Statistique Descriptive : Concepts Fondamentaux

Komlan Doigté KATAKOU

March 25, 2025

Un peu de vocabulaire

Qu'est-ce que la statistique ?

La statistique est une discipline qui étudie les phénomènes à travers la collecte, le traitement, l'analyse et l'interprétation des données, afin d'en extraire de l' information pertinente et de faciliter la prise de décision. Elle repose sur un ensemble de méthodes mathématiques permettant de modéliser des données et de les rendre compréhensibles.

Statistique et Data Science

A la base une branche des mathématiques appliquées, la statistique fait aujourd'hui partie intégrante de ce que l'on appelle la science des données (**Data Science**) où elle est utilisée pour explorer, décrire et prédire divers phénomènes à partir de données massives.

Population vs Echantillon

Population

On appelle **population** l'ensemble des individus sur lesquels porte une étude statistique. Par exemple, si l'on se propose d'étudier la taille moyenne des étudiants d'une université, l'ensemble des étudiants de cette université constitue la population.

Echantillon

Lorsque la taille de la population totale est très grande, on se restreint à un **échantillon** qui est un sous-ensemble (représentatif, de préférence) de la population. Par exemple, si l'on souhaite connaître les intentions de vote à l'approche d'une élection et qu'il serait assez coûteux d'interroger toute la population, on se pouura se contenter d'un échantillon de 3000 électeurs sélectionnés aléatoirement dans différentes régions. En revanche, un échantillon de 3000 électeurs de la même région ou de la même catégorie professionnelle n'est pas représentatif de la population et risquerait de biaiser l'étude.

Qualitatif vs Quantitatif

Les variables sont les caractéristiques qui font l'objet de l'étude : `Taille`, `Age`, `Race`, `Poids`, etc. On distingue :

1. **Variables qualitatives ou catégorielles** : Ce sont des caractéristiques non chiffrées. Elles peuvent être de deux types :
   - **Nominales** : Pas d'ordre défini (ex: couleur des yeux, moyen de transport).
   - **Ordinales** : Ordre défini (ex: niveau d'éducation, niveau d'appréciation (Mauvais, Bon, Excellent)).
2. **Variables quantitatives (numériques)** : Elles peuvent être :
   - **Discrètes** : dans ce cas, l'ensemble des valeurs possibles est dénombrable. Ce sont le plus souvent des valeurs entières (ex: nombre d'enfants d'une famille, nombre d'accidents sur une route, ...).
   - **Continues** : ici, l'ensemble des valeurs possibles n'est pas dénombrable. La variable peut prendre n'importe quelle valeur d'un intervalle donné (ex: poids, taille, température, ...).

Méthodologie Statistique

Le travail du statisticien suit une démarche structurée comme suit :

1. **Collecte des données** : Recueillir des données fiables et pertinentes en lien avec la problématique d'intérêt. Elles peuvent provenir d'enquêtes, de mesures expérimentales ou de bases de données préexistantes.
2. **Organisation et présentation** des données (tableaux, graphiques)
3. **Analyse et interprétation** des données
4. **Modélisation statistique** et **Prise de décision**

Mesures de tendance centrale

On s'intéresse à présent à quelques indicateurs numériques d'une variable quantitative X observée sur N individus. On dispose donc des observations $X_i, i \in \{1, \ldots, N\}$.

Tendance centrale

Les **mesures de tendance centrale** permettent de résumer un ensemble de données en une valeur représentative. Elles aident à identifier le point autour duquel les données sont regroupées et sont essentielles pour comparer différentes distributions de manière simple .

Mesures de Tendance Centrale

1. **La moyenne** : C'est un critère de position qui résume l'information sur l'ordre de grandeur de la plupart des observations . Elle est égale au quotient par le nombre d'observations de la somme des données.

$$\bar{X_N} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

2. **La médiane** : C'est la valeur qui divise l'ensemble des données triées en deux parties de même taille. Par exemple, dire que 1500 € est le salaire médian des individus d'une population signifie que 50% des individus ont un salaire inférieur ou égal à 1500€ et 50% ont un salaire supérieur à 1500€.
3. **Le mode** d'une série statistique est la valeur la plus fréquente de la série.

Mesures de dispersion

Dispersion

Les **mesures de dispersion** permettent d'évaluer la variabilité des données autour d'une valeur centrale. Elles sont essentielles pour comprendre la répartition des données et éviter des conclusions biaisées.

Mesures de Dispersion

1. **Variance** (σ^2) : Elle mesure la dispersion des valeurs autour de la moyenne. Une variance élevée signifie que les valeurs sont très dispersées, tandis qu'une faible variance indique que les valeurs sont proches de la moyenne.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X_N})^2$$

2. **Écart-type** (σ) : est la racine carrée de la variance. Il s'exprime dans la même unité que les données et est souvent plus interprétable.

On distingue également d'autres indicateurs de dispersion comme **l'étendue, l'écart interquartile, le coefficient de variation,** etc.

Quantiles

Les **quantiles** sont des valeurs qui divisent l'ensemble des données en intervalles de même fréquence. Les plus utilisés sont les **quartiles** définis comme suit :

- Le 1^{er} quartile : c'est la quantité notée Q_1 telle que 25% des individus ont une valeur inférieure à Q_1 et 75% une valeur supérieure à Q_1.
- Le 2^{e} quartile : 50% de valeurs inférieures et 50% de valeurs supérieures à Q_2.
  Le voyez-vous ? Q_2 est encore la médiane !
- Le 3^{e} quartile : 75% de valeurs inférieures et 25% de valeurs supérieures à Q_3.

Déciles, Percentiles, ...

On définit de même les autres quantiles. Par exemple, les **déciles** sont tels que le **1^{er} décile** (D_1) divise les données en 10% de valeurs inférieures et 90% de valeurs supérieures à D_1.

À suivre ...

Curieux d'en savoir plus ?
Ce qui vous attend prochainement :

- **Représentations graphiques** : Histogramme, Boxplot, QQ-plot, ...
- **Statistique bivariée** : Corrélation, Nuage de points, ...
- **Formes de distributions** : Queues de distribution, Asymétrie, Aplatissement, ...
- **Illustrations en R et Python**

Rendez-vous très bientôt pour poursuivre ce que nous avons commencé !