

Statistique Descriptive

I- Notions préliminaires en bio-statistique

1. Définitions

1.1. Bio-statistique

On s'intéresse à des caractères ou grandeurs **biologiques** mesurées chez les êtres vivants dont la particularité principale est la **variabilité**.

1.2. Variable

Variable = caractère :

C'est une caractéristique ou un facteur susceptible de prendre **une valeur différente** selon les individus étudiés.

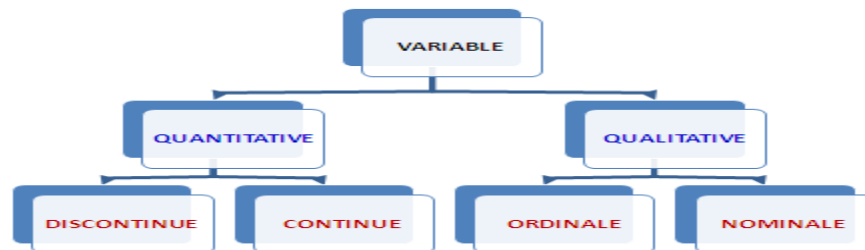
Ex : taille, couleur des cheveux, âge gestationnel...

1.2.1. Modalités : Ce sont les **différentes catégories** que peut présenter une variable

Exemple :

Variables	Modalités
Sexe	masculin, féminin
Fréquence cardiaque	50,75,89,95,105 b/min.
Glycémie	0.65, 1.6, 2.8 g/l
Durée d'incubation d'une maladie	24h, 7 jours, 40 ans.

1.2.2. Les types de variable



a- Variable quantitative

Les modalités s'expriment par des **valeurs numériques**.

Exemples : poids, taille, glycémie, tension artérielle, fréquence cardiaque, nombre d'étudiants par amphi.

a-1- Variable quantitative discontinue

Appelée également **discrète** : les modalités s'expriment par des **nombre entiers**.

Exemple:

Fréquence cardiaque

Nombre de métastases : 0,1,2,3

Rappels de vaccin : 1, 2, 3 injections

a-2- Variable quantitative continue

Peuvent prendre n'importe quelle valeur numérique dans un intervalle d'observations.

- L'ensemble des valeurs possibles appartient à l'ensemble des **nombre réels**.

Exemple:

Taille: 1.72 m

Poids: 60 kg

Age: 24 ans

Cholestérol: 2.22 g/l

b- Variable qualitative

- C'est un caractère dont les modalités s'expriment par **des qualités**.

- **Non mesurable.**

Exemple:

Sexe: masculin, féminin

Groupe sanguin: A, B, AB, O

Maladie: malade, non malade

Profession

b-1- Variable qualitative ordinale

Les modalités sont **ordonnées** selon une échelle de valeur.

Exemples:

Satisfaction du malade (non satisfait, peu satisfait, satisfait, très satisfait)

Intensité de la douleur (légère, moyenne, forte, très forte)

Niveau d'étude (primaire, secondaire, supérieur)

b-2- Variable qualitative nominale

L'ordre des modalités n'est pas important.

Elles sont nommées mais **pas ordonnées**.

Exemples:

Groupe sanguin : A, O, B, AB

Maladie: malade, non malade

Accident : domestique, circulation, sportif

Etat civil : célibataire, marié, divorcé.

1.3. Effectif = Fréquence Absolue

C'est le **nombre d'individus** appartenant à **une modalité donnée**.

Exemple :

La distribution de 40 étudiants selon le sexe.

15 de sexe **masculin** et **25** de sexe **féminin**.

Les effectifs correspondant à chacune des deux modalités sont **15** et **25**.

1.4. Fréquence relative

C'est le **rapport** entre l'effectif d'une modalité de la variable étudiée et la taille de l'échantillon ou de la population étudiée.

Le numérateur fait obligatoirement parti du dénominateur.

La fréquence relative s'exprime en pourcentage.

Exemple :

Pour l'exemple précédent de la distribution de **40** étudiants selon le sexe.

Sexe **masculin** = **15**

Sexe **féminin** = **25**

Calculez la fréquence relative pour les deux sexes:

Fréquence relative (sexe masculin) = $15/40 = 0.375 = 37.5\%$

Fréquence relative (sexe féminin) = $25/40 = 0.625 = 62.5\%$

1.5. Effectifs et fréquences cumulés

- Lorsque les classes d'une variable sont ordonnées (Variable quantitative ou qualitative ordinale) on peut ajouter à l'effectif de chaque classe le total des effectifs des classes inférieures.
- On obtient ainsi **les effectifs cumulés**.
- **Les fréquences cumulées** sont obtenues en divisant les effectifs cumulés par le total de la série.

Exemple : Répartition de 30 sujets selon la composition de leurs fratries

Fratrie	Effectif	Fréquence (%)	Effectif cumulé	Fréquence cumulé (%)
1	11	36,7	11	36,7
2	6	20,0	17	56,7
3	5	16,7	22	73,4
4	4	13,3	26	86,7
>4	4	13,3	30	100,0
Total	30	100,0	/	/

- L'effectif des sujets ayant une fratrie de 3 au plus est de: 22
- La fréquence cumulée des sujets ayant une fratrie de 3 au plus est de : 73,%

1.6. Ratio

C'est le rapport des effectifs de deux modalités d'une même variable.

Sex-ratio = (effectif des garçons)/(effectif des filles)

Exemple

Dans l'exemple précédent de la distribution de **40** étudiants selon le sexe.

Sexe **masculin** = **15**

Sexe **féminin** = **25**

Calculez le *sex-ratio* ? $Sex-ratio = 15 / 25 = 0.6$

1.7. Taux

Probabilité de survenue d'une maladie donnée au **cours du temps**.

S'exprime toujours en fonction d'une certaine **unité de temps**, pour **un lieu géographique donnée** et pour **un groupe de personnes bien défini**.

Numérateur : nombre d'événements (décès, maladie, handicap) survenus au cours d'une certaine période t1-t2.

Dénominateur: population exposée au risque de survenue de cet événement.

Exemple :

Si dans les **24** heures qui suivent un repas à une cantine fréquentée par **300** personnes, **30** présentent des signes d'intoxication alimentaire, le taux de la maladie est égal à : $30 / 300 = 0.10 = 10\% = 100.0$ pour mille

1.8. Série statistique

C'est l'**ensemble des valeurs** prises par une variable **quantitative**

Exemple : les valeurs du poids d'un groupe d'étudiants ou les durées de séjours d'un groupe de malade hospitalisés.

Comment présenter ces données ?

- Classement des données
- Tableaux de présentation
- Graphique
- Résumé des données par des paramètres statistiques (statistique descriptive)

II- Présentation tabulaire des données statistiques

A- Variable qualitative

Considérons l'exemple simple de la distribution de 40 étudiants selon le sexe.

Parmi ces 40 étudiants, 15 étaient de sexe masculin et 25 de sexe féminin.

Tableau 1 : Distribution des étudiants selon le sexe (lieu et temps)

Sexe	Effectif	%
Masculin	15	37,5
Féminin	25	62,5
Total	40	100

B- Variable quantitative

1. Variable quantitative discontinue

On va répartir 65 personnes selon la fratrie

Tableau 2 : Distribution de 65 personnes selon la fratrie (temps et lieu)

Nombre de frère	Effectif	%
0	11	16,9
1	9	13,8
2	12	18,5
3	20	30,8
4	13	20
Total	65	100

2. Variable quantitative continue

On regroupe les valeurs de la variable en classes successives, contigües et ne se recouvre pas.

Exemple

On a relevé la taille de 27 étudiants. L'unité de mesure retenue est le cm et les résultats sont les suivants : 165, 170, 156, 154, 162, 168, 159, 160, 161, 167, 156, 157, 160, 163, 170, 175, 178, 166, 165, 159, 179, 158, 167, 169, 165, 157, 155

Classement par ordre hiérarchique

154, 155, 156, 156, 157, 157, 158, 159, 159, 160, 160, 161, 162, 163, 165, 165, 165, 166, 167, 167, 168, 169, 170, 170, 175, 178, 179

Tableau 3 : Répartition de la taille de 27 étudiants lieu-temps

Taille (cm)	Effectif	%
[150-155[1	03,7
[155-160[8	29,6
[160-165[5	18,5
[165-170[8	29,7
[170-175[2	07,4
[175-180[3	11,1
Total	27	100

Chaque classe est définie par ses :

1.- **limites** : limite supérieure et limite inférieure

2.- **Valeur centrale** = centre de classe (**Limite supérieure + limite inférieure**)/2

3.- **Amplitude** (**Limite supérieure - Limite inférieure**)

III- Représentation graphique

A. Variable qualitative

1. Diagramme en barre

Les différentes modalités d'un caractère sont figurées par des rectangles ou des barres espacées.

La hauteur des barres doit être proportionnelle aux effectifs rencontrés (ou fréquences relatives).

Exemple :

- Distribution de 50 malades selon le sexe : 15 de sexe masculin et 35 de sexe féminin.
- Fréquence relative (sexe masculin) = $15/50 = 0.3 = 30\%$.
- Fréquence relative (sexe féminin) = $35/50 = 0.7 = 70\%$.

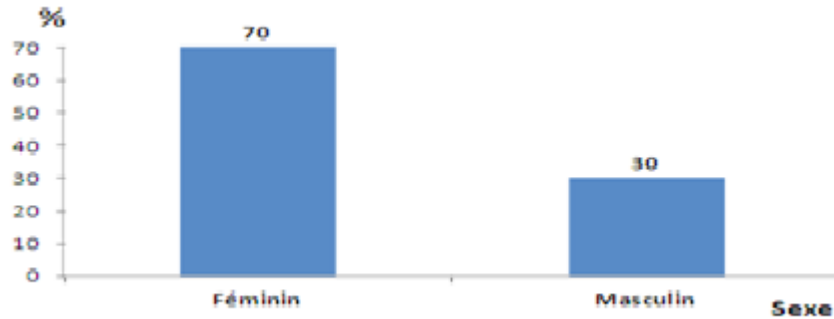


Figure 1 : Répartition des 50 malades selon le sexe (temps et lieu)

2. Diagramme circulaire

Les angles des secteurs sont proportionnels aux fréquences relatives des différentes modalités: 1 % est représenté par $3,6^\circ$.

Nombre de modalité ≤ 6

Exemple :

- Prenons l'exemple simple de la distribution de 50 malades selon le sexe.
- Parmi ces derniers, **15** sont de sexe **masculin** et **35** de sexe **féminin**.

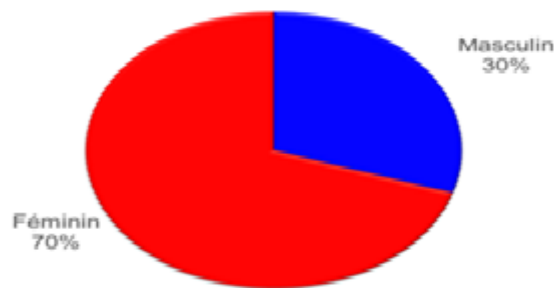


Figure 2 : Répartition des 50 malades selon le sexe (temps et lieu)

B. Variable quantitative

1. Variable quantitative discontinue

a. Diagramme en barre

Les différentes modalités d'un caractère sont figurées par des rectangles ou des barres espacées.

La hauteur des barres doit être proportionnelle aux effectifs rencontrés (ou fréquences relatives).

2. Variable quantitative continue

a. Histogramme :

Diagramme à barres accolées deux à deux.

Les bases doivent être proportionnelles à l'amplitude des classes.

La hauteur doit être proportionnelle à l'effectif de la classe correspondante.

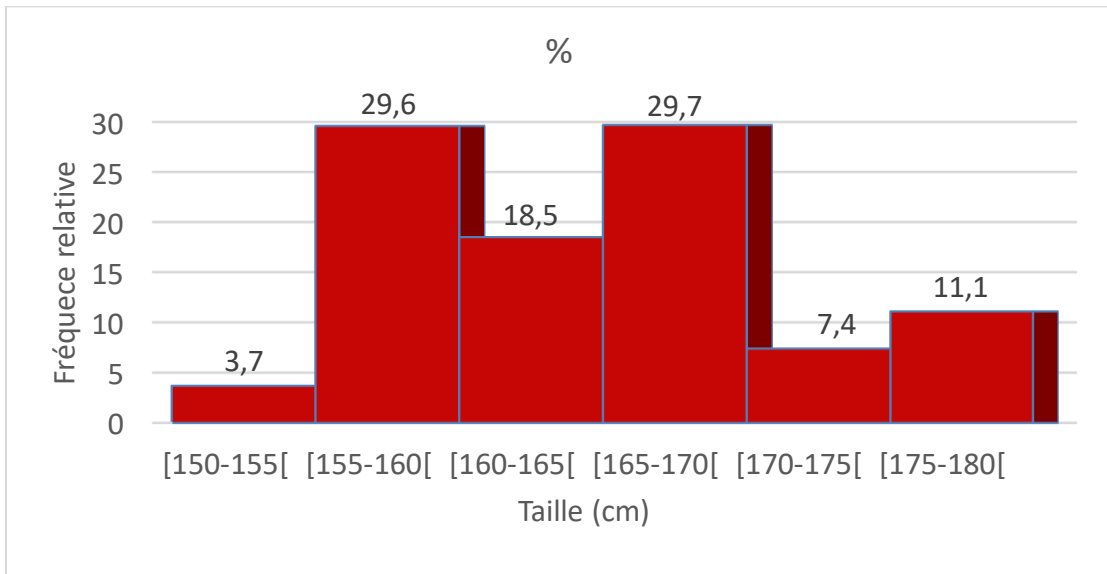


Figure 3 : Répartition de la taille de 27 étudiants (lieu-temps)

c- Polygone de fréquence

- Adapté à la représentation de la distribution d'une variable quantitative continue.
- En ordonnée figure les effectifs ou les fréquences.
- En abscisse figurent les valeurs de la variable quantitative.
- Chaque point du polygone représente l'effectif ou la fréquence pour le point central de la classe de la variable.
- La surface comprise sous le polygone représente 100% des observations.
- L'intérêt principal de ce type de graphe est pouvoir représenter sur **un même graphique plusieurs distributions**

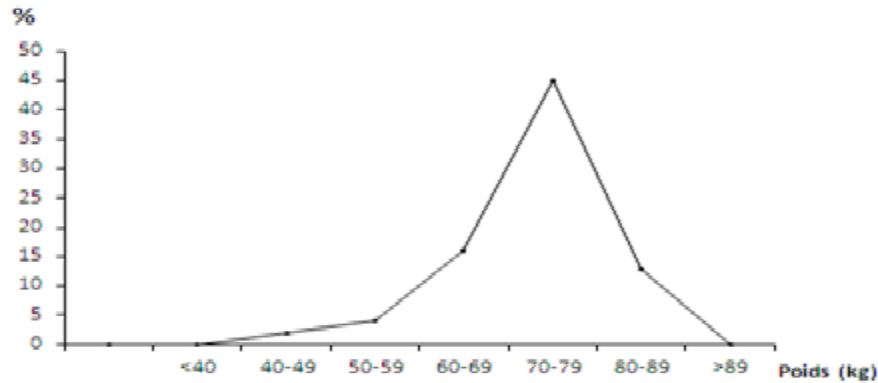


Figure 4 : Distribution du poids de 80 sujets (temps et lieu)

IV- Paramètres de réduction

1- **Paramètres de tendance centrale** : ce sont des mesures qui localisent le centre d'une distribution :

a- Moyenne arithmétique

a.1- Moyenne arithmétique simple : $m = (\text{somme des } x_i) / N$

Exemple : Calculez la moyenne arithmétique de la série statistique suivante:

3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 7, 9

Moyenne arithmétique simple

$$m = (3 + 3 + 3 + 3 + 4 + 4 + 5 + 5 + 5 + 5 + 7 + 9) / 12 = 56 / 12 = 4,7.$$

a.2.- Moyenne arithmétique pondérée : $m = (\text{somme } x_i * n_i) / N$

Exemple**Tableau 3 : Répartition de la taille de 27 étudiants lieu-temps**

Taille (cm)	Centre de classe	Effectif	%
[150-155[152,5	1	03,7
[155-160[157,5	8	29,6
[160-165[162,5	5	18,5
[165-170[167,5	8	29,7
[170-175[172,5	2	07,4
[175-180[177,5	3	11,1
Total	/	27	100

$$m = [(1*152,5) + (8*157,5) + (5*162,5) + (8 * 167,5) + (2*172,5) + (3*177,5)] / 27 = 164,54 \text{ cm}$$

2.- Mode

Appelé également la **valeur modale** ou **dominante** est la valeur de la variable (x_i) qui a l'effectif (n_i) le plus élevé.

Une série peut être unimodale, bimodale ou plurimodale.

Dans la distribution du nombre d'épisodes de syndrome grippal parmi 19 personnes, le mode est égal 1

Tableau 3 : Répartition de la taille de 27 étudiants lieu-temps

Taille (cm)	centre de classe	Effectif	%
[150-155[152,5	1	03,7
[155-160[157,5	8	29,6
[160-165[162,5	5	18,5
[165-170[167,5	8	29,7
[170-175[172,5	2	07,4
[175-180[177,5	3	11,1
Total	/	27	100

Cette série est bimodale

Lorsque les valeurs de la variable sont réparties en classes → **Classe modale**.

C- Médiane

C'est la valeur de la variable qui se trouve au milieu de la série quand les données observées sont rangées par ordre croissant ou décroissant.

Exemple

Déterminez la médiane de la série de la taille des 27 étudiants ?

-Si le nombre des valeurs est **impair**: la médiane correspond à la valeur de la variable dont le rang = $(N+1)/2 = 14$

Classement par ordre hiérarchique

154, 155, 156, 156, 157, 157, 158, 159, 159, 160, 160, 161, 162, **163**, 165, 165, 165, 166, 167, 167, 168, 169, 170, 170, 175, 178, 179

-Si le nombre des valeurs est **pair** : la médiane correspond à l'une des valeurs comprises entre les deux valeurs centrales observées.

On choisit généralement leur demi-somme.

Il est plus judicieux de parler d'intervalle médian.

Classement par ordre hiérarchique des tailles des 26 étudiants

154, 155, 156, 156, 157, 157, 158, 159, 159, 160, 160, 161, **162**, **163**, 165, 165, 165, 166, 167, 167, 168, 169, 170, 170, 175, 178

Lorsque les données sont groupées en classes, la médiane correspond si on utilise les effectifs cumulés, à $N/2$

2- Paramètres de position

a- Quartiles

Les quartiles divisent la série statistique en quatre parties égales comprenant le même nombre de sujets.
Deuxième quartile → Médiane

Si à titre d'exemple notre série statistique est composée de 100 sujets.

Le 1^{er} quartile (quartile inférieur) est la valeur de la variable du 25^{ème} sujet sur 100.

Le 2^{ème} quartile (médiane) est la valeur de la variable du 50^{ème} sujet sur 100.

Le 3^{ème} quartile (quartile supérieur) est la valeur de la variable du 75^{ème} sujet sur 100.

Une fois le rang déterminé, on recherche la valeur correspondante de la variable.

b- Déciles

Les déciles divisent la série statistique en 10 parties égales comprenant le même nombre de sujets.

Ils sont au nombre de 9.

Le 1^{er} décile est la valeur de la variable du 10^{ème} sujet sur 100.

Le 2^{ème} décile est la valeur de la variable du 20^{ème} sujet sur 100, etc.

Le 5^{ème} décile se confond avec le 2^{ème} quartile et la médiane.

c- Percentiles=centiles

Ils sont au nombre de 99.

Divisent la série statistique en 100 parties.

Le 30^{ème} percentile est la valeur de la variable du 30^{ème} sujet sur les 100 sujets.

3- Paramètres de dispersion

Ces paramètres essaient de synthétiser par une seule valeur numérique la dispersion de toutes les valeurs observées.

Considérons les deux séries suivantes:

Série 1: 15, 20, 25, 30, 35

Série 2: 5, 15, 25, 35, 45

Les valeurs de la série 2 sont plus dispersées autour de la moyenne

a- Variance

Paramètre statistique qui étudie la dispersion des observations par rapport à la moyenne.

En cas de données individuelles, la variance

s'écrit: $S^2 = [\text{somme des } (x_i - m)^2] / (N - 1)$

ou :

$$S^2 = [\text{somme } (x_i^2) - \{(\text{somme } x)^2 / N\}] / N$$

En cas de données groupées, la variance s'écrit :

$$S^2 = [\text{somme des } n_i (x_i - m)^2] / N$$

Ou :

$$S^2 = [\text{somme } n_i(x_i^2) - \{(\text{somme } n_i x_i)^2 / N\}] / N$$

b- Ecart type

L'écart type est la racine carrée de la variance : $S = \sqrt{S^2}$

Plus l'écart type est faible, plus les valeurs de la série sont concentrées (moins dispersées) autour de leur moyenne.

Exemple :

Calculez les deux **variances** des deux séries précédentes

$$S^2 (\text{Série 1}) = [(15-25)^2 + (20-25)^2 + (25-25)^2 + (30-25)^2 + (35-25)^2] / (5-1) = 62,5.$$

$$S^2 (\text{série 2}) = [(5-25)^2 + (15-25)^2 + (25-25)^2 + (35-25)^2 + (45-25)^2] / (5-1) = 250,0.$$

Les écarts types correspondants sont:

$$S (\text{série 1}) = \sqrt{62,5} = 7,9$$

$$S (\text{série 2}) = \sqrt{250,0} = 15,8$$

La série 2 est bien évidemment plus dispersée autour de la moyenne.

Le calcul de la variance et de l'écart type permet de quantifier cette dispersion

c- L'étendue

C'est la différence entre les deux valeurs extrêmes.

Etendue= valeur maximale – valeur minimale

d- Coefficient de variation

Le coefficient de variation est un indicateur combinant la moyenne et l'écart type.

Définition : On appelle coefficient de variation (CV), le rapport de l'écart sur la moyenne.

Calcul:

$$CV = s / m * 100$$

l'unité du CV est un nombre sans dimension. il est exprimé en pourcentage.

Propriété:

Le CV exprime le degré de dispersion d'une distribution en fonction de la valeur moyenne, il est utile pour comparer la dispersion de deux variable quantitatives de nature différente (unité différentes).

Références :

1. Zemour L, cours statistique descriptive, faculté de médecine université d'Oran, 2013
2. A.Bezzaoucha, Epidemiologie et Biostatistique: A l'usage des étudiants en sciences médicales, Office des Publications Universitaire
3. T.Ancelle, Statistique Epidémiologie, Maloine, 2002