# Statistiques descriptives

## I. Echantillons

A la base de toute étude statistique, il y a une population, formée d'individus sur lesquels on observe des caractères.

Exemples :

| Les individus | Le caractère |
|---|---|
| Les élèves de la classe | la taille |
| Les Français | la couleur des yeux |
| Les personnes hospitalisées | le groupe sanguin |
| Les lectrices de "Elle" | les réponses à une enquête d'opinion |
| Les galaxies | le nombre d'étoiles |
| Les chromosomes | le nombre de gènes |
| Les villes | la qualité de l'air |
| Les pays | le produit intérieur brut |
| Les films | les recettes |
| Les films | la critique |

Un caractère est dit :

- qualitatif, quand les valeurs ne peuvent être ni ordonnées ni ajoutées :
  la couleur des yeux , le groupe sanguin, éventuellement les réponses à une enquête
- ordinal, quand les valeurs peuvent être ordonnées mais pas ajoutées :
  la qualité de l'air, éventuellement la critique
- quantitatif, quand les valeurs sont numériques :
  la taille, le nombre d'étoiles, le nombre de gènes, le produit intérieur brut, les recettes

Les valeurs prises par un caractère s'appellent les modalités. Pour des raisons de facilité de traitement informatique ou mathématique, on cherche à se ramener à des caractères quantitatifs par un codage. Il faut se souvenir que le codage est arbitraire, et que les résultats numériques que l'on obtient après codage peuvent dépendre de celui-ci.

La statistique intervient quand il est impossible ou inutile d'observer un caractère sur l'ensemble de la population. On l'observe alors sur une sous-population, de taille réduite, en espérant tirer de l'observation des conclusions généralisables à toute la population. Si les données d'un caractère quantitatif sont recueillies sur un ensemble de $n$ individus, le résultat est un $n$-uplet de nombres, entiers ou décimaux, $(x_1, x_2, ..., x_n)$, que l'on appelle échantillon ou série statistique de taille $n$ ( $n$ étant appelé l'effectif total) .
On réserve plutôt le terme d'échantillon au résultat de $n$ expériences menées indépendamment les unes des autres, et dans des conditions identiques (lancers de dés, mesure du poids de nouveau-nés, ...)
On appellera série statistique le résultat de $n$ expériences qui ne sont pas interchangeables. Le cas le plus fréquent est celui où la population est constituée d'instants successifs (relevés quotidiens de températures, chiffres mensuels du chômage,...) On parle alors de série chronologique.

On distingue souvent les caractères discrets (ceux qui ne prennent que peu de modalités distinctes) des caractères continus (pour lesquels les valeurs observées distinctes sont en général en très grand nombre). Pour un caractère continu, on regroupe ses valeurs dans des intervalles appelés classes ; la longueur d'une classe est appelée son amplitude et son milieu son centre.
Une fois recueilli, l'échantillon se présente comme une liste de nombres peu lisible. Le traitement statistique va maintenant consister à le compresser, le résumer par des quantités calculées et des représentations graphiques, afin d'extraire l'information qu'il contient.

En statistique, une fonction d'un échantillon, comme sa moyenne, sa médiane, ses quartiles ou son étendue, par laquelle on cherche à résumer une partie de l'information qu'il contient, s'appelle encore une statistique.
Le mot « statistique » a donc trois sens différents :

- ❒ C'est **un ensemble de données** chiffrées contenant des informations sur un phénomène donné. Par exemple : les statistiques du commerce extérieur, les statistiques du chômage…
- ❒ C'est **une discipline scientifique** dont le but est d'extraire de l'information d'un échantillon en vue d'une prédiction ou d'une décision.
- ❒ C'est **une fonction** d'un échantillon.

## II. Présentation des données

**a) Tableaux de valeurs**

Les données peuvent être présentées dans un tableau des effectifs (sur la première ligne mettre les modalités prises par le caractère et sur la deuxième ligne les effectifs correspondants). Lorsque le caractère est quantitatif, les valeurs sont rangées par ordre croissant.

Les données peuvent aussi être présentées dans un tableau des fréquences en remplaçant les effectifs par les fréquences. La fréquence d'une modalité étant égale au rapport de l'effectif de la modalité par l'effectif total.

La somme des fréquences est égale à 1. On peut encore exprimer les fréquences en pourcentage et dans ce dernier cas la somme est égale à 100.

**b) Représentations graphiques**

Dans le cas où l'échantillon est discret (lorsque le nombre de valeurs différentes est faible devant la taille de l'échantillon), on représentera la série statistique par un diagramme en bâtons (ou en barres) . Il consiste à représenter les valeurs différentes en abscisse, avec au-dessus de chacune une barre verticale de longueur égale à sa fréquence ou à son effectif. Dans le cas où le nombre de valeurs différentes est très faible (inférieur à 10), et surtout pour des échantillons qualitatifs, on utilise aussi des diagrammes circulaires (ou à secteurs ou encore « camembert »), semi-circulaires ou à bandes. Ces représentations consistent à diviser un disque, un demi-disque ou un rectangle proportionnellement aux différentes fréquences.

La représentation correspondant au diagramme en bâtons (dans le cas discret) pour un échantillon considéré comme continu (lorsque presque toutes les valeurs sont différentes), est l'histogramme. Il consiste à représenter les extrémités des intervalles en abscisses et à représenter des rectangles d'aires proportionnelles aux différentes fréquences. Dans le cas simple où les amplitudes des classes sont les mêmes, la hauteur des rectangles est proportionnelle à l'effectif. Dans le cas où les amplitudes sont différentes, la hauteur peut être égale au quotient de l'effectif par l'amplitude de la classe correspondante.

Un nuage de points est l'ensemble des points représentés dont les coordonnées sont les valeurs de la modalité pour les abscisses et les effectifs (ou les fréquences) associés pour les ordonnées.

**c) Effectifs (resp. fréquences) cumulés croissants (resp. décroissant)**

Dans le cas d'une variable quantitative, dans un tableau des effectifs dont les valeurs sont rangées par ordre croissant (resp. décroissant), l'effectif cumulé croissant (resp. décroissant) d'une valeur s'obtient en ajoutant à cet effectif les effectifs des valeurs qui la précédent. Un tel tableau permet de répondre à la question « Quel effectif de la population a une valeur du caractère au plus (resp. au moins) égale à ... ?»

Les effectifs cumulés croissants (resp. décroissants) seront représentés par la courbe (ou polygone) des effectifs cumulés : polygone obtenu en reliant les points dont les coordonnées sont données dans le tableau des effectifs cumulés croissants (resp. décroissants).

On peut de la même façon obtenir le tableau et la courbe des fréquences cumulées croissantes (resp. décroissantes).

## III. Indicateurs de position

**a) Moyenne empirique**

La statistique la plus évidente à calculer sur un échantillon numérique est la moyenne empirique.

1. Définition : La moyenne empirique (arithmétique) d'un échantillon est la somme de ses éléments divisée par leur nombre. Si l'échantillon est noté $(x_1, x_2, ..., x_n)$, sa moyenne empirique est :

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Remarques :

1. Lorsque le caractère est continu, on prend pour valeurs les centres des classes (le centre de la classe $[a_i; a_{i+1}[$ est le nombre $\frac{a_i + a_{i+1}}{2}$).

2. Si $x_1, x_2, ...x_p$ sont les valeurs prises par la variable et $n_1, n_2, ...n_p$ les effectifs correspondants, on parlera de moyenne pondérée :

$$\bar{x} = \frac{1}{n}(n_1x_1 + n_2x_2 + ... + n_px_p) = \frac{1}{n}\sum_{i=1}^{p} n_ix_i$$

3. Moyenne et fréquences Si $f_1, f_2, ..., f_p$ sont les fréquences associées aux valeurs $x_1, x_2, ...x_p$,

$$\bar{x} = f_1x_1 + f_2x_2 + ... + f_px_p = \sum_{i=1}^{p} f_ix_i \quad \text{avec} \quad f_i = \frac{n_i}{n}$$

4. Moyenne élaguée : Un des inconvénients de la moyenne empirique est d'être sensible aux valeurs extrêmes. Une valeur manifestement très différente des autres est souvent qualifiée de valeur aberrante. Qu'elle soit ou non le résultat d'une erreur dans le recueil ou la transcription, on ne peut pas la considérer comme représentative.

Exemple : Supposons que sur un échantillon de 10 valeurs, toutes soient de l'ordre de 10, sauf une, qui est de l'ordre de 1000. La moyenne empirique sera de l'ordre de 100, c'est-à-dire très éloignée de la plupart des valeurs de l'échantillon.

Pour pallier cet inconvénient, on peut décider ne pas tenir compte des valeurs extrêmes dans le calcul de la moyenne. On obtient alors une moyenne élaguée.

**b) Mode**

Pour un caractère discret (ou continu regroupé en classes), le mode (ou la classe modale) de la série statistique est la valeur (ou la classe) qui a la fréquence la plus élevée.

**c) Médiane**

Définition : Pour une série statistique rangée dans l'ordre croissant, la médiane correspond à la valeur de ce caractère qui sépare la population en deux parties de même effectif : il y a autant de valeurs qui lui sont inférieures que supérieures.

Cas d'une série discrète : La série étant ordonnée par ordre croissant, si l'échantillon a une taille impaire (égale à $2k+1$), la médiane est la valeur centrale (la valeur du terme de rang $k+1$) de cette série ordonnée.
Si l'échantillon a une taille paire (égale à $2k$), la médiane est la demie somme des deux valeurs centrales (valeurs des termes de rang $k$ et $k+1$).

Cas d'un caractère continu : On prendra la classe qui contient la valeur et l'on parlera de classe médiane.

Remarque : Si la distribution empirique de l'échantillon est peu dissymétrique, la moyenne et la médiane sont proches. Si l'échantillon est dissymétrique, avec une distribution très étalée vers la droite, la médiane pourra être beaucoup plus petite que la moyenne. Contrairement à la moyenne, la médiane est insensible aux valeurs aberrantes.

**d) Premier et troisième quartiles**

Définitions : Le premier quartile, noté $Q_1$ est la plus petite valeur de la série telle qu'au moins 25% des données soient inférieures ou égales à ce nombre.

Le troisième quartile, noté $Q_3$ est la plus petite valeur de la série telle qu'au moins 75% des données soient inférieures ou égales à ce nombre.

Méthode de calcul : La série de taille *n* étant ordonnée par ordre croissant, le premier quartile est la valeur du terme de rang *p* avec *p* le premier entier supérieur ou égal à $\frac{n}{4}$.

Le troisième quartile est la valeur du terme de rang *p* avec *p* le premier entier supérieur ou égal à $\frac{3n}{4}$.

## IV. Indicateurs de dispersion

**a) Etendue**

Pour mesurer la dispersion, on peut calculer l'étendue, qui est la différence entre la plus petite et la plus grande valeur. Mais cette étendue reflète plus les valeurs extrêmes que la localisation de la majorité des données.

**b) Intervalle interquartile**

L'intervalle interquartile est l'intervalle $[Q_1, Q_3]$ et l'interquartile est le nombre $Q_3 - Q_1$.