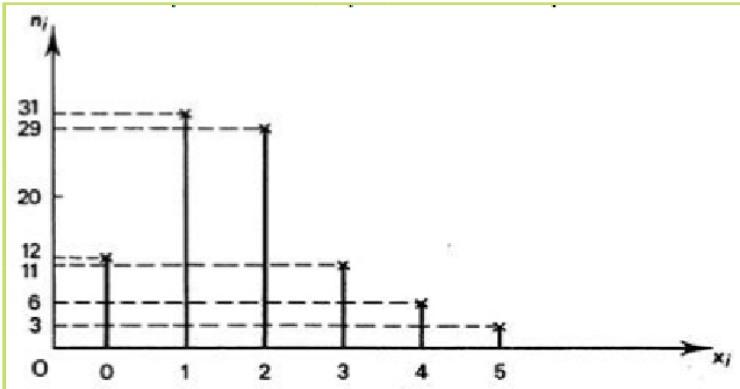
• Il existe différentes représentations graphiques associées au cas mono-variable, nous citons par exemple:

- ✓ Les Diagrammes en bâtons
- ✓ Les Histogrammes
- ✓ Les Graphiques cumulatifs
- ✓ Les Diagrammes en secteur

2. Statistique descriptive

Représentations graphiques

- Diagrammes en bâtons: Un diagramme en bâtons permet de représenter des ***effectifs*** d'une ***variable discrète*** sur deux axes : en abscisses les individus observés et en ordonnées représente l'effectif de chaque individu. On parle aussi de *nuage de points*.
- Exemple : soit la série suivante qui représente le nombre de foyers selon le nombre de personnes par foyer (tableau). Le diagramme à gauche représente graphiquement ce tableau.

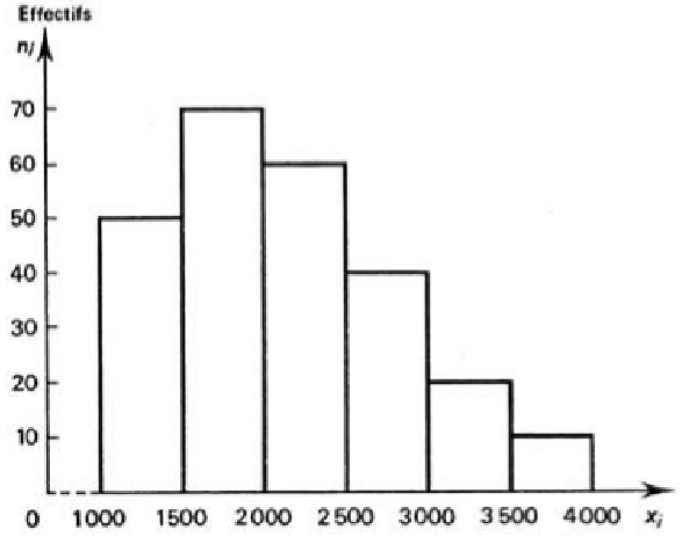


Nombre d'enfants par foyer « x_i »	Nombre de foyer concernés f_i
0	12
1	31
2	29
3	11
4	6
5	3

2. Statistique descriptive

Représentations graphiques

- Histogrammes : Les histogrammes s'adaptent aux cas d'une variable continue *quantitative* dont les valeurs peuvent être classées en intervalles. L'axe des abscisses représente les classes et l'axe des ordonnées représente les valeurs sous forme de rectangles.
- Exemple : soit la série suivante qui représente le nombre de personne selon la tranche de salaire (tableau). Le diagramme à gauche représente graphiquement ce tableau.



Salaire en € : x_i	Effectifs n_i
1000 à 1500	50
1501 à 2000	70
2001 à 2500	60
2501 à 3000	40
3001 à 3500	20
3501 à 4000	10
	250

2. Statistique descriptive

Représentations graphiques

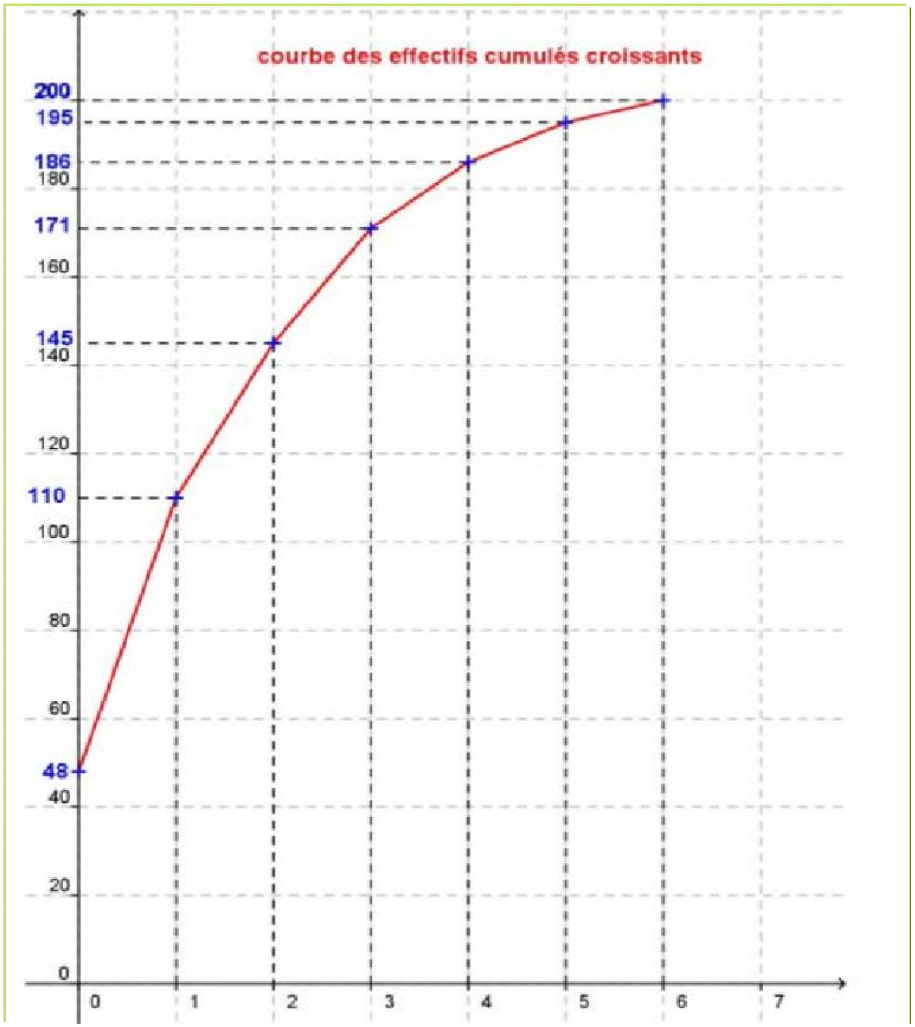
- Graphiques cumulatifs : les graphiques cumulatifs permettent de représenter les cumuls d'effectifs d'une série ordonnée. Si les effectifs initiaux ne correspondent pas à des cumuls, on les calcule d'abord avant de représenter le diagramme.
- Exemple : le tableau suivant représente l'évolution de salaire selon le grade. La deuxième ligne représente le montant d'évolution et la troisième ligne représente le cumul de salaire.

Modalités	0	1	2	3	4	5	6
Effectifs	48	62	35	26	15	9	5
ECC	48	110	145	171	186	195	200

2. Statistique descriptive

Représentations graphiques

- Les Graphiques cumulatifs : Le graphique cumulatif associé est le suivant



2. Statistique descriptive

Représentations graphiques

- Les Diagrammes en secteur :
  - ✓ ce genre de diagramme permet de représenter des proportions d'effectifs par rapport à la totalité.
  - ✓ Avant de le dessiner, on calcule la proportion de chaque valeur de variable (individu).
  - ✓ On calcule ensuite l'angle de chaque valeur de variable.

2. Statistique descriptive

Représentations graphiques

- Exemple : soit le tableau suivant qui donne les effectifs et fréquences des couleurs d'objets observés. Les modalités sont les couleurs bleu, marron, vert et noir. Les fréquences sont traduites en angles en les multipliant par 3,6.

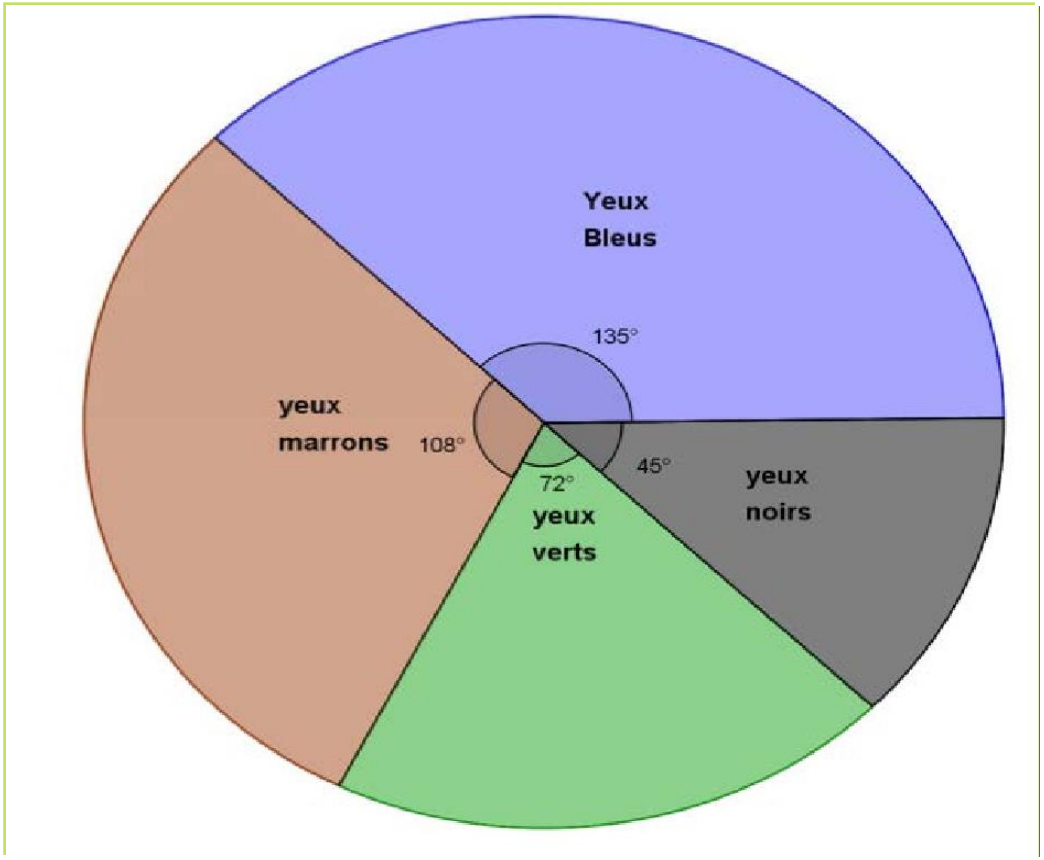
Modalités	bleu	marron	vert	noir	total
Effectifs	15	12	8	5	40
Fréquences	0,375	0,3	0,2	0,125	1
Fréquences en %	37,5	30	20	12,5	100
Angle (en °)	**135**	**108**	**72**	**45**	360

× 3,6

2. Statistique descriptive

Représentations graphiques

- Les Diagrammes en secteur :



3. Statistique multivarié

Cas de deux variables

- Il arrive souvent qu'on ait besoin d'analyser deux variables à la fois et qu'on cherche la relation entre elles.
- Par exemple : la relation entre la taille des enfants et leur âges.
- On note ce deux variables x et y.
- Les notions suivantes sont généralement prises en considération :

3. Statistique multivarié

Cas de deux variables

- La covariance: La covariance décrit la relation entre les changements des deux variables.
  - ✓ Si elle est positive, aux grands écarts d'une variable correspondent de grands écarts de la deuxième et vice-versa.
  - ✓ Par contre, une covariance négative signifie qu'aux grands écarts d'une variable correspondent de petits écarts de la deuxième.
  - ✓ La formule de la covariance est la la suivante:

$$\mathrm{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

3. Statistique multivarié

Cas de deux variables

- **Représentations graphiques :**

✓ Chaque individu est représenté par un point dans un plan.

✓ Les valeurs d'une variable sont placées sur un axe et les valeurs de la deuxième variable sur l'autre axe.

✓ Cette représentation est appelée *nuage de points*.

✓ Elle permet de déduire visuellement s'il y a une relation entre les valeurs des deux variables.

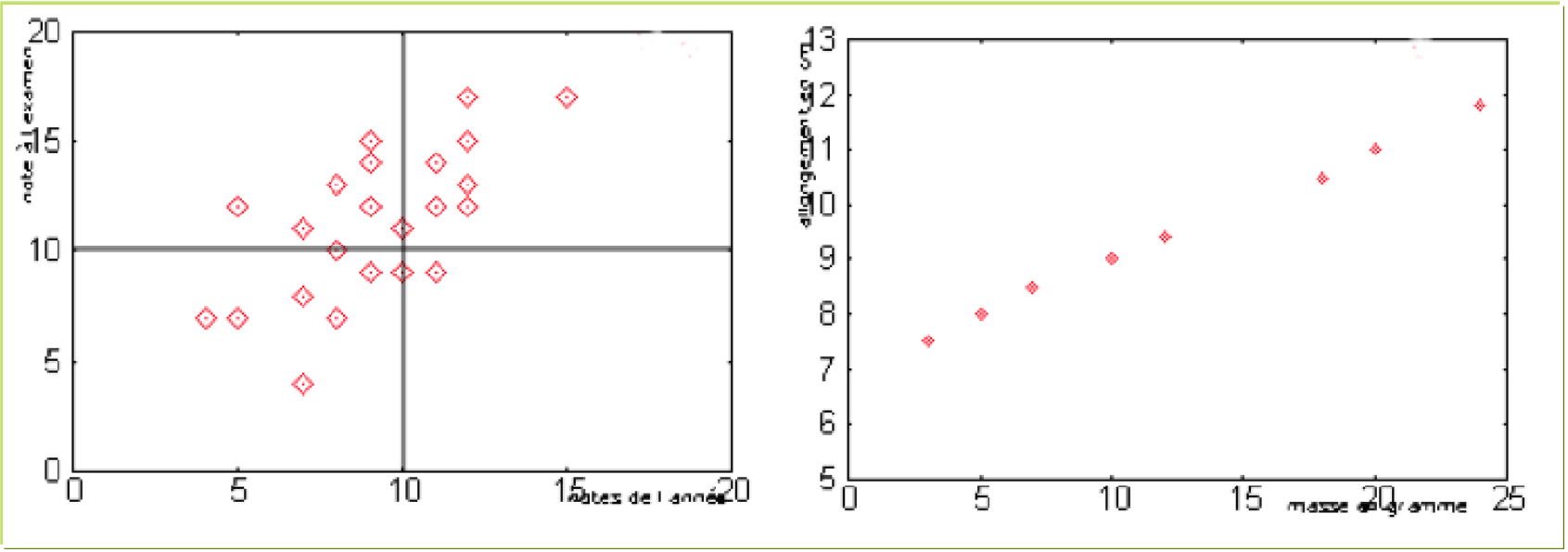
3. Statistique multivarié

Cas de deux variables

- **Représentations graphiques :**

✓Dans les graphiques ci-dessous, le nuage de points à gauche témoigne que les valeurs des d'une variable sont indépendantes des valeurs de l'autre.

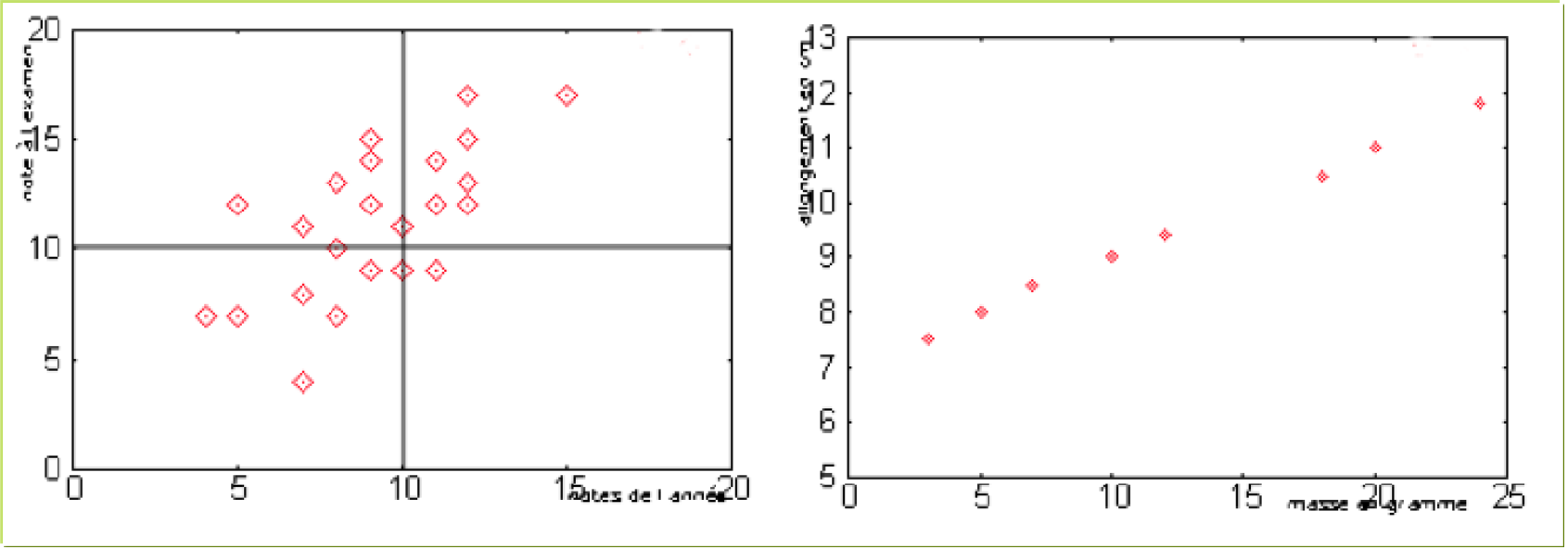
✓ Le nuage à droite montre qu'il se peut qu'il y ait une relation entre les valeurs.



3. Statistique multivarié

Cas de deux variables

- **Représentations graphiques :**



✓ Lorsqu'il peut exister une relation entre les valeurs des deux variables, on procède à l'ajustement

3. Statistique multivarié

Cas de deux variables

- Le coefficient de corrélation linéaire:

  ✓ On appelle *coefficient de corrélation* linéaire des variables X et Y le réel estimé par la covariance de X et Y mais normaliser par leur écart-type respectif, ce qui permettra d'obtenir un indice compris entre -1 et 1 :

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$ avec $$-1 \leq r \leq 1$$

  ✓ Le coefficient de corrélation linéaire mesure l'intensité de la liaison linéaire entre 2 variables pour autant que celle-ci est linéaire ou approximativement linéaire.

3. Statistique multivarié

Cas de deux variables

- Le coefficient de corrélation linéaire
  - ✓ Si $r = 1$: Il existe une relation linéaire entre X et Y soit : $Y = aX + b$ où $a > 0$.

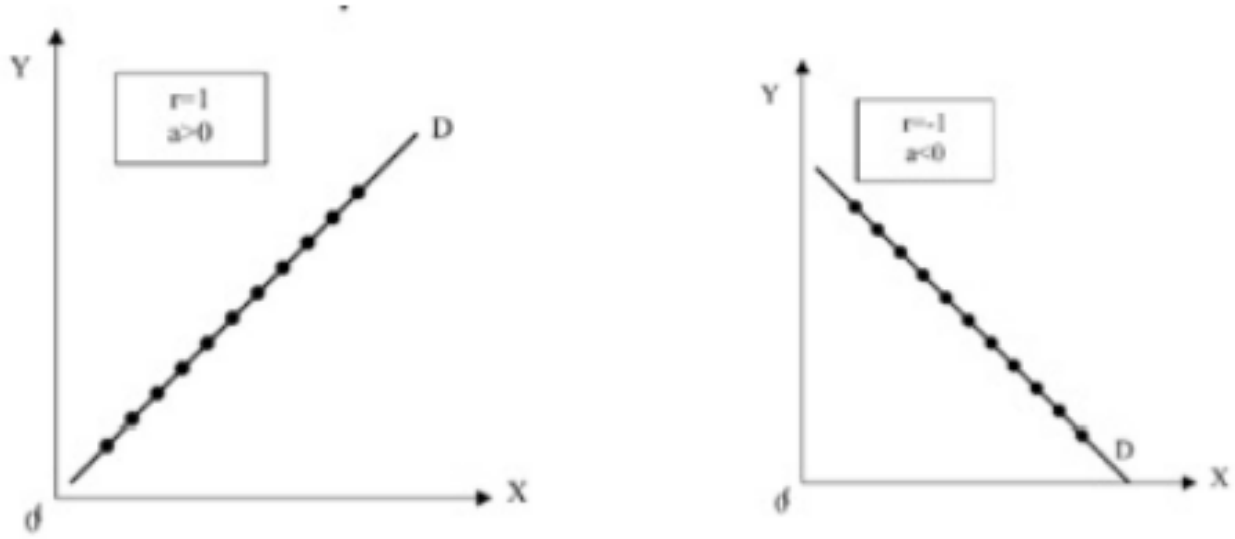


FIGURE 4 – Nuage de point avec $r = 1$ et $r = -1$

3. Statistique multivarié

Cas de deux variables

- Le coefficient de corrélation linéaire

  ✓ Si r = 3/4 : Il existe une liaison forte de type linéaire entre X et Y : des points se situant pratiquement à l'intérieur d'une ellipse de forme très allongée.

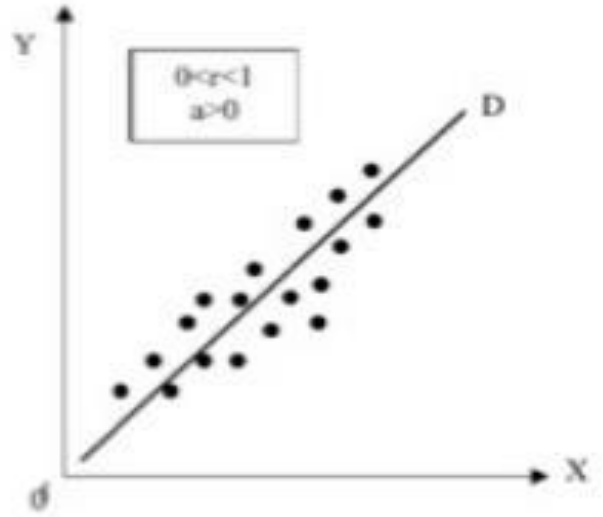


FIGURE 5 – Nuage de point avec $r = \frac{3}{4}$

3. Statistique multivarié

Cas de deux variables

- Le coefficient de corrélation linéaire

✓ r proche de 0 :

Plus r est proche de 0 moins les deux variables sont corrélées linéairement.

✓ Si r = 0 :

Les deux variables X et Y ne sont pas corrélées linéairement. L'ellipse coïncide avec un cercle et les deux droites de régression Y/X et X/Y sont perpendiculaires.

Cependant, r = 0 n'implique pas que les variables X et Y sont indépendantes mais seulement qu'elles ne sont pas corrélées linéairement.

3. Statistique multivarié

Cas de deux variables

- **Le coefficient de corrélation linéaire**

✓ Si r = 0 :

Exemple :

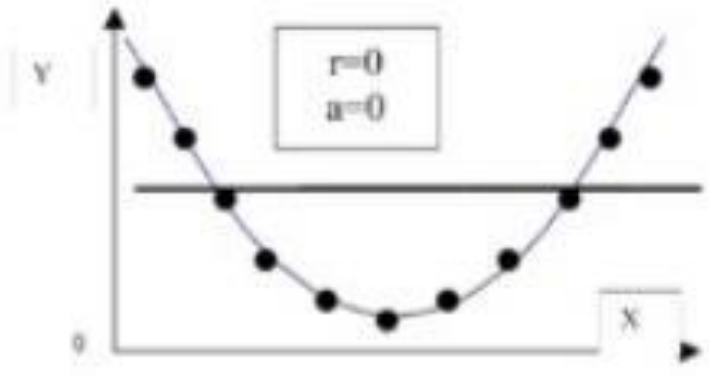


FIGURE 6 – Corrélation non linéaire

3. Statistique multivarié

Cas multidimensionnel

- Le cas multidimensionnel concerne plusieurs variables à la fois.
- La représentation graphique n'est pas adaptée à l'être humaine.
- Le traitement de ces valeurs fait intervenir les méthodes *d'analyse de données*.
- l'analyse de données regroupe une famille de méthodes pour décrire un grand nombre de données avec comme objectif de faire ressortir les relations entre elles ou de comprendre ce qui les rend homogènes

3. Statistique multivarié

Cas multidimensionnel

- Exemples de méthodes:
  - ✓ Analyse en composantes principales (ACP) : réduire p variables corrélées en q variables non corrélées.
  - ✓ Analyse factorielle des correspondances (AFC) : trouver des liens ou correspondances entre deux variables qualitatives (nominales) dans des tableaux de contingence.
  - ✓ Analyse des correspondances multiples (ACM) : L'ACM est l'équivalent de l'ACP pour les variables qualitatives et elle se réduit à l'AFC lorsque le nombre de variables qualitatives est égal à 2.

4. Variables aléatoires

- Quand on fait des analyses (en statistique, en science des données, etc.), on travaille toujours avec des *données* :
  - ✓ des mesures (ex. : température, taille, revenus, etc.)
  - ✓ des observations (ex. : résultats d'expériences, votes, ventes, etc.)
- Ces données ne sont *jamais parfaites* : elles peuvent être incomplètes, bruitées (avec des erreurs), ou varier d'un échantillon à l'autre.
- La théorie des probabilités sert donc à :
  - ✓ exprimer les incertitudes sur les données (Ex.: une mesure peut être "probablement autour de 10" mais pas exactement 10) ;
  - ✓ modéliser les hypothèses qu'on fait sur la façon dont les données sont produites (Ex,: "les hauteurs des personnes suivent probablement une loi normale").

4. Variables aléatoires

- Les données sont considérées comme des variables aléatoires, car elles peuvent prendre différentes valeurs selon les circonstances.
- La valeur d'une variable aléatoire est incertaine avant son observation : on ne peut que la prévoir ou l'estimer.
- La loi de probabilité associée à une variable aléatoire décrit et quantifie cette incertitude en indiquant la probabilité de chaque valeur possible.

4. Variables aléatoires

Expérience aléatoire

- Une *expérience aléatoire*, une expérience dont le résultat n'est pas prévisible *à priori*.

  Ex. : Jet d'une pièce de monnaie, Lancer d'un dé, …

- Un espace d'échantillons (espace universel) Ω est l'ensemble des résultats possibles d'une expérience aléatoire.

- Il est appelé aussi : Espace des épreuves , espace fondamental.

  Ex. : Lancer une pièce de monnaie 2 fois,

  $$\Omega = \{PP, PF, FP, FF\}$$

  Il y a donc 4 issues possibles

4. Variables aléatoires

Evénements

- Un évènement E est réalisé ou non par une épreuve. C'est donc un sous-ensemble de Ω :

  Ex.: La première lancée est une pile: $E = \{PP, PF\}$

- Un événement élémentaire est un événement formé d'une *seule issue*.

  Ex.: les deux lancées piles: $\{PP\}$.

- L'espace des événements $\mathcal{A}$ est l'ensemble de tous les sous-ensembles de Ω, y compris l'événement vide $\emptyset$ (aucune issue) et Ω lui-même (événement certain).

4. Variables aléatoires

Evénement

- C'est l'ensemble des parties de Ω, qu'on appelle aussi la puissance de Ω :

  $\mathcal{A} = P(\Omega) = \{\emptyset, \{PP\}, \{PF\}, \{FP\}, \{FF\}, \{PP, PF\}, \dots, \Omega\}$

- Un espace de probabilité ou espace probabilisé est un triplet $(\Omega, \mathcal{A}, \mathbb{P})$:
  - ✓ Ω: L'univers
  - ✓ $\mathcal{A}$: Un espace d'évènements
  - ✓ Une application de probabilité: $\mathbb{P}: \Omega \rightarrow [0,1]$, appelée probabilité sur Ω, tel que $\mathbb{P}(\Omega) = 1$

4. Variables aléatoires

Probabilités

- Soit un espace de probabilité Ω, et soient A *et* B *deux événements de* Ω .
- On définie sur $(\boldsymbol{\Omega}, \mathcal{A})$:

  ✓ $p(A) \geq 0$ pour chaque A dans $\mathcal{A}$.

  ✓ $p(\Omega) = 1$.

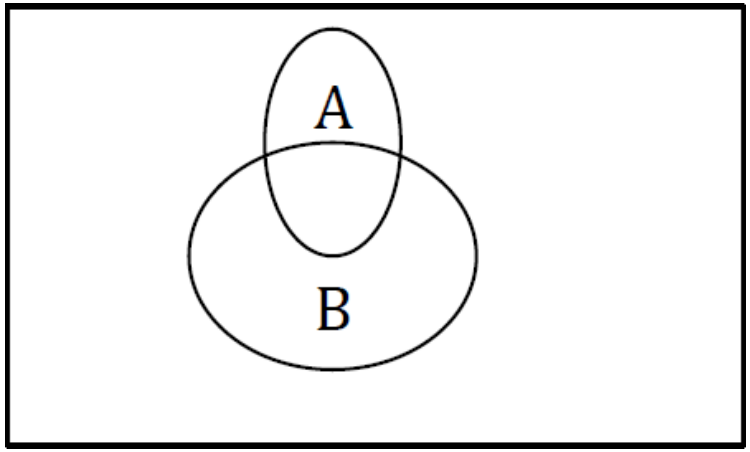
  ✓ $P(A \, U \, B) = p(A) + p(B) - p(A \cap B)$

  ✓ Si A, B sont 2 événements disjoints:

4. Variables aléatoires

Probabilité conditionnelle

- $p(A|B)$ = Fraction où A est vrai sachant que B est vrai aussi. On a alors la loi de Bayes:

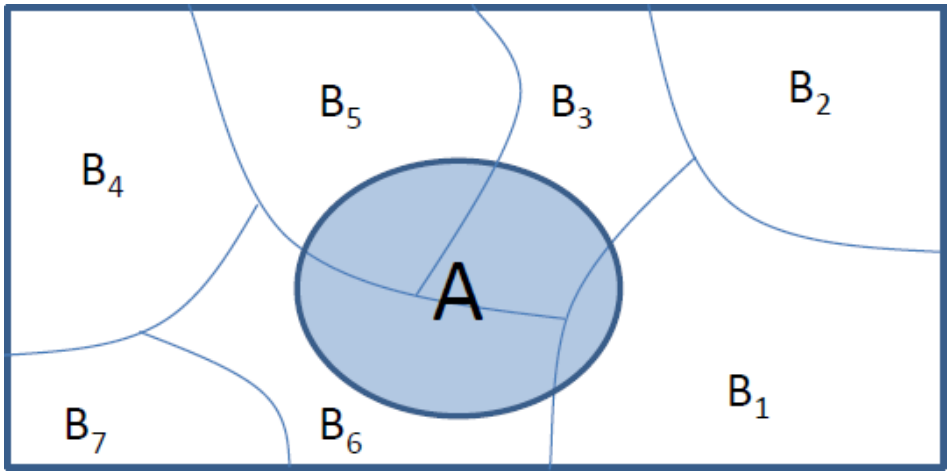


$$p(\mathrm{A}\,|\,\mathrm{B}) = \frac{p(\mathrm{A}\cap\mathrm{B})}{p(\mathrm{B})} \quad \Rightarrow \quad p(\mathrm{A}\cap\mathrm{B}) = p(\mathrm{A}\,|\,\mathrm{B})\,p(\mathrm{B})$$

4. Variables aléatoires

Probabilité marginale

- Soit A un événement et $\{B_1, B_2, \dots, B_n\}$ une partition de Ω.



- On a la probabilité marginale:

$$p(A) = \sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(B_i)p(A|B_i)$$

4. Variables aléatoires

Qu'est-ce qu'une variable aléatoire ?

- Une variable aléatoire est une fonction qui associe un nombre réel à chaque issue d'une expérience aléatoire.
- Elle permet de traduire les résultats de l'expérience en nombres, pour les analyser statistiquement.
- C'est une application de Ω dans IR $(X: \Omega \to \mathbb{R})$., et elle est notée par une lettre majuscule : $X, Y, Z \ldots$
- Exemple : Considérons l'éxpérience aléatoire qui consiste à lancer deux dés équilibrés. $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$. On s'intéresse à la somme des deux résultats.

$$Y(\Omega) = \{2,3,\ldots,12\}$$

4. Variables aléatoires

Types variables aléatoires

- Les variables aléatoires sont des valeurs numériques continues ou discrètes.
- VA discrètes prennent des valeurs dénombrables distinctes.
  - ✓ Exp.: *Nombre de dinars dans mon compte bancaire, Nombre total de face après 100 lancée d'une pièce. (valeurs possibles: 0, 1, 2, …, 100).*
- VA continues prennent des valeurs sur un intervalle de IR non-dénombrables (réelles)
  - ✓ Exp: Temps d'attente avant un bus (10.5, 9.8, …), Revenu d'un client, …

4. Variables aléatoires

Règles de probabilités

Soit x et y deux variables aléatoires discrètes:

x peut prendre les valeurs: $u_1, u_2, \ldots, u_N$

y peut prendre les valeurs: $v_1, v_2, \ldots, v_M$

On aura alors:

- $\sum_{i=1}^{N} p(\mathrm{x} = u_i) = 1$
- $\sum_{j=1}^{M} p(\mathrm{y} = v_j) = 1$
- La probabilité conjointe d'observer $\mathrm{x} = u_i$ et $\mathrm{y} = v_j$ est :

$$p(\mathrm{x} = u_i, \mathrm{y} = v_j)$$

Règles de probabilités

Règle de Bayes pour les variables aléatoires

- Permet d'inverser l'ordre des probabilités conditionnelles.

$$p(\mathrm{y} = v_j|\mathrm{x} = u_i) = \frac{p(\mathrm{x} = u_i|\mathrm{y} = v_j)p(\mathrm{y} = v_j)}{p(\mathrm{x} = u_i)}$$

☞ $p(\mathrm{y} = v_j)$ est appelée probabilité a priori.

☞ $p(\mathrm{y} = v_j|\mathrm{x} = u_i)$ est appelée probabilité a posteriori.

4. Variables aléatoires

Loi de probabilité

- Soit une variable aléatoire $\boldsymbol{X}$ définie sur un univers $\boldsymbol{\Omega}$ et prenant les valeurs $x1, x2, \ldots, xn$.
- La loi de probabilité de $\boldsymbol{X}$ associe à toute valeur xi de $\boldsymbol{X}$ la probabilité: $P(X = xi) = pi$
- Exemple: lancer un dé : $\Omega = \{1,2,3,4,5,6\}$
  - ✓ la variable aléatoire $\boldsymbol{X}\ (\Omega) = \{1,2,3,4,5,6\}$
  - ✓ Chaque issue du lancer de dé est équiprobable et égale à $\mathbf{1/6}$: $\forall\ \omega \epsilon \Omega,\ P(\omega) = \frac{1}{6}$.
  - ✓ Le tableau de la loi de probabilité de la variable aléatoire X:

xi	1	2	3	4	5	6
$P(X = xi\ )$	1/6	1/6	1/6	1/6	1/6	1/6

4. Variables aléatoires

Variable aléatoire discrète

• La *loi de probabilité* d'une variable aléatoire *discrète* est entièrement déterminée par les probabilité pi des événements $\{X = xi\}$, xi parcourant l'univers image $X(\Omega)$.

• La *fonction de masse* de probabilité (PMF) d'une variable aléatoire discrète X est la fonction $fX: \mathbb{R} \to [0,1]$ définit par :

$$fX(x) = P(X = x).$$

- • Elle vérifier deux conditions:
  $fX(x)$ >= 0 et $\sum_{x \in X(\Omega)} fX(x) = 1$
  • Dans notre exemple de lancé un dé:

$$\sum_{x=1}^{6} fX(x) = \ 6 * \frac{1}{6} = 1$$

4. Variables aléatoires

Variable aléatoire discrète

- On utilise un *diagramme en bâtons* pour visualiser la distribution de probabilité d'une variable discrète :

$P(X = xi) = pi$

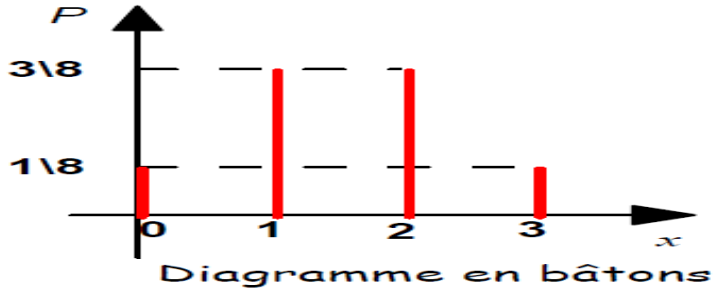
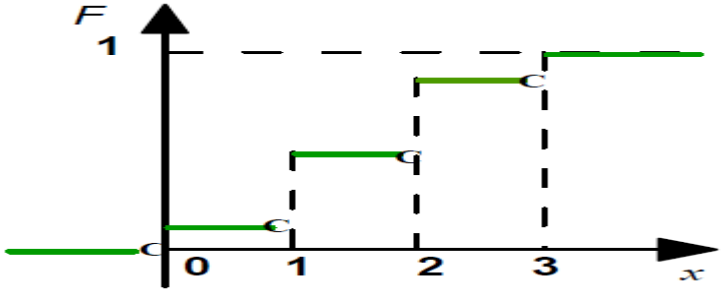


Diagramme en bâtons

- On appelle *fonction de répartition* (fonction de répartition cumulée noté par CDF ou CFF) d'une variable aléatoire X, la fonction F, telle que $F(x) = P(X \leq x)$

$$= \sum_{i=x0}^{x} P(X = i) \ .$$



Fonction de répartition

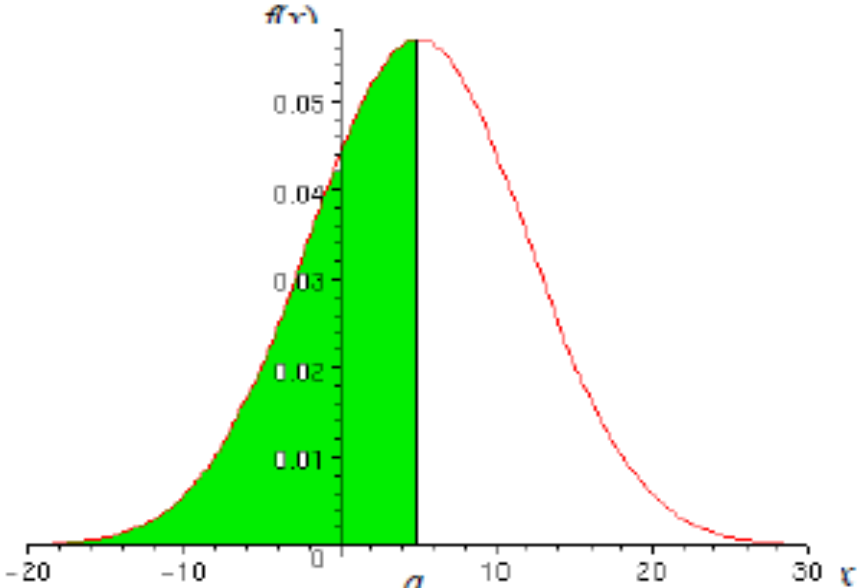
4. Variables aléatoires

Variable aléatoire continue

- Une variable aléatoire est dite continue si elle peut prendre toutes les valeurs dans un intervalle donnée (borné ou non borné) :$X: \Omega \rightarrow X(\Omega)$ avec $X(\Omega) =]a, b[\subset IR$
- La loi de la variable aléatoire continue Xest définit par une fonction f, appelée *densité de probabilité* (*PDF*), tel que pour tous $a, b \in R : P(a \leq X \leq b) = \int_a^b f(x)$, et qui vérifie :

i) $\forall x \in \mathbb{R}, \ f(x) \geq 0$

ii) $\int_{\mathbb{R}} f(x)dx = 1.$



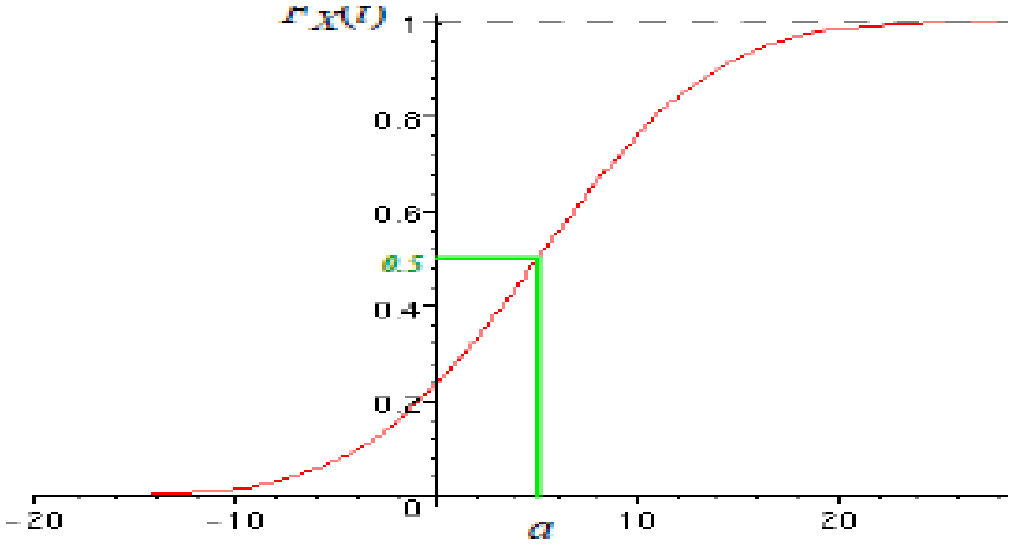
4. Variables aléatoires

Variable aléatoire continue

• On appelle *fonction de répartition* (fonction de répartition cumulée) d'une variable aléatoire continue X, la fonction F ($F(x) = P(X \leq x)$), telle que :

$$F : \mathbb{R} \longrightarrow \mathbb{R}$$

$$x \longrightarrow F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$



4. Variables aléatoires

Moyenne d'une variable aléatoire (VA)

On dénote la moyenne (espérance) d'une VA x par: $\mu = \mathbb{E}(\mathrm{x})$.

☞ Pour VA discrète: $\mathbb{E}(\mathrm{x}) = \sum_{i=1}^{N} u_i \, p(\mathrm{x} = u_i)$

☞ Pour V-A continue: $\mathbb{E}(\mathrm{x}) = \int_{-\infty}^{+\infty} u \, f(\mathrm{x} = u) \, du$

- La variance et l'cart type de la loi de probabilité de X est :

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \big(x_i - E(x)\big)^2 \qquad \sigma(X) = \sqrt{\mathrm{Var}(X)}$$