COMMENT RÉSUMER DES DONNÉES ?
DÉBUT DE RÉFLEXION SUR LES STATISTIQUES DESCRIPTIVES

Une seule variable

MESURES DE TENDANCE CENTRALE

- Mesures du « centre » autour duquel se situent les données

MESURES DE DISPERSION

- Mesures de la dispersion autour de ce « centre »

MESURES DE SYMÉTRIE

- Mesures du degré de symétrie de la disposition des données autour de leur « centre »

Deux variables ou plus

MESURES D'ASSOCIATION

- Mesure du degré d'association entre les variables

COMMENT RÉSUMER DES DONNÉES ?
OBSERVATIONS INDIVIDUELLES
SUR UNE VARIABLE RATIONNELLE OU D'INTERVALLE

Notation

n = nombre d'observations

x_i = valeur de la variable X à la i ème observation

y_i = valeur de la variable Y à la i ème observation

1. Mesures de tendance centrale

- Moyenne : $\mu_x = \left(\frac{1}{n}\right)\sum_i x_i$
- Médiane : c'est la valeur $\breve{x}$ de la variable X telle que 50 % des observations ont des valeurs inférieures à $\breve{x}$, alors que 50 % ont des valeurs qui lui sont supérieures.
- Mode : avec un nombre fini d'observations, c'est la valeur la plus fréquente de la variable X.

2. Mesures de dispersion

- Domaine de variation : valeur minimum et valeur maximum
- Écart inter-quartile
- Variance : $\sigma_x^2 = \frac{1}{n}\sum_i (x_i - \mu_x)^2$

  NOTE : Cette formule est celle qui s'applique à une population, puisque la statistique descriptive ne distingue pas entre population et échantillon.
- Écart-type : $\sigma_x = \sqrt{\sigma_x^2}$
- Coefficient de variation : $C_x = \frac{\sigma_x}{\mu_x}$

COMMENT RÉSUMER DES DONNÉES ?

DISPERSION PAR RAPPORT À LA MOYENNE ?

Petit exemple numérique

	Fréquences		Écarts à la moyenne		
Valeur	**Série A**	**Série B**	**Écart**	**Valeur absolue**	**Écart au carré**
1	0	1	-3	3	9
2	2	0	-2	2	4
3	0	1	-1	1	1
4	4	4	0	0	0
5	0	1	1	1	1
6	2	0	2	2	4
7	0	1	3	3	9

Moyenne :

- Série A : $\mu_A = \frac{1}{8}\left[(2\times 2)+(4\times 4)+(2\times 6)\right] = 4$
- Série B : $\mu_B = \frac{1}{8}\left[(1\times 1)+(1\times 3)+(4\times 4)+(1\times 5)+(1\times 7)\right] = 4$

Somme des écarts à la moyenne :

- Série A : $\left[(2\times -2)+(4\times 0)+(2\times 2)\right] = 0$
- Série B : $\left[(1\times -3)+(1\times -1)+(4\times 0)+(1\times 1)+(1\times 3)\right] = 0$

$$\sum_i \left(x_i - \bar{x}\right) = \sum_i x_i - \sum_i \bar{x}$$
$$= n\bar{x} - n\bar{x} = 0$$

La somme des écarts à la moyenne n'est d'aucune utilité pour mesurer la dispersion !

Somme des écarts à la moyenne en valeur absolue :

- Série A : $[(2 \times 2) + (4 \times 0) + (2 \times 2)] = 8$
- Série B : $[(1 \times 3) + (1 \times 1) + (4 \times 0) + (1 \times 1) + (1 \times 3)] = 8$

Somme des carrés des écarts à la moyenne :

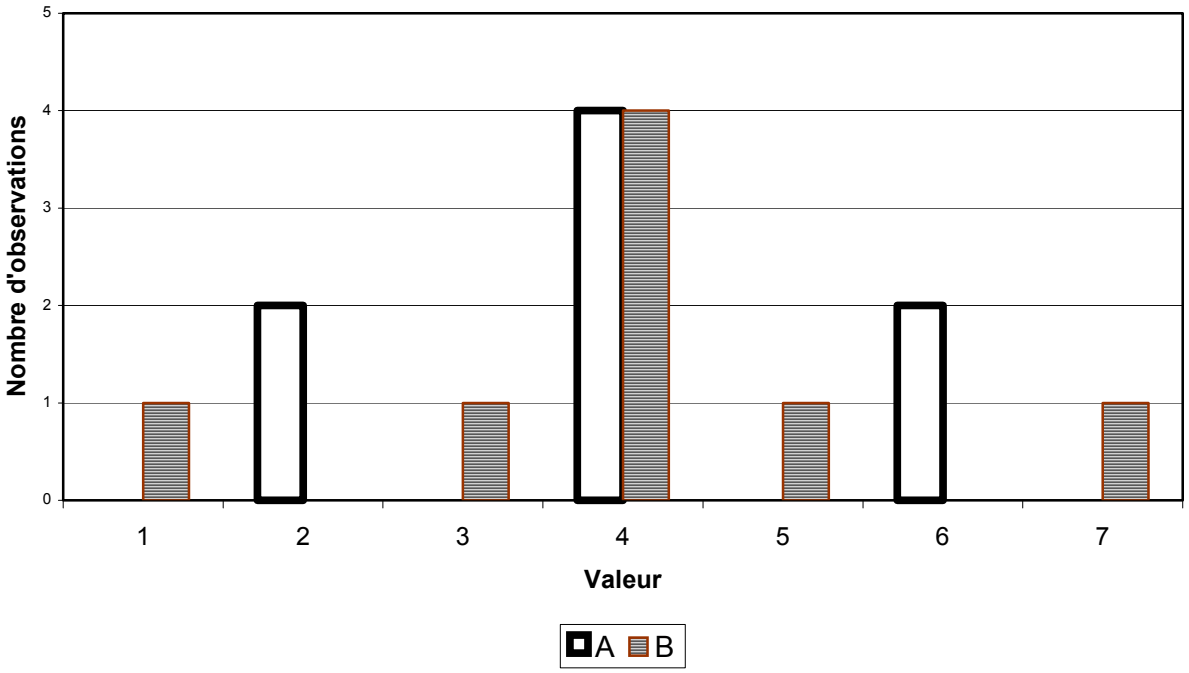
- Série A : $[(2 \times 4) + (4 \times 0) + (2 \times 4)] = 16$
- Série B : $[(1 \times 9) + (1 \times 1) + (4 \times 0) + (1 \times 1) + (1 \times 9)] = 20$

La somme des carrés donne à chaque écart un poids d'autant plus grand que cet écart est grand : elle reflète davantage les grandes différences.

Règles de normalisation :

- Diviser par le nombre d'observations n , pour pouvoir comparer des séries d'observations de différentes tailles.
- Diviser par la moyenne, pour ramener la dispersion de deux séries d'observations au même ordre de grandeur et la rendre indépendante des unités de mesure.

HISTOGRAMME DES SÉRIES



MESURES D'ASSOCIATION ENTRE DEUX VARIABLES

- Covariance

  (1) population : $\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y)$

  (2) échantillon : $s_{xy} = \frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)$

- Coefficient de corrélation simple

  (1) population : $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, avec $-1 < \rho < +1$

  (2) échantillon : $r = \frac{s_{xy}}{s_x s_y}$, avec $-1 < r < +1$