Statistiques, cours, 2nde

F.Gaudon

13 juin 2017

Table des matières

1 Vocabulaire et notations

Définitions et notations :

On considère une série statistique.

- Si la série statistique comporte p valeurs distinctes avec $p \in \mathbb{N}$(c'est à dire p entier naturel) , les valeurs du caractère étudiées sont notées x_i pour i entier naturel allant de 1 à p. Le nombre d'individus pour la valeur x_i, c'est à dire l'effectif pour la valeur x_i est noté n_i.
- Si la série est regroupée en p classes $[a_i; a_{i+1}[$ pour i allant de 1 à p, où a_i sont des réels tels que $a_i < a_{i+1}$, on prend pour valeurs les centres des classes $c_i = \frac{a_i+a_{i+1}}{2}$ pour i allant de 1 à p.
- L'effectif total N est alors égal à $N = n_1 + n_2 + ... + n_p$ ce que l'on note aussi $\sum_{i=1}^{i=p} n_i = N$.

Exemple :

On étudie la hauteur de jeunes plants de riz lors d'une expérience sur une nouvelle variété. La population est alors l'ensemble des plants de riz. Un individu est donc un plant de riz. Le caractère étudié est la hauteur des plants. C'est un caractère quantitatif, c'est à dire un nombre réel. Les valeurs distinctes du caractère x_i sont les différentes hauteurs relevées. Il peut, par exemple y avoir une hauteur $x_1 = 8$ cm pour un effectif de $n_1 = 2$ plants qui ont cette hauteur. L'effectif total N est le nombre total de plants de riz. On peut regrouper les valeurs en classes, par exemple, la classe des plants qui ont une hauteur dans l'intervalle $[20; 25[$ en cm. Le centre de l'intervalle est alors $c_1 = \frac{20+25}{2} = 22,5$.

2 Fréquences, séries cumulées

Définition :

On appelle *fréquence* f_i de la série pour la valeur x_i le nombre réel défini par

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

qui s'exprime aussi en pourcentage en multipliant par 100.

Exemple :

Dans l'exemple des plants de riz, s'il y a 3 plants de taille 18 cm parmi un effectif total de 28 plants, la fréquence pour la valeur $x_i = 18$ est $f_i = \frac{3}{28} \approx 0,107$ soit environ 10,3% de plants de hauteur 18 cm.

Définition :

On appelle :

- *effectif cumulé croissant (ECC)* pour la valeur x_i, la somme des effectifs des valeurs inférieures ou égales à x_i ;
- *fréquence cumulée croissante (FCC)* pour la valeur x_i, la somme des fréquences des valeurs inférieures ou égales à x_i.

Exemple :

On a relevé la hauteur de plants de riz d'une nouvelle variété lors d'une expérimentation :

Hauteur en cm (x_i)	8	12	14	16	17	18	19	20	
Nombre de plans (n_i)	2	2	4	2	2	3	3	4	
Fréquences (f_i)	$\frac{2}{28} \approx 0,0714$	0,0714	0,1429	0,0714	0,0714	0,1071	0,1071	0,1429	0,
Effectifs cumulés croissants	2	4	8	10	12	15	18	22	

3 Caractéristiques d'une série statistique

3.1 Moyenne

Définition :

- Soit x_i les p valeurs distinctes d'une série statistique et n_i les effectifs pour chaque valeur. La *moyenne* pondérée, notée $\bar{x}$, est donnée par :

$$\bar{x} = \frac{n_1x_1 + n2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p}$$

ce qui s'écrit aussi $\bar{x} = \frac{\Sigma_{i=1}^{i=p} n_i x_i}{\Sigma_{i=1}^{i=p} n_i}$

- Dans le cas d'une série où les effectifs des p valeurs x_i sont égaux à 1 (on parlera alors par abus de langage de série "sans effectif") la moyenne non pondérée est donc :

$$\bar{x} = \frac{x_1 + x_2 + ... + x_p}{p}$$

Exemple :

Dans l'exemple des hauteurs de plants de riz, la moyenne est :
$\bar{x} = \frac{2\times8+2\times12+...4\times21+2\times22}{2+2+...+4+4+2} \approx 17,2$ cm

Propriété :
On considère la *distribution des fréquences* d'une série statistique c'est à dire l'ensemble des fréquences des valeurs de la série :

x_1	x_2	$\dots$	x_p
f_1	f_2	$\dots$	f_p

Alors, la moyenne de la série statistique est donnée par :

$$\bar{x} = f_1x_1 + f_2x_2 + ... + f_px_p$$

Preuve :

On a :
$\bar{x} = \frac{n_1x_1}{N} + \frac{n_2x_2}{N} + ... + \frac{n_px_p}{N} = \frac{n_1}{N}x_1 + \frac{n_2}{N}x_2 + ... + \frac{n_p}{N}x_p$
On obtient le résultat en remplaçant $\frac{n_i}{N}$ par f_i pour tout i allant de 1 à p.

Exemple :

Dans l'exemple précédent, on a :
$f_1x_1 + f_2x_2 + \dots + f_px_p \approx 17,2$

3.2 étendue

Définition :
L'étendue d'une série statistique est la différence entre la plus grande valeur et la plus petite valeur de la série.

Exemple :
Dans l'exemple précédent l'étendue est l'étendue est $22 - 8 = 14$ cm.

3.3 Médiane

Définition :

La *médiane* est une valeur du caractère qui sépare la série statistique *ordonnée* en deux sous séries de même effectif.

Méthode de détermination :

- Dans le cas d'un caractère discret d'effectif total N impair, la médiane est la valeur de rang $\frac{N+1}{2}$;
- dans le cas d'un caractère discret d'effectif total N pair, la médiane est la demi-somme des valeurs de rang $\frac{N}{2}$ et de rang $\frac{N}{2}+1$.

Exemples :

- Série avec *effectifs égaux à 1* et effectif total pair : soit la série statistique dont les données sont :
  2 ; 2 ; 3 ; 5 ; 7 ; 7 ; 8 ; 9.
  L'effectif total est $N = 8$. Il est pair donc la médiane est la demi-somme entre les valeurs de rang 4 et 5.
  La médiane est donc $\frac{5+7}{2} = 6$.
- Série avec *effectifs non tous égaux à 1* : on reprend l'exemple des hauteurs de plants de riz.
  On obtient le tableau *d'effectifs* suivant :

Hauteur en cm (x_i)	8	12	14	16	17	18	19	20	21	22
Nombre de plans (n_i)	2	2	4	2	2	3	3	4	4	2
Effectifs cumulés croissants	2	4	8	10	12	15	18	22	26	28

Il y a 9 valeurs distinctes mais 28 données au total. L'effectif total est donc N = 28.
N est pair. La médiane est donc la demi-somme entre les valeurs de rang 14 et 15.
D'après le tableau des effectifs cumulés, les valeurs de rang 14 et 15 sont 18 et 18 donc la médiane est $\frac{18+18}{2} = 18$ cm.

3.4 Quartiles

Définition :

- Le premier quartile noté Q_1 de la série statistique est la plus petite valeur telle qu'au moins 25 % des valeurs de la série lui soient inférieures ou égales ;
- le troisième quartile noté Q_3 de la série statistique est la plus petite valeur telle qu'au moins 75 % des valeurs de la série lui soient inférieures ou égales.

Détermination pratique :

On suppose la série ordonnée dans l'ordre croissant des valeurs du caractère. Soit N l'effectif total.

- Si $\frac{N}{4}$ est un entier alors Q_1 est la valeur de rang $\frac{N}{4}$ et Q_3 est la valeur de rang $\frac{3N}{4}$;
- si $\frac{N}{4}$ n'est pas un entier, alors Q_1 est la valeur dont le rang suit le rang $\frac{N}{4}$ et Q_3 est la valeur dont le rang suit le rang $\frac{3N}{4}$.

Exemple :

Hauteur en cm (x_i)	8	12	14	16	17	18	19	20	21	22
Nombre de plans (n_i)	2	2	4	2	2	3	3	4	4	2
Effectifs cumulés croissants	2	4	8	10	12	15	18	22	26	28

L'effectif total N est 28.

$\frac{N}{4} = 7$ donc le premier quartile est la valeur de rang 7 c'est à dire $Q_1 = 14$ cm.

$\frac{3N}{4} = 21$ donc le troisième quartile est la valeur de rang 21 c'est à dire $Q_3 = 20$ cm.

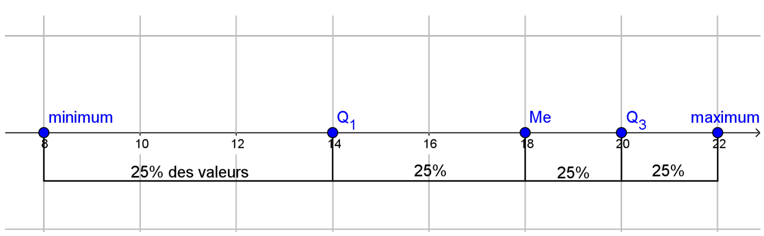
Définition :

On appelle :

- *écart inter-quartiles* la différence $Q_3 - Q1$;
- *intervalle inter-quartiles* l'intervalle $[Q_1; Q_3]$.

Visualisation :

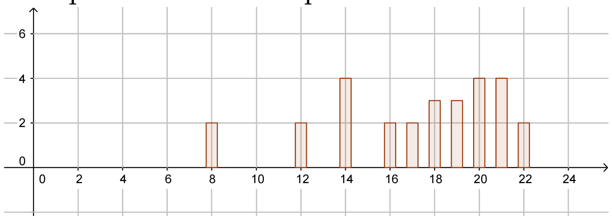
On peut visulaiser la répartition des valeurs en pourcentages de la série sur une droite graduée :

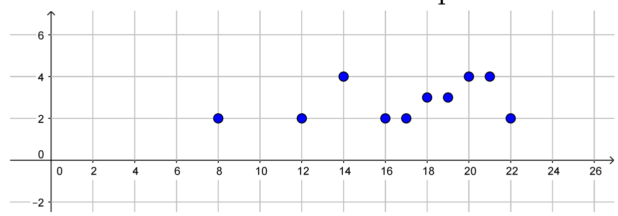


4 Représentations graphiques

4.1 Diagramme en barres ou en bâtons, nuage de points

Il permettent de représenter les valeurs et les effectifs d'une série statistique avec effectifs :





4.2 Polygone des effectifs cumulés croissants

Il permet de représenter les effectifs cumulés et donc de déterminer graphiquement la médiane et les quartiles de la série.