Statistiques univariées

1. Introduction et vocabulaire de travail.

La statistique consiste en un ensemble de méthodes en vue de la collecte des informations de phénomènes[1]. Cette étude a pour objet de traduire les faits naturels ou sociaux avec des nombres, lorsque c'est le cas on parle d'une étude **quantitative**, ou avec des mots, lorsque c'est le cas on parle d'une étude **qualitative**. On s'intéressa dans ce cours aux études quantitatives.
Une étude porte sur une **population**, les éléments de la population sont appelés les **individus** et on étudie des **caractères** de chacun de ces individus. Dans ce chapitre sur les statistiques univariées, on se restreindra à l'étude **d'un seul caractère** d'une population. On ne cherchera pas à étudier la dépendance de deux caractères, cela sera fait en deuxième année.

Dans les études qualitatives, le caractère est donc mesuré par des nombres, qui constituent une variable statistique. Cette variable peut être :

- **Continue** dans le cas où les valeurs de la variable sont des réels d'un intervalle donné. Par exemple : si on étudie le temps de trajet d'un élève pour se rendre en prépa, la variable est un réel de l'intervalle $]0, +\infty[$.
- **Discrète**, lorsque les valeurs de la variable sont des nombres d'un ensemble fini de valeurs, ou bien lorsque les valeurs sont des entiers.
  Par exemple : si on étudie le nombre de fois où un individu a eu un accident de la route, la variable est un entier de $\mathbb{N}$.

Cette année on s'intéressa aux variables discrètes**.**

Exemple : Une étude statistique sur le revenu des Français en 2020 porte sur la population française et le caractère étudié est le revenu des individus (les Français).

La statistique est constituée de deux grands mouvements :

- La **statistique descriptive** qui consiste en la simple description des caractères d'une population d'individus

Exemple : Un statisticien salarié de Netflix (qu'on appelle dans le milieu *data scientist*) va récolter les données de tous les utilisateurs, par exemple il va collecter le temps passé chaque jour par chaque utilisateur à regarder des contenus de type séries. Ces données seront ensuite utilisées pour prendre des décisions.

- La **statistique inférentielle** qui consiste en la prédiction de caractères à partir d'un échantillon d'une population donnée

Exemple : En vue de sonder les intentions de vote des Français à la veille d'une élection, on ne va pas demander aux 49 millions de votants leur intention de vote, mais plutôt à un échantillon, par exemple de 1000 personnes, leur intention, et à partir de ces

[1] Les faits observés

résultats, inférer (c'est-à-dire, déduire, généraliser) les intentions de l'ensemble de la population.
Nous reviendrons en deuxième année sur la statistique inférentielle par le biais des probabilités. En effet choisir un échantillon nécessite un choix *au hasard*.

On appelle série statistique associée à un échantillon l'ensemble des valeurs des caractères des individus.

Par exemple l'étude de l'âge des professeurs agrégés de mathématiques peut être fait à l'aide de l'étude d'un échantillon de 10 individus choisis au hasard. On obtient la série statistique suivante :
28 ; 36 ; 53 ; 50 ; 42 ; 60 ; 28 ; 36 ; 50 ; 29
La population est l'ensemble des professeurs agrégés de mathématiques, les individus sont les professeurs, le caractère est l'âge (qui est quantitatif), on a donc une variable discrète.

2. Étude d'une variable quantitative discrète

On s'intéresse ici uniquement aux séries statistiques discrètes.

On appelle dans la suite Ω la population étudiée, c'est un ensemble dont les éléments sont les individus. On appelle **effectif** son cardinal (le nombre des individus).
L'étude porte sur un caractère quantitatif. On appelle ce caractère la variable statistique, qu'on note habituellement X. L'ensemble des valeurs qu'elle prend se note X(Ω).
Dans l'exemple ci-dessus X(Ω) = {l'ensemble des âges possibles}
Formellement X est une application qui associe à chaque individu un nombre. Elle joue un rôle de modélisation.

Les valeurs prises par X s'appellent les modalités de X. On les note X(Ω)=$\{x_1, \dots, x_p\}$, que l'on décrit généralement dans l'ordre croissant.

On présente généralement une série statistique sous forme de tableau :

Valeur	x_1	x_2	…	x_p
Effectif	n_1	n_2	…	n_p

On appelle **effectif de la modalité** x_i, le nombre entier n_i d'individus dont la modalité est x_i.
On appelle **fréquence** de x_i : $f_i = \frac{n_i}{n}$, qui est un réel de l'intervalle [0,1].
On appelle **fréquence cumulée croissante** associée à x_i, notée f_i^c la fréquence des individus ayant une modalité inférieure ou égale à x_i.

On a donc : $\sum_{i=1}^{p} n_i = n$ et $\sum_{i=1}^{p} f_i = 1$. On a aussi $f_1^c = f_1$ et $f_p^c = 1$.

Définissons quelques indicateurs pertinents :
Définition : Un **mode** d'une série statistique est une valeur ayant le plus grand effectif.

En général, il y en a plusieurs, notamment si chaque valeur a un effectif de 1, chaque valeur est un mode de la série.

Définition : La **médiane** d'une série d'effectif n correspond à :

- Si n est impair, la valeur du milieu (la $\frac{n+1}{2}$ème plus grande valeur de la série).
- Si n est pair, la moyenne des valeurs $\frac{n}{2}$ et $\frac{n}{2}+1$.

Ainsi, la moitié des valeurs de la série sont plus grandes ou égales à la médiane, et la moitié des valeurs sont inférieures ou égales à la médiane.
Par exemple, le salaire médian net mensuel en France en 2022 est de 1789€. Donc au moins la moitié des salariés Français perçoivent moins de 1789€ net par mois.

Définition : Soit $x = (x_i)_{1\leqslant i\leqslant \mathrm{n}}$, une série statistique. On définit la **moyenne** de la série par : $\bar{x} = \frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}} x_i$.

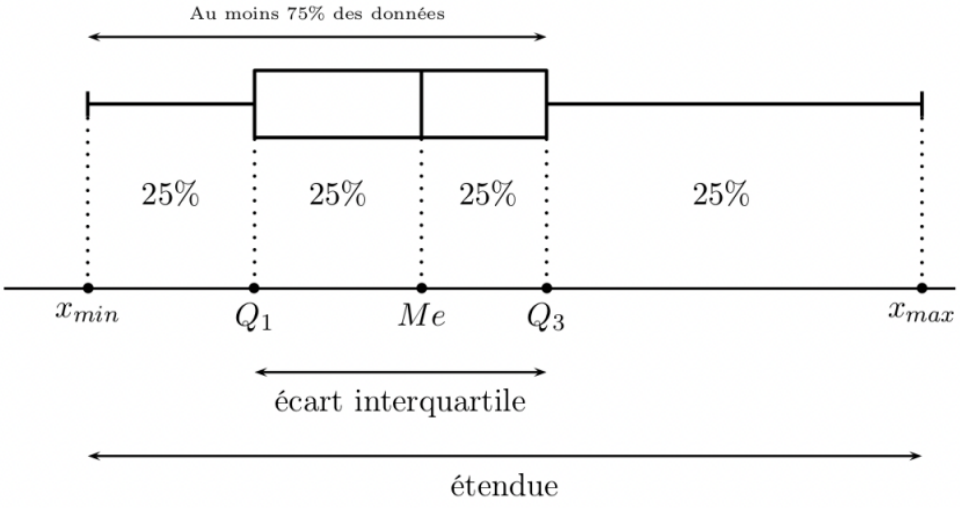
Si $(m_i)_{1\leqslant i\leqslant p}$, $(n_i)_{1\leqslant i\leqslant p}$ et $(f_i)_{1\leqslant i\leqslant p}$ désignent respectivement les modalités, les effectifs et les fréquences on a aussi : $\bar{x} = \frac{1}{\mathrm{n}}\sum_{i=1}^{p} n_i m_i = \sum_{i=1}^{p} f_i m_i$.

Par exemple le salaire moyen net mensuel en France en 2022 est de 2340€. Notons qu'en général, la moyenne n'est pas une valeur de la série.

Définition : Le **premier quartile** est la plus petite valeur Q_1 de la série telle qu'au moins 25% des valeurs de la série soient inférieurs ou égales à Q_1.
Le **troisième quartile** est la plus petite valeur Q_3 de la série telle qu'au moins 75% des valeurs de la série soient inférieurs ou égales à Q_3.
L'intervalle $[Q_1, Q_3]$ est appelé **intervalle interquartile**. Sa longueur est **l'écart interquartile**.

Définition : **Les valeurs extrêmes** d'une série sont les valeurs ayant la modalité minimale et maximale.
L'étendue est la différence entre ces deux valeurs extrêmes.

Définition : On appelle b**oîte à moustache** la représentation suivante :



Propriété (transformation affine) : Si, pour tout $i \in [[1,p]]$, on opère une transformation affine de x en définissant une nouvelle série u : $u_i = ax_i + b$ avec ($a \in \mathbb{R}^*$ et $b \in \mathbb{R}$).
On a alors : $\bar{u} = a\bar{x} + b$.
La médiane de u vérifie également cette propriété.

Cela signifie que si je multiplie toutes les valeurs d'une série par un même nombre et que je leur ajoute un autre même nombre, la moyenne ainsi que la médiane subissent les mêmes transformations. Par exemple, augmenter le revenu de tous les français de 100€ augmente la moyenne nationale de 100€, augmenter le salaire de 10% (donc multiplier par 1,1) augmente la moyenne de 10%.
Attention, cela n'est pas vrai de toutes les transformations, notamment, si je mets toutes les valeurs d'une série au carré, la nouvelle moyenne n'est pas le carré de l'ancienne.
En effet si x=(0,2,3,3), $\bar{x} = 2$ et si je défini u=(0,4,9,9) (les valeurs de x au carré), $\bar{u}$ vaut 5,5 ce qui n'est pas égal à 2^2.

Définition : La variance d'une série statistique est le nombre réel positif, noté s_x^2, tel que

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

L'écart-type est le nombre réel positif, noté σ_x, tel que $\sigma_x = \sqrt{s_x^2}$.

Ces indicateurs mesurent la dispersion des valeurs de la série par rapport à la moyenne.
Une variance nulle correspond à une série où toutes les valeurs sont égales.
Par exemple, l'écart type des notes d'une épreuve de concours mesure les écarts de notes, un écart type fort correspond à une épreuve où il y a beaucoup de notes basses et hautes (par rapport à la moyenne).
En sociologie, on peut mesurer les inégalités sociales en calculant l'écart type de la série des revenus des Français, ou celle des patrimoines.

Propriété (Formule de Koenig) : On a $s_x^2 = \overline{x^2} - \bar{x}^2$

Où on a défini x^2 comme étant la série des valeurs de x au carré : $(x_1^2, \dots, x_n^2)$.

Remarque : C'est cette formule que l'on utilise en pratique (la définition est pénible à manipuler en pratique).

Propriété : Si, pour tout $i \in [[1,p]]$, on opère un changement de variable affine $u_i = ax_i + b$ avec ($a \in \mathbb{R}^*$ et $b \in \mathbb{R}$) alors on a :

$$s_u^2 = a^2 s_x^2 \text{ et } \sigma_u = |a|\sigma_x$$

On comprend que le passage à la racine carrée pour obtenir l'écart type à partir de la variance est une remise à l'échelle.

Notons que le b n'apparaît pas dans cette formule, en effet, augmenter chaque valeur d'un même nombre, n'a pas d'incidence sur la dispersion des valeurs par rapport à la nouvelle moyenne.