CH I : **Statistique descriptive**

Cas unidimensionnel

I) Terminologie de base

- Population ou champ d'étude Ω : c'est l'ensemble concerné par l'étude statistique
- Unité statistique ou individu ω : c'est un élément ou membre de la population
- Échantillon : c'est un sous ensemble de la population sur lequel sont réalisées les observations ou les mesures
- Taille de l'échantillon : est le nombre d'individus de l'échantillon; c'est la cardinalité de l'échantillon

- **Enquête :** opération qui consiste à observer tous les individus d'un échantillon
  - **Recensement :** est une enquête exhaustive; l'échantillon observé est la population entière
  - **Sondage :** est une enquête sur une partie de la population
- **Variable** ou **Caractère :** est une caractéristique observée ou mesurée sur les individus d'une population ou un échantillon
  - Modalité : est une valeur possible prise par une variable.
  - Domaine d'une variable : ensemble des valeurs possibles ou des modalités

1) Les différents types de variables

Une variable peut être quantitative ou qualitative :

- Une variable est dite quantitative quand elle est mesurable. Ses modalités sont des valeurs numériques sur lesquelles on peut effectuer des opérations algébriques
- Une variable est dite qualitative lorsque ses modalités expriment une qualité et non pas une quantité. Une variable qualitative est définie par un ensemble fini de modalités permettant de caractériser les observations.

a) Variables qualitatives

- Variable qualitative nominale **:** les modalités de la variable ne sont pas hiérarchisées ou ordonnées(*exemple: groupe sanguin*).
- Variable qualitative binaire : c'est une variable à deux modalités (*exemple: le succès ou l'échec d'un traitement*). Les deux modalités sont souvent codées 0 et 1, le 1 étant utilisé par convention pour la modalité d'intérêt.
- Variable qualitative ordinale : est une variable dont les modalités sont reliées par une relation d'ordre (*exemple : Niveau d'étude*)

b) Variables quantitatives :

- Variable quantitative continue : on parle de variable continue quand celle-ci peut prendre une quantité indénombrable de valeurs dans un intervalle.
  (exemple : la taille)
- Variable quantitative discrète : sont des variables qui ne peuvent prendre que des valeurs isolées.
  (Exemple : le nombre de frères et sœurs)

2) Tableaux statistiques :

La première étape consiste à organiser les valeurs dans un tableau :

- dans la première colonne du tableau, on liste les valeurs ou les classes distinctes de la série en ordre croissant
- dans les autres colonnes, on présentent les effectifs (ou les fréquences) et les effectifs cumulés (ou les fréquences cumulées) :
  - Effectif d'une modalité : le nombre de répétitions de la modalité, noté $\boldsymbol{n_i}$
  - fréquence d'une modalité : la proportion de l'effectif de la modalité par rapport à l'effectif total, notée $\boldsymbol{f_i}$
  - Effectif cumulé d'une modalité : est le cumul des effectifs des modalités qui lui sont inférieures ou égales, noté $\boldsymbol{N_i}$
  - Fréquence cumulée d'une modalité : est le cumul des fréquences des modalités qui lui sont inférieures ou égales, noté $\boldsymbol{F_i}$

a) Variable quantitative discrète

Ligne des titres

Modalités rangées par ordre croissant

Modalité	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
$\boldsymbol{x_1}$	n_1	n_1	$f_1=n_1/n$	f_1
$\boldsymbol{x_2}$	n_2	n_1+n_2	$f_2=n_2/n$	f_1+f_2
$\boldsymbol{x_3}$	n_3	$n_1+n_2+n_3$	$f_3=n_3/n$	$f_1+f_2+f_3$
⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x_{k-1}}$	n_{k-1}	$n_1+\cdots+n_{k-1}$	$f_{k-1}=n_{k-1}/n$	$f_1+\cdots+f_{k-1}$
$\boldsymbol{x_k}$	n_k	n	$f_k=n_k/n$	1
Total	$\sum_{i=1} n_i = n$		$\sum_i f_i = 1$	

Ligne des totaux

tableau statique

Exemple 1:

119,5	119,9	119,9	119,5
120,0	119,0	120,1	119,8
119,2	119,4	120,3	119,9
120,1	120,1	119,4	119,4
119,2	119,2	119,4	119,5

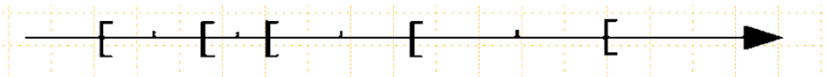
Série statistique

x_i	n_i	N_i	f_i	F_i
119,0	1	1	0,05	0,05
119,2	3	4	0,15	0,20
119,4	4	8	0,20	0,40
119,5	3	11	0,15	0,55
119,8	1	12	0,05	0,60
119,9	3	15	0,15	0,75
120,0	1	16	0,05	0,80
120,1	3	19	0,15	0,95
120,3	1	20	0,05	1
Total	20		1	

Tableau statistique

b) Variable quantitative continue

- ✓ On discrétise une variable quantitative continue en découpant son domaine de variation en classes définies par des intervalles semi ouverts à droite sans discontinuité ni chevauchement.



- ✓ On regroupe les valeurs observées par classes pour cela il faut définir :
  - le nombre de classes k
  - les limites des classes (bornes des intervalles) : $a_1, a_2, ..., a_{k+1}$
  - les centres de classes $x_1, ..., x_k$: $x_i = \frac{a_i + a_{i+1}}{2}$
  - On appelle amplitude de la classe i la longueur de la classe i :

$$l_i = (a_{i+1} - a_i)$$

- ✓ On calcule l'effectif de chaque classe : le nombre d'observations appartenant à la classe

Tableau statistique :

Classe	Centre	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée	Densité des effectifs (ou des fréquences)
$[\boldsymbol{a_1}, \boldsymbol{a_2}[$	x_1	n_1	n_1	f_1=n_1/n	f_1	$d_1 = n_1/(\boldsymbol{a_2} - \boldsymbol{a_1})$
$[\boldsymbol{a_2}, \boldsymbol{a_3}[$	x_2	n_2	$n_1 + n_2$	f_2=n_2/n	$f_1 + f_2$	$d_2 = n_2/(\boldsymbol{a_3} - \boldsymbol{a_2})$
$[\boldsymbol{a_3}, \boldsymbol{a_4}[$	x_3	n_3	$n_1 + n_2 + n_3$	f_3=n_3/n	$f_1 + f_2 + f_3$	$d_3 = n_3/(\boldsymbol{a_4} - \boldsymbol{a_3})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$[\boldsymbol{a_{k-1}}, \boldsymbol{a_k}[$	x_{k-1}	n_{k-1}	$n_1 + \cdots + n_{k-1}$	f_{k-1}=n_{k-1}/n	$f_1 + \cdots + f_{k-1}$	⋮
$[\boldsymbol{a_k}, \boldsymbol{a_{k+1}}[$	x_k	n_k	n	f_k=n_k/n	*1*	$d_k = n_k/(\boldsymbol{a_k} - \boldsymbol{a_{k-1}})$
Total		$\sum_{i=1} n_i = n$		$\sum_i f_i = 1$		

Exemple 2:

On mesure la taille en centimètres de 50 élèves d'une classe :

152 152 152 153 153 154 154 154 155 155 156 156 156 156 156 157 157 157 158 158 159 159 160 160 160 161 160 160 161 162 162 162 163 164 164 164 164 165 166 167 168 168 168 169 169 170 171 171 171 171.

- Nous allons grouper les valeurs en cinq classes de même amplitude =4 :
  [151,5 ; 155,5[, [155,5 ; 159,5[, [159,5 ; 163,5[, [163,5 ; 167,5[, [167,5 ; 171,5[

Tableau Statistique

$[a_i ; a_{i+1}[$	x_i	n_i	N_i	f_i	F_i	d_i
[151,5 ; 155,5[	153,5	10	10	0,20	0,20	2,50
[155,5 ; 159,5[	157,5	12	22	0,24	0,44	3
[159,5 ; 163,5[	161,5	11	33	0,22	0,66	2,75
[163,5 ; 167,5[	165,5	7	40	0,14	0,80	1,75
[167,5 ; 171,5[	169,5	10	50	0,20	1,00	2,50
Total		**50**		**1,00**		

Tableau statique

Nombre de classes

En générale les classes sont déterminées selon la nature du caractère étudier. Mais il existent des formules pour calculer de manière théorique le nombre de classes :

- La règle de Sturge : $k = [1 + 3{,}3 \log_{10}(n)]$
- La règle de Yule : $k = [2{,}5\sqrt[4]{n}]$
- L'amplitude commune des classe est obtenue de la manière suivante :
  $l = \frac{1}{k} \quad (x_{max} - x_{min})$
  x_{max} *et* x_{min} désignent la plus grande et la plus petite valeur observées.

Tableau statistique

Exemple 3

119,5	119,9	119,9	119,5
120,0	119,0	120,1	119,8
119,2	119,4	120,3	119,9
120,1	120,1	119,4	119,4
119,2	119,2	119,4	119,5

Si on applique la régle de Yule on trouve :

$k = [2{,}5\sqrt[4]{20}] = 5$

$a_1 = 118{,}95$ *et* $a_6 = 120{,}35$

$l = 0{,}2(120{,}35 - 118{,}95) = 0{,}28$

Classes	x_i	n_i	N_i	f_i	F_i	d_i
[118,95 , 119,23[	119,09	4	4	0,20	0,20	14,29
[119,23 , 119,51[	119,37	7	11	0,35	0,55	25
[119,51 , 119,79[	119,65	0	11	0,00	0,55	0
[119,79 , 120,07[	119,93	5	16	0,25	0,80	17,86
[120,07 , 120,35[	120,21	4	20	0,20	1,00	14,29
Total		20		1		

Tableau statistique

II) REPRÉSENTATIONS GRAPHIQUES

- ✓ Le choix de la représentation graphique est fortement lié à la nature des variables étudiées.
- ✓ Les représentations les plus fréquemment utilisées sont :
  - ➢ Représentations des effectifs (ou des fréquences):
    - diagramme en bâtons pour les variables discrètes
    - histogramme pour les variables continues
  - ➢ Représentations des effectifs cumulés (ou des fréquences cumulées)
  - ➢ Représentations de la dispersion des valeurs : boîte à moustaches

1) Distribution à caractère quantitatif discret

Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir ci-dessous).

Pour l'illustration, nous prenons l'exemple précédent de départ (nombre d'enfants par famille). Nous rappelons le tableau statistique associe.

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

Diagramme à bâtons

On veut représenter cette répartition sous la forme d'un diagramme en bâtons. À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés

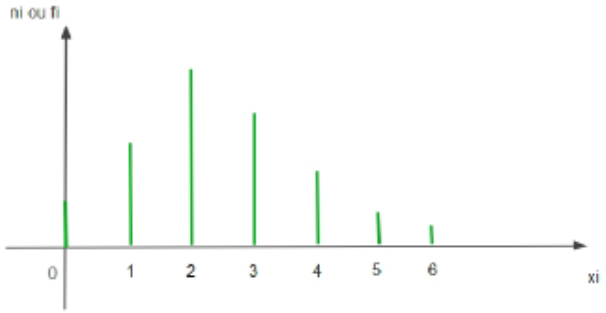


Diagramme à bâtons

Représentation sous forme de courbe et fonction de répartition

Soit la fonction $F_x: \mathbb{R} \longrightarrow [0;1]$ définie par

$F_x :=$ Pourcentage des individus dont la valeur du caractère est $\leq x$.

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

Proposition

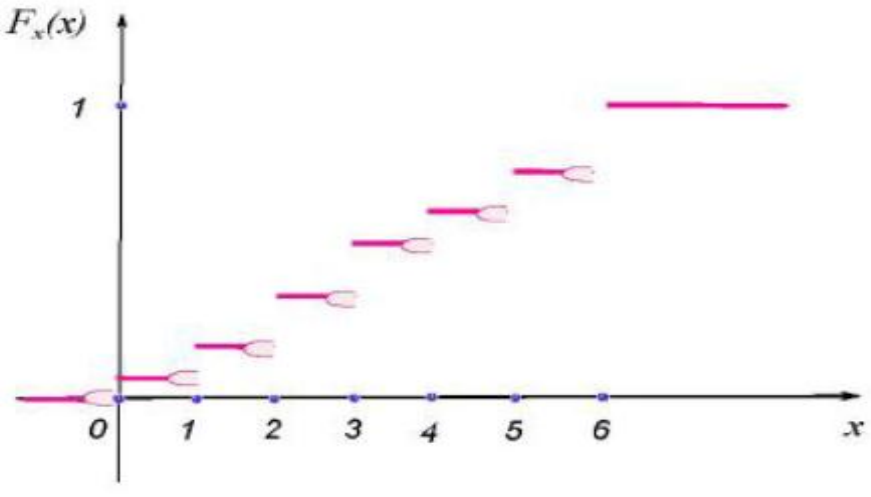
La fonction de répartition satisfait, pour $i\epsilon\{1, \dots, n\}$,

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

Exemple

Soit le tableau suivant

xi	ni	Ni	fi	Fi
0	18	18	0,09	0,09
1	32	50	0,16	0,25
2	66	116	0,33	0,58
3	41	157	0,205	0,785
4	32	189	0,16	0,945
5	9	198	0,045	0,99
6	2	200	0,01	1



Représentation d'une variable quantitative discrète par la courbe cumulative.

2) Distribution à caractère quantitatif continue

Fonction de répartition

Définition

La fonction $F_x : \mathbb{R} \to [0,1]$ *définie par* $F_x(x)$ *représente le pourcentage des individus tel que la valeur de leur caractère est inférieure ou égale à* x. *Elle est donnée par*

$$F_x(x) = \begin{cases} 0, & si \quad x < a_0, \\ \frac{f_1}{h}(x - a_0), & si \quad a_0 \leq x < a_1, \\ F_i + \frac{f_{i+1}}{h}(x - a_i), & si \quad a_i \leq x < a_{i+1}, \\ 1, & si \quad x \geq a_n, \end{cases}$$

et elle s'appelle la fonction de répartition de X.

Avec $h = a_{i+1} - a_i$ fixe.

La courbe de F_x est nulle avant a_0, constante égale à 1 après a_n et joint les points $(a_0, 0)$, (a_1, F_1),...,$(a_n, 1)$ par des segments de droites

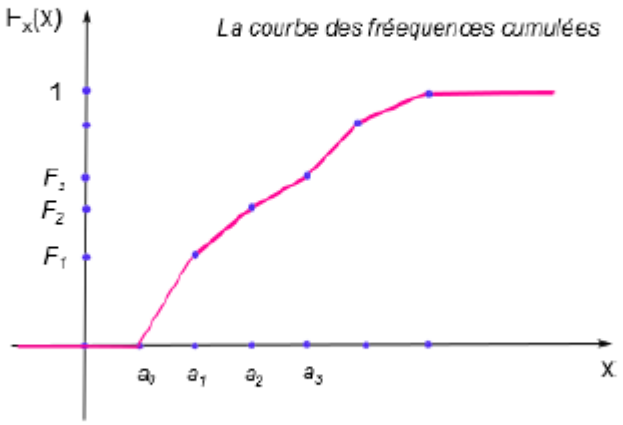
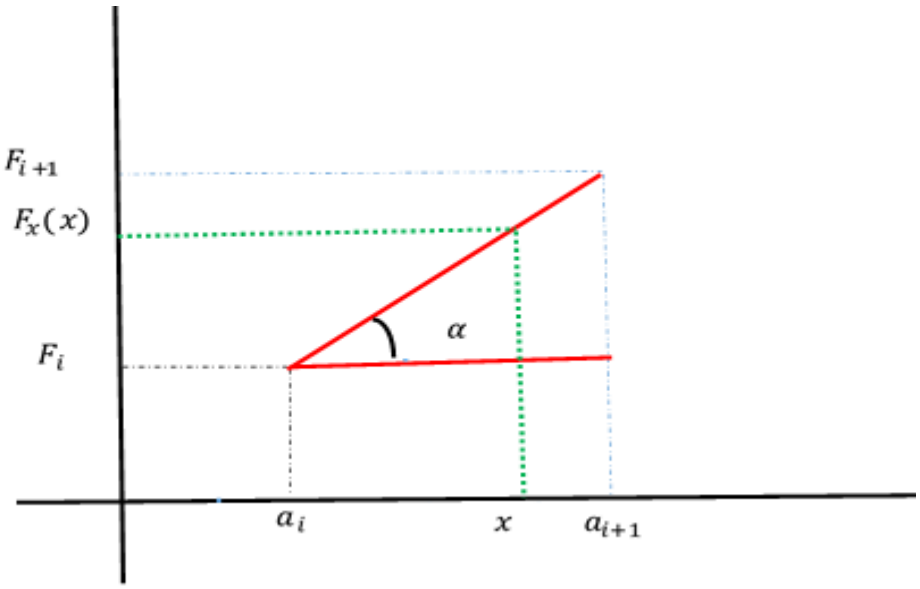


FIGURE : La courbe des fréquences cumulées.

Maintenant on va calculer F_x par extrapolation :

Soit la restriction de F_x sur l'intervalle $[a_i, a_{i+1}]$



On a $$\tan(\alpha) = \frac{F_{i+1} - F_i}{a_{i+1} - a_i} = \frac{F_x(x) - F_i}{x - a_i}$$

D'où la formule de la fonction de répartition est

$$F_x(x) = \frac{f_{i+1}}{h}(x - a_i) + F_i$$

III) VALEURS CARACTÉRISTIQUES

Valeurs caractéristiques

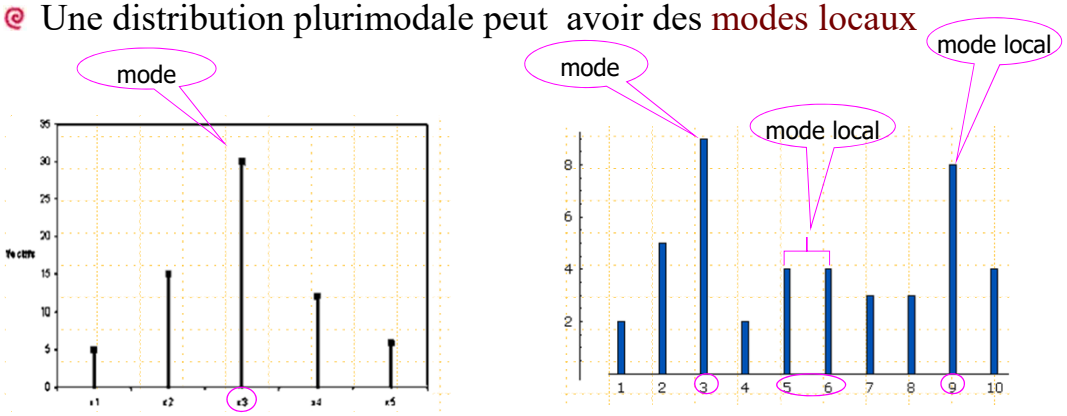
Trois types de valeurs caractéristiques

- Indicateurs de position
- Indicateurs de dispersion
- Indicateurs de forme

1) Indicateurs de position

a) Le Mode

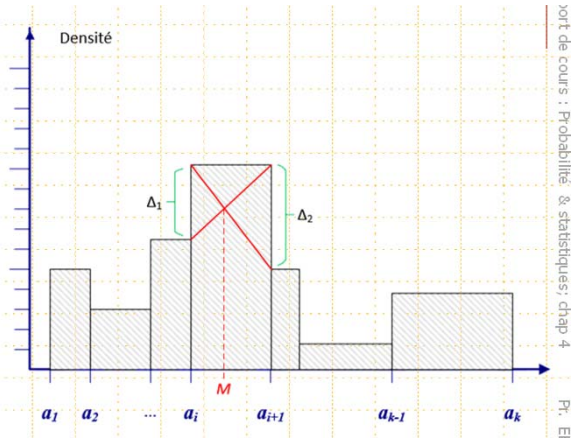
- Est la modalité qui admet le plus grand effectif, la plus grande fréquence
- Une distribution peut avoir plusieurs modes : plurimodale
- Une distribution plurimodale peut avoir des modes locaux



- Le mode est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.
- Lorsque cette valeur est unique, on dit que la distribution est ***unimodale***, dans le cas contraire on dit que la distribution est **plurimodale.**
- Une variable des modes locaux (modalités dont la fréquence est supérieure ou égale aux fréquences adjacentes). Cette situation est intéressante : elle met en évidence l'existence de plusieurs sous-populations, donc l'hétérogénéité de la population étudiée.

Pour une distributions regroupées en classes,

1. On détermine d'abord la classe modale : c'est la classe qui possède la plus forte densité.
2. Puis on calcule la valeur du mode par interpolation linéaire :
   si [a_i, a_{i+1} [est la classe modale alors le mode M vérifie :
   $$Mo = a_i + \frac{\Delta_1}{\Delta_1+\Delta_2}(a_{i+1} - a_i)$$



b) La médiane

On appelle médiane la valeur Me de la V.S X qui vérifie la relation suivante :

$$F_x(M_e^-) < 0.5 \leq F_x(M_e^+) = F_x(M_e)$$

Exemple

Soit le tableau statique suivant :

xi	ni	Ni	fi	Fi
0	18	18	0,09	0,09
1	32	50	0,16	0,25
2	66	116	0,33	0,58
3	41	157	0,205	0,785
4	32	189	0,16	0,945
5	9	198	0,045	0,99
6	2	200	0,01	1

Dans l'exemple précédent, la relation

$$F_x(0) = 0 < 0.5 \leq F_x(0^+) = 0.09$$

n'est pas satisfaite. Donc, la médiane est différente de 0. *Par contre, nous avons*

$$F_x(2^-) = 0.25 < 0.5 \leq F_x(2^+) = F(2) = 0.58.$$

Donc, $Me = 2$.

Distribution à caractère quantitatif continue

Pour une distributions regroupées en classes, le calcul de la médiane se fait en deux étapes :

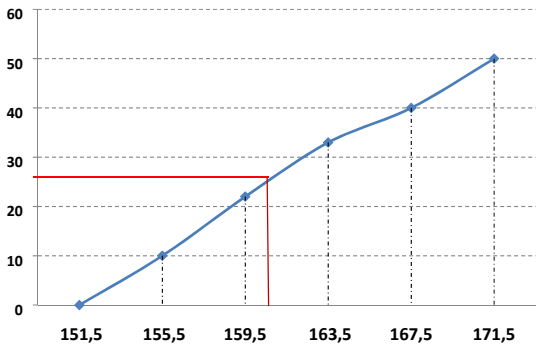
1. On détermine la classe qui contient la médiane : c'est la classe dont la fréquence cumulée vérifie $F_i \leq 0{,}5 \; et \; F_{i+1} > 0{,}5$
2. On calcule la médiane par interpolation linéaire :
   si $[\, a_i \,, a_{i+1} \,[$ est la classe de la médiane, alors

$$Me = a_i + (a_{i+1} - a_i)\frac{0{,}5 - F_i}{F_{i+1} - F_i}$$

Tableau Statistique

$[a_i ; a_{i+1}[$	n_i	N_i	f_i	F_i
[151,5 ; 155,5[	10	10	0,20	0,20
[155,5 ; 159,5[	12	22	0,24	0,44
[159,5 ; 163,5[	11	33	0,22	0,66
[163,5 ; 167,5[	7	40	0,14	0,80
[167,5 ; 171,5[	10	50	0,20	1
Total	**50**			

Courbe cumulative des fréquences



$$Me = a_i + (a_{i+1} - a_i)\frac{0{,}5 - F_i}{F_{i+1} - F_I} = 159{,}5 + 4\frac{0{,}5 - 0{,}44}{0{,}66 - 0{,}44} = 160{,}59$$

Noue pouvons la déterminer graphiquement ou par calcule.

1. **Première méthode** : Graphiquement à partir de la formule

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{0.5 - F(L_i)}{Me - L_i}.$$

   Plus précisément, dans la figure 3.8, nous mettons $F(x) = 0.5$ et $x = Me$.

2. **Deuxième méthode** : En utilisant directement la fonction de répartition donnée par

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

   Nous retrouvons donc

$$0.5 = \frac{f_{i+1}}{h}(Me - L_i) + F_i.$$

c) La Moyenne

- La moyenne est égale à la somme de toutes les valeurs observées divisée par le nombre d'observations
- Pour une distribution groupée en classes, le calcul de la moyenne utilise les centres des classes
- Le calcul de la moyenne s'effectue directement sur les valeurs observées brutes ou bien en utilisant les effectifs ou les fréquences des observations.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \; ; \quad \bar{X} = \sum_{i=1}^{k} f_i x_i$$

Exemple de calcul de la moyenne

Tableau I

x_i	n_i	f_i	$n_i * x_i$	$f_i * x_i$
1	5	0,10	5	0,1
2	9	0,18	18	0,36
3	15	0,30	45	0,90
4	11	0,22	44	0,88
5	3	0,06	15	0,3
6	7	0,14	42	0,84
Total	**50**	**1**	**169**	**3,38**

Tableau II

$[a_i ; a_{i+1}[$	x_i	n_i	f_i	$n_i * x_i$	$f_i * x_i$
[151,5 ; 155,5[	153,5	10	0,20		30,7
[155,5 ; 159,5[	157,5	12	0,24		37,8
[159,5 ; 163,5[	161,5	11	0,22		35,53
[163,5 ; 167,5[	165,5	7	0,14		23,17
[167,5 ; 171,5[	169,5	10	0,20		33,9
Total		**50**	**1**		**161,1**

Colonnes ajoutées pour calculer la moyenne

d) Les quartiles

Trois valeurs *Q1* , *Q2* , *Q3* qui partagent la série ordonnée des observations en 4 groupes d'effectifs égaux

- Le premier quartile *Q1* est obtenu lorsqu'on a cumulé 25% de la population
- Le second quartile *Q2* est obtenu lorsqu'on a cumulé 50% de la population : c'est la médiane
- Le troisième quartile *Q3* est obtenu lorsqu'on a cumulé 75% de la population

➢ Pour déterminer Q_1, Q_2 *et* Q_3 *dans le cas Discret*

- On appelle Le premier quartile la valeur Q_1 de la V.S X qui vérifie la relation suivante :

$$F_x(Q_1^-) < 0.25 \leq F_x(Q_1^+) = F_x(Q_1)$$

- On appelle Le premier quartile la valeur Q_2 de la V.S X qui vérifie la relation suivante :

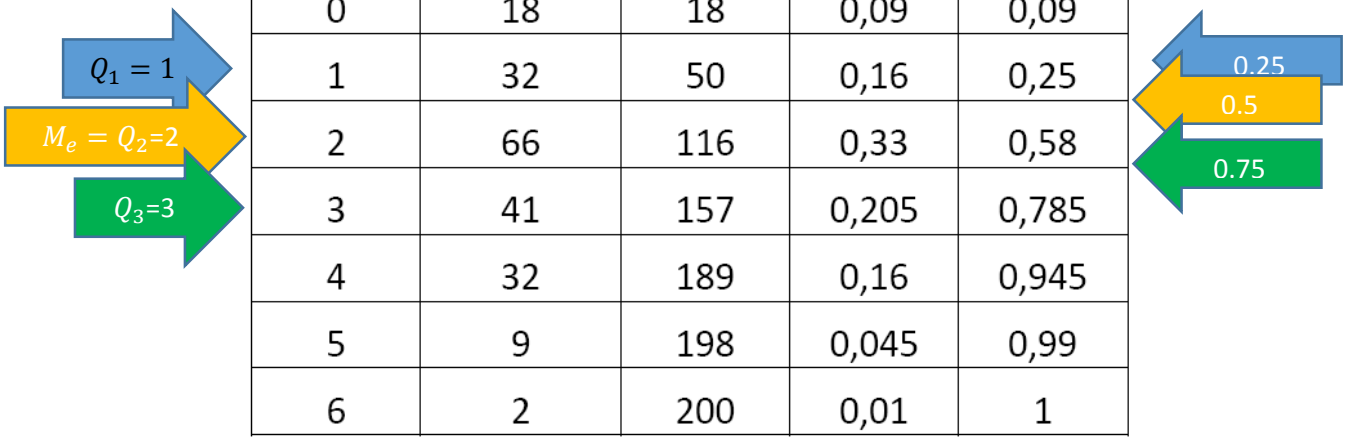
$$F_x(Q_2^-) < 0.5 \leq F_x(Q_2^+) = F_x(Q_2) = F_x(M_e)$$

- On appelle Le premier quartile la valeur Q_3 de la V.S X qui vérifie la relation suivante :

$$F_x(Q_3^-) < 0.75 \leq F_x(Q_3^+) = F_x(Q_3)$$

Exemple

Soit le tableau statique suivant :



xi	ni	Ni	fi	Fi
0	18	18	0,09	0,09
1	32	50	0,16	0,25
2	66	116	0,33	0,58
3	41	157	0,205	0,785
4	32	189	0,16	0,945
5	9	198	0,045	0,99
6	2	200	0,01	1

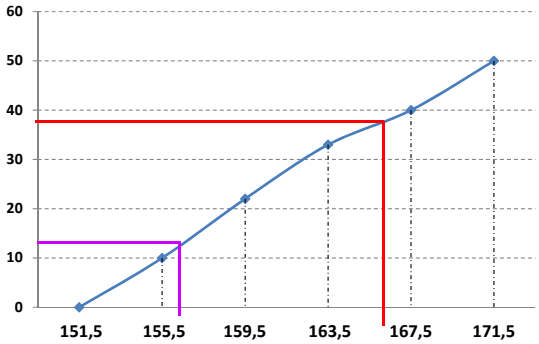
- Pour une distributions regroupées en classes, le calcul des quartiles Q_1 et Q_3 se fait en deux étapes :

1. On détermine la classe qui contient le quartile Q_1 (resp. Q_3) : c'est la classe dont la fréquence cumulée vérifie
   $F_i \leq 0{,}25 \; et \; F_{i+1} > 0{,}25$ (resp. $F_i \leq 0{,}75 \; et \; F_{i+1} > 0{,}75$)
2. On calcule la médiane par interpolation linéaire :
   si $[\, a_i, a_{i+1} \,[$ est la classe de Q_1 (resp. Q_3) alors

$$Q_1 = a_i + (a_{i+1} - a_i)\frac{0{,}25 - F_i}{F_{i+1} - F_I} \text{ (resp. } Q_3 = a_i + (a_{i+1} - a_i)\frac{0{,}75 - F_i}{F_{i+1} - F_I})$$

Exemple

$[a_i ; a_{i+1}[$	n_i	N_i	f_i	F_i
[151,5 ; 155,5[	10	10	0,20	0,20
[155,5 ; 159,5[	12	22	0,24	0,44
[159,5 ; 163,5[	11	33	0,22	0,66
[163,5 ; 167,5[	7	40	0,14	0,80
[167,5 ; 171,5[	10	50	0,20	1
Total	**50**			



$$Q1 = a_i + (a_{i+1} - a_i)\frac{0{,}25 - F_i}{F_{i+1} - F_I} = 155{,}5 + 4\frac{0{,}25 - 0{,}20}{0{,}44 - 0{,}20} = 156{,}33$$

$$Q3 = a_i + (a_{i+1} - a_i)\frac{0{,}75 - F_i}{F_{i+1} - F_I} = 163{,}5 + 4\frac{0{,}75 - 0{,}66}{0{,}80 - 0{,}66} = 166{,}07$$

Indicateurs de dispersion

- Variance et écart type
- Coefficient de variation
- Écart interquartiles

2) Indicateurs de dispersion

a) Variance

On appelle variance de cette série statistique *X*, le nombre

$$Var(X) = \sum_{i=1}^{n} f_i(\overline{x} - x_i)^2$$

On dit que la variance est la moyenne des carrés des écarts à la moyenne $\overline{x}$. Les « écarts à la moyenne » sont les $(\overline{x} - x_i)$.

Théorème

Soit (x_i, n_i) *une série statistique de moyenne* $\overline{x}$ *et de* variance $Var(X)$. *Alors,*

$$Var(X) = \sum_{i=1}^{n} f_i x_i^2 - \overline{x}^2$$

b) Ecart type

C'est la quantité

$$\sigma_x = \sqrt{Var(X)}$$

Le paramètre σ_x *mesure la distance moyenne entre* $\overline{x}$ *et les valeurs de X.*
Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.
– Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).
– Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).

c) coefficient de variation

- Le coefficient de variation (CV) est le rapport de l'écart-type à la moyenne

$$CV = \frac{\sigma_x}{|\overline{x}|}$$

- Le coefficient de variation n'a pas d'unité, Ce coefficient est souvent exprimé sous forme de pourcentage.
- Remarque : Plus le coefficient de variation est faible, plus la série d'observations est homogène. Empiriquement, on peut dire qu'un coefficient de variation inférieur à 15 % semble être, dans bien des cas, une indication d'une bonne homogénéité de la distribution des observations.

d) intervalle et écart interquartile

- On appelle intervalle interquartile l'intervalle $[Q1\,;\,Q3]$;
- L'écart interquartile est l'amplitude de l'intervalle $[Q1\,;\,Q3]$, c'est-à-dire le nombre $(Q3-Q1)$;
- L'écart interquartile est utilisé comme indicateur de dispersion. Il correspond à 50% des effectifs situés dans la partie centrale de la distribution
- Toute valeur n'appartenant pas à l'intervalle $[Q1-1{,}5(Q3-Q1)\,,\;\;Q3+1{,}5(Q3-Q1)]$ est considérée comme aberrante (Valeurs atypiques)

Chapitre 2

Étude d'une variable statistique à deux dimensions

Dans les chapitres précédents, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salaries en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit Ω une population et

$$\boldsymbol{Z = (X, Y) \rightarrow \Omega}$$
$$\boldsymbol{w \rightarrow (X(w), Y(w))}$$

Dans ce cas, Z est dite variable statistique à deux dimensions avec $Card(\Omega) = N$, avec N un entier fini. Le couple (X, Y) est appelé le couple de la variable statistique.

Exemple :

- On observe simultanément sur un échantillon de 200 foyers, le nombre d'enfants X et le nombre de chambre Y .
- On observe sur un échantillon de 20 foyers, le revenu mensuel X en Dh et les dépenses mensuelles Y .

1) Représentation des séries statistiques à deux variables

Soit la variable statistique Z donnée par le couple (X, Y). Soient $x_1, x_2, \dots x_k$ et $y_1, y_2, \dots y_l$ les valeurs prises respectivement par X et Y. Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à k et pour j allant de 1 à l, $Z_{ij} = (x_i, y_j)$.

a) tableau de contingence

$\mathbf{X \setminus Y}$	$C'_1 = [L'_1, L'_2[$ **ou** y_1	...	$C'_l = [L'_l, L'_{l+1}[$ **ou** y_l	**Marginale % à X**
$C_1 = [L_1, L_2[$ **ou** x_1	n_{11} **ou** f_{11}	...	n_{1l} **ou** f_{1l}	$n_{1\bullet}$ **ou** $f_{1\bullet}$
$C_2 = [L_2, L_3[$ **ou** x_2	n_{21} **ou** f_{21}	...	n_{2l} **ou** f_{2l}	$n_{2\bullet}$ **ou** $f_{2\bullet}$
$C_3 = [L_3, L_4[$ **ou** x_3	n_{31} **ou** f_{31}	...	n_{3l} **ou** f_{3l}	$n_{3\bullet}$ **ou** $f_{3\bullet}$
$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$
$C_k = [L_k, L_{k+1}[$ **ou** x_k	n_{k1} **ou** f_{k1}	...	n_{kl} **ou** f_{kl}	$n_{k\bullet}$ **ou** $f_{k\bullet}$
Marginale % à Y	$n_{\bullet 1}$ ou $f_{\bullet 1}$	...	$n_{\bullet l}$ ou $f_{\bullet l}$	N

Où :

n_{ij} est l'effectif qui représente le nombre d'individus qui prennent en même temps la valeur x_i et y_j,

$f_{ij} = \frac{n_{ij}}{N}$ avec $N = Card(\Omega)$ c'est-à-dire $N = \sum_{i=1}^{l} \sum_{j=1}^{k} n_{ij}$

Nous avons la propriété suivante : $\sum_{i=1}^{l} \sum_{j=1}^{k} f_{ij} = 1$

Lois marginales :

Effectifs et fréquences marginale par rapport à Y nous avons, pour j = 1...l, :

$$n_{\blacksquare j} = \sum_{i=1}^{k} n_{ij} \quad \text{et} \quad f_{\blacksquare j} = \sum_{i=1}^{k} f_{ij}$$

Effectifs et fréquences marginale par rapport à X nous avons, pour j = 1...k, :

$$n_{i\blacksquare} = \sum_{j=1}^{l} n_{ij} \quad \text{et} \quad f_{i\blacksquare} = \sum_{j=1}^{l} f_{ij}$$

Remarque

Nous avons les propriétés suivantes

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = N \qquad et \qquad \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1.$$

Exemple :

Nous considérons 10 salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes sont données dans le tableau suivant,

Salaire	6000	7400	7500	8200	8207	8900	9100	9900	9950	10750
Age	*15*	*26*	*20*	*43*	*47*	*37*	*52*	*34*	*50*	*44*

On considère que l'âge et le salaire forment respectivement des classes de pas de 10 ans et de 1000 Dh.

Donc pour le salaire on a les classes suivantes :

$$[15, 25[, \quad [25, 35[, \quad [35, 45[, \quad [45, 55[,$$

et

$$[6, 7[, \quad [7, 8[, \quad [8, 9[, \quad [9, 10[, \quad [10, 11[,$$

pour le salaire $(\times 1000)$.

Cette série statistique est représentée par le tableau suivant,