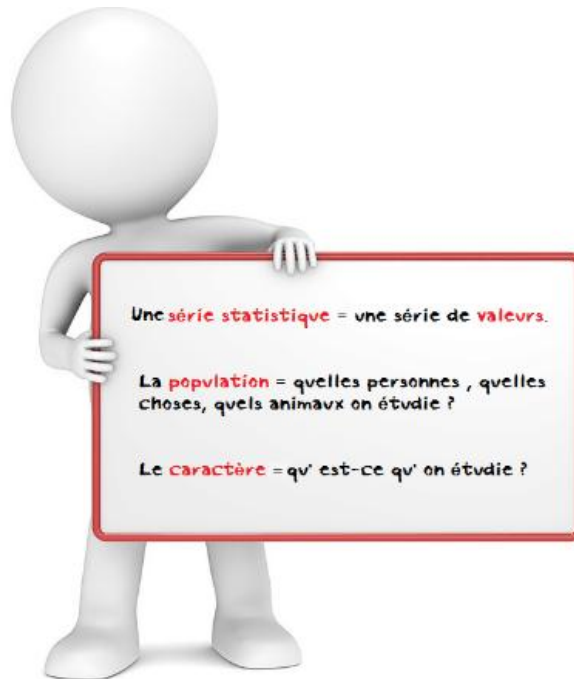


Chapitre 0 : Rappels sur les statistiques descriptives

Qu'est-ce qu'une série statistique ?

Une **série statistique** est simplement une liste de mesures obtenues généralement lors d'une étude ou de relevés de mesures. Pour avoir une série statistique, il suffit de produire une liste de nombres, ou de mesures.



Qu'est-ce qu'une valeur ?

Chaque nombre ou mesure d'une série statistique s'appelle une **valeur**. Attention, une valeur peut être un nombre, mais ce n'est pas obligatoirement le cas.

Les **valeurs extrêmes** sont les valeurs minimum et maximum dans la série.

La population d'une série statistique : c'est quoi ?

La **population** est l'ensemble des personnes ou des animaux ou des choses qui sont étudiées. (Donc, pas forcément des êtres humains !).

On se pose la question : «**Sur qui porte l'étude ? Quelle personne, quel animal ou quelle chose ?**».

Qu'est-ce que le caractère d'une série ?

Le **caractère** d'une série statistique c'est ce qui est mesuré ou étudié pour la population.

On se pose la question «**Sur quoi porte l'étude ? Qu'est-ce qui est étudié ?**»

- Si la valeur du caractère ne peut prendre qu'une série de valeurs isolées, définies à l'avance, on dit que le caractère est **discret**.
- Si la valeur du caractère peut prendre une infinité de valeur comprise entre 2 nombres, on dit que le caractère est **continu**.
- Un caractère peut être **quantitatif** si ses valeurs sont numériques.
- Un caractère peut être aussi **qualitatif** si ses valeurs ne sont pas numériques.

Exemple : une étude est réalisée sur la pointure des chaussures des jeunes Algériens de 10 à 20 ans.

Quelle est la population étudiée ? Quel est le caractère étudié ? Quel est son type ?

Mesures de tendance centrale et de dispersion

On considère sur un échantillon de N individus la variable statistique définie par le tableau de valeurs (rangées dans l'ordre croissant) suivant :

valeur	x_1	x_2		X_p
effectif	n_1	n_2		N_p

Valeur Effectif L'effectif total est $N = \sum_{i=1}^p n_i$

Indicateurs de Tendance centrale

Les mesures de **tendance centrale** permettent de résumer un ensemble de données relatives à une variable quantitative. Elles permettent de déterminer une valeur «typique» ou **centrale** autour de laquelle des données ont **tendance** à se **rassembler**.

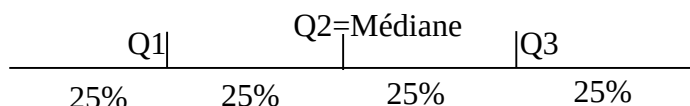
1. Moyenne : L'indicateur le plus couramment utilisé est la moyenne empirique ou moyenne arithmétique. L'inconvénient principal de la moyenne empirique comme indicateur de tendance centrale est d'être assez sensible à la présence de valeurs «aberrantes». Un indicateur de tendance centrale plus robuste est donné par les moyennes tronquée : on retire un certain nombre de valeurs à la fin comme au début (ce sont les valeurs extrêmes).

$$\bar{x} = \frac{1}{n} \sum x_i$$

2. Quartiles : Les quantiles permettent de donner des indications du type «1 personne sur 10 a moins de tel âge».

La médiane est un indicateur de tendance centrale (plus robuste que la moyenne empirique) qui divise la population en deux parties qui ont le même nombre d'individus. Autrement dit-elle sépare l'échantillon en deux parties égales.

3. Médiane : Si la série est de taille impaire la médiane est la donnée située au milieu de la liste (son rang est $\frac{N+1}{2}$) Si la série est de taille paire on prendra comme médiane la moyenne des deux valeurs situées au milieu de la liste (c'est-à-dire celles de rang $N/2$ et $N/2 + 1$).



4. Mode et classe modale. Le mode d'une série statistique est la valeur la plus fréquente. Dans le cas d'une répartition en classe la classe la plus fréquente sera dite modale.

Indicateurs de dispersion

Comme le nom l'indique les indicateurs de dispersions permettent de mesurer comment les données se «répartissent».

On peut définir deux types de mesure de dispersions :

- Les mesures définies par la distance entre deux valeurs représentatives de la distribution.
- Les mesures calculées en fonction de la déviation par rapport à une valeur centrale.

1.Étendue :

L'étendu d'une série statistique est l'écart entre sa plus grande valeur et sa plus petite.

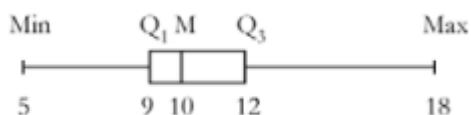
Ce dernier indicateur est très peu robuste. On lui préférera souvent l'intervalle interquartile :

2. Intervalle interquartile :

L'intervalle interquartile est la différence entre le troisième et le premier quartile. On peut remarquer que cet intervalle contient 50% des données. Un premier moyen de mesurer la dispersion des données autour de la moyenne est l'écart moyen absolu.

Diagramme en boîte : Dans un repère orthonormé, on trace trois traits verticaux de centres alignés, dont les abscisses sont Min, Q₁, Me, Q₃ et Max. on fera une boîte de Q₁ à Q₃ qu'on relira aux traits du Min et du Max D'un seul coup d'œil je peux voir indicateurs position me faire une idée concernant l'intervalle interquartile. Entre deux traits verticaux (Min, Q₁, Me, Q₃ et Max) il y a environ 25% de l'effectif.

Exemple



3. Variance empirique et Ecart type :

On appelle variance empirique de la série statistique X la moyenne des carrés des écarts des valeurs par rapport à la moyenne.

La variance :

Pour une population $\text{Var}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2$

Pour un échantillon $\text{Var}(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)$

L'écart-type

Définition écart type. L'écart type noté est la racine carrée de la variance, il permet de donner «une sorte» de moyenne entre les valeurs et leur moyenne.

$$\sigma = \sqrt{\text{var}(x)} = \sqrt{V(X)}$$

L'écart-type représente la dispersion des données autour de la moyenne

Variance et Espérance :

L'espérance d'une variable aléatoire est, lorsqu'elle existe, la moyenne des valeurs de cette variable, pondérées par leurs probabilités de réalisation. On voit bien comment traduire cette définition informelle dans le cas d'une variable aléatoire discrète X en posant :

$$E(x) = \sum x_i \cdot p_i$$

La variance de X vaut, si on a posé $m = E(X)$:

$$V(x) = p_1(x_1 - m)^2 + \dots + p_n(x_n - m)^2 = E(X - m)^2$$

Exemple :

x _i	-2	0	3	7
p _i	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

$$E(x) = -2 \cdot \frac{2}{8} + 0 \cdot \frac{3}{8} + 3 \cdot \frac{2}{8} + 7 \cdot \frac{1}{8} = \frac{9}{8} = 1.125$$

$$V(x) = E(x - 1.125)^2 = \frac{2}{8}(-2 - 1.125)^2 + \frac{3}{8}(0 - 1.125)^2 + \frac{2}{8}(3 - 1.125)^2 + \frac{1}{8}(7 - 1.125)^2 = 8.109375$$

p _i	x	x _i -1.125	(x _i -1.125) ²	p _i *(x-1.125) ²
0.25	-2	-3.125	9.765625	2.44140625
0.375	0	-1.125	1.265625	0.474609375
0.25	3	1.875	3.515625	0.87890625
0.125	7	5.875	34.515625	4.314453125
1	Var(x)			8.109375

Covariance

Pour une série statistique à 2 variables x_i, y_i :

Pour une population
$$\text{Cov}(x) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}$$

Pour un échantillon

$$\text{Cov}(x) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Ou
$$\text{Cov}(x) = \frac{1}{n-1} (\sum x_i y_i - n \bar{x} \bar{y})$$

Application :

Soit la série statistique à deux variables :

x_i	10	11	14	13	12
y_i	25	5	15	10	20

Calculer la covariance ?

$$\bar{x} = 12$$

$$\bar{y} = 15$$

$$\text{Var}(x) = 2.5$$

$$\text{Var}(y) = 62.5$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
10	25	-2	10	-20
11	5	-1	-10	10
14	15	2	0	0
13	10	1	-5	-5
12	20	0	5	0
				-15

somme	60	75
moyenne	12	15

$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
4	100
1	100
4	0
1	25
0	25

10	250
2.5	62.5

Var

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5-1} (-15) = -\frac{15}{4} = -3.75$$

Ou d'une autre façon :

	xi	yi	xi*yi
	10	25	250
	11	5	55
	14	15	210
	13	10	130
	12	20	240
somme	60	75	885
moyenne	12	15	

$$\text{Cov}(x, y) = \frac{1}{n-1} (\sum x_i y_i - n \bar{x} \bar{y}) = \frac{1}{5-1} (885 - 5 * 12 * 15) = \frac{1}{4} (885 - 900) = -\frac{15}{4} = -3.75$$