Chapitre 3 _ Statistiques et Probabilités.

I _ Les séries statistiques à une variable

1_ Définition :

On appelle série statistique X, l'ensemble de toutes les observations x_i, $i = 1 : N$, liées à un caractère donné. Ces observations x_i peuvent être des valeurs discrètes ou continues, comme elles peuvent exprimer des variables quantitatives ou qualitatives.

Une observation x_i peut être observée une, deux ou plusieurs fois le long de la série statistique; ce nombre de fois où est observée la variable x_i est appelé effectif de x_i et il est noté n_i. Dans ce cas la fréquence d'observation de la variable x_i est notée $f_i = \frac{1}{n_i}$

NB₁ : Ces données sont collectées, puis dépouillées, puis classées pour qu'elles puissent être traitées.

exemple 1 : Soit la série statistique X suivante, donner la fréquence d'observation de la valeur 7 et celle de la valeur 1 ?

Réponse : $X = \{7, 6, 0, 1, 5, 0, 7, 1, 1, 6, 3, 1, 7, 5, 2, 1, 0, 3\}$

$x = 7$ est observée 3 fois $\Rightarrow n_7 = 3 \Rightarrow f_7 = \frac{1}{n_7} = \frac{1}{3}$ ou 0.33

$x = 1$ est " 5 fois $\Rightarrow n_1 = 5 \Rightarrow f_1 = \frac{1}{n_1} = \frac{1}{5}$ ou 0.20

NB₂ : Les données de toute série statistique peuvent être résumées par un tableau.

exemple 2 : Soit la série statistique X de l'exemple 1, représenter la sous forme de tableau.

Réponse. X peut être résumé par un tableau simple à deux colonnes : la première est celle correspondant aux valeurs observées, la deuxième concerne les effectifs de ces dernières valeurs, comme c'est indiqué ci-dessous.

x_i	n_i
0	3
1	5
2	1
3	2
5	2
6	2
7	3
	$\Sigma = 18$

x_i est l'ensemble des valeurs observées

n_i est l'effectif des valeurs observées

Dans ce cas la somme des effectifs est égale à $N = \sum_i n_i$ et est appelée : Taille de la série qui est égale à 18 dans le présent exemple, et la fréquence (~~effectif f~~) de x_i est égal à $\frac{n_i}{N} = f_i$ (fréq. Globale)

2. Représentation d'une série statistique.

L'ensemble des données d'une série statistique peut être résumé et représenté de différentes manières selon le cas et selon les besoins.

2.1. Sous forme de tableaux : ces derniers comprennent deux colonnes au moins l'une pour le caractère x_i, l'autre pour les effectifs n_i correspondants.

- Pour les variables qualitatives les x_i sont codés
- Pour les variables quantitatives les x_i sont classés par ordre croissant.

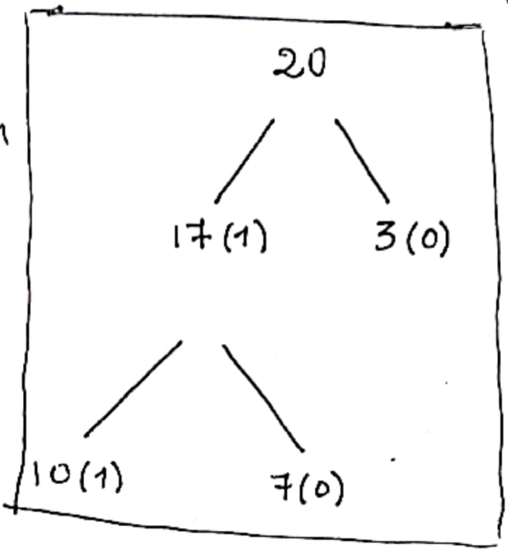
2.2. Sous forme d'arbres :

Cette forme de représentation permet de visualiser les résultats d'un questionnaire, par exemple quand il y a différents choix possibles par question.

Exemple : On pose deux questions à 20 personnes et on leur demande de répondre par OUI (code 1) ou NON (code 0). On obtient 17 personnes qui ont répondu OUI à la 1ère question et 10 qui ont répondu OUI aux 2 questions. Combien de personnes ont, sur les 17, répondu Non à la 2ème question ? représenter sous la forme adéquate.

Réponse : Sur les 20 personnes il y a 17 qui ont répondu OUI il reste 3 personnes pour la réponse NON de la 1ère question. Parmi ces 17 personnes 10 vont encore répondre OUI à la 2ème question, ce qui laisse 7 personnes pour la réponse NON des 17 personnes. La représentation la plus adéquate est celle sous forme d'arbre comme suit :

Titre : Représentation sous forme d'arbre de la série donnée



(Nombre de personnes interrogées)

↓ 1ère question 2 choix (OUI : 1 , NON : 0)

17 ont répondu OUI et 3 Non

↓ 2ème question 2 choix (OUI : 1 , NON : 0)

10 ont répondu OUI et 7 Non

donc le nombre de personnes qui ont répondu OUI à la 1ère question et OUI à la 2ème question est égal à 10, tandis que le nombre de personnes qui ont répondu OUI à la 1ère question et NON à la 2ème question est égal à 7.

2.3. Sous forme de graphiques :

C'est la forme de représentation la plus courante, dans ce cas quelque soit le graphique utilisé, il faut définir ce qu'il représente (le titre), indiquer de façon claire la signification des axes et choisir une unité de représentation adaptée au support utilisé.

Selon le cas, nous pouvons avoir des séries statistiques discrètes ou continues.

2.3.1. Cas d'une série discrète :

Dans ce cas on utilise des diagrammes en bâtons où la hauteur du bâton est proportionnelle à l'effectif n_i (en ordonnées) correspondant à la valeur du caractère x_i considéré (en abscisses).

Exemple : Soit la série $X = \{4, 1, 3, 4, 4, 2, 1, 3, 4, 4, 1, 2, 3, 3, 4, 4, 2, 4, 4, 3, 2, 2, 3, 3,\}$

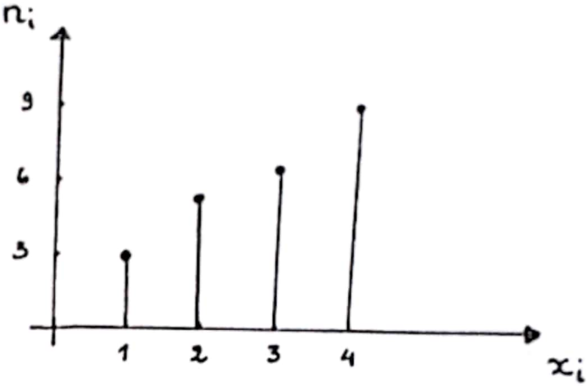
, tracer le graphique qui la représente ?

Réponse : Le graphique qui peut représenter la série X est celui qui représente les effectifs n_i en ordonnées en fonction des valeurs de X en abscisses.

le passage par le tableau des valeurs et leurs effectifs est obligatoire.

nous obtenons :

x_i	n_i
1	3
2	5
3	7
4	9



- Diagramme en bâtons de X -

NB : Il faut noter les informations portées sur le graphique : le titre, la désignation des axes et puis l'échelle.

2.3.2. Cas d'une série continue :

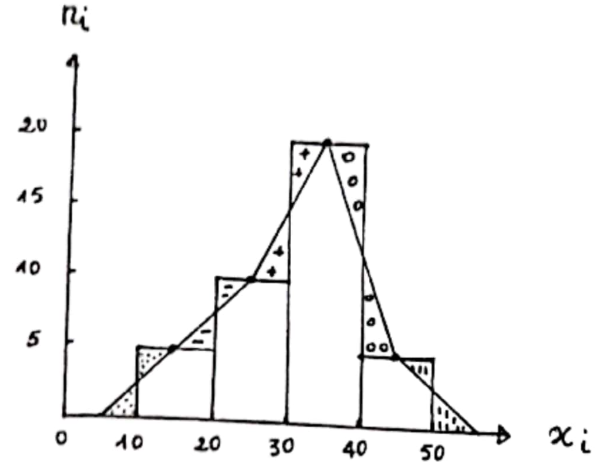
Si la série de données est continue, il faut la diviser en plusieurs intervalles dans ce cas le graphique est appelé Histogramme. On porte en abscisses les valeurs x_i et en ordonnées les effectifs correspondants. Pour chaque classe ou intervalle de x_i on élève un rectangle dont la base est proportionnelle à l'amplitude de la classe x_i et dont la hauteur est proportionnelle à l'effectif n_i.

Exemple : Soit la distribution des effectifs suivante tracer son histogramme des effectifs ?

x_i	n_i
10 - 20	5
20 - 30	10
30 - 40	20
40 - 50	5

Réponse :

Il faut observer l'égalité entre les triangles se trouvant au dessus et au dessous de la courbe des effectifs.

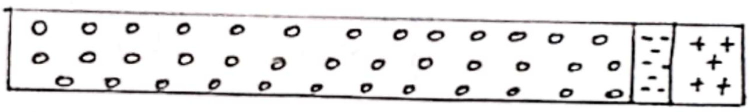


2.3.3. Cas d'une série à variable qualitative :

Dans ce cas on utilise soit des diagrammes à bandes, soit des diagrammes circulaires. La légende des couleurs s'impose.

Exemple : Une usine compte 85% de travailleurs, 05% de cadres supérieurs et 10% de cadres moyens. Représenter son effectif travailleur sous forme de bandes et sous forme de diagramme circulaire

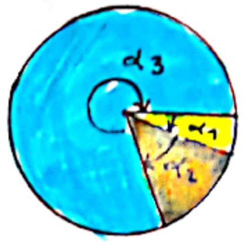
Réponse : • Diagramme à bande : il faut choisir l'échelle adéquate et donner un titre et une légende.



1cm : 10%

Légende : $\boxed{\circ\circ}$: travailleurs, $\boxed{-}$: cadres supérieurs, $\boxed{+}$: cadres moyens.

• Diagramme circulaire : il faut savoir que l'angle qui fait le cercle égale 360° et il faudra calculer les angles α_1 correspondant à 5%, α_2 correspondant à 15% et α_3 correspondant à 85%. Ensuite il faut colorer ou désigner une légende pour la lecture du diagramme sans oublier de lui attribuer un titre.

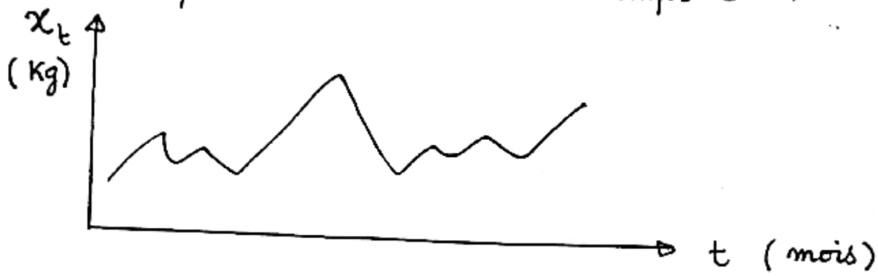


2.4. Autres graphiques :

Il existe d'autres types de graphiques pour représenter les séries statistiques entre autres :

2.4.1. Graphes chronologiques :

Si les observations sont enregistrées au cours du temps, cela veut dire qu'elles sont indicées dans le temps tel que x_t est l'observation d'un caractère x dans le temps t. Dans ce cas l'ensemble de toutes les observations est appelée série chronologique, pour la représenter on porte en abscisses le temps t et en ordonnées x_t correspondant.



- Graphe d'une serie chronologique -

2.4.2. Graphes à coordonnées polaires :

On suppose que le temps s'écoule de manière circulaire, dans le sens des aiguilles d'une montre. Chaque période est représentée par un axe sur lequel on portera la valeur du caractère. L'angle entre les axes dépendra du nombre de périodes.

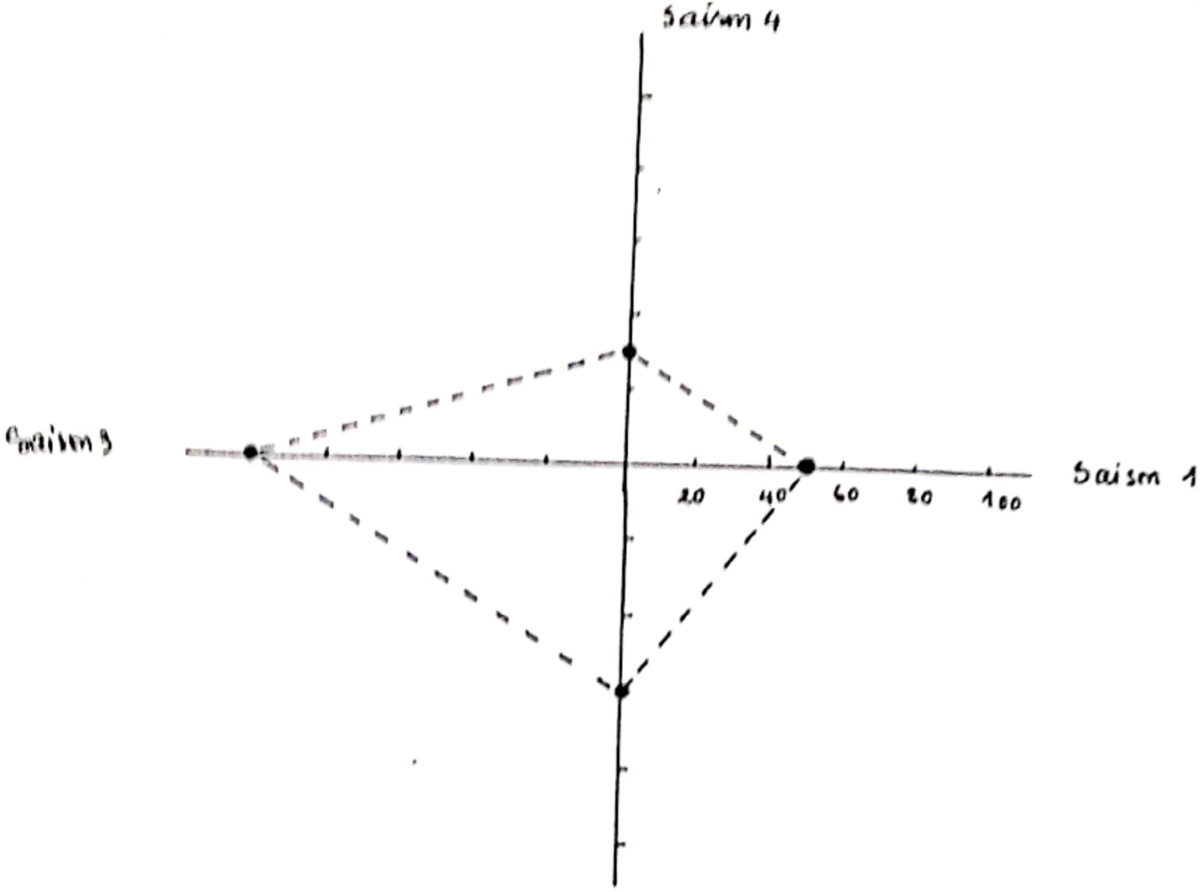
Exemple : Soit la série des bovins nés dans une ferme pendant les quatres saisons de 2008. Donner la représentation polaire de ces naissances au cours des 4 saisons.

Réponse :

x_t	50	60	100	30
t	1	2	3	4

t : saison : 1 : 4

x_t : nombre de bovins nés dans la saison t



- Graphe polaire des bovins nés en 2008 ds la ferme indiqué.

NB: La saison est désigné par le nbre de bovins qui sont nés à son courant. et le graphique permet de visualiser les variations saisonnières de cette naissance bovine.

2.4.3. Graphes à echelle logarithmique :

L'echelle logarithmique est une échelle graduée proportionnellement au Log decimal. Elle représente des phénomènes de croissance à taux constant par une droite et est generalement utilisé quand la variable croit Très rapidement.

3. Effectifs et fréquences :

Il est déja connu que l'effectif n_i d'une valeur est le nombre de fois où elle apparaît dans la série étudié. Soit N l'effectif total. N est la somme de toutes les valeurs des effectifs n_i, tel que :

$$N = \sum_i n_i$$

la fréquence f_i correspondante : $f_i = \frac{1}{n_i}$

avec la frequence globale : $F_i = \frac{n_i}{N}$ / $\sum_i F_i = 1 = 100\%$

3.1. Effectif cumulé croissant (ECC) et effectif cumulé décroissant (ECD) :

L'effectif cumulé croissant (ECC) est le cumul des effectifs dans le sens croissant du début vers la fin des effectifs partiels. L'effectif cumulé décroissant c'est le chemin inverse de la fin du cumul vers le début de celui-ci.

Exemple : Soit la série des effectifs suivants calculer son ECC et son ECD ?

x_i	1	2	3	4	5
n_i	12	10	5	20	15

Réponse : on calcule et on reporte dans le tableau suivant les valeurs de l'ECC et l'ECD de la série considérée.

x_i	n_i	ECC	ECD
1	12	12	62
2	10	22	50
3	5	27	40
4	20	47	35
5	15	62	15
	$\Sigma = 62$		

- Il faut noter que la valeur la plus grande de l'ECC et l'ECD est égale à la somme des effectifs partiels n_i.
- Et que cette valeur la plus grande est la dernière valeur pour l'ECC et la première valeur pour l'ECD.

Dans le cas d'une série continue, les valeurs de x_i sont donnés sous forme de classes ou intervalles. Dans ce cas :

- ECC correspondant à une classe donnée $[a, b[$ indique le nombre de fois où le caractère prend une valeur strictement inférieure à b.
- ECD correspondant à une classe donnée $[a, b[$ indique le nombre de fois où le caractère prend une valeur supérieure ou égale ($\geq$) à a.

3.2. Fréquence cumulée croissante (FCC) et fréquence cumulée décroissante (FCD) :

La même procédure de calcul est poursuivie pour les fréquences. On remplace dans les définitions "le nombre de fois" par "pourcentage" dans l'ECC et l'ECD pour obtenir FCC et FCD.

4. Représentation Numérique d'une série statistique :

Après avoir exposé les différentes manières de représentation graphique d'une série statistique, dans le présent paragraphe, nous allons nous intéresser à la représentation numérique des séries statistiques. Cette représentation résume par le biais de quelques nombres les caractéristiques statistiques des séries. La première classe s'intéresse à la position centrale de la série et les paramètres qui la déterminent, la seconde classe s'intéresse à la dispersion autour de cette position centrale et les paramètres qui la désignent

4.1. Les paramètres de la position centrale (Moyenne)

4.1.1. La moyenne arithmétique :

Soit (x_i, n_i) une série statistique. La moyenne arithmétique est désignée par $\overline{X}$ telle que :

$$\overline{X} = \frac{1}{\sum_i n_i} \sum_i n_i x_i$$

• pour les séries continues, x_i est le centre des classes.

Exemple : Soit la série suivante, calculer sa moyenne arithmétique.

Réponse :

x_i	n_i	x_i'	$n_i x_i'$
[10 : 20[	100	15	1500
[20 : 30[	80	25	2000
[30 : 50[	120	40	4800
[50 : 100[	60	75	4500
[100 : 150[	10	125	1250
	$\sum_i = 370$		$\sum_i = 14050$

$$\overline{X} = \frac{1}{\sum_i n_i} \sum_i n_i x_i'$$

AN : $\overline{X} = \frac{14050}{370} = 38$ unité donc la moyenne arithmétique est $\boxed{\overline{X} = 38 \text{ unité}}$

N.B : $\overline{X} \rightarrow$ Valeur Centrale

4.1.2. La moyenne harmonique (H) :

$$\frac{1}{H} = \frac{1}{N} \sum_i n_i \frac{1}{x_i}$$

$N = \sum_i n_i$

H : moyenne harmonique des x_i.

4.1.3. La moyenne géométrique (G) :

On l'utilise quand les valeurs sont liées entre elles par une liaison de croissance

$$G = \left(\Pi_i \, x_i^{n_i} \right)^{\frac{1}{N}}$$

Π : produit

$N = \sum_i n_i$

G : moyenne géométrique des x_i.

Exemple : une plante augmente de 10% pendant 2 jours, puis de 5% pendant 4 jours et de 12% pendant 6 jours. Calculer le taux de croissance moyen sur la periode ?

Réponse : Soit τ le taux de croissance moyen sur cette période. il est donné par la moyenne géométrique et peut donc être calculé selon la formule :

$$\tau = \left(1.10^2 \times 1.05^4 \times 1.12^6 \right)^{\frac{1}{12}}$$

AN $\boxed{\tau = 9.2\ \%}$

4.1.4 Le mode (m_o) :

C'est la valeur du caractère x_i qui correspond à l'effectif le plus élevé.

4.1.5. La médiane (Me) :

Pour une série (x_i, n_i), $i:1:N$, la médiane est la valeur du caractère x_i qui partage les effectifs en deux parties égales.

- Pour les séries discrètes, si le nombre d'observations est pair la médiane peut être indéterminée.

Exemple : Soit les 2 séries suivantes qui présentent un nbre d'observations pair,
- déterminer la médiane pour chaque série ?

Réponse :

1)

x_i	n_i	
0	2	||
1	3	|||
2	4	||||
3	3	|||
	$N = \Sigma = 12$	

- ici $N = 12$ (il est pair)

$N/2 = 6$

L'observation x_i qui divise les effectifs n_i en deux parties égales se situe entre la 6eme et 7eme observation donc la médiane :

$me = 2$

2)

x_i	n_i	0
0	2	||
1	3	|||•
2	3	|||
3	2	||
	$N = \Sigma = 10$	

- ici $N = 10$ (il est aussi pair)

$N/2 = 5$

la médiane x_i qui divise les effectifs n_i en deux parties égales est dans ce cas indéterminée.

- Pour les série continues, la médiane se calcule par la formule suivante :

$$\boxed{\frac{me - a}{b - a} = \frac{N/2 - ECC_{-1}}{n}} \qquad me \in [a, b[$$

La médiane me correspond à l'effectif $N/2$. On repère dans la colonne ECC, à quelle classe appartient cet effectif. Soit $[a, b[$ cette classe, n son effectif et ECC_{-1} l'effectif cumulé croissant de la classe précédente.

Exemple : Soit la série (x_i, n_i) suivante, déterminer sa médiane.

Réponse :

x_i		n_i	ECC
500	1000	6	6
1000	2000	12	18
2000	3000	30	48
3000	4000	38	86
4000	5000	31	117
5000	6000	26	143
6000	8000	5	148
+ 8000		2	150
		$N = 150$	

$N/2 = 150/2 = 75$

$75 \in [3000 \ 4000[\Rightarrow me \in [3000, 4000[$

$$\frac{me - 3000}{4000 - 3000} = \frac{75 - 48}{38}$$

$me = 3710$ unités

4.2. Les paramètres de dispersion autour de la moyenne

4.2.1. La variance (Var ou σ^2)

La variance est défini comme la moyenne des carrés des écarts à la moyenne. Soit e_i cet écart par rapport à la moyenne $\overline{X}$ des observations x_i $\left(\overline{X} = \frac{1}{\sum n_i} \cdot \sum_i n_i x_i\right)$

$$e_i = x_i - \overline{X}$$

la moyenne des carrés des e_i est :

$$\sigma^2 = \frac{1}{N} \sum n_i e_i^2$$

$$\boxed{\sigma^2 = \frac{1}{N} \sum n_i (x_i - \overline{X})^2}$$

C'est la variance de la serie (x_i, n_i)

Rq: La formule la plus connue pour la variance est celle de Köening. C'est generalement la formule la plus utilisée.

$$\sigma^2 = \frac{1}{N} \sum n_i x_i^2 - \overline{X}^2$$

4.2.2. L'écart-type σ :

C'est la racine carrée de la variance. Il mesure l'écart moyen autour de la moyenne. Ainsi une serie statistique qui possède un faible écart-type est une série où les valeurs sont peu dispersées autour la moyenne qui présente alors un indicateur plus faible.

4.2.3. Le coefficient de variation (Cv)

$$\boxed{C_V = \frac{\sigma}{\overline{X}}}$$

NB: La pratique et l'expérience montrent que pour une distribution unimodale et symétrique, 68% des valeurs $\in [\overline{X} - \sigma, \overline{X} + \sigma]$ et
95% " $\in [\overline{X} - 2\sigma, \overline{X} + 2\sigma]$

ceci est important dans le calcul des bornes de confiance (ou limites de confiance).