Partie II : Etude d'une variable statistique discrète

La variable statistique discrète peut prendre un nombre fini raisonnable de valeurs (notes, nombre d'enfants,...). Dans ce cas, le caractère étudié est appelé caractère discret.

Nous considérons dans tous ce qui suit la situation suivante :

$$X: \Omega \rightarrow \{x_1, x_2, x_3, \ldots\ldots x_n\}$$

Avec card Ω=N (nombre d'individus dans notre étude)

Exemple 1: *Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille. On note X le nombre d'enfants. Les résultats sont donnés dans le tableau suivant :*

$X = x_i$	0	1	2	3	4	5	6
$n_i(effect)$	18	32	66	41	32	9	2

Nous avons :- Ω ensemble des familles

- w une famille

- X nombre d'enfants par famille

On note $X: w \rightarrow X(w)$

On lit à la famille w on associe $X(w)$ *nombre d'enfants de cette famille*

II.1 Effectif partiel (fréquence absolu)

Définition : *Pour chaque valeur* x_i *on associe le nombre d'individu correspondant qu'on appelle effectif partiel ou fréquence absolue.*

Exemple 2*: Dans l'exemple précédent 66 représente le nombre de famille (effectif partiel ou absolu) qui ont deux enfants*

II.2 Effectif cumulé

Définition : *Pour chaque valeur* x_i *on calcule l'effectif cumulé qui est la somme de l'effectif de cette valeur et de toutes les valeurs précédentes et on le note* N_i

Exemple 3*: Calculons les effectifs cumulés correspondants à l'exemple précédent*

x_i	0	1	2	3	4	5	6
N_i	18	50	116	157	189	198	200

N_i *est le nombre d'individus dont la valeur du caractère est inferieur ou égale à* x_i *. De ce fait l'effectif total est donné par :* $N = \sum_{i=1}^{n} n_i$

II.3 Fréquence partiel

Définition : *La fréquence partielle notée* f_i *d'une valeur* x_i *est le rapport de l'effectif de cette valeur sur l'effectif total. Soit* $f_i = \frac{n_i}{N}$

Remarque *: On peut remplacer* f_i *par* $f_i x100$ *qui représente alors un pourcentage*

Exemple 4*: voir l'exemple précédent*

Proposition *: Soit* f_i *définie comme précédemment. Alors,*$\sum_{i=1}^{n} f_i = 1$

Démonstration :

Nous avons $N = \sum_{i=1}^{n} n_i$ *ceci implique :* $\sum_{i=1}^{n} f_i = \sum \frac{n_i}{N} = \frac{1}{N} \sum \quad n_i = \frac{N}{N} = 1$

II.4 Fréquence cumulée

Définition *: pour chaque valeur*x_i*, on pose par définition*

$F_i = f_1 + f_2 + \cdots + f_i$

La quantité F_i *s'appelle la fréquence cumulée de* x_i

Exemple 5*: Dans l'exemple précédent 0.785 représente 78.5% de familles dont le nombre d'enfants est inferieur ou égale à 3*

Exemple 6 *: on s'intéresse aux nombres d'erreurs d'assemblage sur un ensemble d'appareils*

Nombre d'erreurs	*Nombre d'appareils*	*Fréquences cumulées*
0	*101*	*0.26*
1	*140*	*0.61*
2	*92*	*0.84*
3	*42*	*0.94*
4	*18*	*0.99*
5	*3*	*1*

Nous avons 94% des appareils qui ont un nombre d'erreurs d'assemblage inferieur ou égale à 3.

Les tableaux sont un moyen souvent indispensable et très utile dans la classification et la présentation des unités d'une population statistique. Dans ce qui suit nous allons voir *comment on traduit ses tableaux en graphique permettant de résumer d'une man ière visuelle les données.*

II.5 Représentation graphique des séries statistiques

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette dernière. Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe).

Le graphique est un support visuel qui permet :

- *La synthèse : visualiser d'un seul coup d'œil les principales caractéristiques*
- *La découverte : met en évidence les tendances*
- *Le contrôle : on aperçoit mieux les anomalies sur un graphique que dans un tableau*
- *La recherche des régularités : régularité dans le mouvement, répétition du phénomène*

II.5.1 Distribution à caractère qualitatif

A partir de l'observation d'une variable qualitative, deux diagrammes permettent de représenter cette variable : diagramme en bandes (tuyaux d'orgues) et le diagramme à secteurs circulaires (camembert).

***Exemple** 7: le tableau suivant donne les chaines de télévisions préférées par des élevés de terminales*

Chaines TV	*TF1*	*F2*	*F3*	*C+*	*M6*
effectifs	*4*	*3*	*0*	*1*	*2*
F_i					

Diagramme en bandes

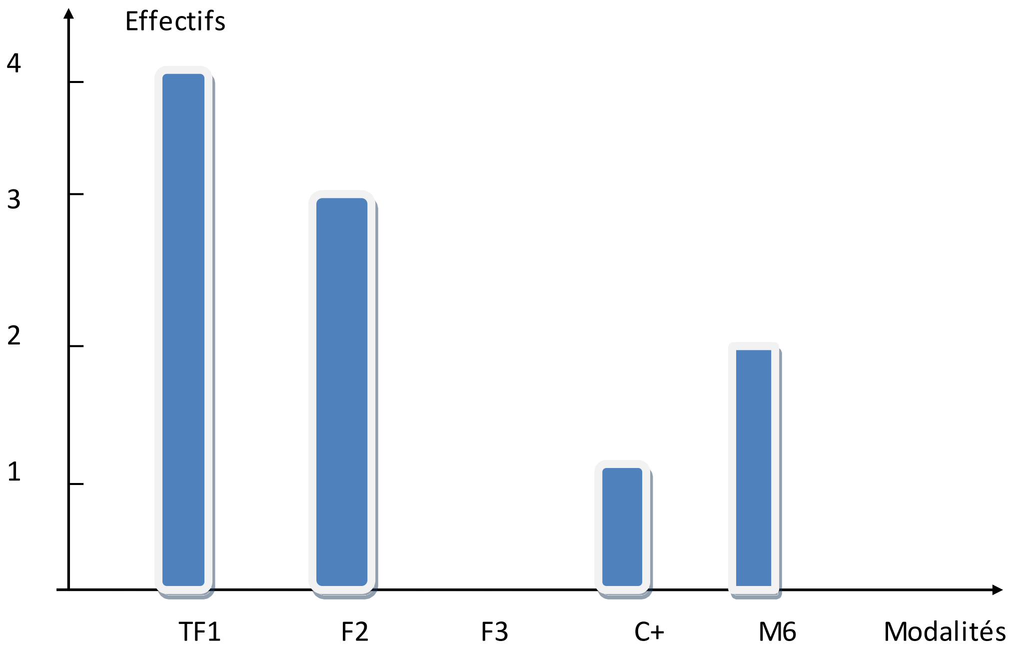


Diagramme en bandes

***Exemple 8**: on garde les mêmes données pour un digramme circulaire sauf qu'ici il faut calculer l'angle correspondant à chaque modalité selon la formule suivante :*

Le degré d un secteur est déterminé à l aide d'une règle de trois que voici :

$$N \rightarrow 360^{\circ}$$

$$n_i \rightarrow \alpha_i$$

D'où $\alpha_i = \frac{n_i X 360}{N}$

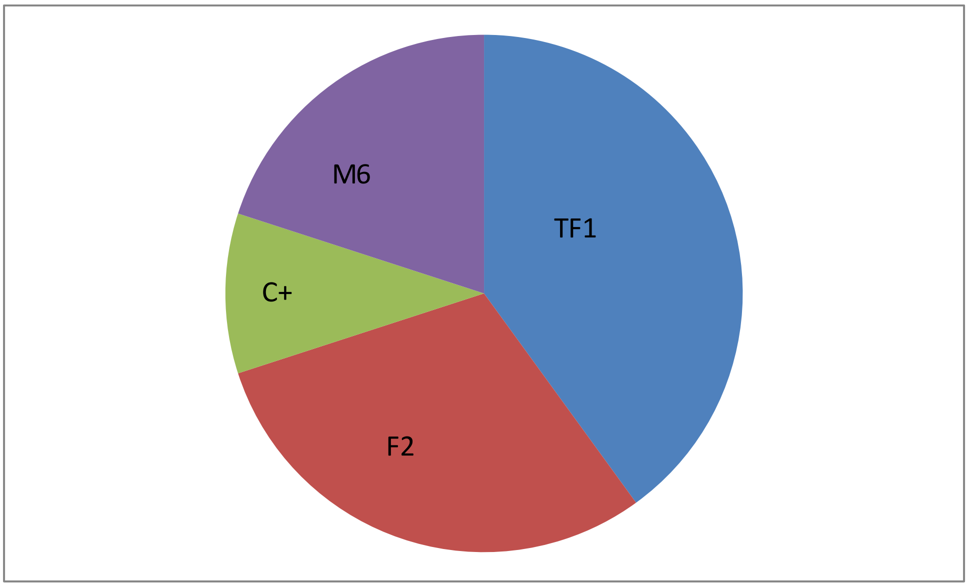


Diagramme circulaire (diagramme en camembert)

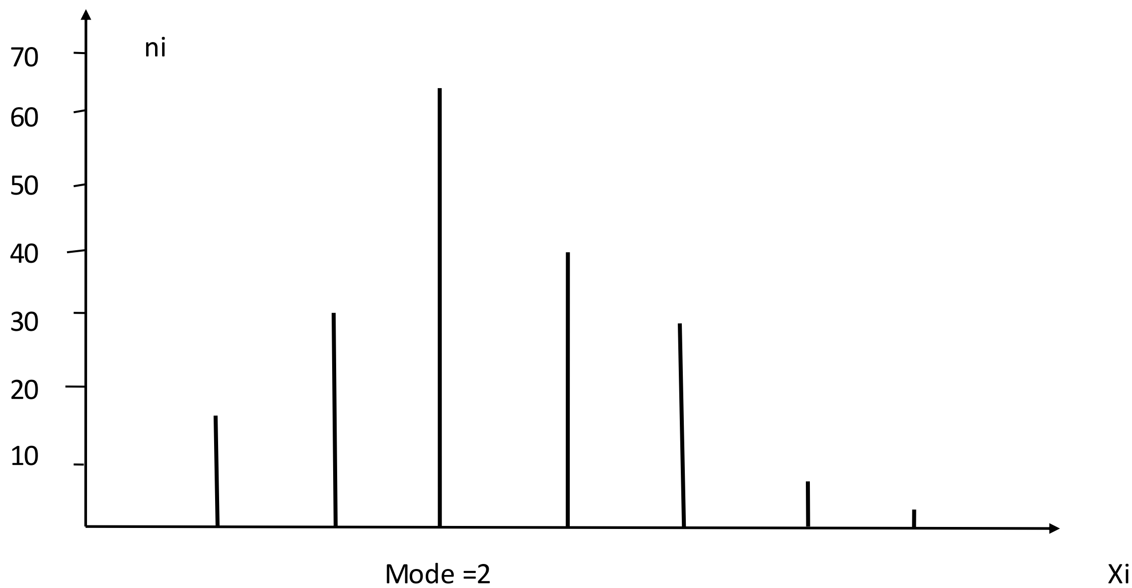
II.5.2 Distribution à caractère quantitatif discret

Une variable discrète est représentée par deux diagrammes à savoir le diagramme en bâtons et le diagramme cumulatif

***Exemple** 9: voir les données de l'exemple vu ci dessus*

$X = x_i$	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

Diagramme en bâtons



Nous allons exploiter les valeurs cumulées dont on avait parlé auparavant pour introduire la notion de fonction de répartition. Cette notion ne concerne que les variables quantitatives.

II.5.3 Représentation sous forme de courbe et fonction de répartition

Soit la fonction $F_x: IR \rightarrow [O,1]$

$F_x(x) = porcentage\ des\ individus\ dont\ la\ valeur\ est\ \leq x$

Cette fonction s'appelle la fonction de répartition de caractère X

Exemple 10*: voir le tableau de l'exemple ci-dessus*

- *si* $x < 0, alors\ F_x(x) = 0$

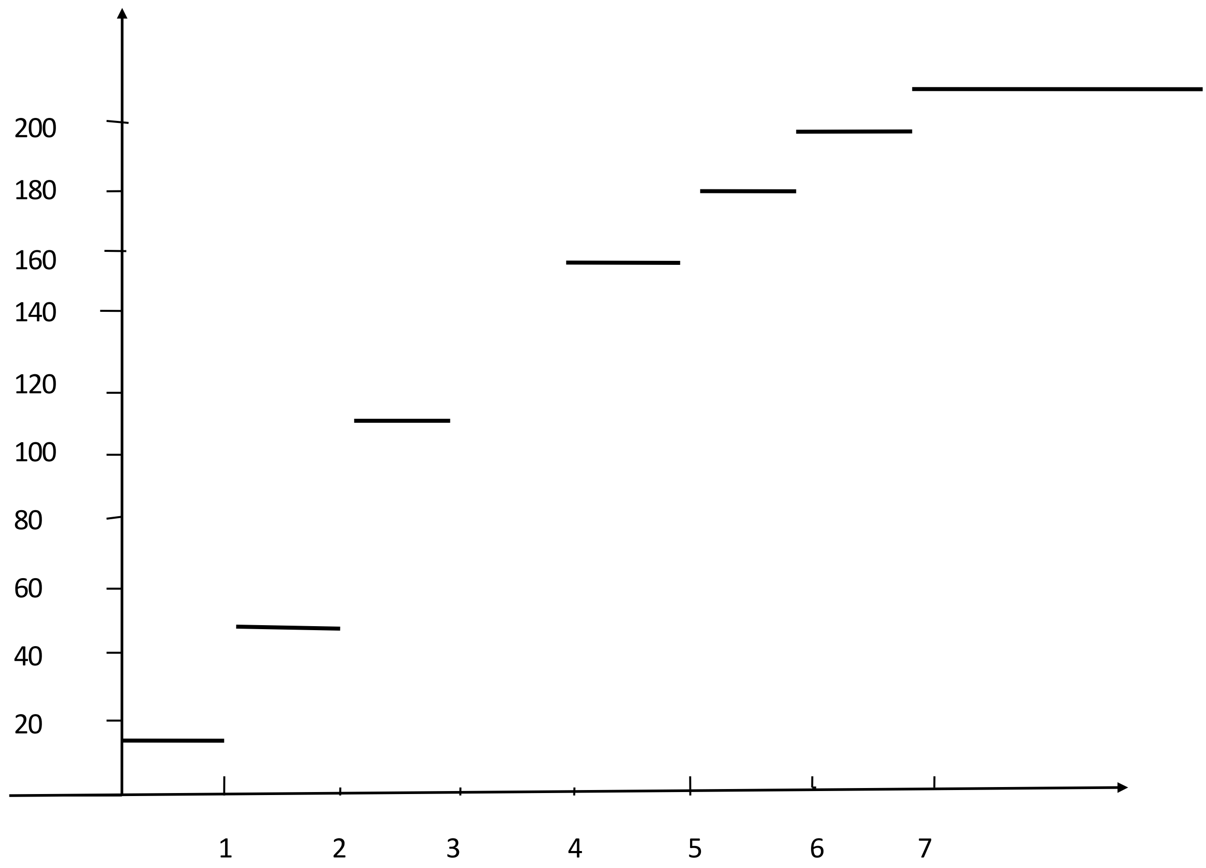
- *si* $x \in [0,1], alors\ F_x(x) = 0.09$

-

-

- *si* $x \geq 6, alors\ F_x(x) = 1$

Cette courbe s'appelle courbe cumulative des fréquences, cette courbe est une courbe en escalier représentant les fréquences cumulées relatives.



Courbe cumulative en escalier

Proposition *: la fonction de répartition satisfait, pour tout* $i \in \{1,2,\ldots\ldots n\}$

- l'égalité $F_x(x_i) = F_i$

-l'expression, $F_x(x) = \begin{cases} 0, & si\, x < x_1 \\ F_1, & si\, x_1 \leq x < x_2 \\ F_i & si\, x_i \leq x < x_{i+1} \\ 1 & si \quad x \geq x_n \end{cases}$

II.6 Paramètres de position (paramètres de tendance centrale)

Les indicateurs statistiques de tendance centrale dit aussi paramètres de position considérés fréquemment sont : la moyenne, la médiane et le mode

II.6.1 Le mode *: Le mode d'une variable statistique est la valeur de la variable qui correspond au plus effectif (la plus grande fréquence), il est noté* M_O

Exemple 11: *Voir les données de l'exemple 1, (diagramme en bâtons,* $M_O = 2$)

Graphiquement il se détermine avec le diagramme en bâtons

Remarque *: On peut avoir plusieurs modes ou rien*

II.6.2 La médiane *: La médiane noté* M_e *est la valeur qui correspond à 50% des observations, elle divise la série en deux parties égales. La détermination de la médiane dans le cas discret se fait de deux manières :*

1^er^ cas : Nombre d'observation pair (n pair)

On doit tout d'abord ranger les valeurs de la série par ordre croissant ou décroissant ensuite on applique la formule suivante : $M_e = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$

Exemple :

$X = x_i$	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2
N_i	18	50	116	157	189	198	200

On n=200 (pair) donc $M_e = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{100} + X_{101}}{2} = 2$

2eme cas : Nombre d'observation impair (n impair)

La même chose on doit ranger les valeurs d'abord, ensuite déterminer la médiane par la formule suivante : $M_e = X_{\frac{n+1}{2}}$

Exemple *: Soit la série suivante : 14, 15, 6,7, 27*

Avant de calculer la médiane il faut ranger la série par ordre croissant : 6, 7, 14, 15, 27 Alors la médiane dans ce cas est égale : $M_e = X_{\frac{n+1}{2}} = X_{\frac{5+1}{2}} = X_3 = 14$

II.6.3 La moyenne

Lorsque X désigne la variable statistique, la valeur moyenne ou moyenne de la série se note m ou $\overline{X}$ *.*

- *1^er^ cas : Si les observations ne sont pas groupées (la série est dite non*

classés) : $\bar{X} = \frac{1}{n}\sum_1^n X_i avec\ n :$
$effectif\ total\ et\ X_i : ieme\ valeur\ de\ la\ variable$

Exemple *: les notes d'un élève aux examens sont les suivantes : 10, 6, 4, 7*

Sa moyenne est : $\bar{X} = \frac{1}{n}\sum_1^n X_i = \frac{1}{4}(10 + 6 + 4 + 7) = 7.75$

- *2eme cas : Si les observations sont groupées (la série est dite classée)*

$\bar{X} = \frac{1}{N}\sum_1^n X_i\, n_i = \sum_1^n f_i X_i, \qquad N = \sum_1^N n_i$

Avec : $X_i : ieme\ valeur\ de\ la\ valeur$

$n_i : ieme\ effectif\ de\ la\ variable$

$f_i : ieme\ fréquence\ de\ la\ variable$

$N : effectif\ total$

Il s'agit en fait d'une moyenne arithmétique pondérée

Exemple *: Un même article est vendu dans différents magasins aux prix indiqués dans le tableau ci-dessous :*

Prix	*Nbre. magasin*	$X_i n_i$
10	*20*	*200*
12	*40*	*480*
14	*30*	*720*
15	*10*	*150*
Total	***100***	***1250***

La moyenne dans ce cas est : $\bar{X} = \frac{1}{N}\sum_1^n X_i\, n_i = \frac{1250}{100} = 12.5$

Même chose avec la formule avec les f_i

II.6.3 Les quartiles

Dans le cas discret la détermination du 1er, 2eme et 3eme quartile se fait a l'aide de la formule suivante :

$$Q_\alpha = \begin{cases} \dfrac{X_{\frac{\alpha n}{4}} + X_{\frac{\alpha n}{4}+1}}{2} & si\ \dfrac{\alpha n}{4} \in IN \\ X_{\left[\frac{\alpha n}{4}\right]+1} & \dfrac{\alpha n}{4} \in IN \end{cases}$$

II.5.3 Paramètres de dispersion

Les indicateurs de dispersion les plus importants sont la variance et l'écart type

1. L'étendue : *noté e, représente la différence entre les valeurs extrêmes de la série, soit* $e = X_n - X_1$, *ainsi la série* $\{2,6,10,14,18\}$ *a pour étendue e=18-2=16*

2. Intervalle interquartile : *noté I, est la différence entre le 3eme et le 1er quartile, soit* $I = Q_3 - Q_1$

Cet intervalle contient 50% des observations, cette caractéristique est nettement meilleure que l'étendue.

3. La variance *: la variance de la variable statistique X noté V(X) est la moyenne des carrés des écarts à la moyenne arithmétique. C'est la caractéristique de dispersion la plus utilisée avec l'écart type.*

Remarque *: la variance est toujours un nombre positif.*

1er cas *: série classée*

$$V(X) = \frac{1}{n}\sum_{1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{1}^{n} {X_i}^2 - \bar{X}^2$$

2eme cas *: série classée*

$$V(X) = \frac{1}{n}\sum_{1}^{n} n_i(X_i - \bar{X})^2 = \frac{1}{n}\sum_{1}^{n} n_i {X_i}^2 - \bar{X}^2 = \sum f_i {X_i}^2 - \bar{X}^2$$

4. L'écart type *: l'écart type noté* $\sigma(X)$ *est la racine carré de la variance*

Et donc $\sigma(X) = \sqrt{V(X)}$