II- Analyse univariée

1- introduction

Les premières statistiques correctement élaborées ont été celles des recensements démographiques. Ainsi le vocabulaire statistique est essentiellement celui de la démographie. Les ensembles étudiés sont appelés **population**. Les éléments de la population sont appelés **individus** ou **unités statistiques**. La population est étudiée selon un ou plusieurs caractères.

1- introduction (suite) :

Pour recueillir des informations sur une population statistique, l'on dispose de deux méthodes :

- la méthode exhaustive ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
- la méthode des sondages ou échantillonnage qui conduit à n'examiner qu'une fraction de la population, un échantillon.

L'**échantillonnage** représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité.

2-Notion de base :

Définition

1. *On appelle effectif de la modalité x_i, le nombre n_i des individus ω tel que $X(\omega) = x_i$ (Cas discret)*
2. *On appelle effectif de la classe $[x_{i-1}, x_i[$, le nombre n_i des individus ω tel que $X(\omega) \in [x_{i-1}, x_i[$ (Cas continu)*

Remarque

On a $\sum_{i=1}^{p} n_i = N$. *l'effectif total.*

Définition (Série statistique)

Une série statistique est l'ensemble des couples $(x_i; n_i)$ *ou* $([x_i; x_{i+1}[; n_i)$.

2-Notion de base (suite) :

Définition

On appelle fréquence de la modalité x_i ou de la classe $[x_{i-1}, x_i[$, le nombre f_i tel que $f_i = \frac{n_i}{N}$

Définition

1. *On appelle effectif cumulé en x_i, le nombre* $\sum_{j=1}^{i} n_j$.
2. *On appelle fréquences cumulées en x_i, le nombre F_i tel que* $F_i = \sum_{j=1}^{i} f_j$.

2-Notion de base (suite) :

Remarque

On peut noter que $\sum_{j=1}^{p} n_j = N$ *, taille de l'échantillon*

$$\sum_{j=1}^{p} f_j = 1$$

en effet $\sum_{j=1}^{p} f_j = \sum_{j=1}^{p} \frac{n_i}{N} = \frac{1}{N} \sum_{j=1}^{p} n_i = \frac{1}{N} \times N = 1$

2-Notion de base (suite) :

En général une série statistique à caractère discret se présente sous la forme :

Valeurs	x_1	x_2	$\cdots\ \cdots$	x_p
Effectifs	n_1	n_2	$\cdots\ \cdots$	n_p
Fréquences	f_1	f_2	$\cdots\ \cdots$	f_p

TABLE : caractère discret

et pour un caractère continue, on a la représentation suivante :

classes	$[x_0; x_1[$	$[x_1; x_2[$	$\cdots\ \cdots$	$[x_{p-1}; x_p]$
effectifs	n_1	n_2	$\cdots\ \cdots$	n_p
fréquences	$f_1 = \dfrac{n_1}{N}$	$f_1 = \dfrac{n_2}{N}$	$\cdots\ \cdots$	$f_p = \dfrac{n_p}{N}$

TABLE : caractère continu

2-Notion de base (suite) :

Exemple (caractère discret)

Soit un échantillon de 64 familles. Le caractère étant le nombre d'enfants par famille (tableau 2.1)

x_i	0	1	2	3	4	5
n_i	16	18	14	11	3	2
$f_i = n_i/n$						
E.C.C						
E.C.D						

Avec : x_i : Nombre d'enfants, n_i : Nombres de familles,
$f_i = n_i/n$: Fréquence
E.C.C : Effectifs cumulés croissants
E.C.D : Effectifs cumulés décroissants

2-Notion de base (suite) :

Exemple (caractère discret)

Soit un échantillon de 64 familles. Le caractère étant le nombre d'enfants par famille (tableau 2.1)

x_i	0	1	2	3	4	5
n_i	16	18	14	11	3	2
$f_i = n_i/n$	0.25	0.281	0.218	0.172	0.047	0.031
E.C.C	16	34	48	59	62	64
E.C.D	64	48	30	16	5	2

Avec : x_i *: Nombre d'enfants,* n_i *: Nombres de familles,*
$f_i = n_i/n$ *: Fréquence*
E.C.C : Effectifs cumulés croissants
E.C.D : Effectifs cumulés décroissants

2-Notion de base (suite) :

Exemple (caractère continu)

Soit un echantillon de 80 personnes d'une collectivité portant sur la taille. On adopte un intervalle de classe de 0,05 m. (tableau 3.1)

classe	Effectif	E.C.C	E.D.C
[1.55 ;1.60[	3		
[1.60 ;1.65[	12		
[1.65 ;1.70[	18		
[1.70 ;1.75[	25		
[1.75 ;1.80[	15		
[1.80 ;1.85[	5		
[1.85 ;190[	2		

2-Notion de base (suite) :

Exemple (caractère continu)

Soit un echantillon de 80 personnes d'une collectivité portant sur la taille. On adopte un intervalle de classe de 0,05 m. (tableau 3.1)

classe	Effectif	E.C.C	E.C.D
[1.55 ;1.60[	3	3	80
[1.60 ;1.65[	12	15	77
[1.65 ;1.70[	18	33	65
[1.70 ;1.75[	25	58	47
[1.75 ;1.80[	15	73	22
[1.80 ;1.85[	5	78	7
[1.85 ;190[	2	80	2

3- Paramètres statistiques :

1- Paramètres de position

a/ Dominante ou mode

Définition

Lorsque la variable est discrète, **une dominante** *ou* **mode** *est une valeur du caractère qui correspond à un effectif maximum.*
La série est unimodale, bimodale $\cdots$ *lorsque le nombre de modes est 1, 2* $\cdots$*.*
Lorsque la variable est continue, une ***classe modale*** *correspondra à un effectif maximum.*

Exemples

1. Considérons la série statistique de l'exemple 2 : le mode est 1 enfant puisqu'il est associé à l'effectif maximum 18.
2. Considérons la série statistique de l'exemple 4 : la classe modale est $[1.70 - 1.75[$ puisqu'il est qui corespond à l'effectif maximum 25.

3- Paramètres statistiques (suite) :

b/ Moyenne arithmétique

Définition

Lorsque la variable est discrète la moyenne arithmétique $\overline{X}$ de la série statistique est la moyenne pondérée

$$\overline{X} = \frac{\sum_{i=1}^{p} n_i x_i}{\sum_{i=1}^{p} n_i} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N}$$

Lorsque la variable est continue la moyenne est :

$$\overline{X} = \frac{\sum_{i=1}^{p} n_i c_i}{\sum_{i=1}^{p} n_i}$$

où les $c_i = \dfrac{x_i + x_{i+1}}{2}$ *sont les centres des classes.*

3- Paramètres statistiques (suite) :

Théorème

- $\overline{X+b} = \overline{X} + b$
- $\overline{aX} = a\overline{X}$

Exemples

1)Le nombre moyen d'enfants par famille de la série statistique de l'exemple 2 est :

$$\begin{aligned} \overline{X} &= \\ &= \end{aligned}$$

2)Pour la série statistique de l'exemple 4, la taille moyenne d'une personne est

$$\begin{aligned} \overline{X} &= \\ &= \end{aligned}$$

Exemples

1)Le nombre moyen d'enfants par famille de la série statistique de l'exemple 2 est :

$$\begin{aligned} \overline{X} &= \frac{1}{64}(16 \times 0 + 1 \times 18 + 2 \times 14 + 3 \times 11 + 4 \times 3 + 5 \times 2) \\ &= \frac{90}{64} \end{aligned}$$

2)Pour la série statistique de l'exemple 4, la taille moyenne d'une personne est

$$\begin{aligned} \overline{X} &= \frac{3 \times 1.575 + 12 \times 1.625 + 18 \times 1.675 + 25 \times 1.725}{80} \\ &\quad + \frac{15 \times 1.775 + 5 \times 1.825 + 2 \times 1.875}{80} \\ &= \frac{137}{80} \end{aligned}$$

3- Paramètres statistiques (suite) :

c/ Médiane

Définition

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.

Remarques :

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

- si n est impair, alors $n = 2m + 1$ et la médiane est la valeur du milieu $Me = x_{m+1}$.
- si n est pair, alors $n = 2m$ et une médiane est une valeur quelconque entre x_m et x_{m+1}. Dans ce cas $Me = \dfrac{x_m + x_{m+1}}{2}$.

▶ Dans le cas où les valeurs prises par le caractère étudié sont groupées en classe, on cherche la classe contenant le $\dfrac{n^e}{2}$ individu de l'échantillon. En supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur,la position exacte du $\dfrac{n^e}{2}$ individu est déterminée de la façon suivante (par interpolation linéaire) :

3- Paramètres statistiques (suite) :

$$\frac{M_e - x_m}{\frac{N}{2} - N_m} = \frac{x_{m+1} - x_m}{N_{m+1} - N_m} \quad \text{donc}$$

$$M_e = x_m + (x_{m+1} - x_m)\left(\frac{\frac{N}{2} - N_m}{N_{m+1} - N_m}\right)$$

avec $m \in \mathbb{N}$ telle que $N_m \leq \frac{N}{2} < N_{m+1}$ et N_m effectif cumulé de la classe $[x_m, x_{m+1}[$

Remarque

Dans le cas d'une série groupée en classes, la médiane est représentée aussi par la valeur de la variable correspondant à l'intersection du polygône des effectifs cumulés avec l'horizontale représentant l'effectif moitié

Exemples :

1) Les données suivantes représentent le capital social en 10^3 de 17 sociétés marocaines créées entre le 20 et 24-10-1995 (les valeurs du caractère étudié sont classé par ordre croissant) :

$$\overbrace{10;\ 10;\ 20;\ 20;\ 30;\ 50;\ 50;\ 50};\quad \overset{\texttt{médian}}{\overset{\downarrow}{90}}\ ;\underbrace{100;\ 100;\ 100\ ;\ 100\ ;\ 200\ ;\ 200\ ;\ 200\ ;\ 300}.$$

Le nombre de valeurs observées est 17 (impaire). Dans ce cas, le capital médian est le neuvième (c.à.d 90) car il divise le nombre de sociétés en 2 ensembles égaux : 8 sociétés sont créées avec un capital inférieur à 90 000 DH et 8 autres sont créées avec un capital supérieur à 90 000 DH.

Donc Me=90 000

Exemples :[suite]

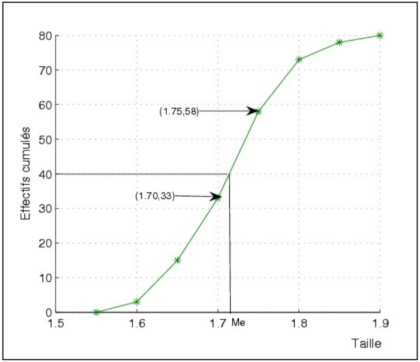
2) Les capacités de production de 10 sucreries au Maroc en 1993 (en 10^3 tonnes) sont les suivants :

$\overbrace{30;\ 35;\ 35;\ 45;}\ 50 \underset{;}{\overset{\text{médian}}{\downarrow}} 51;\ \overbrace{78;\ 80;\ 90\ ;\ 500}$. Le nombre de valeurs observées est 10 (paire) ; la médiane est, en effet, située ente la cinquième capacité (50) et la sixième (51). Dans ce cas-la, on peut estimer la production médiane par : $\dfrac{50+51}{2} = 50.5$

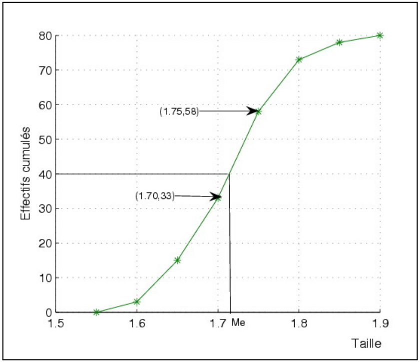
3) Médiane associée à la série statistique de l'**exemple2** : on a $N = 64$, les valeurs centrales sont la $32^{ème}$ valeure et la $33^{ème}$ valeure qui sont égales à 1. Donc la médiane est égale à 1.

4)Médiane associée à la série statistique de l'**exemple4** : on a $N = 80$ chercher la médiane revient à trouver la taille de la $40^{ème}$ personne. On procède par interpolation linéaire en utilisant le polygône des effectifs cumulés.

Exemples : [suite]



Exemples : [suite]



On a $\dfrac{Me - 1.70}{40 - 33} = \dfrac{1.75 - 1.70}{58 - 33}$.

$$\Longrightarrow Me = 1.70 + \frac{7\,(1.75 - 1.70)}{25} = 1.71\,.$$

3- Paramètres statistiques (suite) :

Caractéristiques de disperssion
Les paramètres de position sont insuffisants pour caractériser complètement une série.
Par exemple, deux séries différentes ayant la même moyenne, ne se repartissent pas nécessairement de la même manière autour de cette moyenne. Elles sont plus ou moins étalées, ce qui sera décrit par les caractéristiques de dispersion.
Un paramètre de dispersion se rapporte a la différence de deux valeurs du caractère alors qu' un paramètre de position représente une valeur du caractère

3- Paramètres statistiques (suite) :

a/ Ecart moyen arithmétique

Définition

Écart moyen arithmétique est la moyenne arithmétique des écarts par rapport à la moyenne arithmétique $\overline{X}$ des valeurs du caractère

$$\overline{E(X)} = \frac{1}{N}\sum_{i=1}^{p} n_i|x_i - \overline{X}|$$

Lorsque la variable est continue l'écart moyen arithmétique est :

$$\overline{E(X)} = \frac{1}{N}\sum_{i=1}^{p} n_i|c_i - \overline{X}|$$

où les $c_i = \dfrac{x_i + x_{i+1}}{2}$ *sont les centres des classes.*

3- Paramètres statistiques (suite) :

b/ Variance

Définition

La variance d'une série de valeurs du caractère est la moyenne arithmétique des carres des écarts de ces valeurs par rapport à leur moyenne arithmétique.

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2$$

Lorsque la variable est continue l'écart moyen arithmétique est :

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(c_i - \overline{X})^2$$

où les $c_i = \dfrac{x_i + x_{i+1}}{2}$ *sont les centres des classes.*

3- Paramètres statistiques (suite) :

Théorème

- $V(X) = \dfrac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2$
- $V(X+b) = V(X)$
- $V(aX+b) = a^2 V(X)$

c/ Ecart-type

Définition

L' écart-type (ou écart quadratique moyen) est la racine carrée de la variance $\sigma = \sqrt{V}$

3- Paramètres statistiques (suite) :

Définition

L'étendue d'une série *est la difference entre la plus grande et la plus petite valeur du caractère.*

d/ D'autres définitions

Soit $(x_i, n_i)_{i\in[1,p]}$ ou $(x_i, f_i)_{i\in[1,p]}$ une série statistique discrète.

Soit N son effectif total et supposons que pour tout indice i, x_i est strictement positif. On appelle

- **moyenne harmonique** de la série est le nombre h définie par

$$\frac{1}{h} = \frac{1}{N}\sum_{i=1}^{p}\frac{n_i}{x_i} = \sum_{i=1}^{p}\frac{f_i}{x_i} \quad \left(\text{moyenne de } \left(\frac{1}{x_i}\right)\right)$$

3- Paramètres statistiques (suite) :

- **moyenne géométrique** de la série est le nombre g définie par :

$$g = \left(\prod_{i=1}^{p} x_i^{n_i}\right)^{1/N} = \prod_{i=1}^{p} x_i^{f_i}.$$

- **Moment d'une série statistique :** on appelle moment d'ordre q par rapport à x_0 la moyenne arithmétique des puissances $q^{ièmes}$ des déviations des valeurs du caractère par rapport à x_0 notée :$m_q = \frac{1}{N}\sum_{i=1}^{p} n_i\,(x_i - x_0)^q$.

- Si $x_0 = 0$ et $q = 1$, le moment n'est rien d'autre que la moyenne.
- Si $x_0 = \overline{x}$ et $q = 2$, le moment n'est rien d'autre que la variance.