Cours Biostatistique L2 Biologie

Statistique descriptive

Année universitaire 2019/2020

➢ **Le centre de classe :**

$$C_i = \frac{(a_i + a_{i+1})}{2}$$

Le nombre de classe

➢ **La règle de STURGES :**

a) Cette règle est utilisée pour déterminer le nombre de classe à utiliser pour représenter une variable statistique continue. Le tableau regroupe en classe est souvent appelé **distribution groupée.**

Le nombre de classe est égal à l'entier le plus proche de la quantité $1 + 3.3\log(n)$(cours probabilité-statistique L2 Biologie, USDB, 2010/2011).

Exemple : pour un échantillon de taille N=200, on doit utiliser$1 + 3.3\log 200 = 1 + (3.3 * 2.3) = 8.59 = 9\ classe$.

L'amplitude : constante de ces classes sera égale à :

$$a = \frac{X_{max} - X_{min}}{Nombre\ de\ classe}$$

➢ **Règle de Yule: J=2.5$\sqrt[4]{n}$**

4) Caractéristiques Numériques D'une Série Quantitative

4.1) Caractéristiques de Position

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser l'ordre de grandeur des observations ***(Saïd Chermak ;2012)***.

a) La moyenne

1) La moyenne arithmétique : La moyenne arithmétique est notée $\hat{X}$

- Série brute $X_1, X_2, \ldots\ldots\ldots, X_n$

$$\hat{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Série groupée

Valeurs de la variable	Effectifs	Fréquence
x_1	n_1	$f_1 = n_1/n$
…….	……	……
x_i	n_i	$f_i = n_i/n$
…….	……	…….
x_k	n_k	$f_k = n_k/n$

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

- Série classée

classe	Effectifs	Fréquences	Centres de classe
$[e_1 - e_2[$	n_1	f_1	$X_1 = (e_1 + e_2)/2$
$[e_2 - e_3[$	n_2	f_2	$X_2 = (e_2 + e_3)/2$
……..	……	……..	……..
$[e_k - e_{k+1}[$	n_k	f_k	$X_k = (e_k + e_{k+1})/2$

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

2) Moyenne géométrique

Utilisée dans le cas de phénomènes multiplicatifs (taux de croissance moyen)

$$\boldsymbol{G = \sqrt[n]{x_1^{n_1}\, X_2^{n_2} \ldots\ldots\ldots\ldots X_k^{n_k}}}$$

3) Moyenne Harmonique

$$\boldsymbol{H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}}$$

4) Moyenne quadratique

$$\boldsymbol{q = \sqrt[2]{\sum \frac{n_i {x_i}^2}{n}}}$$

b) Le mode : désigné par **Mo** est la valeur de la variable statistique la plus fréquente. Le mode correspond à la classe de fréquence maximale dans la distribution des fréquences. On peut identifier le mode comme la valeur médiane de la classe de fréquence maximale ou bien effectuer une interpolation linaire pour obtenir la valeur exacte du mode comme suit ***Saïd Chermak ;2012)***.

$$M_o = x_m + \frac{i\Delta i}{\Delta s + \Delta i}$$

Avec :

X_m **:** limite inférieure de la classe d'effectif maximal.

i **:** intervalle de classe ($x_{m+1} - x_m$)

Δi : Ecart d'effectif entre la classe modale et la classe inférieure la plus proche.

Δs : Ecart d'effectif entre la classe modale et la classe supérieure la plus proche.

Exemple:

Caractère X : x_i : longueur de la rectrice bornes des classes	[140-145[	[145-150[	[150-155[	[155-160[	[160-165[	[165-170[	[170-175[
Valeur médiane des classes, x_i	142,5	147,5	152,5	157,5	162,5	167 ,5	172,5
n_i : nombre d'individu par classe de taille x_i	1	1	9	17	16	3	3
f_i : fréquence relative	0,02	0,02	0,18	0,34	0,32	0,06	0,06
$f_i cum.$: fréquence relative cumulée	0,02	0,04	0,22	0,56	0,88	0,94	1

(Saïd Chermak ;2012)

Dans le cas de la distribution de la longueur de la rectrice centrale de la gélinotte huppée, la valeur du mode est :

Valeur approchée : La classe de fréquence maximale est [155,160[avec ni= 17 d'où Mo= 157,5 mm

Valeur exacte :

$$M_o = 155 + \frac{5*8}{(1+8)} = 159.44$$ D'où Mo= 159,4 mm

Avec x_m= 155 mm, Δi= 17-9 = 8, Δs = 17-16 = 1 et i= 5mm.

Remarque :

- Une distribution de fréquences peut présenter un seul mode (distribution **unimodale**) ou plusieurs modes (distribution **bi** ou **trimodale**).
- Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur de la moyenne arithmétique.

  Mo $\approx \hat{x}$.

c) La médiane

La médiane, **Me**, est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0,5 ou 50%. Elle correspond donc au centre de la série statistique classée par ordre croissant, ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures ***(Saïd Chermak ;2012)***.

Dans le cas où les valeurs prises par le caractère étudié **ne sont pas regroupées en classe,**

- **Si n est impair :** Me est le terme de rang $\left(\frac{N+1}{2}\right)$
- **Si n est pair :** la médiane est le centre de l'intervalle formé par le rang $\frac{N}{2}$ **et** $\frac{N}{2} + 1$

Dans le cas où les valeurs prises par le caractère étudié sont groupées en classe, on cherche la classe contenant le $n^e/2$ individu de l'échantillon.

$$M_e = x_m + (x_{m+1} - x_m)\left(\frac{\frac{n}{2} - N_i}{n_i}\right)$$

Avec

x_m : limite inférieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).

x_{m+1} : limite supérieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).

n_i : effectif de la classe médiane

N_i : effectif cumulé inférieur à x_m

n : taille de l'échantillon

N.B il faut ordonner dans l'ordre croissant ou décroissant.

Exemple :

Dans le cas de la distribution de la longueur de la rectrice centrale de la gélinotte hupée, la valeur de la médiane est :

- **Cas des données non groupées:**

n= 50 donc Me ∈[x25, x26]

soit Me ∈[158mm, 159mm] ou Me=158,5mm

➢ **Cas des données groupées :**

n=50, la 25^{éme} valeur se situe dans la classe [155-160[qui contient les individus de 12 à 28. d'où avec L_m= 155 mm, f_m= 17 individus, $f_m$cum. = 11 individus et i = 5mm

$$M_e = 155 + \frac{5}{17}\left(\frac{50}{2} - 11\right) = 159.11$$

N.B : Si la distribution des valeurs est symétrique, la valeur de la médiane est prochede la valeur de la moyenne arithmétique. Me ≈$\hat{x}$.

d) Quartiles : ce sont les valeurs de X_i qui partagent la série statistique a quatre parties égalées.

- **Le premier Quartile** d'une série statistique est la plus petite valeur Q_1 telle qu'au moins 25 % des valeurs sont inférieures ou égales à Q_1.
- **Le troisième Quartile** d'une série statistique est la plus petite valeur Q_3 telle qu'au moins 75 % des valeurs sont inférieures ou égales à Q_3.
- **Déciles et percentiles**

Les 9 déciles sont les nombres réels qui partagent l'étendue en dix intervalles de même effectif.

Les 99 percentiles sont les nombres réels qui partagent l'étendue en cent intervalles de même effectif. (cours probabilité-statistique L2 Biologie, USDB, 2010/2011). .

4.2) Caractéristiques de Dispersion

a) Etendue : E = $x_{max} - x_{min}$

b) Intervalle interquartile : IQ = Q_3 - Q_1

c) Variance

- **Série brute :** $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$
- **Série groupée ou classée** $S^2 = \frac{1}{n}\sum_{i=1}^{n} n_i(x_i - \hat{x})^2 = \sum_{i=1}^{n} f_i(x_i - \hat{x})^2$

d) Ecart-type : $s = \sqrt{S^2}$

e) Coefficient de variation :

$$C_V = \frac{s}{\bar{x}} * 100$$

Le CV permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations. Il donne une bonne idée du degré d'homogénéité d'une série. Il faut qu'il soit le plus faible possible (<15% en pratique)

4.3) Paramètres de forme

a) Symétrie

- **Coefficient d'asymétrie de Pearson**

$$\delta = \frac{\bar{X} - Me}{s}$$

On a $-1 \leq \delta \leq 1$

$\delta = 0$ symétrie parfaite

$\Delta < 0$ série étalée à gauche

$\delta > 0$ série étalée à droit

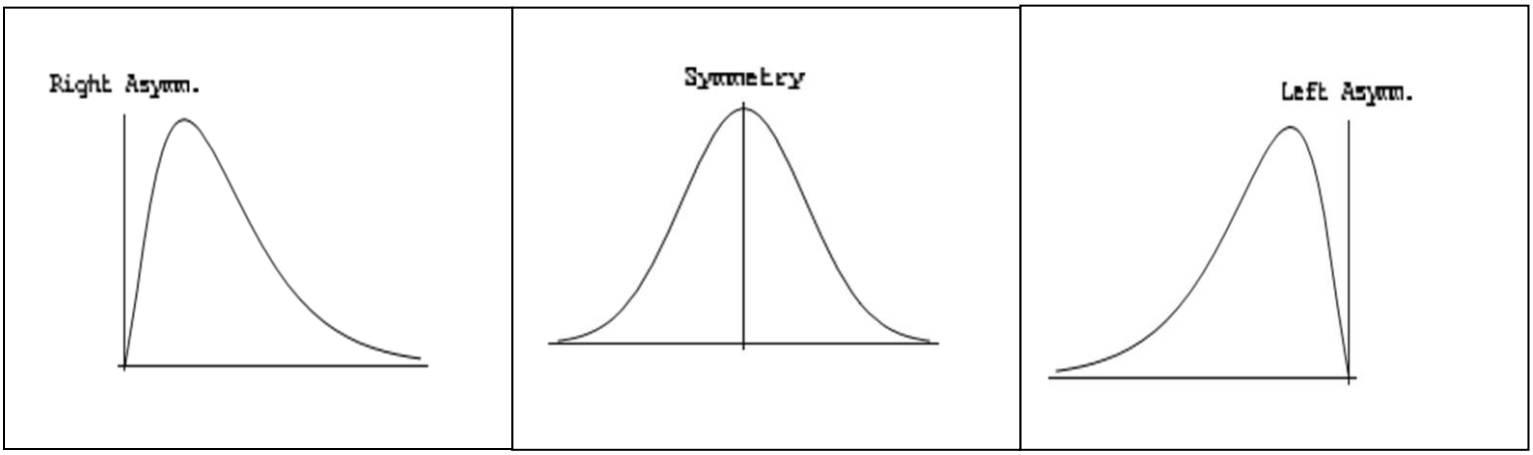
- **Coefficient de Yule**

$$q = \frac{Q_3 + Q_1 - 2Me}{Q_3 - O_1}$$

q =0⇒ symétrie parfaite

q <0⇒ série étalée à gauche

q >0 ⇒ série étalée à droite



(Saïd Chermak ;2012).

b) **Aplatissement**

Une distribution est plus ou moins aplatie selon que les fréquences des valeurs voisines des valeurs centrales diffèrent peu ou beaucoup les unes par rapport aux autres.

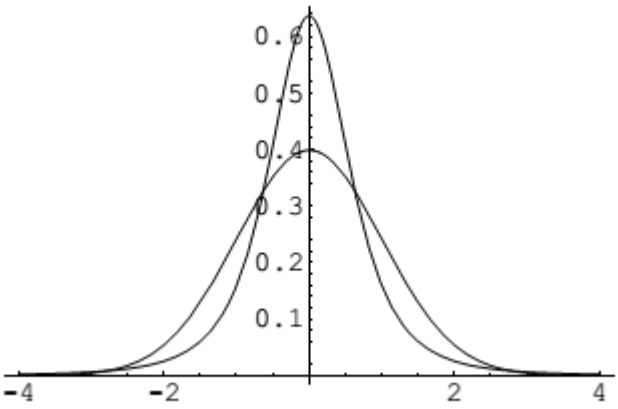
- **Coefficient d'aplatissement de Fisher**

$$a = \frac{m_4}{\sigma} \qquad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

a =3 pour une distribution qui suit une loi normale centrée réduite. *f*

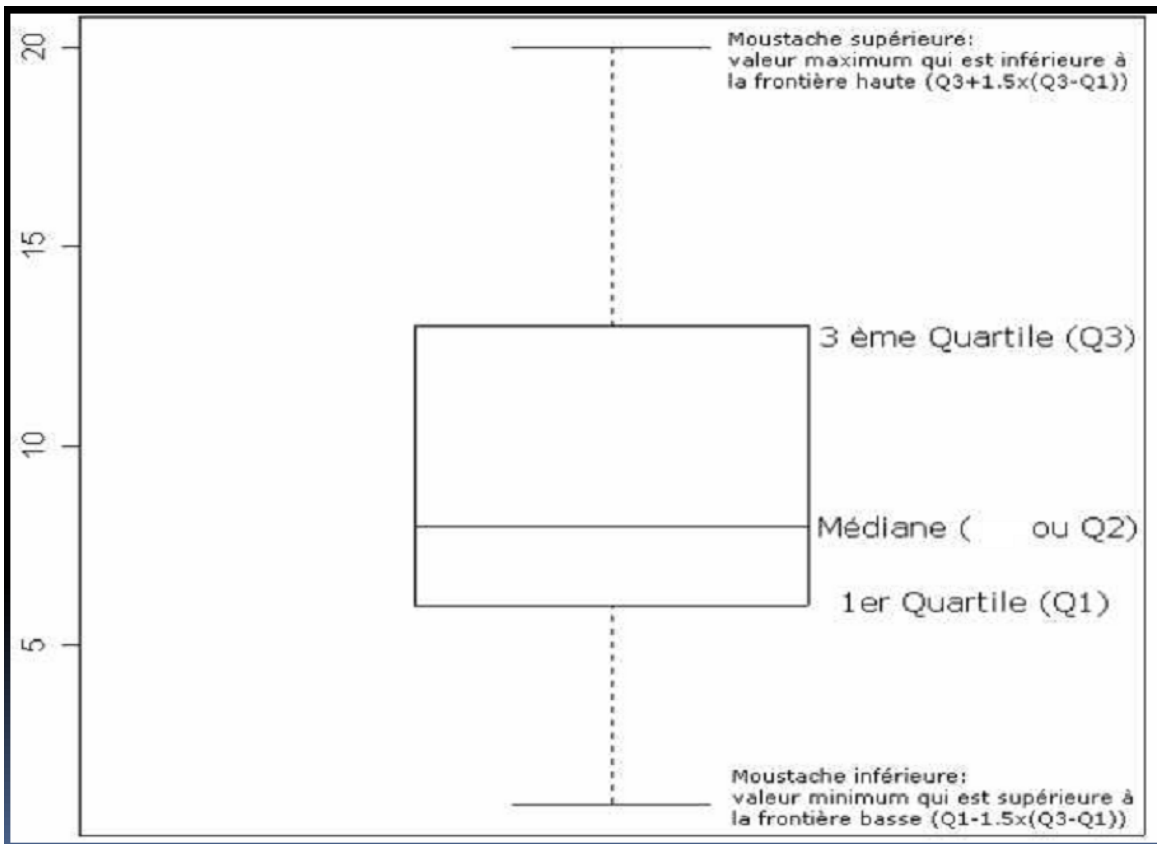
Si a >3, la concentration des valeurs de la série autour de la moyenne est forte: la distribution n'est pas aplatie *f*

Si a<3, la concentration des valeurs autour de la moyenne est faible: la distribution est aplatie



5) **La boîte à moustaches :** est un diagramme simple qui permet de représenter la distribution d'une variable.

Construction :



(www.facebook.com/DomaineSNV/).

On repère sur la boîte à moustaches d'une variable:

- l'échelle des valeurs de la variable, située sur un axe vertical ou horizontal.
- la valeur du 1er quartile Q1 (25% des effectifs), correspondant au trait inférieur de la boîte,

- la valeur du 2ème quartile Q2 (50% des effectifs), représentée par un trait à l'intérieur de la boîte,
- la valeur du 3ème quartile Q3 (75% des effectifs), correspondant au trait supérieur de la boîte,
- les 2 « moustaches» inférieure et supérieure, représentées ici par des traits de part et d'autre de la boîte. Ces 2 moustaches, délimitent les valeurs dites adjacentes qui sont déterminées à partir de l'écart interquartile (Q3-Q1).
- les valeurs dites extrêmes, atypiques, exceptionnelles, (outliers) situées au-delà des valeurs adjacentes sont individualisées. Elles sont représentées par des marqueurs (carré, ou étoile, etc.).