Chapitre II : Statistique descriptive à un caractère

La statistique descriptive basée sur à un caractère permet grâce au calcul de certains indices et paramètres, d'obtenir des éléments d'information sur la population étudiée. La caractérisation d'une population sur la base d'une propriété donnée est effectuée selon les étapes suivantes :

a- Spécification du type de caractère : qualitatif ou quantitatif.

b- Mesure du caractère (observation).

Les étapes a et b sont appelées : expérience.

c- Regroupement des données.

d- Traitement : cette étape est basée sur l'utilisation de méthodes mathématiques qui permettent d'obtenir des résumés numériques (statistiques) pour avoir une connaissance ou un jugement sur la population étudiée.

- Statistique descriptive à un caractère quantitatif

La première opération consiste à effectuer une ordination (arrangement) des données ou des résultats obtenus.

1- Etendue de la série statistique : L'étendue d'une S.S. se calcule en faisant la différence entre la plus grande valeur et la plus petite valeur de la série. $e = x_{max} - x_{min}$

2- Moyenne arithmétique : Elle est calculée en faisant la somme des valeurs de la S.S. et en divisant par le nombre des valeurs de la série. C'est la valeur représentative de la S.S.

Si : $x_1, x_2, \ldots\ldots X_n$ sont les éléments de la S.S. la moyenne arithmétique est:

$$\bar{X} = \frac{\sum_{i=1}^{n} xi}{n}$$

Au lieu de considérer toutes les données de la population, on utilise la moyenne arithmétique comme valeur (base) de jugement pour la caractérisation de cette population.

3- Médiane : C'est la valeur qui occupe la position médiane (milieu) dans la S.S. Elle partage la série en 2 parties contenant chacune 50% de l'ensemble. La première partie renferme les valeurs qui sont inférieures à la médiane et la deuxième partie renferme les valeurs qui sont supérieures à la médiane. Elle se détermine en fonction du nombre de valeur de la S.S. (nombre pair ou impair).

* Si n est un nombre pair : Med = $X_{med} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$

Exp : Donner la médiane de la série suivante : 1.2 ; 4.3 ; 5.1 ; 0.8 ; 1.9 ; 2.1 ; 2 ; 3.5.

Solution : la S.S ordonnée : 0.8 ; 1.2 ; 1.9 ; 2 ; 2.1 ; 3.5; 4.3 ; 5.1.

n=8, $X_{n/2} = X_4$; $X_{n/2+1} = X_5$ donc : $X_{med} = \frac{X4+X5}{2} = \frac{2+2.1}{2} = 2.05$.

* Si n est un nombre impair : $X_{med} = X_{\frac{n+1}{2}}$

Exp : Donner la médiane de la série suivante : -1 ; 0 ; 2 ; 3.5 ; 5 ; 9.1 ; 10.2 ; 15 ; 16.2 ; 18.4 ; 20.5 ; 26 ; 30.

Solution : n= 13, n+1=14 donc : Med= $X_{med} = X_{14/2} = X_7 = 10.2$.

- Dans le cas d'une S.S. la moyenne et la médiane constituent des indices ou des valeurs de position centrale ou médiane. Ce sont des indices d'échelle de location ou de position car ils ont la même unité de mesure que les valeurs observées.

4- variance : Le degré de fluctuation ou de variation des valeurs mesurées est obtenue par l'indice d'échelle de dispersion appelé **variance**. Ce paramètre fournit le carré de l'écart moyen entre les valeurs mesurées et leur moyenne.

Si : $x_1, x_2, \ldots .. X_n$ sont les éléments de la S.S., alors la variance est :

$Var(x) = \sigma^2(x) = \frac{1}{n} \sum_{i=1}^{n} (xi - \bar{x})^2$; Cas d'une S.S. ordonnée.

$Var(x) = \sigma^2(x) = \frac{1}{N} \sum_{i=1}^{n} ni\ (xi - \bar{x})^2$; Cas d'une série en classes.

Si la valeur de la variance est très grande, cela indique que les valeurs de la série sont très éloignées ou dispersées de la moyenne. On peut considérer alors que le caractère étudié est très fluctuant dans la population et par conséquent que la population est **hétérogène** par rapport à ce caractère.

Si la valeur de la variance est petite, cela indique que les valeurs sont bien concentrées autour de la moyenne et donc que la population est **homogène** par rapport à ce caractère.

Mais l'indice d'homogénéité de la population relativement à un caractère s'exprime par le **coefficient de variation (CV %)**.

$CV\ \% = \frac{\sigma}{\bar{x}} . 100$

Selon la formule, cet indice exprime le rapport entre l'échelle de dispersion σ et l'échelle de tendance centrale ou de position ou de location $\bar{x}$. Il est exprimé en pour cent et selon la valeur obtenue, la population est jugée comme homogène ou hétérogène. En général si :

- $CV \leq 20\ \%$, cela indique que la population est **homogène** par rapport au caractère étudié.
- $CV > 20\ \%$, cela indique que la population est **hétérogène** par rapport au caractère étudié.

- **La deuxième étape dans la statistique descriptive** consiste à structurer les valeurs de la S.S. Cette opération consiste à subdiviser (partitionner) l'ensemble des valeurs de la série en classes. Chaque classe contient des valeurs égales ou proches les unes des autres. Les individus dont les valeurs sont proches ou égales seront considérés comme identiques et forment donc une même classe.

Il existe plusieurs méthodes pour subdiviser un ensemble de valeurs en classes. La méthode qui permet d'obtenir la plus grande information est celle qui donne un nombre de classes k.

$K = \log_2^N = \frac{\text{Log N}}{\text{Log 2}}$

Si toutes les classes de la répartition ont la même amplitude a, alors: $a = \frac{e}{k}$.

Une classe est définie par un intervalle avec une borne inférieure appelée **limite inférieure** de la classe et une borne supérieure appelée **limite supérieure** de la classe. L'amplitude de la classe est égale à la longueur de l'intervalle présenté par la classe.

a= longueur de l'intervalle = limite supérieure - limite inférieure

Le centre de la classe est le milieu de l'intervalle.

Centre de la classe = $\frac{\lim \sup + \lim inf}{2}$

1- Représentation des données en classes (répartition ou distribution)

La répartition en classes est un tableau dans le quel il figure les classes et les effectifs correspondants.

Intervalle de classes	a_1- b_1	a_2- b_2	……..	……..	……..	a_k- b_k	Total
Effectif ni	n_1	n_2	……..	……..	……..	n_K	N

L'effectif d'une classe qu'on désigne par ni est le nombre de valeurs qui sont comprises dans la classe.

N nombre des valeurs dans la série ; $N = \sum_{i=1}^{k} ni$

Par définition, une distribution est formée de classes disjointes ou exclusives (une valeur n'appartient qu'à une seule classe).

On peut symboliser les classes soit par les limites des intervalles, soit par le centre de la classe.

2- Caractéristique d'une distribution

- La fréquence relative simple d'une classe : représente le poids en % de cette classe par rapport à l'ensemble de classe de la distribution est donnée par la formule : $fi = \frac{ni}{n}$; $0 < fi \leq 1$.

Pour une distribution donnée, la somme des fréquences est égale à 1. $\sum_{i=1}^{k} fi = 1\ ou\ 100\ \%$.

- Les fréquences cumulées : Sont obtenus par ajout (addition) des fréquences relatives simples. Elles sont symbolisées par F. On distingue les fréquences cumulées ascendantes (croissantes) qui sont obtenus par addition des fréquences relatives simples par passage d'une classe à une autre et les fréquences cumulées descendantes (décroissantes) obtenues par soustraction de fréquences relatives simples par passage d'une classe à une autre.

-La fréquence cumulée ascendante de la classe j : $F_a(j) = \sum_{i=1}^{j} fi = F_a\ (j-1) + fj$

-La fréquence cumulée descendante de la classe j : $F_d(j) = F_d\ (j-1) - fj$

$= 100\ \% - F_a(j)$

$= 100\ \% - \sum_{i=1}^{j} fi$

Exemple : on vous donne les classes de la taille de 400 individus.

Taille	135	140	145	150	155	160	165	**Total**
ni	60	70	80	65	45	50	30	400
fi (%)	15	17.5	20	16.25	11.25	12.5	7.5	100
Fa (%)	15	32.5	52.5	68.75	80	92.5	100	/
Fd (%)	85	67.5	47.5	31.25	20	7.5	0	/

3- Représentation graphique d'une distribution : IL existe deux formes de représentation graphique d'une distribution

- Histogramme pour une variable continue **:** c'est une représentation en fonction des fréquences. On représente sur l'axe des X les valeurs de la variable qui sont soit les centres des classes ou les limites des classes et sur l'axe des Y, les valeurs des fréquences correspondantes.

- Courbe cumulative : graphe obtenue par la représentation des valeurs de la variable et les fréquences cumulées croissantes ou décroissantes.

4- Paramètres et indices d'une distribution : Ce sont des paramètres et indices qui fournissent une indication sur la variable étudiée. On distingue les paramètres de position, de dispersion et des paramètres de forme de la courbe.

1- paramètres de position : valeurs qui occupent une position spéciale parmi les valeurs observées.

a) La moyenne pondérée : Elle indique la valeur centrale et son calcule est comme suit :

$\bar{x} = \frac{\sum_{i=1}^{k} nixi}{N}$ où $N = \sum_{i=1}^{k} ni$; donc: $\bar{x} = \sum_{i=1}^{k} fi\ xi$

b) la médiane : Dans une S.S. simple, la médiane est la valeur de la variable qui occupe la position médiane et se détermine en fonction du nombre de valeur de la S.S. : nombre pair ou impair.

Pour une distribution en classe, la détermination de la médiane se fait de deux façons:

- Méthode analytique (par calcul) : le principe est le suivant :

On détermine dans le tableau de distribution la classe qui contient la médiane en parcourant la fréquence cumulée et on s'arrête à la classe pour laquelle on a une fréquence cumulée ascendante ≥ 0.5 ; puis on applique la formule pour le calcul.

$$\text{Médiane} = X_{med} = \text{limite inférieure de la classe médiane} + a\ \frac{0.5 - Fa_{\text{classe mediane}-1}}{fi_{\text{classe médiane}}}$$

$$Med = \lim inf_{cl\ med} + a\ \frac{0.5 - Fa_{cl\ med-1}}{fi_{cl\ med}}$$

$Fa_{\text{classe médiane}-1}$: Fréquence relative cumulée ascendante de la classe qui précède la classe médiane.

$fi_{\text{classe médiane}}$: Fréquence relative simple de la classe médiane.

a : l'amplitude = limite supérieure - limite inférieure.

Ex. : Soit la distribution suivante :

Classe : 0-5 ; 5-10 ; 10-15 ; 15-20 ; 20-25

fi : 0.25 ; 0.28 ; 0.32 ; 0.09 ; 0.06

Fa : 0.25 ; 0.53 ; 0.85 ; 0.94 ; 1

$Fa_{\text{classe médiane}} \geq 0.5$; Donc la classe médiane est la classe : 5-10

$$\text{Médiane} = X_{med} = 5 + 5\ \frac{0.5 - 0.25}{0.28} = 9.46$$

Donc 50% des valeurs de X sont ≤ 9.46 dans cette distribution.

- Méthode graphique : Sur le graphe, on représente la courbe cumulative puis en trassant une droite parallèle à l'axe des X et d'équation : Y = 0.5. Cette droite va couper le graphe de la courbe cumulative en un point dont l'abscisse représente la médiane.

c) les quartiles : Ce sont des valeurs de la variable qui partagent l'ensemble de la distribution en quatre intervalles (quatre parties) contenant chacun le quart (25%) de l'ensemble de la distribution. Il y a trois quartiles appelés Q_1 (Q_{25}), Q_2 (Q_{50}), Q_3 (Q_{75}).

Le 1er quartile Q_1 est la valeur qui est supérieure au 1er quart des observations (Q_1>25%). Le 2ème quartile Q_2 est égal à la médiane, et il est > à 50% des valeurs de la variable. Le 3ème quartile Q_3 est > à 75% des valeurs de la variable.

Détermination de Q_1 et Q_3 :

Calcul de Q_1 :

$$Q_1 = \text{limite inférieure de la classe } Q_1 + a\ \frac{0.25 - Fa_{\text{classe } Q_1 - 1}}{fi_{\text{classe Q1}}}$$

$$Q_1 = \lim inf_{cl\ Q1} + a\ \frac{0.25 - Fa_{cl\ Q1-1}}{fi\ cl_{Q1}}$$

Calcul de Q_3 :

$$Q_3 = \text{limite inférieure de la classe } Q_3 + a\ \frac{0.75 - Fa_{\text{classe } Q_3 - 1}}{fi_{\text{classe } Q_3}}$$

$$Q3 = \lim inf_{\text{cl } Q3} + a\ \frac{0.75 - Fa_{\text{cl } Q3 - 1}}{fi_{\text{cl } Q_3}}$$

$Fa_{\text{classe } Q_1 - 1}$: Fréquence relative cumulée ascendante de la classe qui précède la classe Q_1.

$Fa_{\text{classe } Q_3 - 1}$: Fréquence relative cumulée ascendante de la classe qui précède la classe Q_3.

$fi_{\text{classe } Q1}$: Fréquence relative simple de la classe qui contient Q_1.

$fi_{\text{classe}}\ Q_3$: Fréquence relative simple de la classe qui contient Q_3.

$\lim inf_{\text{cl } Q1}$: Limite inférieure de la classe qui contient Q_1.

$\lim inf_{\text{cl } Q3}$: Limite inférieure de la classe qui contient Q_3.

- L'intervalle [Q_1, Q3] est appelés intervalle interquartile est égale à, Q_3- Q_1.

IIQ = Q_3- Q_1

d) Les déciles : ce sont des valeurs de la variable qui partagent l'ensemble de la distribution en 10 intervalles (parties) contenant chacun 10% ou 1/10 de l'ensemble de la distribution (d_1, d_2,……, d_9).

e) Les centiles : ce sont des valeurs de la variable qui partagent l'ensemble de la distribution en 100 intervalles (parties) contenant chacun 1% ou 1/100 de l'ensemble de la distribution (c_1, c_2,……, c_{99}).

f) Le mode : C'est la valeur de la variable qui correspond à la plus grande fréquence observée de la distribution. Dans une distribution observée, le mode est égal au centre de la classe (milieu de l'intervalle) à laquelle correspond la plus grande fréquence.

Ex. : Donner le mode de la distribution des 400 individus selon leur taille.

Solution : Le mode est égal à 145 cm puisque c'est le milieu de la classe [142.5-147.5] et cette classe correspond la fréquence relative : fi = 0.20 = 20%

2- paramètres de dispersion : ce sont des valeurs qui expriment le degré de dispersion ou de fluctuation ou de variabilité des valeurs de la variable étudiée.

a) Ecart moyen arithmétique : c'est la moyenne arithmétique des écarts par rapport à la moyenne arithmétique des valeurs du caractère.

$\overline{E} = \frac{1}{N} \sum_{i=1}^{n} ni\ |xi - \bar{x}|$

b) Variance, écart type: la variance d'une série de valeurs du caractère est la moyenne arithmétique des carrés des écarts de ces valeurs par rapport à leur moyenne arithmétique.

$Var(x) = \sigma^2(x) = \frac{1}{N} \sum_{i=1}^{n} ni\ (xi - \bar{x})^2$

L'écart type (écart quadratique moyen) est la racine carré de la variance : $\sigma = \sqrt{var}$.

C'est le plus significatif de tous les paramètres de dispersion.

C) Moments simples d'une distribution: on appelle moments simples d'ordre r d'une distribution la quantité m_r : $m_r = \frac{\sum_{i=1}^{k} ni\ xi^r}{N} = \frac{1}{N} \sum_{i=1}^{k} nixi^r$

- pour r = 1; $m_1 = \frac{\sum_{i=1}^{k} nixi}{N} = \frac{1}{N} \sum_{i=1}^{k} nixi$ = moyenne pondérée des valeurs.

- pour r = 2; $m_2 = \frac{\sum_{i=1}^{k} nixi^2}{N} = \frac{1}{N} \sum_{i=1}^{k} nixi^2$ = moyenne pondérée des carrés des valeurs.

- pour r = 3; $m_3 = \frac{\sum_{i=1}^{k} nixi^3}{N} = \frac{1}{N} \sum_{i=1}^{k} nixi^3$ = moyenne pondérée des cubes des valeurs.

Ex : calculer le moment simple d'ordre 1, 2 et 3 de la distribution de la taille de 400 individus.

d) Moments centrés : on appelle moments centrés d'ordre r d'une série ou une distribution la quantité notée μ_r : $\mu_r = \frac{1}{N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^r$

- pour r =1; $\mu_1 = \frac{1}{N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^1 = 0$

- pour r = 2; $\mu_2 = \frac{1}{N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^2$ = variance = moyenne carrée des écarts centrés.

- pour r = 3; $\mu_3 = \frac{1}{N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^3$ = moyenne des cubes des écarts centrés.

- pour r = 4; $\mu_4 = \frac{1}{N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^4$ = moyenne des puissances 4 des écarts centrés.

3- paramètres de forme : la courbe de distribution ou histogramme peut être caractérisé par plusieurs indices ou critères relatifs (la forme et la régularité de cette courbe).

a) indice de symétrie: on appelle indice ou coefficient de symétrie de la courbe, la quantité définie par γ_1 : $\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{1}{\sigma^3 N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^3$

γ_1 peut être nulle, négative ou positive.

- Si $\gamma_1 = 0$ (proche de 0) ; la courbe est symétrique.

- Si $\gamma_1 < 0$ (donc $\mu_3 < 0$) ; la courbe est asymétrique (dissymétrique) ; la partie de la courbe située à gauche de la moyenne est plus grande que la partie située à droite de la moyenne.

- Si $\gamma_1 > 0$ (donc $\mu_3 > 0$) ; la courbe est asymétrique (dissymétrique) ; la partie de la courbe située à droite de la moyenne est plus grande que la partie située à gauche de la moyenne.

b) indice d'aplatissement: il indique le degré d'aplatissement de la courbe par rapport à celui de la courbe de Gauss (courbe normale). L'expression de ce coefficient est donnée par la formule :

$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{1}{\sigma^4 N} \sum_{i=1}^{k} ni\ (xi - \bar{x})^4 - 3$

Pour la courbe de Gauss : $\frac{\mu_4}{\sigma^4} = 3$ c.-à-d. que $\gamma_2 = 0$

- Si $\gamma_2 = 0$; la courbe étudiée à la même allure (même aplatissement) que la courbe de Gauss.
- Si $\gamma_2 < 0$; la courbe étudiée est plus aplatie que la courbe de Gauss.
- Si $\gamma_2 > 0$; la courbe étudiée est moins aplatie (plus pointue) que celle de Gauss.