Statistiques

Chapitre 15

Vocabulaire

Définition — Termes de base

- Une *population* est l'ensemble étudié.
- Un élément de la population est un *individu*.
- On étudie un *caractère* (ou *variable*) commun aux individus. Il peut être :
  - *qualitatif* (couleur, sexe, sport pratiqué) ;
  - *quantitatif* : valeurs numériques. Discret (nombre d'enfants par famille, note) ou continu (taille, poids).
- L'*effectif* n_i d'une valeur x_i est le nombre d'individus correspondant.
- L'*effectif total* est $N = \sum_i n_i$.
- La *fréquence* de la valeur x_i est $f_i = \frac{n_i}{N}$, souvent exprimée en pourcentage.

Exemple. Note d'une classe de 25 élèves :

$$8, 12, 15, 10, 12, 14, 9, 11, 13, 12, 16, 10, 11, 12, 14, 13, 9, 15, 11, 12, 13, 10, 14, 12, 11.$$

Effectif total $N = 25$. Effectif de la note 12 : $n_{12} = 6$. Fréquence : $f_{12} = \frac{6}{25} = 0.24 = 24\%$ (soit $0,24$).

Organisation des données

Tableau d'effectifs et de fréquences

Une série discrète se présente sous forme de tableau :

x_i	8	9	10	11	12	...
n_i	1	2	3	4	6	...
f_i	0.04	0.08	0.12	0.16	0.24	...

Données regroupées en classes

Pour une variable continue, on regroupe les valeurs en *classes* (intervalles). On note $[a, b[$ une classe de centre $\frac{a+b}{2}$ et d'amplitude $b - a$.

Effectifs / fréquences cumulés

L'effectif cumulé croissant en x_i est la somme des effectifs des valeurs $\leq x_i$. Idem pour les fréquences cumulées.

Représentations graphiques

- **Diagramme en bâtons** : pour une variable discrète, hauteur proportionnelle à l'effectif (ou la fréquence).
- **Histogramme** : pour une variable continue, aire de chaque rectangle proportionnelle à l'effectif de la classe.
- **Polygone des fréquences** : ligne brisée joignant les milieux des hauteurs des bâtons / sommets des rectangles.
- **Diagramme circulaire** : angles proportionnels aux fréquences.
- **Courbe des effectifs cumulés croissants** : utile pour lire la médiane et les quartiles.

Mesures de tendance centrale

Mode

Définition
Le *mode* est la (ou les) valeur(s) du caractère ayant l'effectif le plus élevé. Pour des données regroupées, on parle de *classe modale*.

Médiane

Définition — Médiane (série discrète)
La *médiane* Me est la valeur qui partage la série rangée en deux parties d'effectif égal :

- Si N est impair : la médiane est la valeur de rang $\frac{N+1}{2}$.
- Si N est pair : la médiane est la moyenne des valeurs de rang $\frac{N}{2}$ et $\frac{N}{2}+1$.

Exemple. Série : $5, 8, 10, 12, 14$ ($N = 5$). Médiane $= 10$ (3e valeur).

Série : $5, 8, 10, 12, 14, 18$ ($N = 6$). Médiane $= \frac{10+12}{2} = 11$.

Moyenne arithmétique

Définition
La *moyenne* d'une série $(x_1, x_2, ..., x_p)$ d'effectifs $(n_1, ..., n_p)$ est :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{N} = \sum_{i=1}^{p} f_i x_i.$$

Pour des classes, on utilise les centres de classe.

Exemple. Si $x_i \in \{8, 12, 16\}$ avec effectifs $(2, 5, 3)$ et $N = 10$: $\overline{x} = \frac{2\times8+5\times12+3\times16}{10} = \frac{16+60+48}{10} = 12.4$ (soit $12,4$).

Mesures de dispersion

Étendue

Définition
L'*étendue* est la différence entre la plus grande et la plus petite valeur observée.

Variance et écart-type

Définition — Variance et écart-type
La *variance* de la série est :

$$V = \left(\frac{1}{N}\right) \sum_{i=1}^{p} n_i (x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2.$$

L'*écart-type* est $\sigma = \sqrt{V}$.

L'écart-type mesure la dispersion autour de la moyenne, dans la même unité que les données.

Théorème — Formule de König-Huygens

$$V = \left(\frac{1}{N}\right) \sum_{i=1}^{p} n_i x_i^2 - \overline{x}^2.$$

Cette formule est souvent plus pratique pour le calcul.

Exemple. Pour la série $x_i \in \{8, 12, 16\}$ d'effectifs $(2, 5, 3)$ et $\overline{x} = 12.4$: $\sum n_i x_i^2 = 2 \times 64 + 5 \times 144 + 3 \times 256 = 128 + 720 + 768 = 1,616$. $V = \frac{1616}{10} - 12.4^2 = 161.6 - 153.76 = 7.84$. $\sigma = \sqrt{7.84} \approx 2.8$.

Écart absolu moyen

Définition
L'*écart absolu moyen* est :

$$\mathrm{EAM} = \left(\frac{1}{N}\right) \sum_{i} n_i \, |x_i - \overline{x}|.$$

Quartiles, écart interquartile

Définition
Le 1er quartile Q_1 est la plus petite valeur telle qu'au moins 25% des observations sont $\leq Q_1$. Le 3e quartile Q_3 est défini de manière analogue avec 75%. L'*écart interquartile* est $Q_3 - Q_1$.

Interprétation des indicateurs

- La **moyenne** est sensible aux valeurs extrêmes ; la **médiane** l'est moins. Pour des salaires d'entreprise, la médiane est souvent plus représentative que la moyenne.
- L'**écart-type** est petit quand les données sont concentrées autour de la moyenne, grand quand elles sont dispersées.

Exemple. Deux séries A et B ont la même moyenne 10, mais $\sigma_A = 1$ et $\sigma_B = 5$. La série B est beaucoup plus dispersée — par exemple, $A : 9, 10, 11$ et $B : 5, 10, 15$.