# Statistiques descriptives

## I. Vocabulaire

### a) Population

C'est l'ensemble des individu ou objet sur lesquels porte une étude statistique P

Ex : *logement d'une ville, personnel d'une entreprise, animaux d'un parc naturel, …*

### b) Échantillon

C'est une partie de la population étudiée sur laquelle porte une étude statistique.

- Une étude statistique portant sur un échantillon est appelée « sondage »
- Une étude statistique portant sur la population tout entière est appelée « recensement »

### c) Caractères et modalités

Une étude statistique porte sur un ou plusieurs caractères communs à tous les individus d'une population.

Un caractère est aussi appelé variable.

Ex : type de logement, salaire mensuel, régime alimentaire, …

- Les modalités sont des différentes valeurs que peut prendre un caractère *(une variable)*
  - Type de logement : Studio, maison individuel, T2, T3, T4, …
  - Salaire mensuel : $\mathbb{R}+$
  - Régime alimentaire : carnivore, herbivore, omnivore, …

On distingue 2 types de caractères :

- Qualitatifs : Un caractère est dit qualitatif si ses modalités sont des attributs qualitatifs.

  Ex : *type de logement et régime alimentaire*
- Quantitatif : Un caractère est dit quantitatif si ses modalités sont des attributs numériques.

  Ex : *salaire mensuel*

Les caractères quantitatifs sont de 2 types :

- Discrète : une variable est dite discrète si ses modalités appartiennent à un ensemble discret.

Ex : *N, ℤ, toute partie de N ou ℤ…*

- Continues : une variable est dite continue si ses modalités prennent des valeurs réelles.

  Ex : salaire mensuel

Remarque : pour étudier une variable continue, on constitue des classes de valeur possibles. Ces classes sont des intervalles, donc l'étude égale ou différentes de ces classes constitues alors de nouvelles modalités de la variable.

Attention : Le découpages en classe des modalités d'une variable continue peut influencer le résultat de l'étude statistique. Le découpage doit être fin dans des zones de forte densité et large dans des zones de faible densité.

### d) Effectif et fréquence

- L'effectif relatif à une modalité ou une classe de modalité est le nombre d'individus correspondant à cette modalité ou cette classe de modalité.

  On note « $n_i$ » si cela correspond à la $i^{ème}$ modalité (ou classe).

- Si le nombre d'individus étudiés est $n \in N$, la fréquence de la $i^{ème}$ modalité est :

$$f_i = \frac{n_i}{n}$$

Remarque : Si une variable X comporte K modalité $(x_1, x_2, \dots, x_k)$, dont les effectifs sont $(n_1, n_2, \dots, n_k)$, et les populations $(f_1, f_2, \dots, f_k)$, alors on a :

- $\sum_{i=1}^{k} n_i = n$
- $\sum_{i=1}^{k} f_i = 1$

## II. Représentation graphique

Il existe différentes façons de représenter graphiquement les observations d'une variable.

### a) Variables qualitatives

- Diagramme circulaire : c'est un diagramme dans lequel chaque modalité et représentée par un secteur angulaire proportionnel à sa fréquence. Si $(f_1, f_2, \dots, f_k)$ sont les fréquences des modalités de X, les secteurs angulaires correspondant sont :

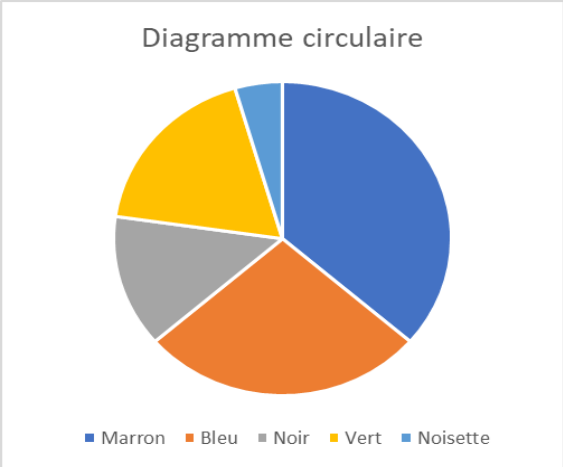
$$\alpha_i = 360 * \alpha_i, 1 \leq i \leq k$$

  Remarque : $\sum_{i=1}^{k} \alpha_i = 360$

- Diagramme en tuyau d'orgue : c'est un diagramme formé de rectangles, tous de mêmes largeurs, et de hauteur proportionnelle aux fréquences.

Ex : *Soit X une variable décrivant la couleur des yeux dans un groupe de 40 individus. Les résultats sont donnés dans le tableau suivant :*

| Modalités | $n_i$ | $f_i$ | $\alpha_i$ |
|---|---|---|---|
| Marron | 16 | 0.364 | 131.04 |
| Bleu | 12 | 0.233 | 98.28 |
| Noir | 6 | 0.136 | 48.96 |
| Vert | 8 | 0.182 | 85.52 |
| Noisette | 2 | 0.045 | 16.2 |
| TOTAL | 44 | 1 | 360 |




### b) Variables qualitatives

Il existe deux sortes de représentation :

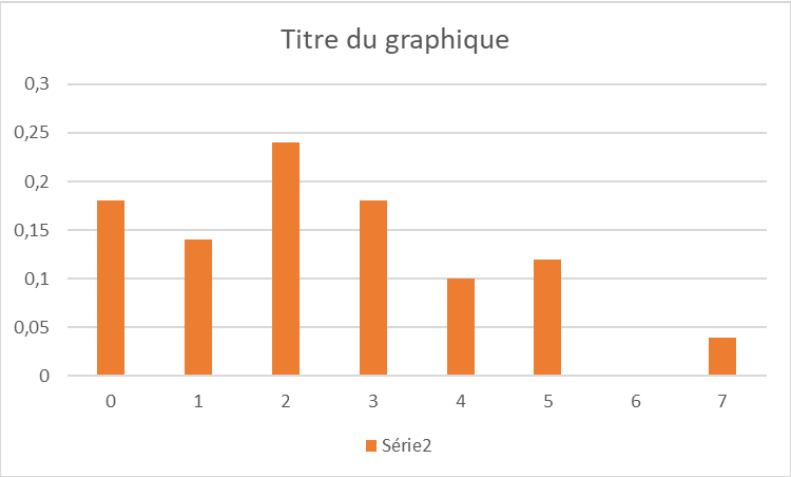
- Diagrammes différentiels :
  - Diagrammes à bâtons (cas discret)
  - Histogramme (cas continue)
  - Polygone statistique (cas continue)
- Diagrammes intégraux :
  - Courbes cumulatives (dans les deux cas => discrets et continu)

#### 1) Cas d'une variable discrète

Dans le cas d'une variable discrète le diagramme utilisé est le diagramme à bâtons ou la hauteur de chaque bâton et proportionnel à la fréquence de la modalité correspondante.

Ex : *nombre d'enfant dans un échantillon de 50 familles*

| xi (modalités) | ni | f1 | Fi |
|---|---|---|---|
| 0 | 9 | 0,18 | 0,18 |
| 1 | 7 | 0,14 | 0,32 |
| 2 | 12 | 0,24 | 0,56 |
| 3 | 9 | 0,18 | 0,74 |
| 4 | 5 | 0,1 | 0,84 |
| 5 | 6 | 0,12 | 0,96 |
| 6 | 0 | 0 | 0,96 |
| 7 | 2 | 0,04 | 1 |
| TOTAL | 50 | 1 | |

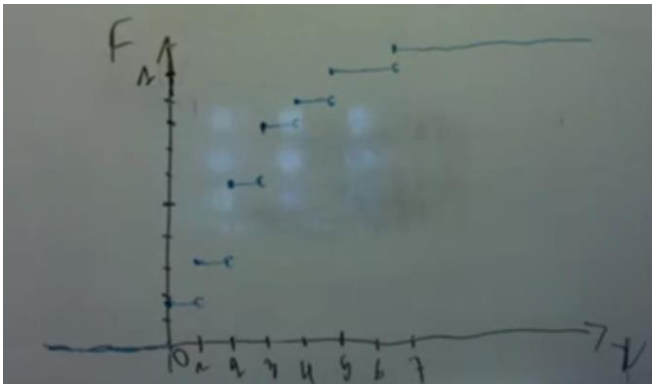



**Définition** : Fonction cumulative :
La fonction cumulative est une fonction définie sur $\mathbb{R}$ par :

$$\forall x \in [x_i, x_i + 1[\ , F(X) = \sum_{j=1}^{i} f_j$$

**Remarque** :

- Si $x < x_1$ alors $F(x) = 0$
- Si $x >= x_k$ alors $F(x) = 1$



### 2) Cas d'une variable continue

Si $e_{i-1}$ et $e_i$ sont les extrémités de la $i^{ème}$ classe, on note $C_i$ le centre de cette classe et ai son amplitude. (1 <= i <= k).

Le diagramme différentiel est **l'histogramme.**

Histogramme : Un histogramme est une représentation graphique où chaque classe est représentée par un rectangle de base proportionnelle à son amplitude et de surface proportionnelle à sa fréquence.

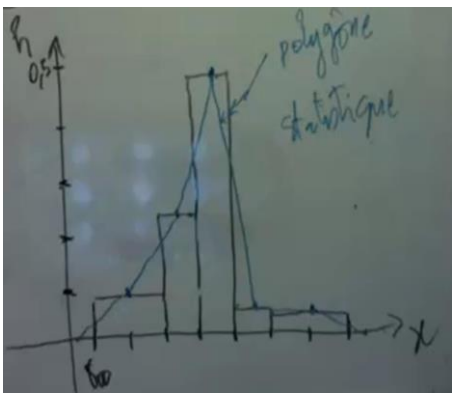
Remarque :

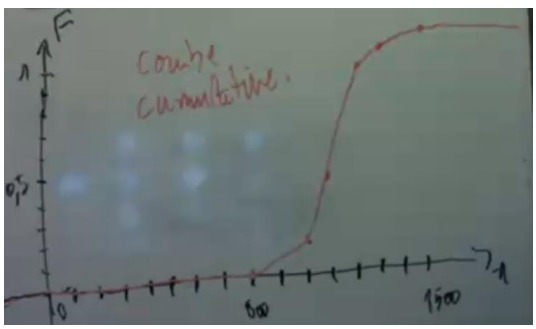
La hauteur associée à la classe numéro i est $h_i = \frac{f_i}{\alpha_i}$

Polygone statistique : Un polygone statistique est un polygone reliant les milieux des bases supérieures des rectangles de l'histogramme.

Ex : *Salaire mensuel du personnel d'une entreprise*

| | | | | *100 (pour retirer des 0) | |
|---|---|---|---|---|---|
| classes | ni | fi | ai | hi | Fi |
| 800 - 1000 | 26 | 0,186 | 200 | 0,093 | 0,186 |
| 1000 - 1100 | 33 | 0,236 | 100 | 0,236 | 0,421 |
| 1100 - 1200 | 64 | 0,457 | 100 | 0,457 | 0,879 |
| 1200 - 1300 | 7 | 0,050 | 100 | 0,050 | 0,929 |
| 1300 - 1500 | 10 | 0,071 | 200 | 0,036 | 1,000 |
| TOTAL | 140 | 1 | 700 | | |




## III. Description numérique d'une variable

### a. Paramètres de position

Définition : La **médiane** $\mu$ est la valeur de la variable X pour laquelle la moitié au moins des observations lui sont supérieures ou égales et la moitié au moins des observations inférieures ou égales.

Remarque : Dans le cas d'une variable continue on a $F(\mu) = 0.5$.

Ex : *(cas du nombre d'enfants dans 50 familles)*

Dans le cas continu, on détermine d'abord la classe médiane.

$$\mu \in [1100,1200]$$

On détermine ensuite u par la méthode d'interpolation linéaire.

(1100 ; 0.421), (u ; 0.5), (1200 ; 0.879)

$$\leftrightarrow \frac{u - 1100}{0{,}5 - 0{,}421} = \frac{(1200 - 1100)}{(1200 - 1100)} = \frac{100}{0.458}$$

$$u = 1117$$

### b. Mode

**Définition** : Le **mode** est la valeur de la variable X ayant la plus grande fréquence.

Remarque :

- Certaine série statistique peuvent avoir plusieurs modes dans le cas d'une variable continue on parle de classe modale.
- La classe modale est la classe ayant la plus grande hauteur $h_i = \frac{f_i}{\alpha_i}$.

Ex : *Salaire d'une entreprise. La classe modale est* $M_0$ = [1100,1200[

### c. Moyenne

Si la variable X prend les valeurs $x_i, \ldots, x_k$ avec les fréquences $f_i, \ldots, f_k$, alors la moyenne de $\bar{X}$ est :

$$\bar{X} = \sum_{i=1}^{k} f_i x_i$$

Et comme $f_i = \frac{n_i}{n}$ , $1 \leq i \leq k$ , on a aussi :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$$

Dans le cas continu on a :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} f_i c_i = \frac{1}{n} \sum_{i=1}^{k} c_i n_i \text{ , ou } c_i n_i \text{ est le millieu de la classe n° } = i$$

- Propriéte :

1) Linéarité : Soit X une variable distinguée prenant les valeurs $x_i$, ..., $x_k$ avec les fréquences $f_i$, ..., $f_k$. Si on pose $y = ax + b$ ou a et b sont des paramètres réels alors :

$$y = a\bar{X} + b$$

- $\sum_{i=1}^{k} n_i(x_i - \bar{X}) = \sum_{i=1}^{k} n_i * x_i - n_{\bar{X}} = n_{\bar{X}} - n_{\bar{X}} = 0$

- On considère la fonction :

$$h(a) = \sum_{i=1}^{k} n_i (x_i - a)^2$$

La fonction h admet son minimum au point $a = \bar{X}$.

En effet,

$$h'(a) = -2\sum_{i=1}^{k} n_i(x_i - a)^2 = 0$$

$$= \sum_{i=1}^{k} n_i x_i - a\sum_{i=1}^{k} n_i = 0$$

$$= \sum_{i=1}^{k} n_i x_i - a * n = 0$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \bar{X}$$

Et $h''(a) = 2\sum_{i=1}^{k} n_i = 2n > 0$, donc $a = \bar{X}$ est un minimum.

- On considère une variable X mesurée sur 2 populations.
    - P1 : effectif = $n_1$, moyenne = $\bar{X}_1$
    - P1 : effectif = $n_2$, moyenne = $\bar{X}_2$

La moyenne globale est alors :

$$\bar{X} = \frac{n_1\bar{X}_1 + n_1\bar{X}_2}{n_1 + n_2}$$

- Paramétre de dispersion :
    - L'étendue

**Définition** : L'étendue d'une variable statistique X est donné par : $e = \max\{x_i\} - \min\{x_i\}$.

Remarque : C'est un paramètre grossier car il ne dépend que du maximum et minimum de la série statistique.

- Intervalle interquartile

Définition : Le premier quartile noté $Q_1$ est la valeur de la variable X pour laquelle au moins ¼ des observations lui son inférieur ou égale et au moins ¾ des observations lui son supérieur ou égale.

Le premier quartile noté $Q_3$ est la valeur de la variable X pour laquelle au moins ¾ des observations lui son inférieur ou égale et au moins ¼ des observations lui son supérieur ou égale.

[Q1, Q3] => Intervalle interquartile

Remarque : Q2 = la médiane.

*Ex* : *(cas du nombre d'enfants dans 50 familles)*

$Q_1 = 1$ / $Q_3 = 4$

Donc ici l'intervalle interquartile = [1, 4[

- Ecart moyen
  - Ecart moyen absolu

Soit X une variable prenant les valeurs ses valeurs $x_i, \ldots, x_k$ avec les fréquences $f_i, \ldots, f_k$. L'écart moyen absolu est alors :

$$E = \sum_{i=1}^{k} f_i |x_i - \bar{X}| = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{X})$$

Remarque : L'écart moyen absolu donne une image conforme de la réalité mais pose des problèmes techniques de manipulation à cause de la valeur absolue.

  - Ecart type et variance

Soit X une variable prenant les valeurs les valeurs $x_i, \ldots, x_k$ avec les fréquences $f_i, \ldots, f_k$. La variance est alors :

$$\text{var}(X) = \sum_{i=1}^{k} (x_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{X})^2$$

Si X est une variable continue, alors :

$$\text{Var}(X) = \sum_{i=1}^{k} (c_i - \bar{X})^2, \text{ou ci est le millieu de la classe n°} = i$$

**Définition** : L'écart type est définis par :

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

- Propriétés

1) $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{k} n_i {x_i}^2 - \bar{X}^2$
2) Si **y = aX + b**, alors **Var(y) = a² Var(X)**
3) On considère deux populations :
   - P1 : effectif = $n_1$, variance = $\mathrm{Var}_1(X)$
   - P1 : effectif = $n_2$, variance = $\mathrm{Var}_2(X)$

   Alors la variance de X de la population globale est $P = P_1 \cup P_2$ est donné par :

$$\mathrm{Var}(X) = \frac{n_1\mathrm{Var}_1(X) + n_2\mathrm{Var}_2(X)}{n_1 + n_2} = \frac{n_1(X_1\text{-}\bar{X})^2 + n_2(X_2\text{-}\bar{X})^2}{n_1 + n_2}$$

$$= \text{moyenne des variance} + \text{variance des moyennes}$$

Ici $\bar{X}$ est la moyenne de X sur P.

- Coefficient de variation

**Définition** : Le coefficient de variation d'une variable X est définis par :

$$cv = \frac{\sigma_X}{\bar{X}}$$

Remarque : Le coefficient de variation est une grandeur sans unité de mesure. On peut alors comparer les dispersions de séries exprimées dans des unités de mesure différente.

- Les moments

Le **moment** d'une variable X prenant ses valeurs $x_i, \ldots, x_k$ avec les fréquences $f_i, \ldots, f_k$ est défini par :

$$ml = \sum_{i=1}^{k} f_i {x_i}^l = \sum_{i=1}^{k} n_i {x_i}^l$$

Le **moment centré** d'ordre l de X est définis par :

$$ml = \sum_{i=1}^{k} f_i (x_i\text{-}\bar{X})^l = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i\text{-}\bar{X})^l$$

Remarque :

- $m_0 = \mu_0 = 1$
- $m_1 = \bar{X}$
- $\mu_1 = 0$
- $\mu_2 = Var(X) = m_2 - m_1^2$

- Caractéristique de forme :
  - Coefficient d'asymétrie
    - Coefficient de Fisher :

$$\gamma_x = \frac{\mu_3}{\sigma_X^3}$$

- Si la série X est symétrique, alors = 1
- Si A => B alors B =>A
- Si $\gamma$ != 0 => la série n'est pas
- Si $\gamma < 0$, alors la série est asymétrique oblique à droite et étalée à gauche