Données et statistiques

M162 — Analyser et modéliser des données

Jérôme Frossard

EPAI

3 septembre 2023

Qu’est-ce que la statistique ?

Dans le cadre de ce cours, on peut définir la statistique comme l’ensemble de techniques permettant de décrire numériquement et graphiquement des ensembles d’éléments. La statistique est donc un moyen de produire de l’information à partir des données.

Les éléments de ces ensembles sont des personnes ou des choses, concrètes ou abstraites, qui peuvent être **distinctement identifiées**

En statistique, un ensemble est appelé une **population** et un élément, un **individu**, même lorsque ce sont des choses. Par exemple :

- Population : Les candidat·e·s au CFC, les contrats d’apprentissage
- Individu : la candidate au CFC no 345683, le contrat d’apprentissage no 425678

Un **échantillon** est un sous-ensemble représentatif que l’on construit lorsque la population est trop grande pour être observée en entier.

Variable statistique (I/III)

Pour décrire les éléments d'un ensemble, il faut d'une certaine manière « mesurer » ces éléments.

Il est impossible de mesurer un élément dans sa totalité, on doit mesurer séparément ses différentes **caractéristiques** (propriétés ou attributs).

Il est impossible de mesurer toutes les propriétés d'un élément, on ne mesure que les propriétés que l'on **juge pertinentes** pour **ce que l'on veut en faire**.

En statistique, une propriété que l'on mesure est une **variable statistique**. C'est une variable, car sa valeur peut varier d'un individu à l'autre.

Chaque valeur mesurée est une **observation** (une données brute).

Variable statistique (II/III)

Une variable peut être :

- Quantitative ou qualitative (catégorique) ou
- Discrète ou continue

Variable quantitative : Variable dont la valeur est un nombre qui représente une quantité. Elle peut être discrète ou continue.

Variable qualitative (catégorique) : Variable dont les valeurs représentent une catégorie, mais pas une quantité. Elle est toujours discrète.

Variable discrète : Variable dont le domaine de valeurs est fini ou dénombrable (chaque élément peut être mis en correspondance avec un nombre entier unique).

Variable continue : Variable dont les valeurs sont des nombres réels. Elle est toujours quantitative.

Variable statistique (III/III)

Exemples de variables qualitatives (catégoriques) :

- Couleur des yeux : bleu, vert, marron, etc.
- Type de véhicule : voiture, moto, camion, etc.
- Nombre d'employés : 1-10, 11-50, 51-200 ..., 1001-10000, plus de 10'000

Exemples de variables quantitatives discrètes :

- Nombre de pages d'un livre
- Nombre de vue d'un article

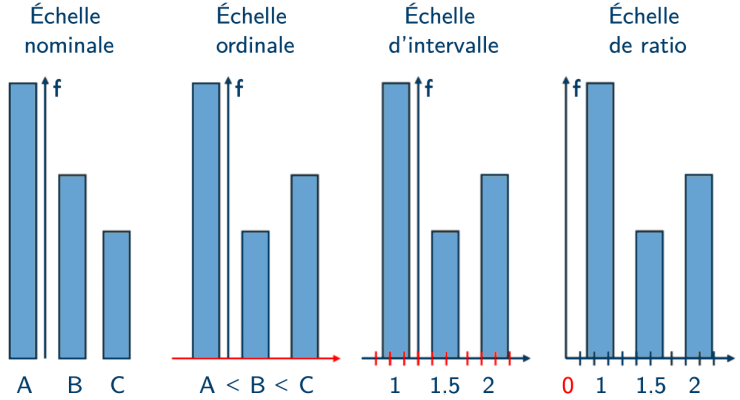
Exemples de variables quantitatives continues :

- Poids
- Température

Données et échelles de mesure

Mesurer une variable consiste à trouver, sur une **échelle de mesure**, la valeur qui correspond le mieux à celle de la variable.

On distingue quatre niveaux d'échelles de mesure :



Échelle nominale

Une échelle nominale se construit en répartissant les observations en plusieurs catégories (ou classes). Par exemple :

- Couleurs des briques LEGO : rouge, jaune, noir, blanc
- Numéro postal : un nombre de quatre chiffres

Opérations possibles :

- Dénombrement : compter les éléments d'une classe
- Fréquences (f) : nombre d'éléments d'une classe / nombre total d'éléments
- Comparaison : équivalent (=) ou différent ($\neq$)
- Mesure de tendance centrale : mode (classe la plus représentée)

Le plus souvent, les catégories d'une échelle nominale sont codées par des noms. Elles peuvent aussi être codées par des nombres, mais même dans ce cas, **toutes les opérations arithmétiques sont interdites**.

Échelle ordinale

Comme pour l'échelle nominale, les observations sont réparties dans différentes catégories (ou classes), mais il existe une relation d'ordre entre les classes. Par exemple :

- Grades d'un militaire : général > … > sergent > … > soldat
- Classe d'un hôtel : 4 étoiles > 3 étoiles > 2 étoiles > 1 étoile

Opérations possibles :

- Toutes celles des échelles nominales
- Comparaison : plus grand que (>) et plus petit que (<)
- Mesure de tendance centrale :
  médiane (valeur qui coupe l'ensemble en deux parties égales)

Là encore, même lorsque les catégories sont codées par des nombres, **toutes les opérations arithmétiques sont interdites**.

Échelle d'intervalle (ou relative)

Une échelle d'intervalle est une échelle ordinale numérique pour laquelle la distance entre les valeurs successives est toujours la même. Par exemple :

- Température en °C : 0 et 100 sont, respectivement, la température de fusion et d'ébullition de l'eau.
- Quotient intellectuel : la valeur moyenne de 100 est arbitraire.

Opérations possibles :

- Toutes celles des échelles ordinales
- Opérations arithmétiques :
  - différence entre deux valeurs (—),
  - addition d'une valeur et d'une différence (+)
- Mesure de tendance centrale : moyenne arithmétique

Échelle de ratio (ou absolue)

Une échelle de ratio (ou absolue) est une échelle d'intervalle sur laquelle le zéro représente l'absence de la quantité mesurée. Par exemple :

- Taille en mètre : 0 m représente une longueur nulle
- Masse en kg : 0 kg représente une masse nulle

Opérations possibles :

- Toutes celles de l'échelle d'intervalle
- Opérations arithmétiques :
  - addition ou soustraction de deux valeurs (+/—),
  - produit d'une valeur et d' un nombre sans dimension (×)
  - rapport de deux valeurs (÷)
- Mesure de tendance centrale : moyenne géométrique

Mesure de tendance centrale

Une mesure de tendance centrale est une mesure statistique qui caractérise les éléments d'un ensemble. Il en existe plusieurs :

- Mode : classe la plus représentée dans l'ensemble
- Médiane : valeur qui divise l'ensemble en deux parties égales
- Moyenne arithmétique (la plus connue)