CHAPITRE IV

VARIABLE STATISTIQUE CONTINUE

Dans la première partie de ce cours (définitions et terminologie) nous avons exposé six exemples dont deux déjà traitent de variables statistiques continues :

L'exemple 2 où on s'intéresse au temps de réaction au son d'un certain nombre de personnes.

L'exemple 4 où il est question de pesées effectuées sur les éléments d'un échantillon de poissons prélevé dans lac.

Dans ces deux exemples, le caractère est de par sa nature (temps ou poids) une variable continue.

D'une manière générale, l'essence même du caractère indique qu'il est continu ou pas. Ainsi, ce qui a rapport au temps, à la vitesse, à la longueur, à la température, ...etc. est continu. On peut bien imaginer un mobile partir du repos et atteindre une vitesse V en étant passé continûment par toutes les valeurs de l'intervalle [0 ; V]. On peut également imaginer l'épuisement d'une durée déterminée t d'une manière continue, de la valeur t jusqu'à la valeur zéro.

Il est possible donc que la variable prenne n'importe quelle valeur de l'intervalle considéré. Ce qui n'est pas le cas de certaines variables (les variables discrètes) qu'on a étudié précédemment. Le nombre d'enfants d'une famille, par exemple, ne peut prendre que certaines valeurs isolées.

Ainsi, nous pouvons distinguer naturellement une variable continue d'une autre qui ne l'est pas. Mais en statistique ce n'est pas ceci qui constitue la véritable préoccupation. La distinction entre variables discrètes et variables continues se fait surtout en fonction de la manière dont on veut les traiter

Lorsqu'on considère les différentes valeurs de la variable valeur par valeur et qu'on les traite isolément nous avons, de cette manière admis, que la variable est discrète. Mais si on répartit l'étendue en intervalles et que l'on adopte que les modalités des individus ne sont plus leurs valeurs exactes mais leurs appartenances à tel ou tel intervalle alors la variable est considérée continue.

Exemple

Répartition des âges d'un groupe de 120 personnes

1èr cas

Age (en années)	Nombre de personnes	Age (en années)	Nombre de personnes	Age (en années)	Nombre de personnes	Age (en années)	Nombre de personnes
1	4	6	5	11	6	16	2
1,5	3	6,5	2	11,5	2	16,5	5
2	1	7	9	12	4	17	0
2,5	5	7,5	1	12,5	5	17,5	0
3	8	8	1	13	1	18	3
3,5	0	8,5	1	13,5	3	18,5	2
4	1	9	4	14	3	19	2
4,5	6	9,5	3	14,5	4	19,5	0
5	2	10	2	15	5	20	3
5,5	3	10,5	7	15,5	1	20,5	1

2ème cas

Age (en années)	[1 - 4,5[	[4,5 - 10[	[10 - 12,5[	[12 ,5 - 15[	[15 - 17,5[	[17,5 - 21[
Nombre de personnes	22	37	21	16	13	11

La variable **âge** a été traitée comme une variable statistique discrète dans le premier cas et comme une variable statistique continue dans le deuxième cas.

Sont considérées donc comme continues, en plus des variables qui sont "naturellement" continues, celles dont les valeurs sont trop nombreuses pour être traitées isolement.

IV . A – Description préliminaire

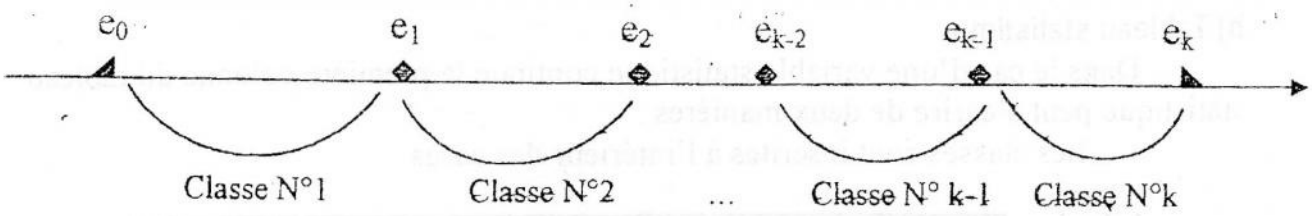
a) Classes

Entre l'étape de la collecte des données et celle de la construction du tableau statistique se fait une étape intermédiaire : la répartition de l'étendue en classes.

On subdivise l'étendue parcouru par la variable statistique en intervalles disjoints qui le recouvrent en entier. Ces intervalles sont les modalités de la variables et on les appelles les **classes**.

La classe numéro i est l'intervalle $[\ e_{i-1}\ ,\ e_i\ [$ (fermé à gauche ouvert à droite).

e_{i-1} et e_i sont appelés les **extrémités** de la classe.



Chaque classe est caractérisée par son centre et son amplitude.
l'amplitude de la classe est la distance séparant ses deux extrémités, c'est-à-dire la longueur du segment les joignant.

Ainsi l'amplitude de la classe numéro *i* est

$$a_i = e_i - e_{i-1}$$

Le **centre** de la classe est le point milieu du segment joignant ses extrémités.
Le centre de la classe numéro *i* est

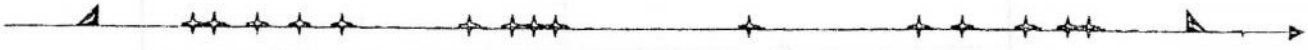
$$c_i = \frac{e_i + e_{i-1}}{2}$$

Remarque importante (α):
Lorsque nous répartissons l'étendue en classes, les valeurs que possédaient les individus sont "perdues". Il ne nous reste comme information que le fait qu'ils appartiennent à telle ou telle classe. Mais comment se dispersent ces individus à l'intérieur de chaque classe ?

Nous allons admettre que la distribution se fait de manière uniforme dans chaque classe. C'est-à-dire comme si les points étaient à égales distances les uns des autres. (Nous allons voir, plus tard, que cette supposition va nous permettre de justifier un certains nombre de développements).

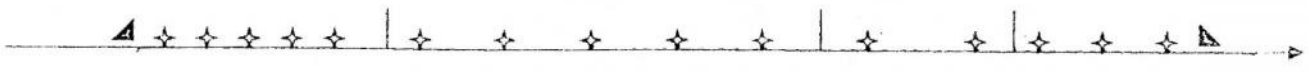
Exemple : (visuel)

Supposons que notre population soit composée de 15 individus dont les valeurs sont indiquées par des points sur l'axe.



Supposons que nous ayons décidé de diviser l'étendue en quatre classes.

Dans la première classe seront affectés 5 individus, dans la deuxième 5, dans la troisième 2 et dans la quatrième 3 individus.
Alors chaque segment sera divisé en parties d'égales longueurs (autant de parties que d'individus). Nous aurons ainsi :



La distribution véritable est celle dessinée sur la première droite ; mais la distribution que nous supposons désormais existante est celle dessinée sur la troisième droite. Nous disons que

les points ont été **uniformément** distribués.

b) Tableau statistique

Dans le cas d'une variable statistique continue la première colonne du tableau statistique peut s'écrire de deux manières :

- Les classes sont inscrites à l'intérieur des cases

Classes	**Effectifs** n_i
$[\ e_0\ ;\ e_1\ [$	n_1
$[\ e_1\ ;\ e_2\ [$	n_2
...	...
$[\ e_{i-1}\ ;\ e_i\ [$	n_i
...	...
$[\ e_{k-1}\ ;\ e_k\ [$	n_k
Total	**N**

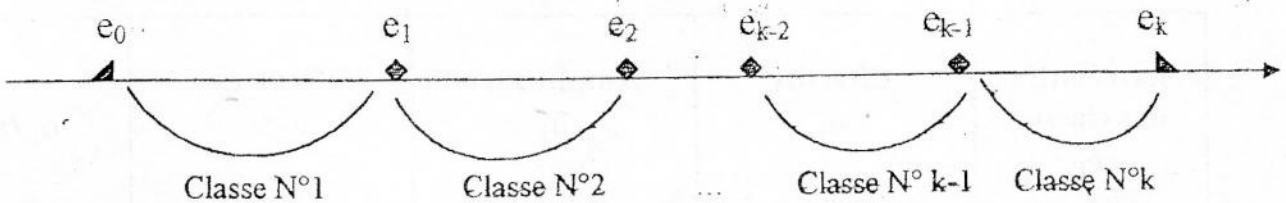
- Les extrémités des classes sont reportées sur les lignes

Extrémités des classes e_i	**Effectifs** n_i
e_0	
	n_1
e_1	
	n_2
e_2	
...	...
e_{i-1}	
	n_i
e_i	
...	...
e_{k-1}	
	n_k
e_k	
Total	**N**

Exemple :

Nous allons reprendre l'exemple 2 du chapitre I (définitions et terminologie) pour illustrer ce que nous venons de voir.

Pagarde Chloé



Chaque classe est caractérisée par son centre et son amplitude.
l'amplitude de la classe est la distance séparant ses deux extrémités, c'est-à-dire la longueur du segment les joignant.

Ainsi l'amplitude de la classe numéro *i* est

$$a_i = e_i - e_{i-1}$$

Le **centre** de la classe est le point milieu du segment joignant ses extrémités.
Le centre de la classe numéro *i* est

$$c_i = \frac{e_i + e_{i-1}}{2}$$

Remarque importante (α):
Lorsque nous répartissons l'étendue en classes, les valeurs que possédaient les individus sont "perdues". Il ne nous reste comme information que le fait qu'ils appartiennent à telle ou telle classe. Mais comment se dispersent ces individus à l'intérieur de chaque classe ?

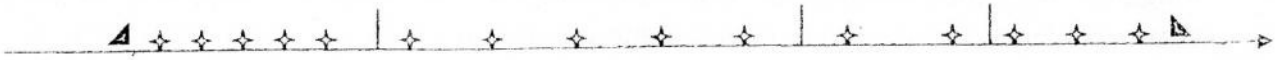
Nous allons admettre que la distribution se fait de manière uniforme dans chaque classe. C'est-à-dire comme si les points étaient à égales distances les uns des autres. (Nous allons voir, plus tard, que cette supposition va nous permettre de justifier un certains nombre de développements).

Exemple : (visuel)

Supposons que notre population soit composée de 15 individus dont les valeurs sont indiquées par des points sur l'axe.

Supposons que nous ayons décidé de diviser l'étendue en quatre classes.

Dans la première classe seront affectés 5 individus, dans la deuxième 5, dans la troisième 2 et dans la quatrième 3 individus.
Alors chaque segment sera divisé en parties d'égales longueurs (autant de parties que d'individus). Nous aurons ainsi :



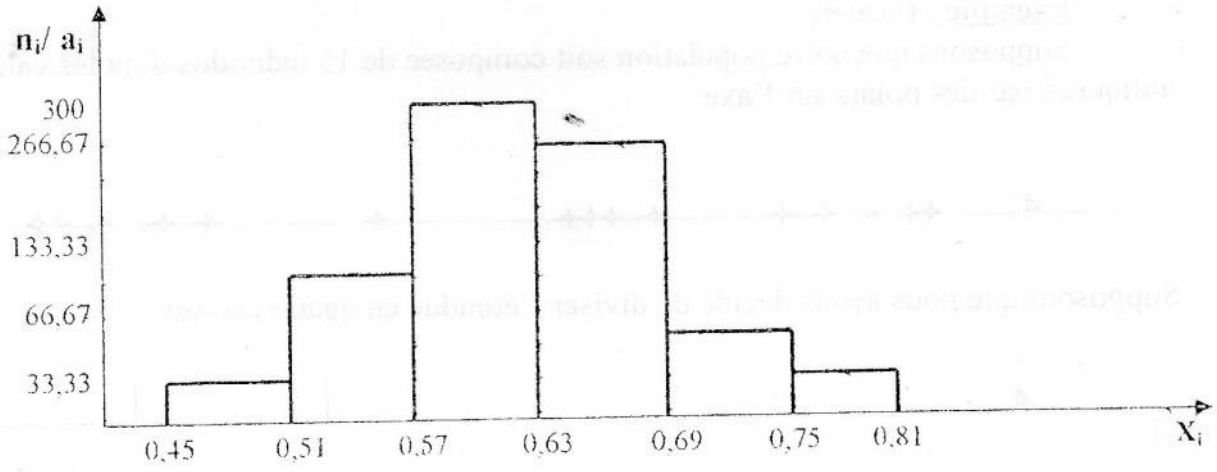
La distribution véritable est celle dessinée sur la première droite ; mais la distribution que nous supposons désormais existante est celle dessinée sur la troisième droite. Nous dirons que

Extrémités des classes e_i	Effectifs n_i	Amplitudes a_i	Centres c_i	n_i / a_i
0,45				
	2	0,06	0,48	33,33
0,51				
	8	0,06	0,54	133,33
0,57				
	18	0,06	0,60	300
0,63				
	16	0,06	0,66	266,67
0,69				
	4	0,06	0,72	66,67
0,75				
	2	0,06	0,78	33,33
0,81				
Total	**50**			

c) Diagramme différentiel

Dans le cas d'une variable statistique continue le diagramme différentiel s'appelle **histogramme**. A chaque classe est associé un rectangle dont la largeur est l'amplitude de la classe et dont la hauteur est le rapport de la fréquence sur l'amplitude (n_i/ a_i ou f_i/ a_i).

Exemple : (temps de réaction au son)



Remarque :

Nous pouvons utiliser comme ordonnées aussi bien n_i/ a_i que f_i/ a_i. La différence ne serait qu'une variation d'échelle. Nous avons fait une remarque analogue lorsqu'on a définit le diagramme en bâtons pour la variable discrète.

Nous avons à noter également un autre détail. Dans le diagramme en bâtons la longueur de chaque segment est égale à la fréquence relative correspondante (ou à l'effectif correspondant). Par conséquent, la somme des longueurs de tous les segments est égale à la somme des fréquences relatives (ou à la somme des effectifs) et donc, elle est égale à l'unité (ou à l'effectif total N)

D'une manière semblable, nous pouvons constater que la somme des surfaces des rectangles composants l'histogramme est égale à l'unité quand on utilise f_i/ a_i. Elle serait egale à l'effectif total N quand on utilise n_i/ a_i.

En effet la surface de chaque rectangle est

$$s_i = (n_i / a_r)\, a_i = n_i$$

et donc,

$$\sum_{i=1}^{k} s_i = \sum_{i=1}^{k} n_i = N$$

Ainsi, ce qui était longueur dans le cas d'une variable discrète devient surface dans le cas d'une variable continue.

A noter également l'apparition de la notion d'amplitude qui peut différencier une classe d'une autre (dans notre exemple les classes sont d'égales amplitudes ; c'est ce qui a fait que l'on obtient la même allure pour la figure en utilisant tout aussi bien n_i / a_i ou simplement n_i, mais ceci n'est qu'un cas particulier ; les classes sont en générale d'amplitudes inégales).

d) Effectifs cumulés, fréquences relatives cumulées et diagramme intégral

Les définitions de ces notions restent les mêmes pour les variables statistiques continues comme pour les variables discrètes.

Nous avons, de ce fait :

L'effectif cumulé correspondant à la valeur X_i est le nombre des individus ayant une valeur inférieure ou égale à X_i.

Autrement dit, c'est la somme des effectifs qui se sont accumulés en atteignant cette valeur. Ce qui s'écrit :

$$N_i = \sum_{p=1}^{i} n_p$$

De même,

La fréquence cumulée correspondant à la valeur X_i est la fréquence des individus ayant une valeur inférieure ou égale à X_i.

$$F_i = \sum_{p=1}^{i} f_p$$

C'est-à-dire :

Exemple : (temps de réaction au son)

e_i	n_i	f_i	N_i	F_i
0,45			0	0
	2	0,04		
0,51			2	0,04
	8	0,16		
0,57			10	0,20
	18	0,36		
0,63			28	0,56
	16	0,32		
0,69			44	0,88
	4	0,08		
0,75			48	0,96
	2	0,04		
0,81			50	1
Total	50	1		

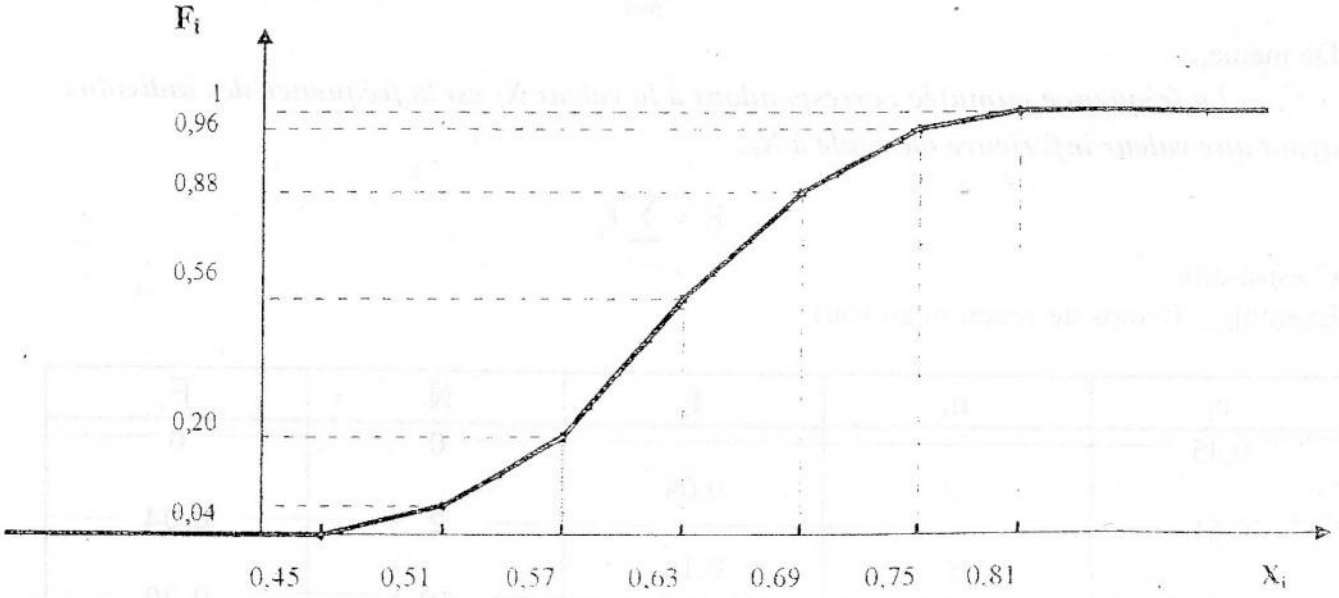
Diagramme intégral :

Avant de définir le diagramme intégral nous avons à apporter quelques précisions importantes. Nous avons vu, lors de l'étude de la variable statistique discrète, que le diagramme intégral n'est autre que la courbe représentative de la fonction cumulative. Nous avons constaté que c'est une courbe en escalier. Ce genre de croissance par sauts est dû au fait qu'entre deux valeurs de la variable il n'y a aucune accumulation. Toute accumulation supplémentaire se fait lorsqu'on dépasse une valeur de la variable.

Dans le cas de la variable continue, cela ne se passe pas de la même manière. En effet, en allant d'une extrémité de la classe à l'autre on accumule les individus de cette classe. Il y a donc croissance de la fonction cumulative à l'intérieur même de la classe et non seulement aux extrémités. Comment se fait cette croissance ? La remarque (**α**) de la page (32) répond parfaitement à la question. ***Nous avons supposé que la distribution est uniforme***. Ce qui signifie que la croissance est constante d'une extrémité de la classe à l'autre. Une croissance constante est synonyme de croissance linéaire ; c'est-à-dire suivant une droite.

Par conséquent, la courbe cumulative est une ligne brisée faite de segments de droite joignants les points dont les coordonnées sont : en abscisse, les extrémités des classes et en ordonnées les fréquences cumulées correspondantes à ces extrémités.

Exemple : (temps de réaction au son)



IV . B – Caractéristiques de position centrale

a) La classe modale

Nous ne pouvons pas parler ici de mode car les modalités sont des classes et non des valeurs. La notion de mode correspond à une idée d'intensité plutôt qu'à une idée d'effectifs, comme on aurait tendance à le croire. C'est vrai que dans le cas de la variable discrète il y a superposition des deux notions. Mais pour la variable continue l'amplitude entre en jeu. Les classes peuvent être d'inégales amplitudes et une classe qui a le plus grand effectif n'est pas nécessairement la classe où le caractère est le plus intense. Il y a plus d'habitants au Sahara qu'à Constantine mais cela ne veut pas dire que la densité y est plus grande.

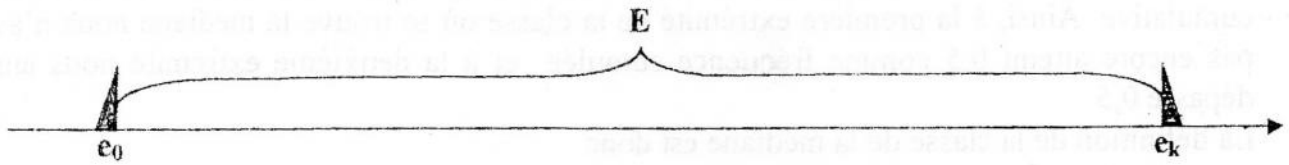
Nous définirons donc la classe modale comme suit :

C'est la classe qui correspond au plus grand rapport n_i / a_i

a) L'étendue : (noté E)

C'est la longueur de l'intervalle sur lequel se disperse la variable. C'est-à-dire :

$$\mathbf{E = e_k - e_0}$$



Ainsi, pour notre exemple nous avons : $\mathbf{E} = e_k - e_0 = 0{,}81 - 0{,}45 = \mathbf{0.36}$

b) L'écart interquartiles : (noté I_Q)

C'est la différence entre le troisième quartile et le premier quartile.

$$\mathbf{I_Q = Q_3 - Q_1}$$

Pour notre exemple nous avons

$$\mathbf{I_Q} = Q_3 - Q_1 = 0{,}666 - 0{,}578 = \mathbf{0{,}088}$$

c) Variance et écart-type

La remarque qui nous a permis d'utiliser les centres de classes pour le calcul de la moyenne nous servira aussi pour développer un raisonnement analogue au sujet de la variance et d'établir que

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i . c_i^{\,2} - \overline{X}^{\,2}$$

ou

$$Var(X) = \sum_{i=1}^{k} f_i . c_i^{\,2} - \overline{X}^{\,2}$$

L'écart-type est la racine carrée de la variance

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

Pour notre exemple nous avons

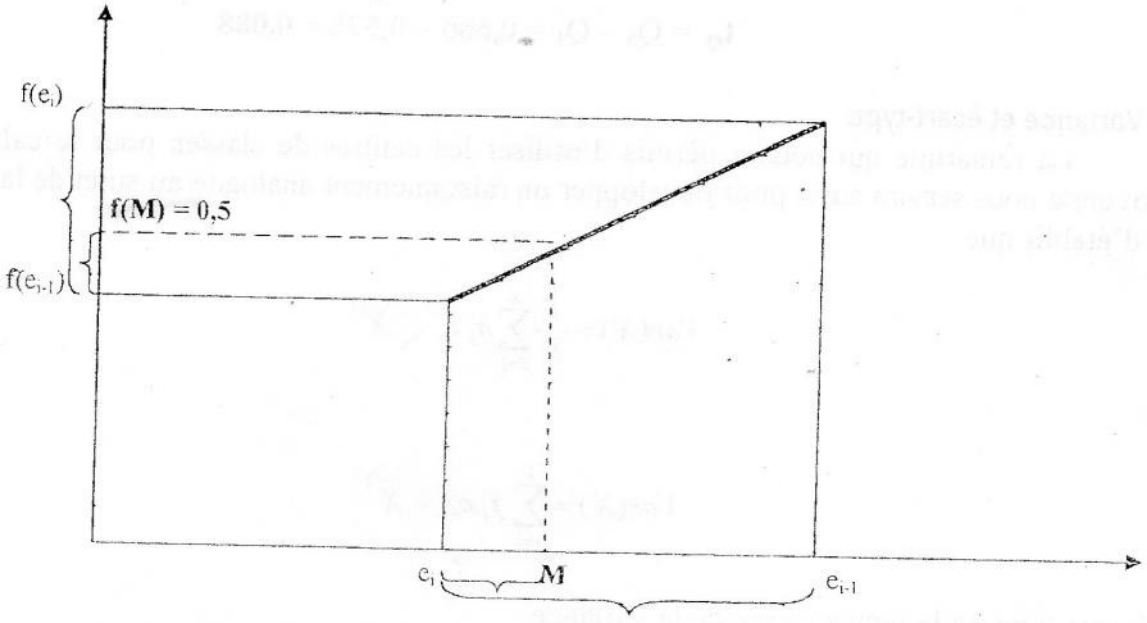
c) Médiane

Pour déterminer la médiane il faut d'abord situer la classe de la médiane.
Nous savons que la médiane est la valeur de la variable statistique qui partage la population en deux effectifs égaux. C'est la valeur M telle que F(M) = 0,5 où F est la fonction cumulative. Ainsi, à la première extrémité de la classe où se trouve la médiane nous n'avons pas encore atteint 0,5 comme fréquence cumulée et à la deuxième extrémité nous aurons dépassé 0,5.
La définition de la classe de la médiane est donc :

C'est la classe $[\, e_{i-1} ; e_i \,[$ ***telle que***

$$\begin{cases} F(e_{i-1}) < 0,5 \\ \quad et \\ F(e_i) > 0,5 \end{cases}$$

Une fois que la classe de la médiane est déterminée, nous pouvons calculer la médiane par interpolation linéaire.



et donc

$$M = e_{i-1} + (e_i - e_{i-1}).\frac{0,5 - F_{i-1}}{F_i - F_{i-1}} = e_{i-1} + a_i.\frac{0,5 - F_{i-1}}{f_i}$$

Pour notre exemple nous avons

$$M = e_{i-1} + a_i.\frac{0,5 - F_{i-1}}{f_i} = 0,57 + 0,06.\frac{0,5 - 0,20}{0,36} = 0,62$$

Remarque :

- Lorsqu'une extrémité e_i possède exactement 0,5 comme fréquence cumulée, c'est-à-dire lorsque $F(e_i) = 0,5$ alors la médiane est justement cette valeur e_i (elle répond exactement à la définition de la médiane et elle s'obtient donc par simple lecture sur le tableau statistique)

d) Quartiles

Pour déterminer les quartiles la procédure est la même que pour la médiane.
Il s'agit d'abord d'identifier les classes où se trouve les quartiles et de les calculer ensuite par interpolation linéaire.
Comme le premier quartile $\mathbf{Q_1}$ cumule le quart de la population alors il se trouve dans la classe $[\ e_{i-1}\ ;\ e_i\ [$ qui vérifie les conditions :

$$\begin{cases} \mathbf{F(e_{i-1}) < 0,25} \\ \quad \textit{et} \\ \mathbf{F(e_i) > 0,25} \end{cases}$$

Pour calculer Q_1 nous allons utiliser la formule de l'interpolation linéaire

$$Q_1 = e_{i-1} + a_i . \frac{0,25 - F_{i-1}}{f_i}$$

pour notre exemple nous avons

$$Q_1 = 0,57 + 0,06 . \frac{0,25 - 0,20}{0,36} \approx 0,578$$

Le troisième quartile $\mathbf{Q_3}$ cumule les trois quarts de la population et donc sa classe est celle qui vérifie

$$\begin{cases} F(e_{i-1}) < 0,75 \\ \quad \textit{et} \\ F(e_i) > 0,75 \end{cases}$$

La valeur de Q_3 est donc

$$Q_3 = e_{i-1} + a_i . \frac{0,75 - F_{i-1}}{f_i}$$

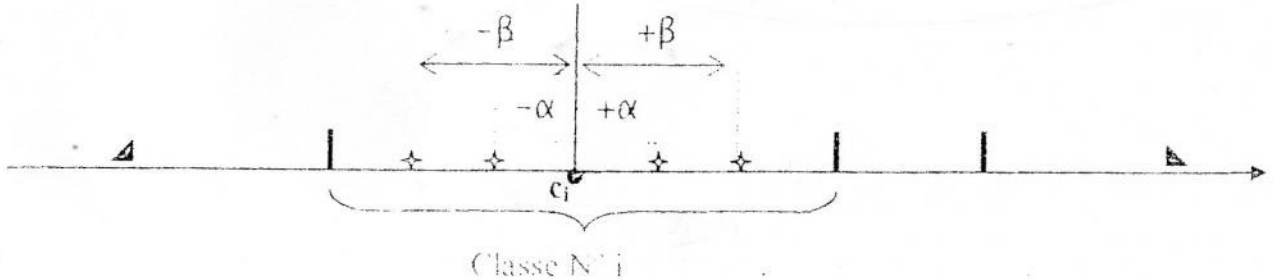
Pour l'exemple on obtient

$$Q_3 = 0,63 + 0,06 . \frac{0,75 - 0,56}{0,32} \approx 0,666$$

e) Moyenne arithmétique

Comment pouvons-nous calculer une moyenne arithmétique alors que nous n'avons pas de valeurs mais des intervalles ?

Souvenons-nous de la supposition que nous avons faite au sujet de la distribution des individus à l'intérieur des classes. Nous avons admis que cette distribution se fait d'une manière uniforme. Les points sont à égales distances les uns des autres. Ce qui signifie qu'ils sont, deux à deux, à égale distance du milieu de l'intervalle.



Classe N° i

e_i	n_i	f_i	c_i	$n_i \cdot e_i$	$f_i \cdot e_i$	$n_i \cdot e_i^2$	$f_i \cdot e_i^2$
0,45							
	2	0,04	0,48	0,96	0,0192	0,4608	0,00921
0,51							
	8	0,16	0,54	4,32	0,0864	2,3328	0,04665
0,57							
	18	0,36	0,60	10,8	0,216	6,48	0,1296
0,63							
	16	0,32	0,66	10,56	0,2112	6,9696	0,13939
0,69							
	4	0,08	0,72	2,88	0,0576	2,0736	0,04147
0,75							
	2	0,04	0,78	1,56	0,0312	1,2168	0,02433
0,81							
Total	**50**	**1**		**31,08**	**0,6216**	**19,5336**	**0,3906**
Total/N				**0,6216**		**0,3906**	

Nous avons donc :

$$\mathrm{Var}(X) = 0{,}3906 - (0{,}6216)^2 = 0{,}0042$$

et pour l'écart-type nous avons

$$\sigma_X = \sqrt{0{,}0042} \approx 0{,}065$$

$$\mathrm{Var}(X) = \sum_i \frac{f_i \cdot c_i^2}{f_i \cdot c_i}$$