Université Mohamed Khider de Biskra
Faculté des Sciences Exactes et SNV
Département de biologie

Module : Biostatistique

Niveau : Licence 2

Prof : Amel Chine

Année universitaire :2023/2024

Chapitre 1

Statistiques descriptives

1.1 Notions de base

Population : C'est un ensemble d'individus ou d'objets tels que : personnes, animaux, plantes et choses.

Echantillon : C'est un sous-ensemble de la population.

Unités ou objets statistiques : sont les éléments de la population ou de l'échantillon.

Caractère ou variable : C'est la caractéristique qui rassemble les membres de la population. On distingue deux types de variables :

- Variable qualitative : Ils n'ont aucune valeur numérique. Ils ont décrit par un mot ou une phrase par exemple : le groupe sanguin (A+,O+, ect), la couleur des yeux (marron, bleu,...), le sexe (homme, femme)
- Variable quantitative : Sont des variables qui sont comptées ou mesurées sur une échelle numérique. Ils ont décrit par un nombre par exemple : les résultats d'un examen (11,15,18,....), le nombre d'enfants dans une famille (0,1,2,..), la taille

des étudiants (160,175,170,....), température moyenne(18,20,25,...). Les données quantitatives peuvent être :

- Discrètes : Ce sont des variables qui sont comptées. Par exemple : nombre d'enfants dans une famille (0,1,2,..). nombre d'habitants dans une ville,.....
- Continues : Sont des variables numériques mesurées sur une échelle continue (intervalle). par exemple : taille des personnes ([160-165[, [165,170[, [170,175])

Modalités : Sont les valeurs prises par la variable, par exemple : nombre d'enfants par famille (0,1,2,4,3.) donc les modalités sont {0,1,3,4}.

Taille de la population ou taille de l'échantillon : C'est le nombre d'unités statistiques dans la population, noté N ou dans l'échantillon et noté n.

Série statistique : On appelle série statistique la séquence de valeurs prises par une variable X sur les unités statistiques. Les valeurs de la variable sont notées

$$X_1, X_2, ..., X_n$$

Example 1 *Dans une étude portant sur* 10 *familles, nous calculons le nombre d'enfants par famille. Les valeurs de la variable sont :*

$$0, 1, 1, 1, 2, 2, 3$$

donc : la variable X est le nombre d'enfants par famille et les modalités sont :{0,1,2,3}. Alors

$$X_1 = 0, X_2 = 1, X_3 = 2, X_4 = 3$$

et la taille de l'échantillon est $n = 10$.

1.2 Variable qualitative

Une variable qualitative a des valeurs distinctes qui ne peuvent pas être ordonnées. On note k le nombre de valeurs ou de modalités distinctes. On appelle effectif d'une modalité le nombre de fois que la modalité i est répétée. On note n_i l'effectif de la modalité X_i. La Fréquence notée f_i est l'effectif divisée par le nombre d'unités statistiques n. Il est défini par :

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

Example 2 *Soit la série statistique suivante sur l'état civil de* 30 *personnes*

M	M	D	C	M	C	V	D	V	D
D	M	M	M	M	C	C	C	M	M
D	D	C	V	C	M	M	M	D	M

ou M : marié (e) ;C : célibataire, D : divorcé and V : : veuf.

Donc la variable qualitative X est : l'état civil et ses modalités sont {M,C,D,V} et $k = 4$ (nombre de modalités). Tla taille de l'échantillon égale à $n = 30$.

Avec cette série statistique nous organisons le tableau de fréquence comme suit :

X_i	n_i	f_i	$f_i.100$
M	13	$\frac{13}{30} = 0.433$	43.3
D	7	$\frac{7}{30} = 0.233$	23.3
C	7	$\frac{7}{30} = 0.233$	23.3
V	3	$\frac{3}{30} = 0.1$	10

Nous concluons que le pourcentage le plus élevé est celui des personnes mariées avec 43,33%.

1.3 Variable quantitative

Comme nous l'avons vu dans la section précédente, une variable quantitative est une variable qui peut être mesurée et comptée et peut prendre deux types : discrète : la variable dans ce

cas est dénombrable et prend des valeurs isolées, ou continue : ses valeurs sont représentées par des intervalles appelés des classes.

1.3.1 Variable quantitative discrète

Effectif, fréquence et effectif cumulé croissant

1- **Effectif :** L'effectif noté n_i est le nombre de fois que les modalités sont répétées.

2- **Fréquence** : La Fréquence notée f_i est l'effectif divisée par la taille de l'échantillon ou le nombre d'unités statistiques. Il est défini par :

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

Remark 3 *1.*

$$\sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} f_i = 1.$$

2.

$$0 < f_i < 1$$

3- **Effectif cumulé :** L'effectif cumulé est l'effectif de la première modalité qui est ajoutée à l'effectif de la deuxième modalité, et cette somme est ajoutée à la troisième modalité et ainsi de suite, les effectifs obtenus de cette manière sont appelées effectif cumulé. L'effectif cumulé est utilisée pour connaître le nombre d'observations qui se situent au-dessus (ou en dessous) d'un effectif particulier dans des données statistiques. Il existe deux types d'effectif cumulé croissant et décroissant, noté par $N_i \nearrow$ et $N_i \searrow$ respectivement. Les effectifs cumulés sont definis par :

$$N_i \nearrow = \sum_{j=1}^{i} n_j, \ i = 1, 2, ...k$$

et

$$N_i \searrow = n - \sum_{j=1}^{i} n_j.$$

4- **Fréquence cumulée croissantes et décroissantes :** La fréquence relative cumulée croissante, notée $F_i \nearrow$ est la proportion entre l'effectif cumulée croissant$N_i \nearrow$ et n. La fréquence relative cumulée décroissante, notée $F_i \searrow$ est la proportion entre l'effectif cumulée décroissante $N_i \searrow$ et n.

$$F_i \nearrow = \frac{N_i \nearrow}{n}, \ F_i \searrow = \frac{N_i \searrow}{n}$$

5- **Table des fréquences :** Est donné par :

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$F_i \nearrow$	$F_i \searrow$
X_1	n_1	$f_1 = \frac{n_1}{n}$	n_1	$n_1 + n_2 + ... + n_k = n$	$\frac{N_1 \nearrow}{n}$	$\frac{N_1 \searrow}{n}$
X_2	n_2	$f_2 = \frac{n_2}{n}$	$n_1 + n_2$	$n_2 + ... + n_k$	$\frac{N_2 \nearrow}{n}$	$\frac{N_2 \searrow}{n}$
$\vdots$						
X_{k-1}				$n_k + n_{k-1}$	$\frac{N_{k-1} \nearrow}{n}$	$\frac{N_{k-1} \searrow}{n}$
X_k	n_k	$f_k = \frac{n_k}{n}$	$n_1 + n_2 + ... + n_k = n$	n_k	$\frac{N_k \nearrow}{n}$	$\frac{N_k \searrow}{n}$
Total	n	1				

Example 4 *Soit la série suivante sur les notes de 20 étudiants*

11	*12*	*15*	*14*	*14*	*10*	*10*	*12*	*10*	*10*
15	*11*	*11*	*10*	*14*	*11*	*12*	*12*	*15*	*12*

Question : *Remplis le tableau des fréquences*

Réponse : *La variable statistique c'est les notes des étudiants et la taille de l'échantillon* $n = 20$. *Les modalités sont* $\{10, 11, 12, 14, 15\}$ *donc* $k = 5$. *Nous trions d'abord les modalités de la plus petite valeur à la plus grande et nous présentons le tableau des fréquences comme suit*

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$F_i \nearrow$	$F_i \searrow$
$X_1 = 10$	5	$\frac{5}{20} = 0.25$	5	20	$\frac{5}{20} = 0.25$	$\frac{20}{20} = 1$
$X_2 = 11$	4	$\frac{4}{20} = 0.2$	9	15	$\frac{9}{20} = 0.45$	$\frac{15}{20} = \frac{3}{4}$
$X_3 = 12$	5	$\frac{5}{20} = 0.25$	14	11	$\frac{14}{20} = 0.7$	$\frac{11}{20} = 0.55$
$X_4 = 14$	3	$\frac{3}{20} = 0.15$	17	6	$\frac{17}{20} = 0.85$	$\frac{6}{20} = \frac{3}{10}$
$X_5 = 15$	3	$\frac{3}{20} = 0.15$	20	3	$\frac{20}{20} = 1$	$\frac{3}{20} = 0.15$
Total	20	1		0		

1.3.2 Mesures statistiques descriptives

Tendance centrale : Une mesure de tendance centrale est une valeur qui représente l'entrée typique, ou centrale, d'un ensemble de données. Les mesures de tendance centrale sont : la moyenne, le mode, la médiane et les quartiles.

1. **Moyenne :** La moyenne d'un ensemble de données est la somme du produit entre la modalité et son effectif divisée par la taille de l'échantillon n,ou c'est la somme du produit entre la modalité et sa fréquence relative. La moyenne est notée par $\bar{X}$ et définie par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i X_i$$

ou

$$\bar{X} = \sum_{i=1}^{k} f_i X_i$$

Example 5 *Pour les données suivantes sur des âges de 10 personnes :*

$$35, 32, 30, 30, 30, 30, 22, 22, 25, 25$$

Nous calculons la moyenne. Nous trions d'abord les données et déterminons les modalités de la variable. X =âge, $n = 10$, modalitiés sont :$\{22, 25, 30, 32, 35\}$ et $k = 5$. Deuxième étape, nous remplissons le tableau des fréquences, comme suit :

X_i	n_i	f_i	$N_i \nearrow$	$F_i \nearrow$	$n_i X_i$	$f_i X_i$
22	2	0.2	2	$\frac{5}{20} = 0.25$	44	4.4
25	2	0.2	4	$\frac{9}{20} = 0.45$	50	5.0
30	4	0.4	8	$\frac{14}{20} = 0.7$	120	12
32	1	0.1	9	$\frac{17}{20} = 0.85$	32	3.2
35	1	0.1	10	$\frac{20}{20} = 1$	35	3.5
Total	10	1			$\sum_{i=1}^{5} n_i X_i = 281$	$\sum_{i=1}^{5} f_i X_i = 28.1$

donc :$\bar{X} = 28.1$ ou $\bar{X} = \frac{1}{10}(281) = 28.1$.

2. **Mode :** Le mode est la modalité qui a le plus grand effectif, elle est notéeMo. Pour l'ensemble des données suivant :

$$1, 1, 2, 3, 3, 3, 4, 4, 5$$

nous avons 5 modalités $\{1, 2, 3, 4, 5\}$. Le mode est $Mo = 3 = X_3$ car $n_3 = 3$ est la plus grande effectif et $n_1 = 2$, $n_2 = 1$, $n_4 = 2$ et $n_5 = 1$.

Remark 6 *nous pouvons trouver plus que du mode dans un ensemble de données. Pour cet ensemble de données*

$$1, 1, 1, 2, 3, 3, 3, 4, 4, 5$$

on retrouve deux modes$X_1 = 1$ *et* $X_3 = 3$ *parce qu'ils ont le même plus grand effectif* $n_1 = n_3 = 3$. *dans ce cas, l'ensemble de données est appelé bimodal.*

3. **Médiane :** La médiane d'un ensemble de données est la modalité qui se situe au milieu des données lorsque l'ensemble de données est ordonné, ou c'est la modalité qui divise la série statistique en deux parties égales (50 %). Elle est notée par Me. Nous avons deux cas pour déterminer la médiane :
   - Si la taille de l'échantillon n est un nombre impair, la médiane est la modalité en position $\frac{n+1}{2}$

$$Me = X_{\frac{n+1}{2}}.$$

     Par exemple. L'ensemble de données suivant de 9 nombres

$$1, 1, 2, 3, 3, 3, 4, 4, 5$$

     nous observons que la taille de l'échantillon $n = 9$, donc la valeur médiane est$Me = X_{\frac{9+1}{2}} = X_5 = 3$.
   - Si la taille de l'échantillon n est un nombre pair, la médiane est la moyenne des deux valeurs médianes aux positions $\frac{n}{2}$et$\frac{n}{2} + 1$

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}.$$

Nous prenons l'exemple suivant. Cet ensemble de données de 10 nombres

$$1, 1, 2, 3, 3, 4, 4, 5, 5, 6$$

La valeur médiane est :

$$\begin{aligned} Me &= \frac{X_{\frac{10}{2}} + X_{\frac{10}{2}+1}}{2} \\ &= \frac{X_5 + X_6}{2} \\ &= \frac{3+4}{2} \\ &= 3.5 \end{aligned}$$

4. **Quartiles** : Le quartile est une mesure qui divise l'ensemble de données ou la série statistique en quatre parties, ou quarts, de taille plus ou moins égale et les données doivent être classées du plus petit au plus grand pour calculer les quartiles. Il existe trois quartiles comme suit :
   - **Le premier quartile Q_1 :** C'est la valeur qui coupe le premier quart de l'échantillon. Il est également appelé quartile inférieur, car 25 % des données se situent en dessous de ce point. Il peut être calculé comme :

$$Q_1 = X_{\frac{n}{4}}.$$

   - **Le deuxième quartile (Q_2) :** Est la médiane d'un ensemble de données.
   - Le troisième quartile (Q_3) : C'est la valeur qui coupe le troisième quart de l'échantillon. Il s'agit du quartile supérieur, car 75 % des données se situent en dessous de ce point. Il peut être calculé comme :

$$Q_3 = X_{\frac{3n}{4}}.$$

   Nous prenons l'exemple suivant. L'ensemble de données suivant de 9 nombres

$$1, 1, 2, 3, 3, 3, 4, 4, 5$$

   alors

$$\begin{aligned} Q_1 &= X_{\frac{10}{4}} = X_{2.25} \backsimeq X_3 = 2 \\ Q_3 &= X_{\frac{30}{4}} = X_{7.5} \backsimeq X_8 = 4 \end{aligned}$$

On écrit la médiane et les quartiles en fonction de la fréquence cumulée comme suit :

$$\begin{aligned} Me &= N^{-1} \nearrow \left(\frac{n}{2}\right) \\ Q_1 &= N^{-1} \nearrow \left(\frac{n}{4}\right) \\ Q_3 &= N^{-1} \nearrow \left(\frac{3n}{4}\right) \end{aligned}$$

1. **Example 7** *Nous prenons l'exemple (5), et selon le tableau des fréquences et la colonne de $N_i \nearrow$ nous déterminons la médiane, Q_1 et Q_3*
   la taille de l'échantillone $n = 10$, alors

$$\begin{aligned} Me &= \frac{N^{-1} \nearrow \left(\frac{10}{2}\right)}{2} \\ &= N^{-1} \nearrow (5) \\ &= 30 \end{aligned}$$

et

$$\begin{aligned} Q_1 &= N^{-1} \nearrow \left(\frac{10}{4}\right) = N_{2.5}^{-1} \backsimeq N_3^{-1} = 25 \\ Q_3 &= N^{-1} \nearrow \left(\frac{30}{4}\right) = N_{7.5}^{-1} \backsimeq N_8^{-1} = 30 \end{aligned}$$

On peut calculer la médiane et les quartiles par la fréquence cumulée et on lit les valeurs dans le tableau de fréquence dans la colonne :$F_i \nearrow$

$$\begin{aligned} Me &= F^{-1} \nearrow (0.5) \\ Q_1 &= F^{-1} \nearrow (0.25) \\ Q_3 &= F^{-1} \nearrow (0.75) \end{aligned}$$

Dans cet exemple, on obtient :

$$\begin{aligned} Me &= F^{-1} \nearrow (0.5) = 30 \\ Q_1 &= F^{-1} \nearrow (0.25) = 25 \\ Q_3 &= F^{-1} \nearrow (0.75) = 30 \end{aligned}$$

Pour déterminer le mode, nous utilisons la colonne n_i, le mode dans notre exemple est $Mo = 30$ car l'effectif le plus élevé est 4.

Mesures de dispersion :

1. **Variance :** La variance est l'écart carré moyen entre chaque modalités et le centre de la distribution représenté par la moyenne. elle est définie par :

$$var(X) = \frac{1}{n} \sum_{i=1}^{k} n_i (X_i - \bar{X})^2$$

ou

$$var(X) = \frac{1}{n} \sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2$$

et car $\frac{n_i}{n} = f_i$, la variance peut être calculée par :

$$var(X) = \sum_{i=1}^{k} f_i X_i^2 - \bar{X}^2$$

2. **Ecart-type** : L'écart type est la racine carrée de la variance. Elle est notée par σ :

$$\sigma = \sqrt{var(X)}$$

Example 8 *Nous prenons le même exemple (5). Dans le tableau des fréquences, nous pouvons ajouter une nouvelle colonne pour calculer* $f_i X_i^2$ *puis nous calculons la variance et l'écart type*

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$n_i X_i$	$f_i X_i$	$f_i X_i^2 = (f_i X_i) X_i$
22	2	0.2	*2*	*10*	44	4.4	$22 * 4.4 = 96.8$
25	2	0.2	*4*	*8*	50	5.0	$25 * 5 = 125.0$
30	4	0.4	*8*	*6*	120	12	$30 * 12 = 360.0$
32	1	0.1	*9*	*2*	32	3.2	$32 * 3.2 = 102.4$
35	1	0.1	*10*	*1*	35	3.5	$35 * 3.5 = 122.5$
Total	10	1		*0*	$\sum_{i=1}^{5} n_i X_i = 281$	$\sum_{i=1}^{5} f_i X_i = 28.1$	$\sum_{i=1}^{5} f_i X_i^2 = 806.7$

and $\bar{X}^2 = 789.61$, *donc*

$$\begin{aligned} var(X) &= 806.7 - 789.61 \\ &= 17.09 \end{aligned}$$

et l'écart type :

$$\begin{aligned} \sigma &= \sqrt{17.09} \\ &= 4.134 \end{aligned}$$

3. **Et :endue** : la plage d'un ensemble de données est la différence entre la valeur maximale et la valeur minimale dans l'ensemble. Il est défini par :

$$Range = X_{\max} - X_{\min}$$

Example 9 *D'après l'exemple (5), l'etendue est :*

$$\begin{aligned} I &= 35 - 22 \\ &= 13 \end{aligned}$$

4. **Ecart interquartile :** L'écart interquartile d'un ensemble de données est la différence entre le premier quartile Q_1 et le troisième quartile Q_3. Il est défini par :

$$IQ = Q_1 - Q_3$$

Remark 10 *Contrairement à l'etendue et l'écart interquartile, la variance est une mesure qui prend en compte la dispersion de toutes les valeurs d'un ensemble de données.*

1.4 Paramètre de la forme

Le premier paramètre de forme c'est le paramètre d'asymétrie qui nous permet de mesurer et conclut si la distribution est symétrique ou non, il existe plusieurs coefficients pour étudier l'asymétrie comme le coefficient de Pearson qui compare la moyenne ave le mode ou la médiane ainsi que la skewness. Le deuxième paramètre de forme c'est le paramètre d'applatissment qui corspond au Kurtosis. Dans cette section on va parler seulement sur le paramètre d'asymétrie.

Pour une distribution parfaitement symétrique, on a :

$$\bar{X} = Me = Mo.$$

Coefficient d'asymétrie de Pearson : Ce coefficient est défini par :

$$\delta = \frac{\bar{X} - Me}{\sigma_X}.$$

où

$$-1 \leq \delta \leq +1.$$

- Si $\delta = 0$, cela indique une distribution parfaitement symétrique où les données sont uniformément équilibrées des deux côtés de la moyenne.
- Si $\delta > 0$, cela suggère une distribution positivement asymétrique où la queue du côté droit est plus longue, et la majorité des points de données sont concentrés sur le côté gauche de la moyenne.
  Dans ce cas on dit que la queue de la distribution est étalée à droite.
- Si $\delta < 0$, cela indique une distribution asymétrique négative où la queue du côté gauche est plus longue, et la majorité des points de données sont concentrés sur le côté droit de la moyenne. Dans
  ce cas on dit que la queue de la distribution est étalée à gauche.

1.5 Paramètre d'homogénéité

Coefficient da la variation Ce coefficient permet d'apprecier la représentativité de la moyenne par rapport à l'ensemble des observation. Il donne une bonne idée du dégré d'homogénéité d'une série, il faut qu'il soit le plus faible possible (<15% en pratique), et on peut l'obtenir en appliquant la formule suivante

$$CV = \frac{\sigma}{\bar{X}} \times 100.$$

Example 11 *D'après l'exemple (5), le coerficient d'asymétrie égale à :*

$$\begin{aligned} \delta &= \frac{28.1 - 30}{4.134} \\ \delta &= -0.459\,6 \end{aligned}$$

:On a : $\delta = -0.459\,6 < 0$ *la queue de la distribution est étalée à gauche.*
Pour le coefficient de variation :

$$\begin{aligned} CV &= \frac{4.134}{28.1} \times 100 \\ CV &= 14.\,712 \end{aligned}$$

On a : $CV = 14.\,712\% < 15\%$ *alors la série est homogène.*