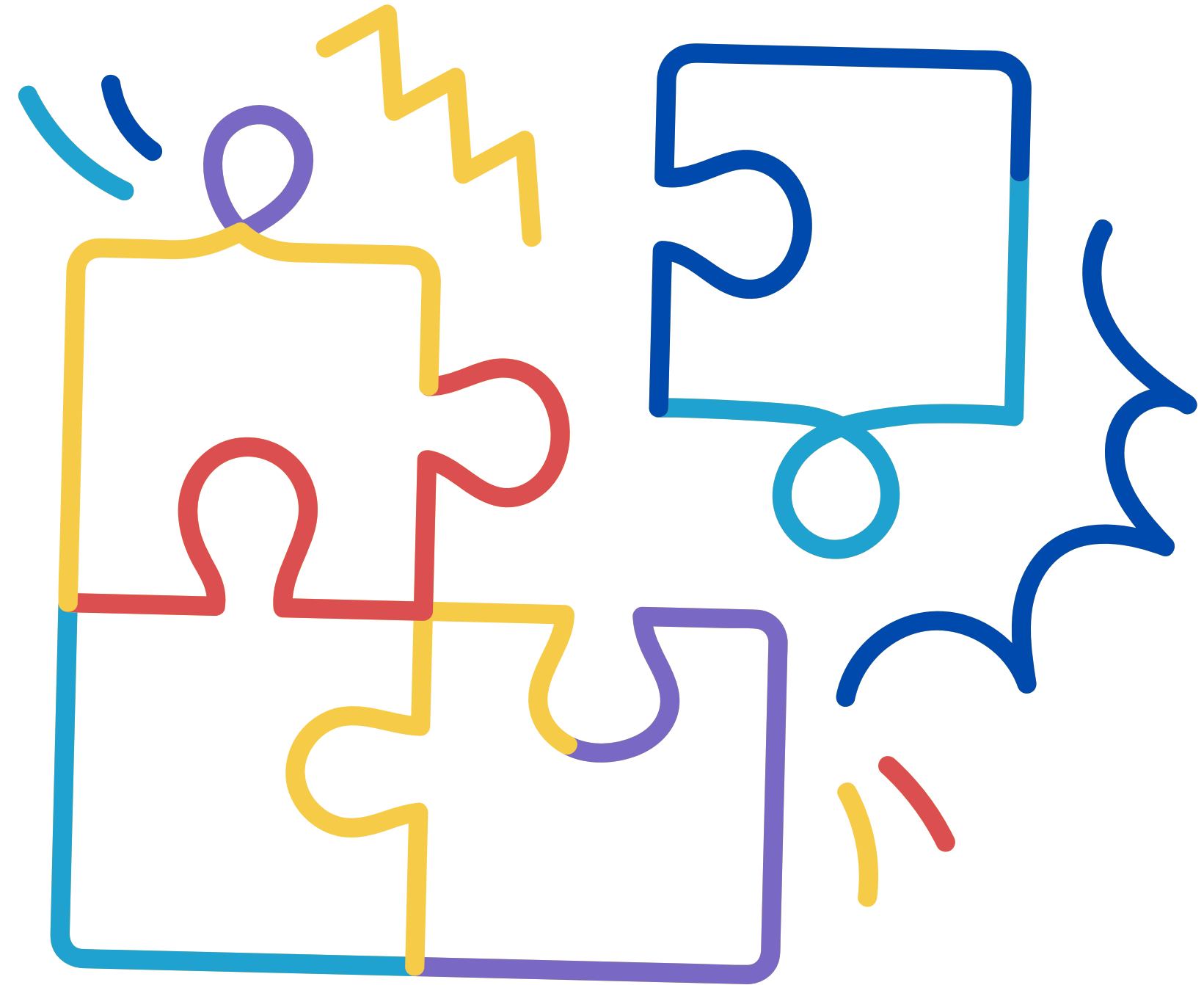


Octobre 2025

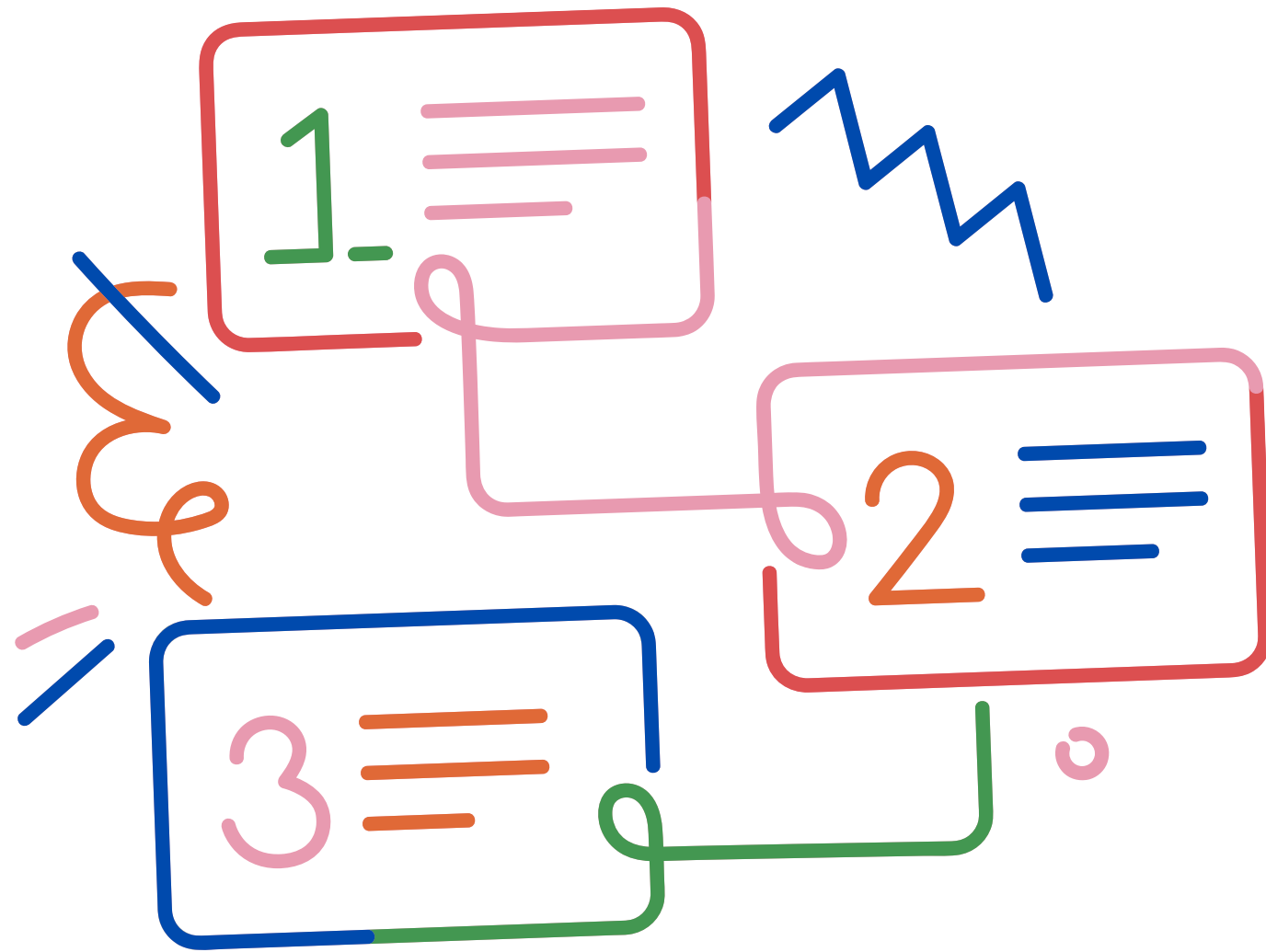
Module : Analyse de données et
classification

Le B.A.-BA de l'Analyse de données

Pr. Khadija LEKDIOUI



Sommaire



- 01. C'est quoi l'analyse de données
- 02. Généralités et vocabulaires de base
- 03. Univers – Phénomènes - Modélisation
- 04. Variables Continues et Variables Discretes
- 05. L'Analyse Technique
- 06. L'Analyse Univariée
- 07. L'Analyse Multivariée

C'est quoi une donnée ?



Les données

Chaque jour, nous voyons, entendons ou faisons des choses qui peuvent être enregistrées ou mesurées.

→ Ces petites informations enregistrées, ce sont des données.



Exemple Concret

quand tu prends ta température, le thermomètre affiche $37,8^{\circ}\text{C}$ → c'est une donnée.

Quand tu regardes ta note d'examen, tu vois 14/20 → c'est une donnée.

Quand ton téléphone compte 8 000 pas dans la journée, ce chiffre est une donnée.

Quand tu choisis la couleur de tes yeux dans un formulaire, c'est aussi une donnée.



Une donnée

C'est une information simple, qu'on peut observer, mesurer ou enregistrer.

De la donnée à l'information

La donnée est un élément brut.

Quand on la regroupe, nettoie, et interprète, elle devient une information.

	Description	Exemple
Donnée	Observation brute	Température = 38 °C
Information	Interprétation	Le patient a de la fièvre
Connaissance	Règle ou conclusion	Une température > 37.5°C indique un état fébrile



Analyse de données

L'analyse de données sert justement à transformer des données en informations, puis en connaissances utiles.

Pourquoi analyser les données ?



1. Pour comprendre le monde qui nous entoure

Les données sont partout : dans nos téléphones, les réseaux sociaux, les entreprises, les hôpitaux, les écoles... Mais sans les analyser, elles ne servent à rien.

Analyser les données, c'est transformer des chiffres bruts en informations utiles.

exemple

1. Une entreprise analyse les ventes pour savoir quels produits plaisent le plus.
2. Un médecin analyse les données de santé pour détecter une maladie plus tôt.



2. Pour prendre de meilleures décisions

Les décisions basées sur les données sont plus objectives et fiables.

Au lieu d'agir selon son intuition, on agit selon les faits.

exemple

1. Une école analyse les notes des élèves pour repérer ceux qui ont besoin d'aide.
2. Une application comme Uber analyse les trajets pour prévoir la demande et ajuster les prix.



3. Pour prédire et améliorer

L'analyse de données ne se limite pas à comprendre le passé.

Elle permet aussi de prédire l'avenir et d'optimiser les performances.

exemple

1. Netflix prédit quel film tu aimeras regarder ensuite.
2. Une usine prévoit quand une machine risque de tomber en panne pour intervenir à temps.

C'est quoi l'analyse des données ?



l'analyse de données

Les données sont partout et elles sont indispensables pour comprendre le monde. Mais pour en tirer une information utile, il faut passer par une étape clé : l'analyse de données.

En d'autres mots, analyser les données, c'est faire parler les chiffres.



Définition

L'analyse de données est l'ensemble des méthodes et outils permettant de transformer des données brutes en informations compréhensibles et exploitables.

On y applique des techniques **statistiques, mathématiques et informatiques** pour :
décrire les données,
détecter des relations,
faire des prévisions,
ou aider à la prise de décision.



Exemple concret

Une école collecte les notes des étudiants dans différentes matières.

En analysant ces données, elle peut :

- repérer les étudiants en difficulté,
- identifier les matières les plus exigeantes,
- et améliorer les méthodes d'enseignement.

Sans analyse : on a juste des chiffres. **Avec analyse** : on obtient des informations utiles pour agir.

C'est quoi l'analyse des données ?

Les grandes étapes de l'analyse de données

Etape	Description	Exemple
Collecte des données	Rassembler les informations utiles	Notes, ventes, températures..
Nettoyage / Préparation	Supprimer les erreurs, valeurs manquantes, doublons	Enlever les notes manquantes
Exploration (EDA)	Observer, décrire, visualiser les données	Moyenne, médiane, histogramme
Analyse statistique / Modélisation	Trouver des relations ou modèles	Corrélation entre âge et note
Interprétation des résultats	Traduire les chiffres en décisions	Adapter la méthode d'évaluation
Communication / Visualisation	Présenter les résultats clairement	Graphiques, tableaux, rapports

Ces étapes forment un cycle complet d'analyse.

Lien entre analyse de données et statistiques



l'analyse de données / Statistiques

Quand on fait de l'analyse de données, on cherche à comprendre ce que les données racontent.

Mais pour cela, il faut un langage commun : celui des statistiques.

En d'autres mots :
Les statistiques sont les outils essentiels qui permettent de décrire, résumer et interpréter les données.



Statistiques

L'analyse de données s'appuie donc sur les méthodes statistiques pour :

- repérer des tendances (ex : la moyenne des ventes augmente),
- mesurer la dispersion (ex : certaines valeurs sont très éloignées de la moyenne),
 - comparer des groupes,
 - et visualiser les résultats.



Exemple concret

Si tu disposes des notes de 100 étudiants,
les statistiques te permettent de calculer la moyenne, la médiane, ou encore de tracer un histogramme.

Ces outils t'aident à résumer un grand volume d'informations et à mieux comprendre la répartition des données.

Généralités et vocabulaires de base

Statistique :

La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leurs analyses, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensible.

✓ La démarche statistique est la suivante :

✓ Recueil des données :

- Population
- Echantillon

✓ Traitement des données

- Calcul
- Classement
- Compression

✓ Interprétation et analyse

- Présentation
- Organisation
- Réflexion

Généralités et vocabulaires de base

Statistique descriptive :

La statistique descriptive concerne la collecte, le résumé et la présentation des données. Elle inclut des mesures comme :

Les mesures de tendance centrale (moyenne, médiane, mode)

Les mesures de dispersion (variance, écart-type, étendue)

la visualisation des données (histogrammes, diagrammes en barres, camemberts, etc.)

Objectif : Résumer les données de manière informative.

Statistique inférentielle :

La statistique inférentielle vise à tirer des conclusions sur une population à partir d'un échantillon de données.

Objectif : Faire des prévisions ou des généralisations à partir d'un échantillon de données.

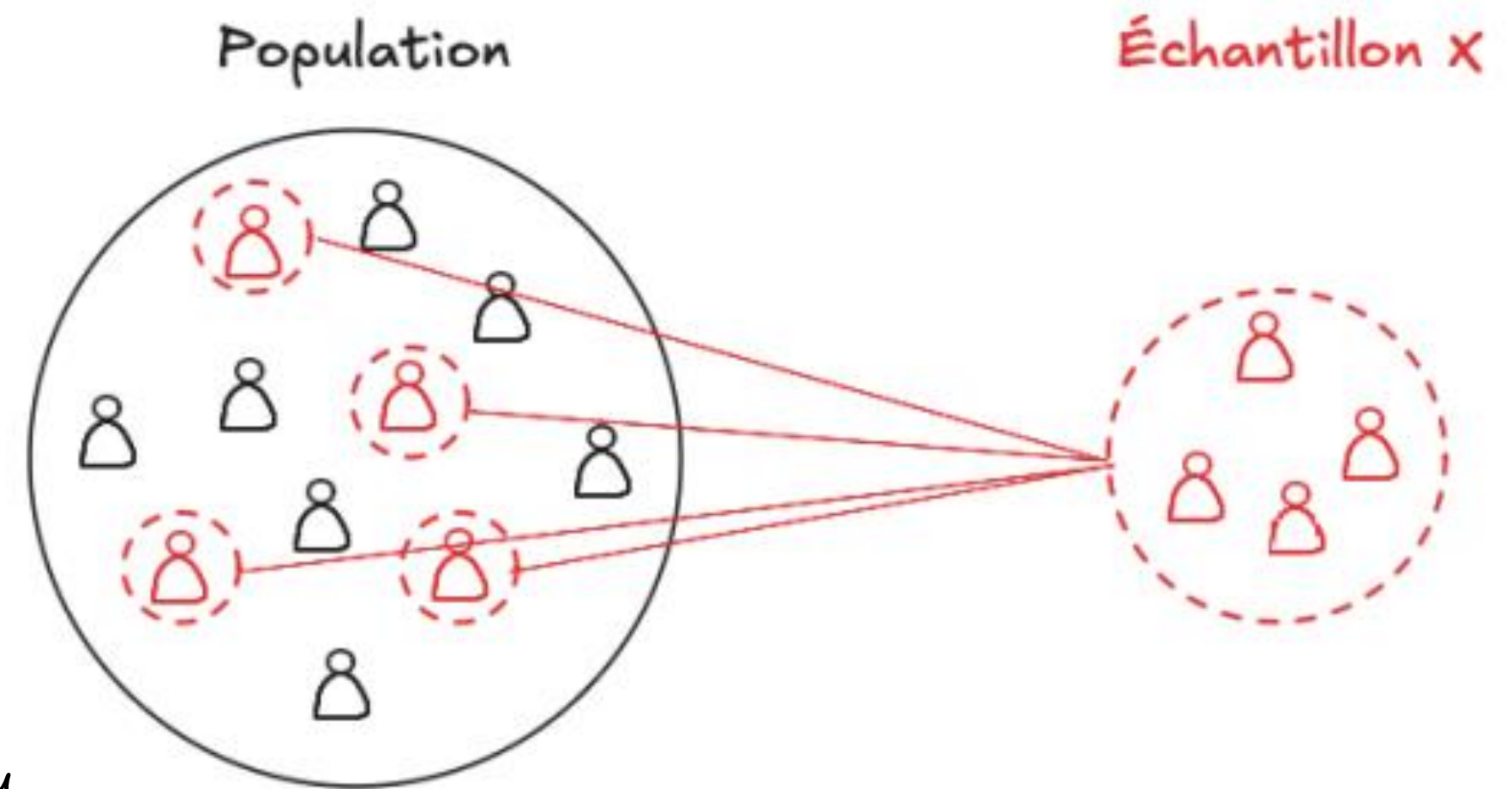
Statistiques Descriptives

Population et échantillon :

En statistique, on distingue toujours la population l'échantillon.

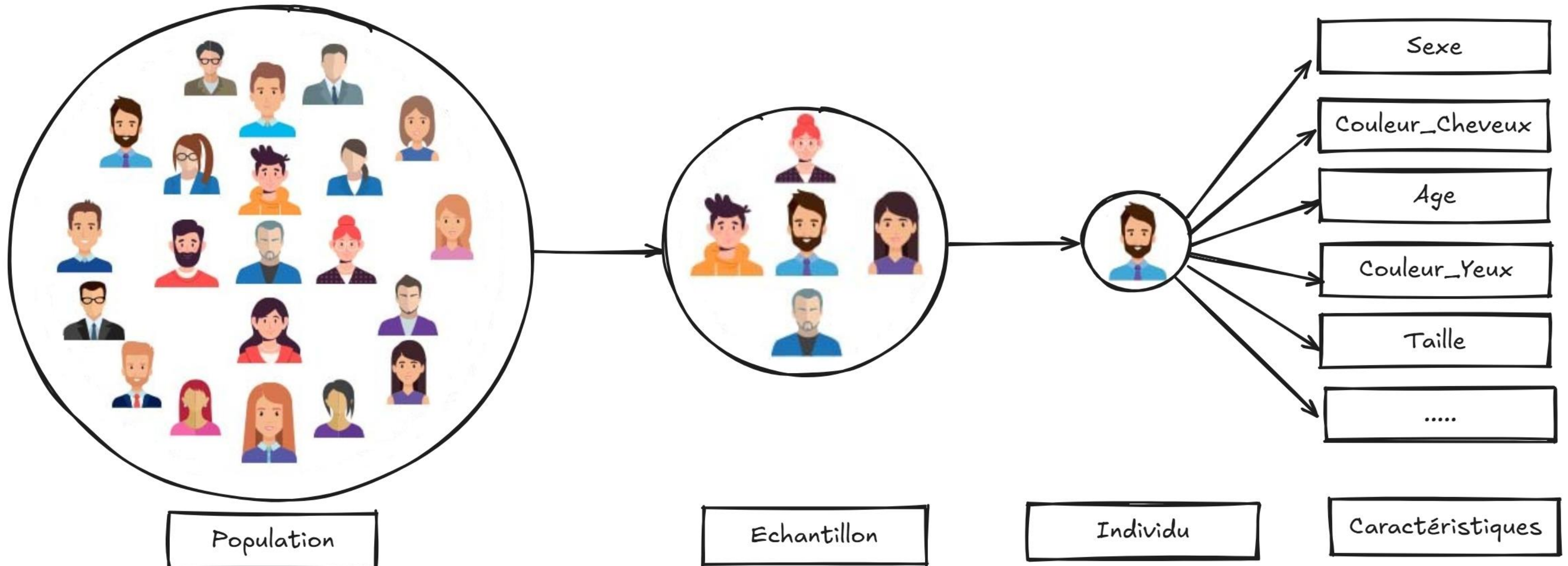
Population : La population est l'ensemble complet des individus, objets ou éléments que l'on souhaite étudier. Elle est souvent trop grande pour être analysée entièrement. Par exemple, tous les étudiants d'une université, tous les arbres d'une forêt, ou tous les citoyens d'un pays.

Échantillon : Un échantillon (sample en anglais) noté X est un sous-ensemble de la population, choisi de manière à être représentatif de celle-ci. On utilise un échantillon pour inférer des propriétés de la population. Par exemple, un sondage sur 1000 citoyens pour estimer l'opinion politique d'un pays.





Exemple :

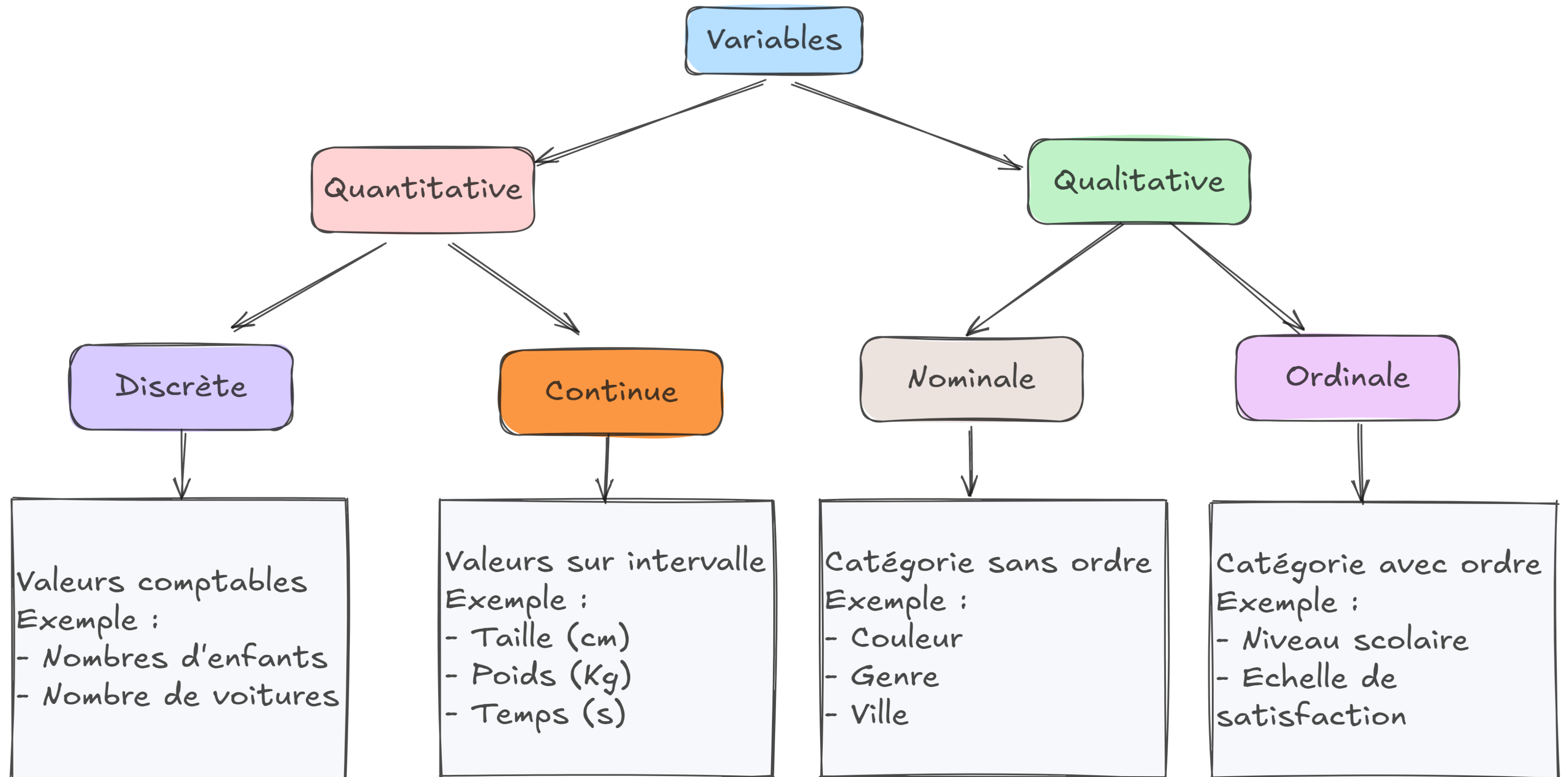




Exemple :

ID	Sexe	Couleur_Cheveux	Age	Couleur_Yeux	Taille (cm)
1	Femme	Roux	28	Marron	160
2	Homme	Brun	30	Marron	175
3	Homme	Noir	18	Noir	165
4	Femme	Noir	31	Noir	158
5	Homme	Gris	65	Bleu	182

Les types de variables et comment les décrire



Indicateur de position

Moyenne :

Les moyennes sont des indicateurs de position permettant de représenter une valeur centrale caractéristique d'un ensemble de données. Il en existe plusieurs types selon le contexte d'analyse. Les trois plus courantes sont la moyenne arithmétique, la moyenne harmonique et la moyenne géométrique.

Moyenne arithmétique :

La moyenne arithmétique est la forme la plus classique de moyenne. Elle est obtenue en additionnant toutes les valeurs d'un ensemble puis en divisant par le nombre total de valeurs. Elle est adaptée lorsque toutes les observations ont le même poids.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Indicateur de position

Moyenne harmonique :

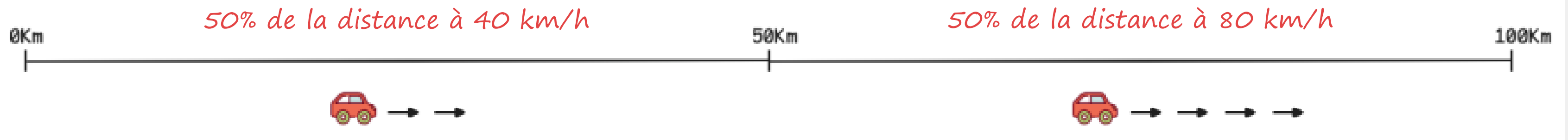
La moyenne harmonique est utilisée lorsque les données représentent des rapports ou des vitesses (par exemple, des vitesses sur une même distance). Elle est définie comme l'inverse de la moyenne des inverses des valeurs. Un exemple bien connu en classification est le score F1, qui correspond à la moyenne harmonique entre la précision et le rappel.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Indicateur de position

Exemple — Calcul de la vitesse moyenne correcte (moyenne harmonique) :

Considérons un trajet de 100 km divisé en deux portions de distance égale (50 km chacune). La première moitié est parcourue à une vitesse de 40 km/h, et la seconde à 80 km/h. Il s'agit d'un cas typique où la moyenne harmonique est nécessaire pour obtenir la vitesse moyenne réelle du trajet.



Commençons par calculer le temps réellement passé sur chaque portion :

Pour la première portion : $\frac{50}{40} = 1.25 \rightarrow 1\text{h}15$

Pour la deuxième portion : $\frac{50}{80} = 0.62 \rightarrow 38 \text{ min}$

Le temps total de trajet est donc : $1.25 + 0.62 = 1.875$ (1h52)

Indicateur de position

La vitesse moyenne réelle sur l'ensemble du trajet est obtenue en divisant la distance totale par le temps total:

$$v = \frac{100}{1.875} = 53.3 \text{ km/h} \neq 60 \text{ km/h}$$

On peut aussi retrouver ce résultat directement avec la moyenne harmonique des deux vitesses, puisque les distances sont égales :

$$\bar{x}_H = \frac{2}{\frac{1}{40} + \frac{1}{80}} = 53.3 \text{ km/h}$$

Ce résultat coïncide avec celui obtenu par le calcul du temps total, ce qui confirme que la moyenne harmonique est ici la méthode correcte. À l'inverse, si l'on utilisait la moyenne arithmétique des deux vitesses (40 km/h et 80 km/h), on obtiendrait :

$$\bar{x}_H = 60 \text{ km/h}$$

Ce résultat est trompeur : une vitesse moyenne de 60 km/h impliquerait un temps de trajet de $100 \div 60 = 1.66$ (1h40min), ce qui est inférieur au temps réel de 1h52 min. On sous-estimerait donc la durée du trajet.

Conclusion : Lorsque l'on parcourt des distances égales à des vitesses différentes, c'est la moyenne harmonique qui fournit la vraie vitesse moyenne. Elle prend en compte le temps passé à chaque vitesse, ce que ne fait pas la moyenne arithmétique.

Indicateur de position

Médiane :

La médiane est une mesure de tendance centrale qui correspond à la valeur centrale d'une série de données ordonnées.

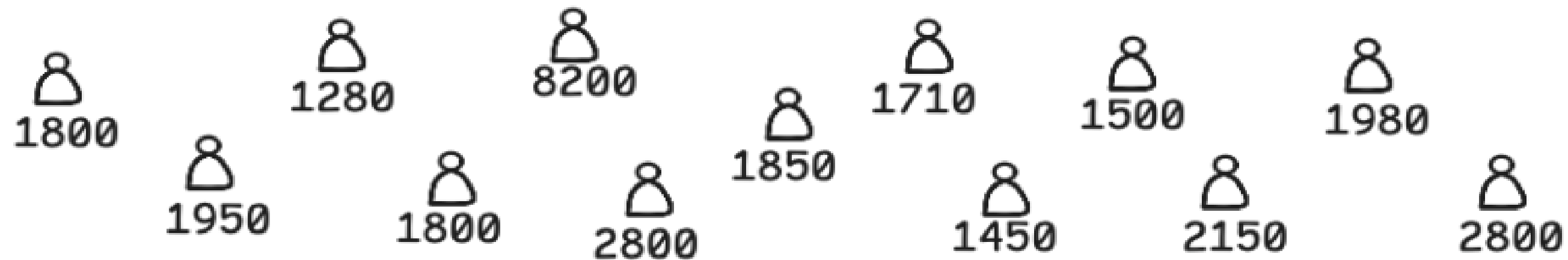
- Si le nombre d'observations n est impair, la médiane est la valeur située en position $(n - 1)/2$ (en comptant à partir de 0).
- Si n est pair, elle est la moyenne des deux valeurs centrales, en positions $n/2$ et $(n/2) - 1$ (en comptant à partir de 0).

La médiane divise la distribution en deux parties égales : 50% des observations sont inférieures ou égales à la médiane, et 50% sont supérieures ou égales à celle-ci. Elle est également robuste aux valeurs extrêmes (contrairement à la moyenne).

Indicateur de position

Exemple :

Soit l'échantillon suivant, constitué d'individus et leur salaire, répartissez les individus en ordre croissant et calculez la médiane.



1280	1450	1500	1710	1800	1800	1850	1950	1980	2150	2800	2800	8200
------	------	------	------	------	------	------	------	------	------	------	------	------

Maintenant, si on enlève une personne, on se retrouve avec un nombre impaire d'individus. Calculez la médiane sur ce nouvel échantillon.



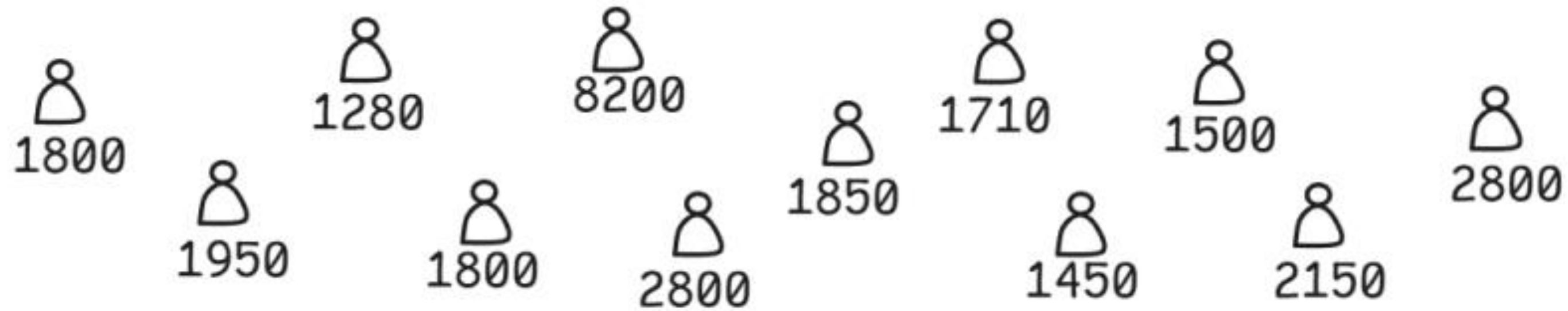
1280	1450	1500	1710	1800	1800	1850	1950	2150	2800	2800	8200
------	------	------	------	------	------	------	------	------	------	------	------

1825

Indicateur de position

Exemple :

Différence avec la Moyenne :



$$\bar{x}=2440$$

La moyenne est tirée vers le haut du fait des valeurs extrêmes, alors que la médiane reste insensible à ces valeurs.

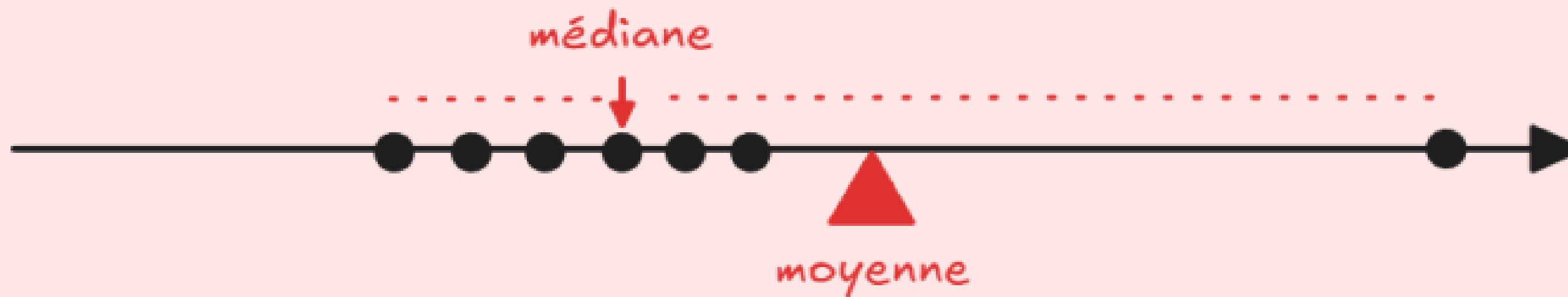
Indicateur de position

Important :

Il ne faut surtout pas penser que la médiane est plus représentative que la moyenne !

Ces 2 mesures se complètent !

- *La moyenne représente le point d'équilibre de vos données, comme le barycentre d'une balançoire*
- *la médiane divise vos données en 2 parties, avec 50% en dessous et 50% au dessus.*



Indicateur de position

Quartiles, déciles, percentiles :

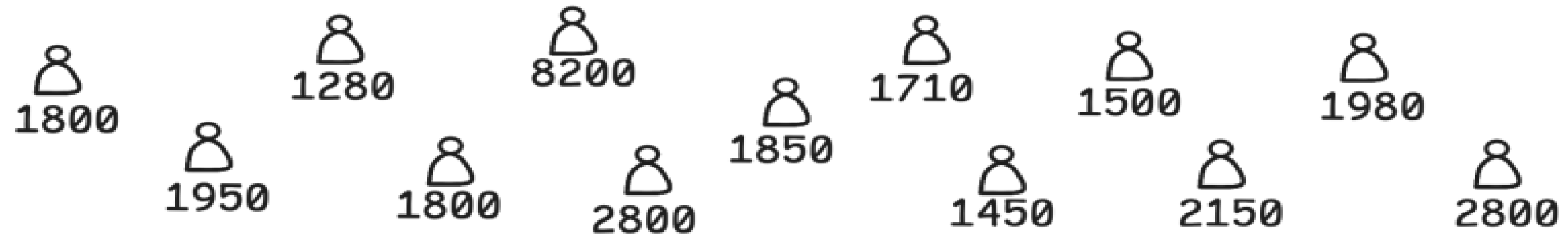
Ces mesures de position permettent de diviser un ensemble de données ordonnées en intervalles contenant un nombre égal d'observations.

- Quartiles : divisent les données en 4 parts égales.
 - Q1 (1er quartile) : 25% des données sont inférieures ou égales.
 - Q2 (2e quartile) : c'est la médiane (50%).
 - Q3 (3e quartile) : 75% des données sont inférieures ou égales.
- Déciles : divisent les données en 10 parties égales.
 - D1 : 10% des données sont inférieures ou égales.
 - D9 : 90% des données sont inférieures ou égales.
- Percentiles : divisent les données en 100 parts égales.
 - Le k-ième percentile correspond à la valeur sous laquelle se trouvent k% des données.
 - Exemple : le 90e percentile est la valeur sous laquelle se trouvent 90% des observations.

Ces indicateurs sont très utilisés pour analyser la position relative d'une donnée dans une distribution

Indicateur de position

Soit l'échantillon suivant, calculez les quartiles :



1280	1450	1500	1710	1800	1800	1850	1950	1980	2150	2800	2800	8200
------	------	------	------	------	------	------	------	------	------	------	------	------

$$Q1 = 1500 + 1710 / 2$$

$$Q2$$

$$Q3 = 2150 + 2800 / 2$$

Indicateur de dispersion

Les indicateurs de dispersion permettent de mesurer la variabilité ou l'étalement des données autour d'un indicateur de position (comme la moyenne ou la médiane). Ils sont indispensables pour évaluer la fiabilité des valeurs centrales et détecter la présence d'hétérogénéité dans un jeu de données. Les principaux indicateurs sont :

- **L'étendue** : différence entre la plus grande et la plus petite valeur d'un ensemble de données.

C'est une mesure simple mais sensible aux valeurs extrêmes.

- **La variance** : mesure la dispersion moyenne des données par rapport à la moyenne.
- **L'écart-type** : racine carrée de la variance, il permet de quantifier l'écart moyen des données à la moyenne, tout en restant dans la même unité que les données initiales.
- **L'écart interquartile (IQR)** : défini comme la différence entre le troisième et le premier quartile ($El = Q3 - Q1$), il mesure la dispersion centrale en étant insensible aux valeurs extrêmes.

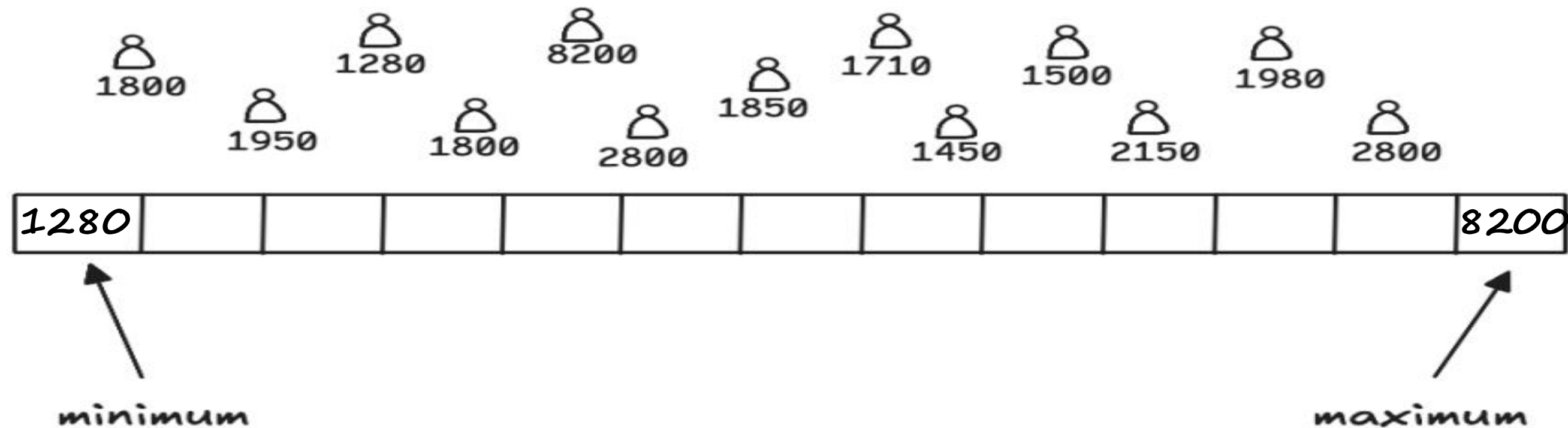
Indicateur de dispersion

Étendue :

L'étendue est la mesure de dispersion la plus simple. Elle correspond à la différence entre la plus grande et la plus petite valeur d'un ensemble de données.

$$\text{Etendue} = \max - \min$$

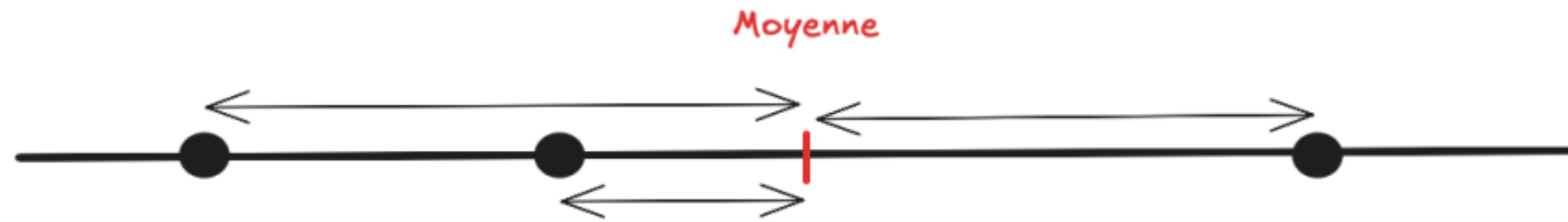
Cette mesure est très sensible aux valeurs extrêmes (outliers), ce qui peut limiter son utilité dans certaines analyses. Calculez l'étendue de l'échantillon suivant, en triant préalablement les données.



Indicateur de dispersion

Variance :

La variance mesure la dispersion des données autour de la moyenne. Elle est définie comme la *moyenne* des carrés des écarts à la moyenne.



$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Indicateur de dispersion

Écart-type (std) :

L'écart-type est la racine carrée de la variance. Il mesure la dispersion dans les mêmes unités que les données (contrairement à la variance qui est en unités au carré).

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Indicateur de dispersion

Pourquoi calculer la variance ?

Voici les données des deux classes :

Classe A :

- Notes des élèves : 11, 13, 12, 14, 10, 12, 13, 12, 11
- La moyenne = 12
- $s^2 = 1.33$
- $s = 1,15$

Classe B :

- Notes des élèves : 5, 5, 6, 8, 10, 14, 16, 18, 20
- La moyenne = 12
- $s^2 = 30$
- $s = 5,47$

- Les notes des élèves sont assez proches les unes des autres.
- L'écart-type est faible (1,15), ce qui confirme que les performances sont homogènes : la majorité des élèves a des notes proches de la moyenne.

- Les notes sont plus dispersées, avec des écarts importants entre les élèves.

Indicateur de dispersion

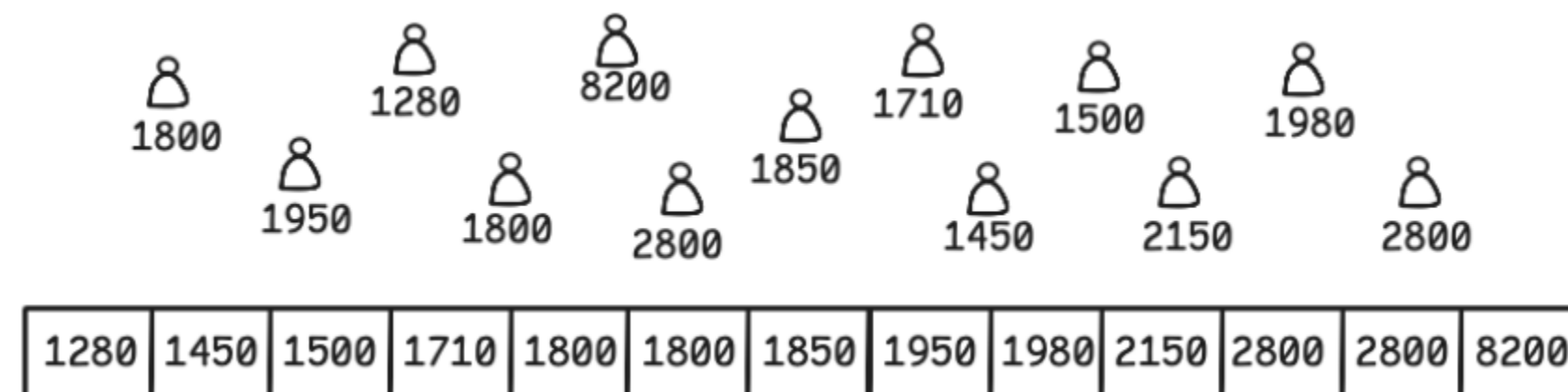
IQR (Intervalle interquartile) :

L'IQR (Interquartile Range) est une mesure de dispersion qui représente l'écart entre le troisième quartile (Q_3) et le premier quartile (Q_1) d'un ensemble de données. Il mesure la dispersion des 50% des données centrales, ce qui le rend robuste aux valeurs extrêmes.

$$IQR = Q_3 - Q_1$$

- Q_1 (premier quartile) : La valeur telle que 25% des données sont inférieures ou égales à Q_1 .
- Q_3 (troisième quartile) : La valeur telle que 75% des données sont inférieures ou égales à Q_3 .

Soit l'échantillon suivant, calculez l'IQR.



Distributions

En plus de pouvoir caractériser les échantillons par leurs indicateurs de position et de dispersion, il est courant de visualiser comment les données sont réparties dans l'espace, c'est ce qu'on appelle observer leur distribution.

Box Plots :

Le box plot (ou boîte à moustaches) est une représentation graphique synthétique des données quantitatives. Il permet de visualiser :

- la médiane (Q_2),
- le premier (Q_1) et le troisième quartile (Q_3),
- l'intervalle interquartile (IQR),
- les valeurs minimales et maximales dans une plage définie,
- les valeurs extrêmes (outliers), si représentées.

Distributions

Box Plots :

Le box plot se compose :

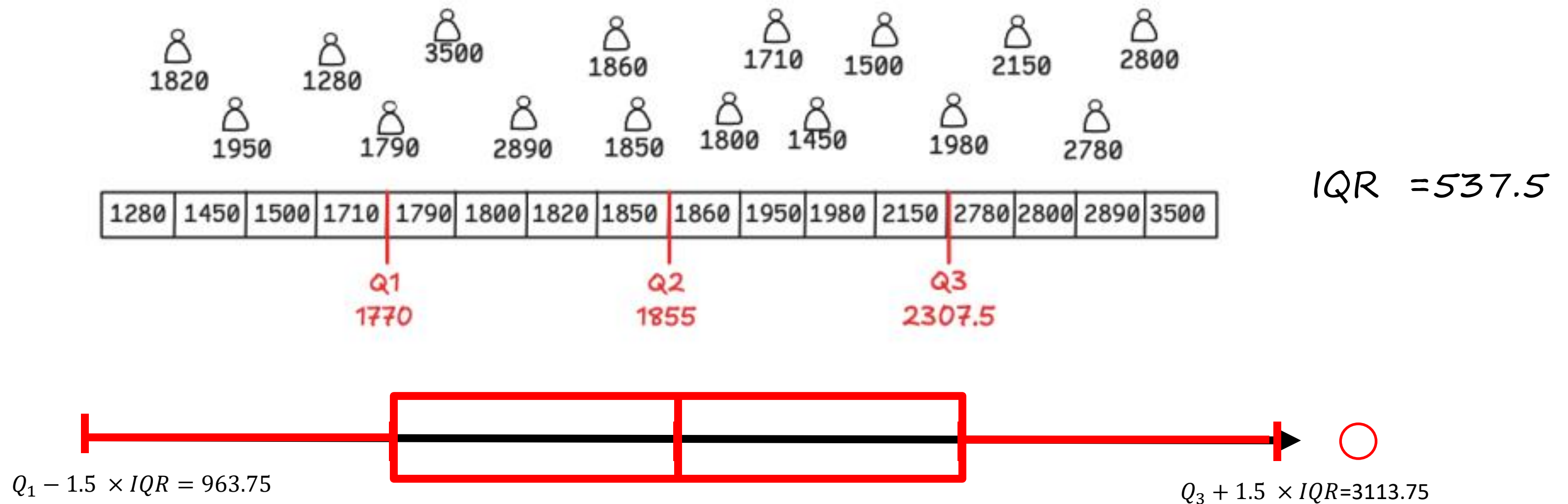
- d'un rectangle allant de Q_1 à Q_3 ,
- d'un trait vertical à l'intérieur de la boîte marquant la médiane,
- de segments (les « moustaches ») s'étendant généralement jusqu'aux valeurs minimales et maximales non considérées comme outliers,
- de points isolés pour les valeurs aberrantes souvent définies comme en dehors de:

$$Q_1 - 1.5 \times IQR; Q_3 + 1.5 \times IQR$$

Distributions

Box Plots :

Le box plot est particulièrement utile pour comparer plusieurs distributions côte à côte et visualiser rapidement leur symétrie, leur étalement, et la présence de valeurs extrêmes. Soit la population suivante, calculez les indicateurs nécessaires pour construire ensuite la box plot.



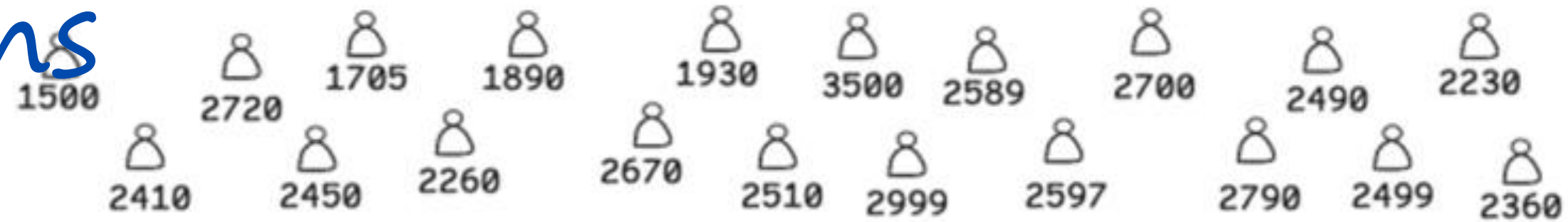
Distributions

Histogrammes:

L'histogramme est une représentation graphique de la distribution d'une variable quantitative continue ou discrète.

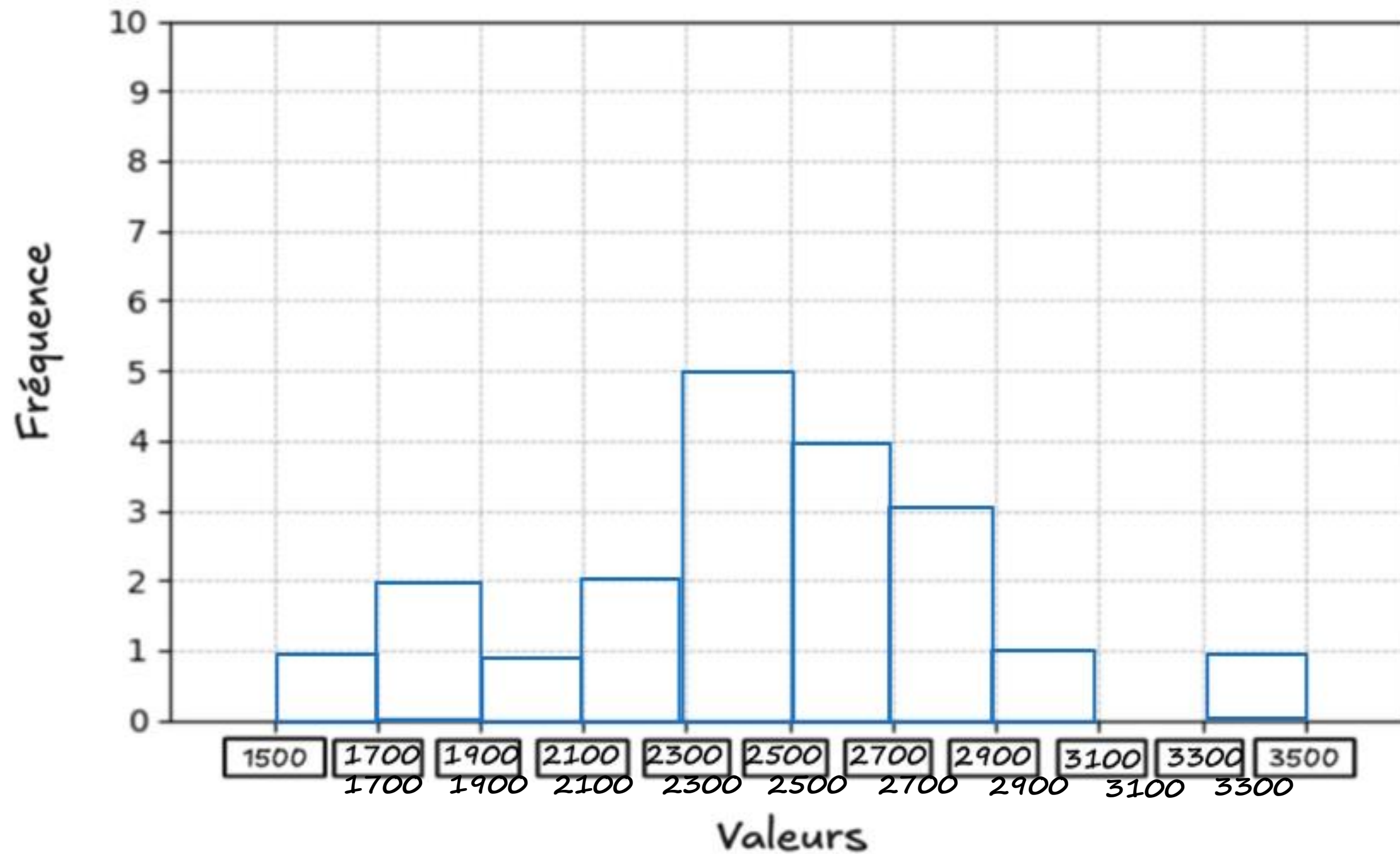
- L'axe horizontal (abscisses) représente les classes (intervalles de valeurs).
- L'axe vertical (ordonnées) représente l'effectif ou la fréquence de chaque classe. Les barres sont adjacentes (collées) pour indiquer la continuité des données (contrairement aux diagrammes en bâtons pour les données discrètes ou qualitatives).
- La hauteur de chaque barre est proportionnelle à la fréquence (ou à l'effectif).

Distributions



1500	1705	1890	1930	2230	2260	2360	2410	2450	2490	2499	2510	2589	2597	2670	2700	2720	2790	2999	3500
------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

Soit la population suivante, découpez les valeurs en 10 intervalles, puis tracez l'histogramme



Univers – Phénomènes – Modélisation

Notre univers est rempli de phénomènes aléatoires, c'est à dire dont les issues ne peuvent pas être prédites avec une parfaite certitude.

Lancé de dé



durée de vie d'une étoile

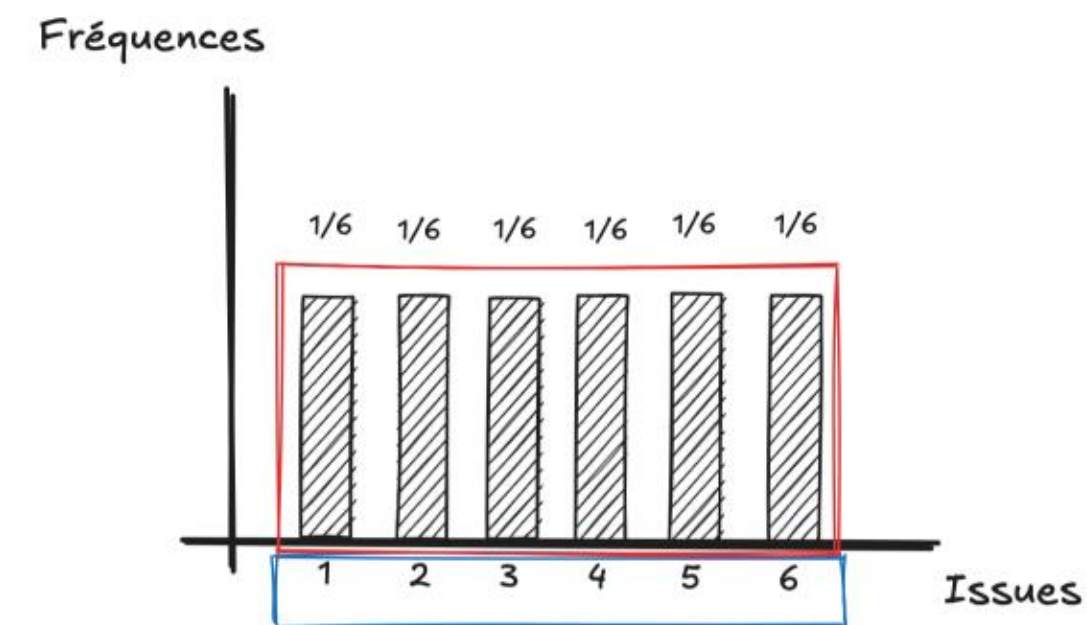
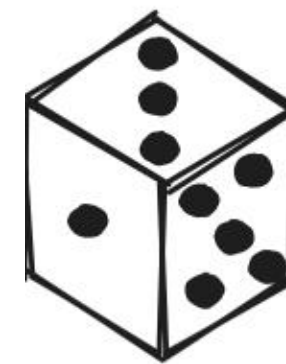
L'heure à laquelle vous allez vous réveiller demain

Le nombre de voitures qui circulent chaque jour dans votre rue

La température de votre café

En mathématique, on modélise ces phénomènes avec des variables aléatoires (V.A). Une V.A est une application qui se caractérise par deux choses :

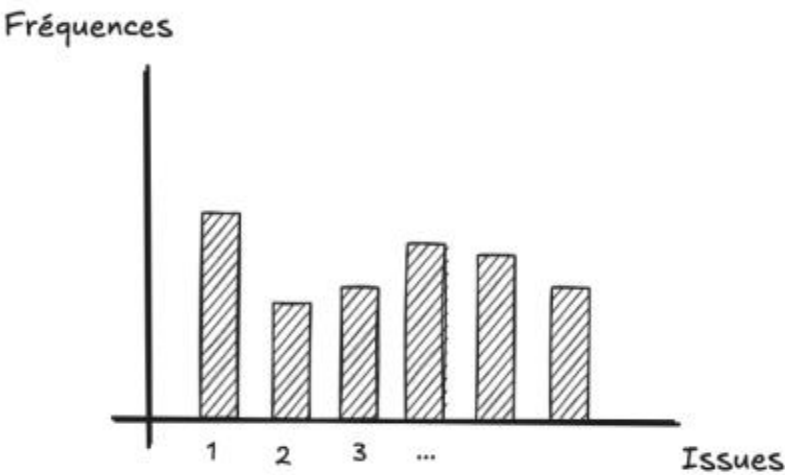
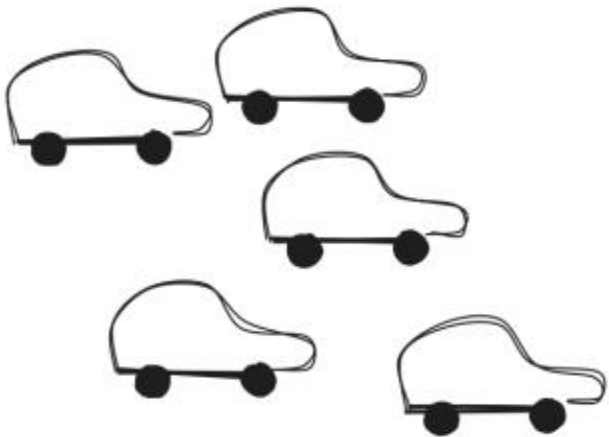
- **Un domaine** : L'ensemble des résultats possibles qui peuvent émerger d'un phénomène
- **Une loi de probabilité** : La fréquence d'apparition de chacun de ces résultats



Univers – Phénomènes – Modélisation

Dans la pratique, on ne connaît que très rarement le véritable domaine et la Loi de Probabilité qui caractérisent les phénomènes qui nous entourent.

Quelle est la Variable Aléatoire qui décrit le nombre de voitures qui circulent chaque heure dans votre rue ?



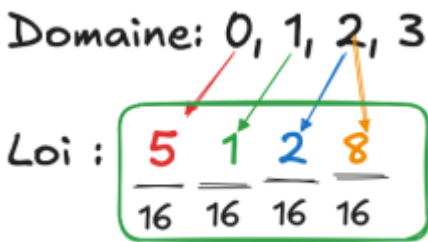
En réalité, les phénomènes qui nous entourent sont semblables à des boites noires, dont on ignore le fonctionnement réel. Tout ce qu'on peut faire, c'est observer un échantillon de ce que ces boites peuvent produire Et tenter d'en tirer des conclusions.

Prenons l'exemple de la boîte magique, qui génère un nombre aléatoire toutes les secondes.



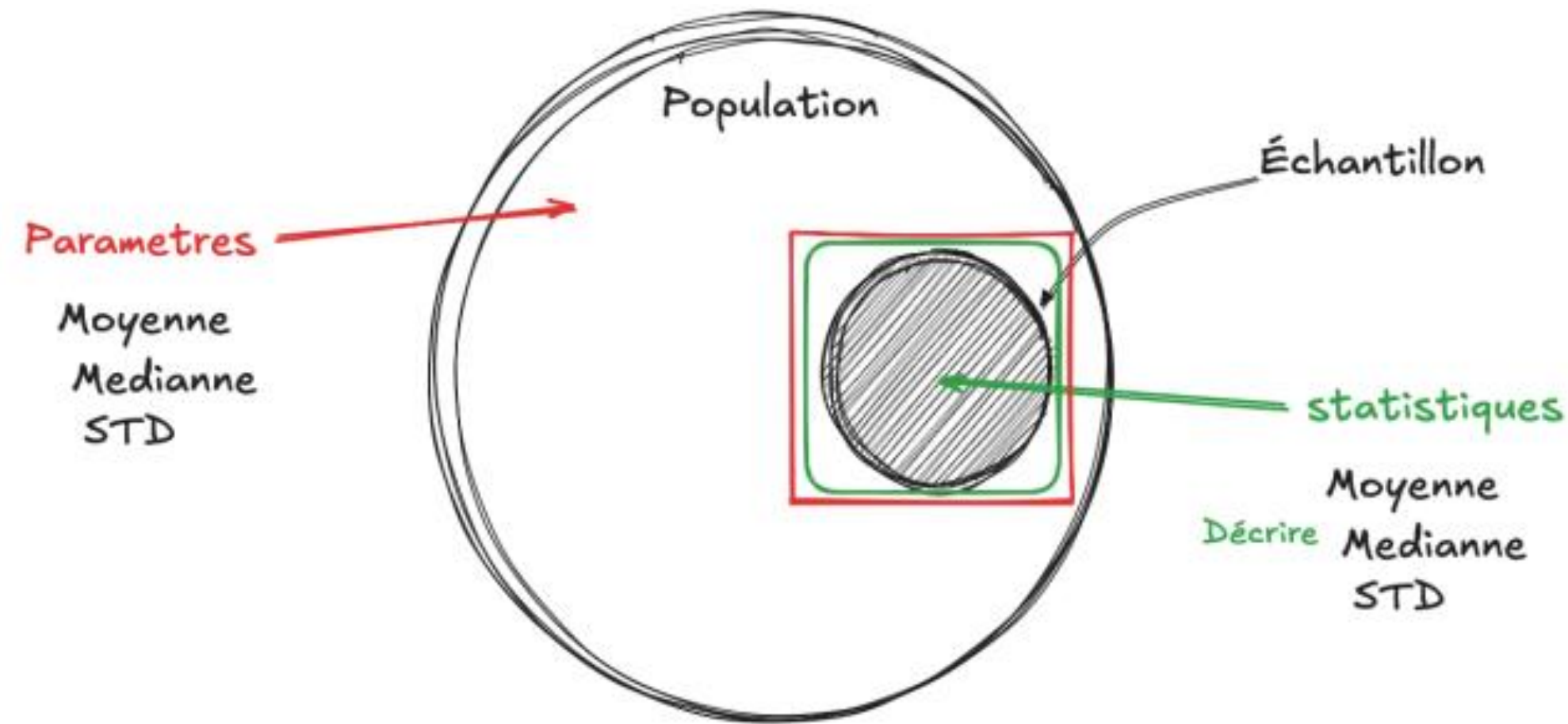
3, 0, 0, 3, 0, 3, 2, 3, 3, 1, 0, 2, 0, 3, 3, 3,

Domaine : ?
Loi : ?



Univers – Phénomènes – Modélisation

Mais ceci restera toujours une estimation. C'est notre théorie, hypothèse. C'est en fait le but même des Statistiques de chercher à approximer ce que peuvent être ces V.A.



Data Science & Machine Learning

- ✓ On travaille toujours avec un échantillon (une petite partie des données).
- ✓ La population complète est trop vaste (parfois infinie) pour être observée.
- ✓ Exemple : ici, on n'a vu que 16 résultats, alors qu'il en existe une infinité possibles.

Univers – Phénomènes – Modélisation

Rôle du Data Scientist (avant le ML)

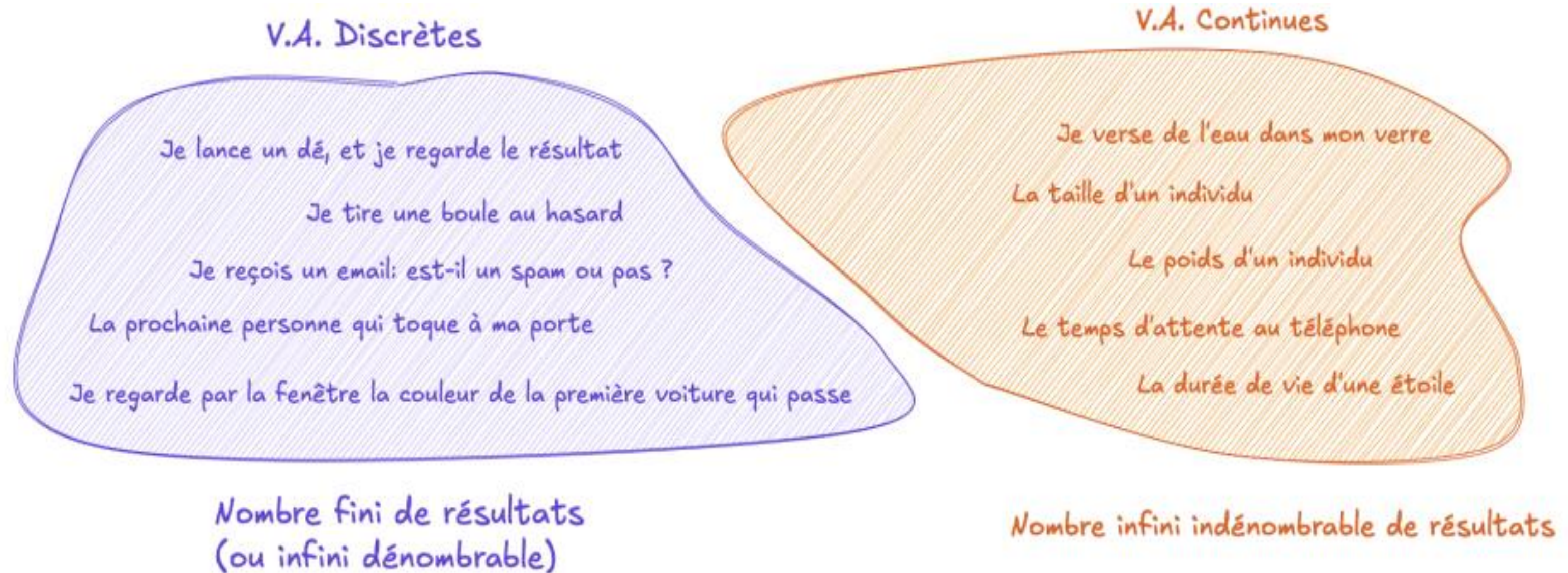
Votre première mission (avant de faire du ML), est en fait d'être capable d'estimer du mieux possible ces 3 choses :

- ✓ Le domaine d'une V.A
- ✓ Sa Loi de Probabilité
- ✓ La confiance que vous avez dans votre estimation !

Pour cela, vous devez développer votre capacité à :

- ✓ Formuler des hypothèses (les bonnes)
- ✓ Les tester rigoureusement
- ✓ Conclure avec confiance pour apporter de la valeur et continuer d'avancer

Variables Continues et Variables Discrètes



En Data Science on considère qu'il existe uniquement deux grands types de Variables Aléatoires :

- ✓ Celles dont le domaine contient un nombre fini ou infini dénombrable de résultats : *discrètes*
- ✓ Celles dont le domaine contient un nombre infini indénombrable de résultats : *continues*

Variables Continues et Variables Discrètes

Variable discrète vs continue

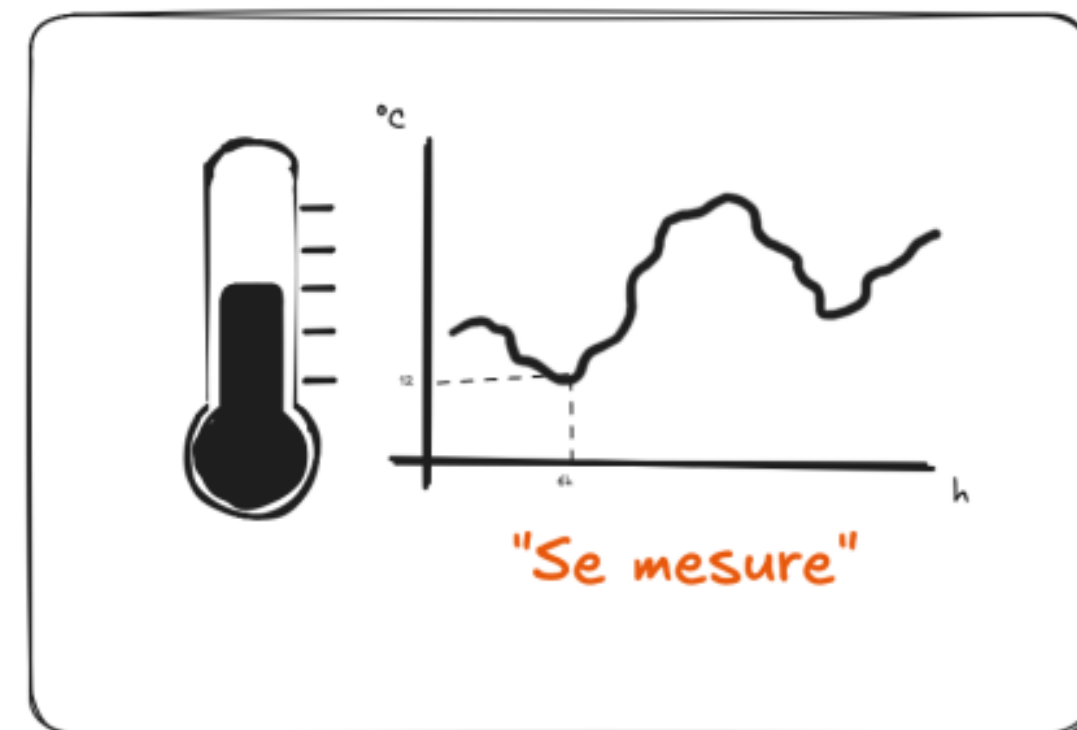
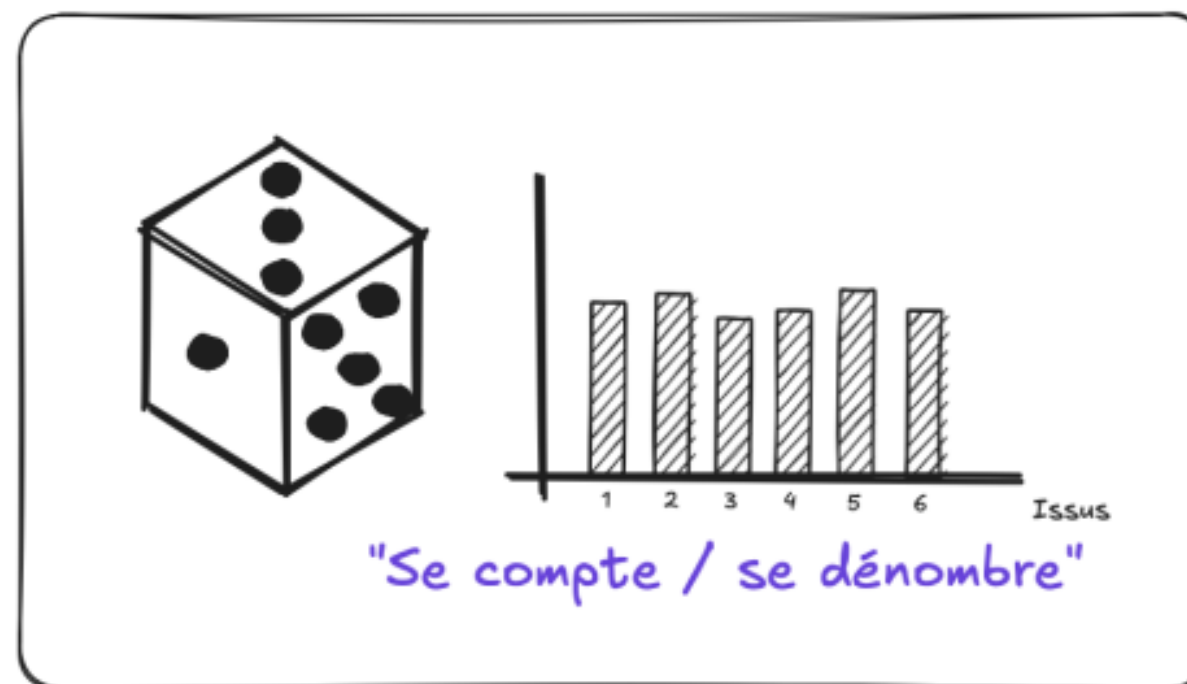
❖ Variable discrète

- ✓ Se compte ou se dénombre.
- ✓ Exemple : nombre de lancers de dé, nombre d'élèves dans une classe.

❖ Variable continue

- ✓ Se mesure.
- ✓ Exemple : quantité d'eau dans un verre, taille d'un individu.

➔ En résumé : le discret se compte, le continu se mesure.

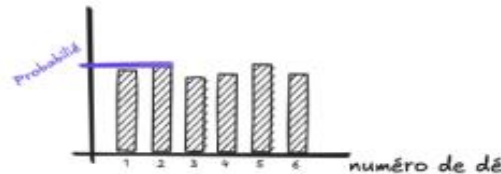
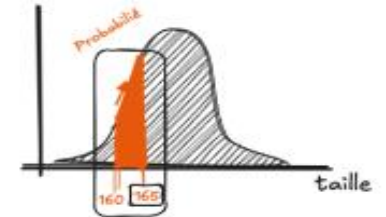


Variables Continues et Variables Discrètes

On peut ainsi dresser un tableau qui compare le domaine et la loi de probabilité d'une variable aléatoire en fonction du type de la variable, entre continue et discret.

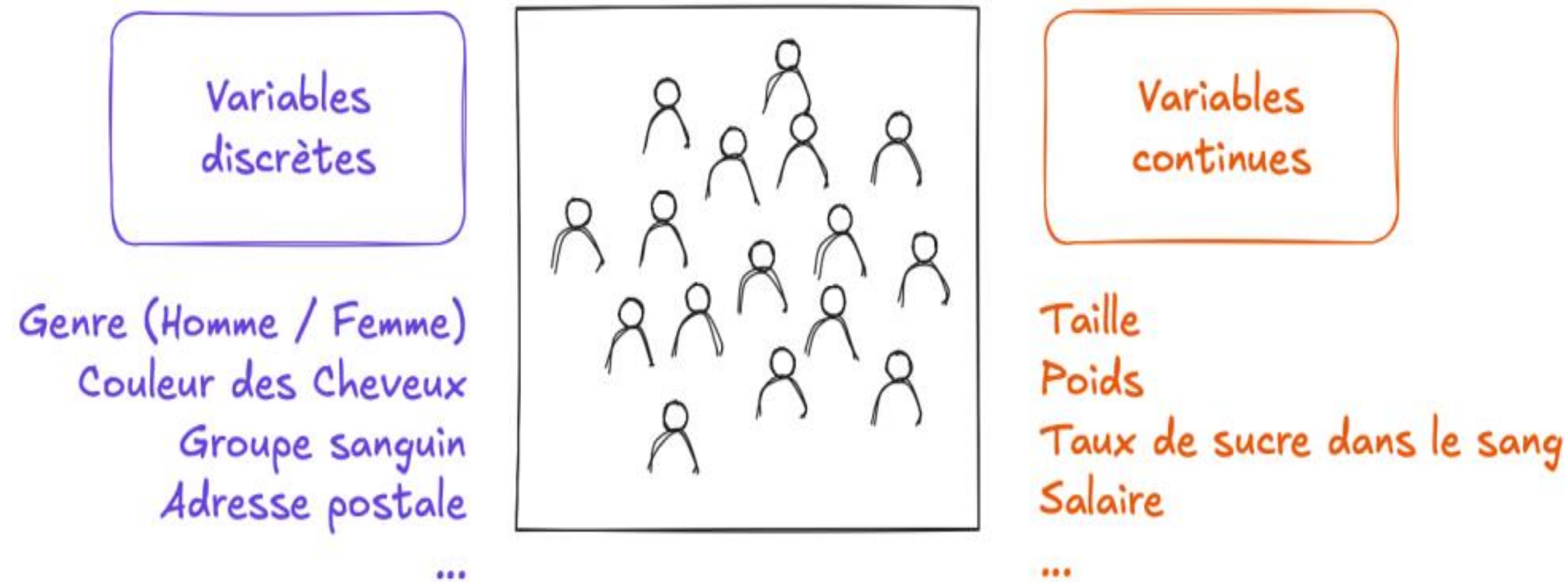
Domaine et loi de probabilité

- ✓ Pour une variable aléatoire discrète, le domaine est constitué d'une liste de valeurs, finie ou infinie dénombrable. La loi de probabilité se représente par une fonction de masse, ce qui permet de connaître directement la probabilité de chaque valeur.
- ✓ Pour une variable aléatoire continue, le domaine est un intervalle de valeurs, infini et indénombrable. La probabilité d'obtenir une valeur exacte est quasiment nulle, donc on utilise une fonction de densité. On calcule alors la probabilité sur un intervalle en mesurant l'aire sous la courbe. Par exemple, pour une personne mesurant 1m65, on calcule la probabilité qu'elle mesure entre 1m64 et 1m66.

	V.A. Discrète	V.A. Continue
Domaine	fini ou infini dénombrable liste de valeurs [1, 2, 3, 4]	infini indénombrable intervalle de valeurs [0, 100]
Loi de Probabilité	Fonction de Masse 	Fonction de Densité 

L'Analyse Technique

- ❖ Faire une analyse de données doit rester simple. Si cela n'est pas le cas, c'est que ce n'est pas la bonne méthode.
- ✓ En pratique, il y a 5 points essentiels à analyser.
- ✓ Exemple : un groupe d'individus décrit par des variables discrètes et continues.



La première question que nous allons nous poser est : Que pouvons-nous faire avec ces 2 types de variables ?

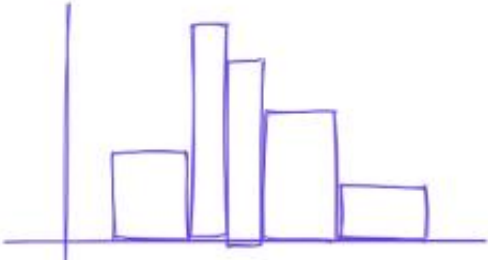


L'Analyse Technique

- ❖ On commence par l'étude des variables une par une, indépendamment. C'est par là qu'il faut commencer l'analyse. Cela constitue l'analyse univariée.
 - ✓ Analyse **discrète** : analyser les variables **discrètes** des données, comme par exemple le nombre d'individu avec les cheveux blanc, blond, etc
 - ✓ Analyse **continue** : analyser les variables **continues** des données, comme par exemple calculer la taille moyenne des individus, ou encore l'individu le plus grand
- ❖ Ensuite on peut s'intéresser aux relations entre nos variables. Ce qui constitue l'analyse multivariée.
 - ✓ Analyse **discrète-discrète** : on analyse la relation entre deux variables discrètes, par exemple analyser le genre d'une personne par rapport à sa couleur de cheveux
 - ✓ Analyse **discrète-continue** : on analyse la relation entre une variable discrète et une variable continue (et inversement), par exemple la relation entre la taille et le genre des individus, et savoir si les femmes sont plus grandes que les hommes, etc
 - ✓ Analyse **continue-continue** : on analyse la relation entre deux variables continues, par exemple la relation entre la taille et le poids d'un individus.

L'Analyse Univariée

❖ L'analyse univariée est la première étape de toute analyse de données. Elle consiste à :

- ✓ Analyser les variables une par une
- ✓ Identifier leurs statistiques principales
- ✓ Représenter les résultats avec des graphiques adaptés

	Variable discrète	Variable continue
Statistiques	Calcul des effectifs <code>value_counts()</code>	<ul style="list-style-type: none">- Moyenne- Médiane- Variance- Écart-type- Minimum- Maximum- Q1, Q3 <code>describe()</code>
Graphiques	Bar chart  <code>value_counts().plot(kind="bar")</code>	Histogramme  Boîte à moustaches 

L'Analyse Univariée

Analyser une variable avec Python

✓ Variable discrète

- On utilise `value_counts()` de pandas
- Permet de calculer les effectifs (nb d'individus par classe)
- Exemple : pour la variable `genre` → nb de Femmes et nb d'Hommes
- Visualisation : Bar Chart pour comparer les catégories

✓ Variable continue

- On utilise `describe()` de pandas
- Donne les statistiques principales : moyenne, médiane, variance, Q1, Q3...
- Visualisation :
 - Histogramme → forme de la distribution
 - Boîte à moustaches → dispersion et valeurs extrêmes

L'Analyse Univariée

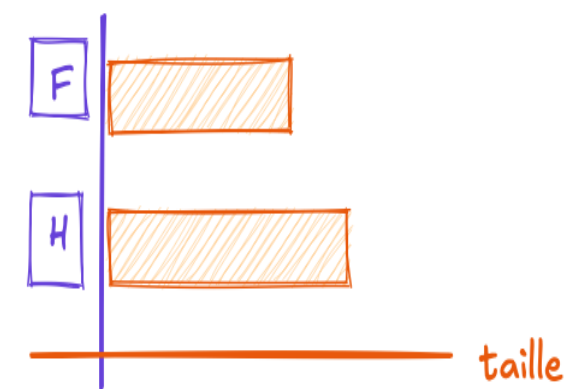
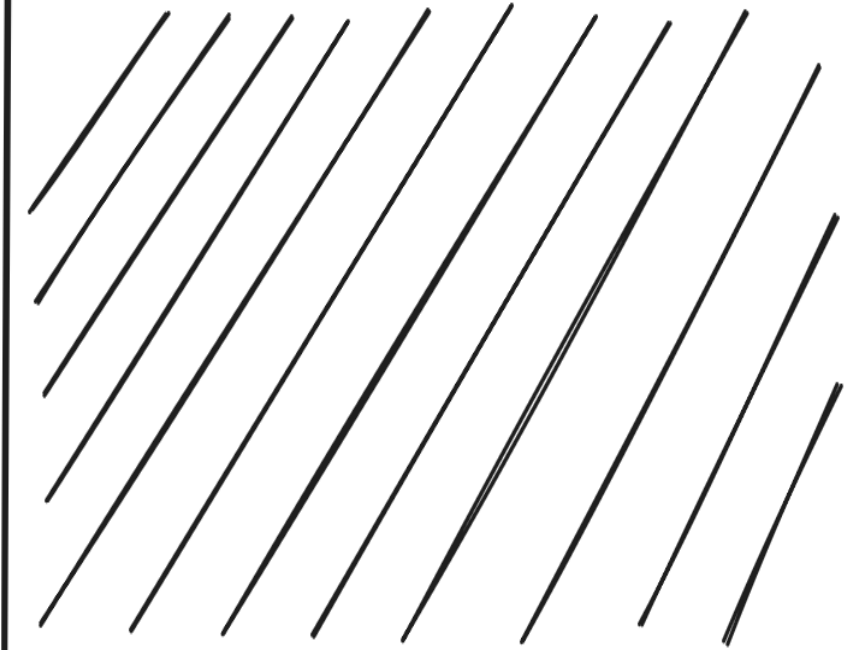
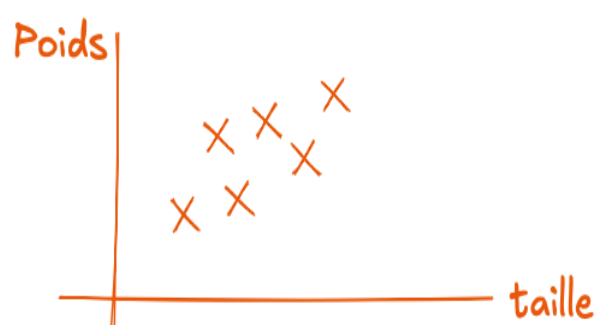
Outils python

- [La documentation pandas de value counts\(\)](#)
- [La documentation seaborn de countplot\(\)](#)
- [Tous les graphiques possibles avec Pandas](#)
- [Générer un Histogramme avec Seaborn](#)
- [Générer une BoxPlot avec Pandas](#)
- [Générer une BoxPlot avec Seaborn \(Catplot\)](#)

L'Analyse Multivariée

Analyse multivariée

- Étudie les relations entre les variables
 - Trois cas principaux :
 - ✓ Deux variables **continues**
 - ✓ Deux variables **discrètes**
 - ✓ Une variable **discrète** et une variable **continue**
- ➔ **Objectif** : comprendre comment les variables interagissent entre elles.

	Variable discrète	Variable continue												
Variable discrète	<table><tr><td></td><td>Brun</td><td>Blond</td><td>Noir</td></tr><tr><td>H</td><td>35</td><td>1</td><td>15</td></tr><tr><td>F</td><td>19</td><td>9</td><td>14</td></tr></table> <p>Table de contingence</p>		Brun	Blond	Noir	H	35	1	15	F	19	9	14	 <p><code>groupby(Variable discrète)[Variable continue]</code></p>
	Brun	Blond	Noir											
H	35	1	15											
F	19	9	14											
Variable continue		 <p>Scatter Plot</p>												

L'Analyse Multivariée

Outils python

- [Table de contingence avec pandas](#)
- [La documentation pandas de Groupby\(\)](#)
- [La documentation seaborn de Catplot\(\)](#)

Exercices

Exercice 1 :

Un élève a obtenu les notes suivantes : 4;6;3;9;10;8;12;10;19;12;20;12;18 . Calculer sa moyenne et son écart-type.

Exercice 2 :

La température est relevée chaque heure pendant 4 jours dans une forêt. Les 97 résultats obtenus ont été triés et sont rassemblés dans le tableau suivant :

Température	12	14,5	15	15,5	16	16,5	17	17,5	18	18,5	19	19,5	22
Nombre de fois où cette température a été relevée	2	3	7	10	12	15	10	10	9	7	7	3	2

Déterminer la mediane M, les quartiles Q1 et Q3 de celle serie statistique et dessiner le graphe de boîte moustache

Exercice 3 :

Le tableau suivant donne les températures moyennes par mois à Paris et à Pékin en degrés Celsius.

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Pekin	-5	-4	4	15	27	31	31	30	26	20	10	-5
Paris	3	4	7	10	14	17	19	18	16	17	7	6

- 1) Calculer la moyenne, l'étendue, la variance et l'écart-type des températures mensuelles pour chacune de ces villes.
- 2) Comparer et analyser les résultats obtenus.

Bibiographie

Formation ML-PRO par Guillaume saint-cirgue

