

Chapitre 3 : Statistiques permettant de résumer une série

Dernière mise à jour le 16 juillet 2009

0 – [Introduction](#)

1 – [Les statistiques de tendance centrale](#)

A- Le mode

- 1) Définition
- 2) Remarques à propos du mode
 - a) Une série peut avoir plusieurs modes
 - b) Le mode n'existe pas forcément
 - c) Le mode n'est pas forcément la valeur la plus élevée
 - d) Variables et caractères peuvent avoir un mode
 - e) Mettre la série sous forme de distribution pour repérer le mode

B - La moyenne arithmétique

- 1) La moyenne arithmétique simple
- 2) La moyenne arithmétique pondérée
- 3) Données regroupées en catégories
- 4) Décomposition de la moyenne (deux dimensions)

C - La médiane

- 1) Origine du mot, sens géométrique
- 2) Sens du mot en statistique descriptive
- 3) Méthode de calcul
 - a) n est pair
 - b) n est impair

2 – [Les statistiques de dispersion](#)

A - Minimum, maximum, intervalle de variation et rapport de variation

- 1) Minimum et maximum d'une série
- 2) Intervalle de variation ou étendue
- 3) Rapport de variation

B - Intervalles interquartile et interdécile

- 1) Définition et calcul des quartiles
- 2) Définition, calcul et interprétation de l'intervalle interquartile
- 3) Définition et calcul des déciles
- 4) Définition, calcul et interprétation de l'intervalle interdécile
- 5) Note sur les centiles

C - Variance, écart-type et coefficient de variation

- 1) La variance
 - a) Définition
 - b) Exemple
 - c) Utilité de la variance
- 2) L'écart-type
 - a) Définition
 - b) Méthode de calcul
 - c) Utilité de l'écart-type
- 3) Le coefficient de variation
- 4) Décomposition de la variance (deux dimensions)

3 – [Résumé](#)

Annexe 1 : [Méthode alternative pour le calcul des quartiles](#)

Annexe 2 : [Formule de décomposition de la moyenne](#)

Exercices corrigés :

TD3 2008 : [Sujets](#) - [Corrigés](#)

TD3 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#)

0 – Introduction

Il est habituel de distinguer **deux types de statistiques résumées**, qu'on ensuite à les réunir dans des statistiques résumées plus élaborées :

- Les statistiques qui résument la **tendance « centrale »** ou « position » d'une série ou d'une distribution :
 - o **mode,**
 - o **moyenne**
 - o **médiane.**
- Les statistiques qui résument la **dispersion** d'une série ou d'une distribution sans référence à aucune statistique de tendance centrale :
 - o **intervalle interquartile** essentiellement.
- Les statistiques qui résument la **dispersion** d'une série ou d'une distribution qui tiennent aussi compte de sa valeur centrale. C'est le cas de :
 - o **l'écart-type,**
 - o **de la variance**
 - o **du coefficient de variation**¹.

Remarques

1) Bien qu'il soit possible d'effectuer des calculs de statistiques résumées (moyenne, médiane, mode, etc.) sur les données **groupées en catégories**, c'est fortement déconseillé de le faire quand on dispose des données brutes ou regroupées par valeurs ou modalités². **C'est une question de bon sens**. Si l'on effectue les calculs sur des données regroupées par catégories, on obtient des résultats approximatifs et même carrément faux quand on les compare aux calculs effectués sur les données brutes (sans parler des formules abstruses, abscondes et sources de migraines qu'il faut employer pour effectuer les calculs de moyennes, médianes, quartiles ou variances sur des données regroupées en catégories !). Les méthodes de calcul sur les données groupées plutôt que sur les données brutes n'ont plus désormais qu'une place limitée dans la statistique descriptive moderne, dans la mesure où l'informatique se charge des calculs pour un nombre arbitrairement grand de valeurs³.

¹ Il existe aussi des statistiques qui résument la « forme » d'une distribution, mais celles-ci ne sont plus guère utilisées aujourd'hui dans la mesure où il est plus facile d'observer directement le graphique d'une distribution pour en apprécier la forme que de calculer des coefficients dont l'interprétation est fastidieuse.

² Ce qui est normalement le cas, sauf situation particulière, lorsque les données brutes ont été perdues, cas d'école autrefois utilisé pour compliquer malicieusement les exercices d'examen.

³ En revanche, les tableaux qui présentent des données regroupées par classes de valeurs ou de modalités sont très utiles, comme nous l'avons vu dans les chapitres 1 et 2. Le regroupement de données à des fins graphiques s'avère aussi très souvent judicieux, comme on le verra au chapitre 5.

2) Les statistiques résumées comme le mode, la moyenne, la médiane, l'intervalle interquartile ou l'écart-type ne permettent pas toujours de tirer des conclusions définitives concernant la ou les variables (le ou les caractères) étudiés dans la population statistique. En effet, il n'existe pas de relation systématique entre ces différentes statistiques résumées. Par exemple, imaginons que l'on soit confronté aux notes de 3 groupes d'étudiants, A, B et C : que pourrions-nous conclure si c'est dans le groupe A que la note modale est la plus élevée, dans le groupe B que la note moyenne est la plus élevée et dans le groupe C que la note médiane est la plus élevée, alors que les trois groupes ont une dispersion similaire des notes, que cette dispersion soit mesurée par l'intervalle interquartile ou par l'écart-type ?

C'est la raison pour laquelle, en complément des statistiques résumées, un graphique comparatif des distributions étudiées n'est pas considéré comme un luxe (voir le chapitre 5).

1 – Les statistiques de tendance centrale

A – Le mode

1) définition

Le **mode** d'une série est la valeur ou la modalité qui revient le plus fréquemment dans la série ou la distribution.

Exemple : Soit la série {8, 4, 4, 3, 4, 3, 8, 2,5}

La valeur la plus fréquente de cette série est 4. Le mode est donc égal à 4. L'effectif associé à ce mode est 3.

2) Remarques à propos du mode

a) Une série peut avoir plusieurs modes

Soit la série $S = \{4, 0, 1, 1, \mathbf{2}, \mathbf{2}, \mathbf{2}, \underline{3}, \underline{3}, 4, \mathbf{2}, \underline{3}, 4, 5, \mathbf{2}, 1, \underline{3}, \underline{3}, 4, 5\}$, les "2" sont mis en gras et les "3" sont soulignés, car ce sont les valeurs qui reviennent le plus souvent : 5 fois chacune. Cette série a 2 modes, elle est **bimodale**. Ses deux modes sont : 2 et 3. L'effectif associé à chacun de ces modes est : 5. Bien entendu, on peut avoir des séries avec 3, 4, 5, etc. modes. Ce sont alors des **séries multimodales**.

b) Le mode n'existe pas forcément

C'est le cas lorsque toutes les valeurs ont le même effectif comme dans l'exemple suivant : {8, 6, 5, 7, 3,1}. Dans ce cas, on peut aussi dire que toutes les valeurs sont modales.

c) Le mode n'est pas la valeur la plus élevée

Il ne faut pas confondre le mode, qui est la valeur la plus fréquente, avec la valeur la plus élevée de la série. Dans la série {8, 6, 5, 7, 3, 1}, il n'y a pas de mode, mais la valeur la plus élevée est 8. Il peut néanmoins arriver que le mode soit aussi la valeur la plus élevée.

d) Variables et caractères peuvent avoir un mode

Le mode existe aussi bien dans le cas d'une série qui se rapporte à une variable (série de valeurs) que dans le cas d'une série qui se rapporte à un caractère (série de modalités).

Par exemple, le mode de la série {8, 4, 2, 3, 6, 3, 1,0} est la valeur 3, car cette valeur est la seule à revenir 2 fois et aucune autre valeur n'est plus fréquente dans la série.

De la même façon, la série {A,C,C,D,,A,A,C,E,E,B,C} a la modalité C pour mode car c'est la modalité C qui revient le plus souvent (elle revient 4 fois) et aucune autre modalité n'est plus fréquente dans la série.

e) Mettre la série sous forme d'une distribution pour repérer le mode

Pour détecter le mode, il est souvent plus facile de distribuer les éléments de la série par valeurs (ou par modalités). Soit la série suivante :

$S1 = \{8, 4, 5, 3, 3, 2, 1, 1, 1, 1, 2, 1, 6, 1, 1, 0, 1, 4, 2, 2, 0, 4, 4, 4, 3, 2, 2, 4, 0, 0, 0, 0, 5, 0, 0\}$

Cette série peut être mise sous forme d'une distribution par valeurs de la façon suivante :

Valeurs	Effectifs
0	8
1	8
2	6
3	3
4	6
5	2
6	1
7	0
8	1
	35

0 et 1 sont les deux valeurs modales de la distribution (ou la série correspondante), c'est-à-dire les plus fréquentes. Cette distribution (et la série correspondante) est donc bi-modale

Nous voyons alors plus facilement que les deux valeurs modales de la série sont 0 et 1.

B - La moyenne arithmétique

Le mot moyenne a pour origine le latin "médius", mot signifiant "qui est au milieu". "Médius" est aussi l'origine du mot "médiane". Pourtant, en statistique, les deux mots conduisent à des définitions différentes. Ceci nous laisse supposer que la notion de milieu n'est pas toujours facile à définir.

1) La moyenne arithmétique simple

La moyenne arithmétique d'une série ou **moyenne arithmétique simple** se calcule par une formule qui est donnée par l'expression :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

Le "x" surmonté d'un trait désigne classiquement la moyenne. On remarque que la somme va de 1 à n où n désigne le nombre d'unités statistiques de la population.

Appliquons cette définition au calcul de la moyenne de la série suivante = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}. On a donc :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{4+0+1+1+2+2+2+3+3+4+2+3+4+5+2+1+3+3+4+5}{20} = \frac{54}{20} = 2,7$$

2) La moyenne arithmétique pondérée

La moyenne arithmétique d'une distribution ou **moyenne arithmétique pondérée** se calcule par une formule qui est donnée par l'expression :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i \quad (2)$$

La somme varie cette fois de 1 à k , avec k qui représente le nombre de valeurs de la série. Dans le cas où aucune valeur n'est répétée $k=n$. Sinon $k < n$. Remarquons que la somme va de 1 à k , mais que cette somme est divisée par n et non par k .

La notation n_j représente les effectifs ou fréquences absolues des valeurs. Appliquons cette définition au calcul de la moyenne de la distribution :

x_j	n_j
0	1
1	3
2	5
3	5
4	4
5	2

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{(0 \times 1) + (1 \times 3) + (2 \times 5) + (3 \times 5) + (4 \times 4) + (5 \times 2)}{20} = \frac{0 + 3 + 10 + 15 + 16 + 10}{20} = \frac{54}{20} = 2,7$$

Les différences entre la formule de la moyenne donnée en (1) et celle donnée en (2) sont importantes à noter, **quoique les deux formules donnent nécessairement le même résultat.**

- La première différence tient au fait que dans la formule (1) la somme se fait sur les n unités statistiques, alors que dans la formule (2) la somme se fait sur les k valeurs.
- La seconde différence tient au fait que dans la formule (1), chaque élément additionné compte pour un, c'est-à-dire que la pondération est de 1. Dans la formule 2, les pondérations sont différentes. La formule (1) est en fait un cas particulier de la formule (2). En effet si $n = k$ et que $n_j = 1$ pour $j = 1$ à k alors les deux formules sont identiques.

En pratique, si le calcul de la moyenne doit être fait sans machine à calculer ni tableur et que les données sont peu nombreuses (inférieures à 10), mieux vaut utiliser la formule (1). Sinon, la formule (2) peut être préférée.

La moyenne arithmétique ne peut pas être calculée pour un caractère (dimension quantitative). Soit par exemple le caractère sexe, avec les deux modalités "F" et "H", dans une population de 10 personnes. On a la série suivante : {F,H,F,F,H,H, F,F,F,H}. La modalité "Femme" est plus fréquente (6 contre 4 pour la modalité "Homme") : c'est le mode. En revanche, on ne peut pas calculer de moyenne arithmétique. La même chose est vraie si l'on met cette série sous forme d'une distribution :

Sexe	Effectifs
H	4
F	6

On peut calculer les fréquences associées à chaque modalité. On voit alors que la modalité la plus fréquente est "F" (0,6 contre 0,4 pour la modalité "H"). Mais la notion de moyenne arithmétique n'a pas de sens.

3) Calcul de la moyenne sur des données regroupées en catégories

Ainsi que précisé dans l'introduction à ce chapitre, il faut à tout prix éviter de procéder à ce type de calcul. Nous ne le donnons ici qu'à titre d'information. Lorsque l'on veut à toute force calculer la moyenne d'une distribution par classes de valeurs, celle-ci s'obtient en prenant la formule de la moyenne pondérée et en remplaçant dans cette formule " x_j " par " c_j ", où c_j représente le **centre de la classe j**, c'est-à-dire la moyenne arithmétique des extrémités de classe. A défaut d'avoir une valeur x_j on prend " c_j ". Ceci explique que le calcul de la moyenne donne un résultat imprécis. On va le voir dans les deux exemples suivants :

Soit la série déjà utilisée précédemment : {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}. Nous savons que la moyenne arithmétique simple appliquée à cette série est :

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k n_j x_j = \frac{(0 \times 1) + (1 \times 3) + (2 \times 5) + (3 \times 5) + (4 \times 4) + (5 \times 2)}{20} = \frac{0 + 3 + 10 + 15 + 16 + 10}{20} = \frac{54}{20} = 2,7$$

Exemple 1 : Supposons maintenant que seule la distribution par classe de valeurs d'amplitudes égales nous soit donnée :

Classes	n_j
[0-2[	4
[2- 4[	10
[4- 6]	6

Pour calculer la moyenne, nous devons déterminer les centres de classe, puis faire la somme des " $n_j \times c_j$ " et diviser par n . Autrement dit, nous devons appliquer la formule :

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k n_j c_j$$

La notation c_j indique le centre de classe et où k représente le nombre de classes. Comme indiqué précédemment, le centre de classe est égal à la moyenne des extrémités de classe. On a donc le tableau de calcul suivant :

Classes	n_j	c_j (moyenne des extrémités de classe)	$n_j \times c_j$
[0-2[	4	1	4
[2- 4[	10	3	30
[4- 6]	6	5	30
			64

Et finalement :

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k n_j c_j = \frac{1}{20} \sum_{j=1}^3 [(4 \times 1) + (10 \times 3) + (6 \times 5)] = \frac{4 + 30 + 30}{20} = \frac{64}{20} = 3,2$$

Nous avons donc une marge d'erreur non négligeable par rapport à la vraie moyenne, à savoir 2,7. La marge d'erreur en pourcentage est donnée par :

$$\frac{3,2 - 2,7}{2,7} \times 100 = \frac{0,5}{2,7} \times 100 = 18,5\%$$

La marge d'erreur dépend de la définition des classes.

Exemple 2 : Supposons que l'on ait maintenant deux classes d'amplitudes inégales. Le calcul se fait de la même façon, mais on obtient un résultat différent :

Classes	n_i	c_i	$n_i \times c_i$
[0-4[	14	2	28
[4- 6]	6	5	30
			58

La moyenne est donc :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i c_i = \frac{1}{20} \sum_{i=1}^2 [(14 \times 2) + (6 \times 5)] = \frac{28 + 30}{20} = \frac{58}{20} = 2,9$$

On voit donc que chaque fois que l'on change les classes ou que l'on modifie leur amplitude, on exerce un effet sur la moyenne par le jeu de la modification des centres de classe. Il est donc facile de manipuler la moyenne en choisissant les amplitudes de classe. **C'est pourquoi il est recommandé de ne pas calculer la moyenne à partir d'un tableau d'effectifs regroupés par catégories de valeurs.**

4) Décomposition de la moyenne (série bidimensionnelle)

Pour introduire la décomposition de la moyenne en sous-moyennes, considérons l'exemple suivant.

Exemple : soit la distribution des notes obtenues par 50 étudiants dans une matière (les notes sont sur 20). Ces étudiants sont classés par groupe de TD. On a donc deux dimensions (première dimension : la note sur 20 ; deuxième dimension : le groupe). Les données sont présentées dans le tableau ci-après ([voir fichier EXCEL](#)) :

Deuxième dimension : Le groupe, qui est un caractère à 4 modalités (A,B,C et D)				
	A	B	C	D
Première dimension : la notation qui est une variable qui peut prendre des valeurs entre 0 et 20 (y compris des valeurs décimales ici limitées à 0,5)	2,5	2,5	5	4,5
	3	6,5	5,5	4,5
	5,5	8	5,5	7
	8	8	9,5	7,5
	9	8,5	10	9,5
	9,5	12	10,5	10
	10,5	13	10,5	11
	12	14,5	13	11
	14	15	13,5	12
	16	15	16,5	12,5
	16,5	17	17	13
	18		17	15
			18,5	
	A	B	C	D
Moyenne de chaque groupe	10,375	10,9090909	12,7333333	9,79166667
Coeff de pondération	(12 / 50)	(11 / 50)	(15 / 50)	(12 / 50)
Coeff de pondération calculé	0,24	0,22	0,3	0,24
Moyenne x pondération	2,49	2,4	3,82	2,35
On vérifie que $(0,24 \times 10,355) + (0,22 \times 10,909) + (0,3 \times 12,73) + (0,24 \times 9,79) = 11,06$				

On peut naturellement calculer la moyenne totale des 50 notes. On obtient 11,06.

On peut ensuite calculer la moyenne dans chacun des groupes A, B, C et D. On obtient 10,375 pour le groupe A, 10,909 pour le groupe B, 12,73 pour le groupe C et 9,79 pour le groupe D.

Il semble alors naturel de poser la question suivante : Existe-t-il une formule qui relie les 4 moyennes par groupe et la moyenne totale ? La réponse est oui, et la formule est⁴ :

$$\bar{x} = \frac{n_A}{n} \bar{x}_A + \frac{n_B}{n} \bar{x}_B + \frac{n_C}{n} \bar{x}_C + \frac{n_D}{n} \bar{x}_D$$

Avec $n_A+n_B+n_C+n_D=n$

La moyenne totale est égale à la somme de chaque moyenne préalablement pondérée par la part de l'effectif du groupe dans le total. Dans notre exemple, cela donne :

$$(0,24 \times 10,355) + (0,22 \times 10,909) + (0,3 \times 12,73) + (0,24 \times 9,79) = 11,06$$

Cette formule permet de voir que la valeur de la moyenne totale est influencée :

- Par les moyennes de chaque groupe
- Par l'importance relative des effectifs de chaque groupe. Par exemple, comme le groupe C est le plus nombreux (15 étudiants), la moyenne du groupe C aura plus d'influence sur la moyenne totale que la moyenne du groupe B qui ne comprend que 11 étudiants.

On peut alors se demander quelle serait la moyenne totale si le nombre d'étudiants était le même dans chaque groupe. La réponse est simple : s'il y avait autant d'étudiants dans chaque groupe, chaque moyenne compterait pour 1/4 et l'on aurait alors :

	$(0,25 \times 10,355) + (0,25 \times 10,91) + (0,25 \times 12,73) + (0,25 \times 9,79) = 10,95$			
Moyenne x (1/4)	2,59375	2,72727273	3,18333333	2,44791667

La différence entre 11,06 et 10,95, soit environ 0,11, est attribuable à l'importance relative des effectifs de chaque groupe. Dans certains cas, il peut être intéressant d'isoler cette différence, afin de mieux saisir ce qui relève de la différence des moyennes de ce qui relève de la structure de la population (ici, la « structure », c'est les effectifs de chaque groupe).

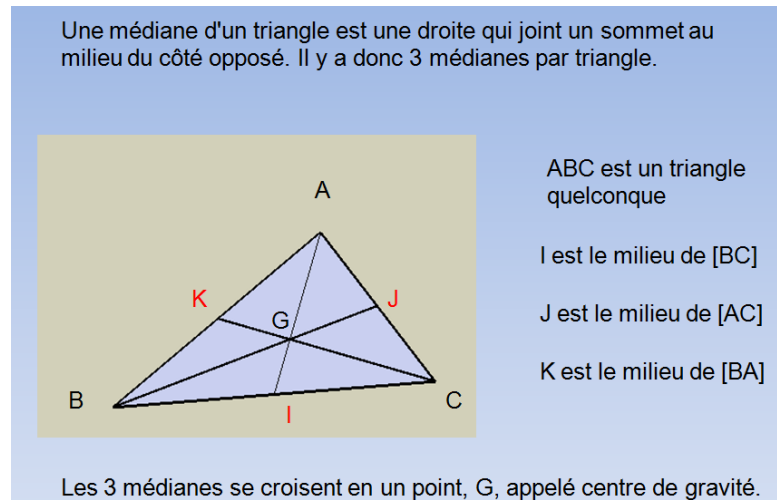
C - La médiane

1) Origine du mot, sens géométrique

Le mot « médiane » a pour origine le latin « médius », mot signifiant « qui est au milieu ». « Médius » est aussi l'origine du mot "moyenne". Pourtant, en statistique, les deux mots conduisent à des définitions différentes. Ceci nous laisse supposer que la notion de milieu n'est pas toujours facile à définir.

⁴ Voir la démonstration dans [l'annexe 2 à ce chapitre](#).

Avant d'avoir une définition statistique, la médiane a d'abord une définition géométrique. On définit ainsi, par exemple, les médianes d'un triangle. Une médiane d'un triangle est une droite qui joint un sommet au milieu du côté opposé. Il y a donc 3 médianes par triangle.



2) Sens du mot en statistique descriptive

La **médiane** est la valeur de la variable (et dans certains cas, la modalité du caractère) qui partage la population, dont les unités statistiques ont préalablement été classées par ordre croissant des valeurs (de la variable considérée), en deux sous populations égales.

3) Méthode de calcul

- **Toujours se ramener à une série** : Si les données se présentent sous forme d'une distribution, convertir la distribution en série.
- **Classer la série** : Une fois la série constituée, l'ordonner en classant les chiffres par ordre croissant. On désignera par k le rang d'une valeur dans la série (voir tableau ci-après).
- **Déterminer si la série comprend un nombre pair ou impair d'unités statistiques** : Soit n le nombre d'unités statistiques. Deux cas peuvent alors se présenter : celui où n est pair et celui où n est impair.

a) n est pair

Dans ce cas la médiane est égale à la moyenne arithmétique de $a(k)$ et de $a(k+1)$ où k est tel que $n = 2k$ [$k=n/2$] et où $a(k)$ et $a(k+1)$ sont les valeurs associées à k et à $k+1$

Exemple : Soit la série $S = \{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5\}$.

On a alors le tableau suivant (où les valeurs sont classées par ordre croissant):

Série classée par ordre croissant		Rang
0		1
1		2
1		3
1		4
2		5
2		6
2		7
2		8
2		9
a(k)	3	10
a(k+1)	3	11
	3	12
	3	13
	3	14
	4	15
	4	16
	4	17
	4	18
	5	19
	5	20

Puisque $n=20$, n est pair. Donc $n=2k$ [soit $k=n/2$] donne ici $k=20/2=10$ et $k+1=11$. Les valeurs associées sont $a(k)=3$ et $a(k+1)=3$. Par conséquent :

$$Me = [a(k) + a(k+1)]/2 = (3+3)/2 = 3$$

On peut aussi appliquer la formule proposée par le tableur EXCEL (cf. la notice technique sur le site de Microsoft : <http://support.microsoft.com/kb/103493/fr>⁵).

Dans ce cas, on doit calculer Me de la façon suivante :

$$Me = a(k) + [f \times [a(k+1) - a(k)]]$$

Pour comprendre cette formule, il faut d'abord définir g :

$$g = [(1/2)(n-1)] + 1.$$

Une fois que l'on a défini g , on peut en déduire k et f :

k représente le rang de la valeur dans le classement du tableau ci-dessus et est égal à la partie entière de g .

f est la partie décimale de g .

$a(k)$ est la valeur associée au rang k

$a(k+1)$ est la valeur associée au rang $k+1$.

⁵ La notice technique sur le site Internet de Microsoft (<http://support.microsoft.com/kb/103493/fr>) présente la formule différemment mais aboutit au même résultat.

Si l'on applique cette formule, on voit que :

$$g = [(1/2)(20-1)]+1=10,5$$

Par conséquent $k= 10$ et $f = 0,5$. On a donc $a(k) = 3$ et $a(k+1)=3$. Dès lors :

$$Me = 3 + [0,5(3-3)] = 3$$

b) n est impair

Dans ce cas la médiane est égale à $a(k+1)$ où k est défini par $(n-1)/2$, soit $n=2k+1$ et où $a(k+1)$ est la valeur associée à $k+1$

Exemple : si l'on prend la série précédente mais que l'on enlève le 20^{ème} élément, on a alors la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4} ou $n = 19$ et le tableau suivant (où les valeurs sont classées par ordre croissant) :

Série classée par ordre croissant		Rang
0		1
1		2
1		3
1		4
2		5
2		6
2		7
2		8
$a(k)$ → 2		9 ← k
$a(k+1)$ → 3		10 ← $k+1$
3		11
3		12
3		13
3		14
4		15
4		16
4		17
4		18
5		19

Ici $k= (n-1)/2= 9$ [$2k+1 = n$ donne donc $2k+1 = 19$, puisque $k=9$]. Donc $k+1 = 10$.

Par conséquent la médiane est égale à $a(k+1)$ soit 3.

On peut, comme dans le cas où n est pair, appliquer la formule proposée par le tableur EXCEL et on obtiendra le même résultat.

2 - Les statistiques de dispersion

Alors que les statistiques résumées précédentes (mode, moyenne arithmétique et médiane) voulaient résumer la tendance centrale, la position ou le milieu d'une série de chiffres, celles que nous allons étudier maintenant entendent résumer le degré de dispersion, de variabilité ou encore de volatilité autour du milieu, des chiffres d'une série.

L'indicateur royal de la dispersion est la variance ou, mieux, sa racine carrée, que l'on appelle l'écart-type. Cependant, son utilisation, quoique très fréquente en science, est pratiquement nulle dans les médias. Ceci est un réel problème, car la communication de certaines informations sur une série de chiffres (qu'il s'agisse de salaires, de prix, de taux de croissance, etc.) passe nécessairement par la connaissance des propriétés de la variance et de l'écart-type. Pour éviter de recourir à l'utilisation de la variance et de l'écart-type, les médias se contentent de commenter les valeurs extrêmes d'une série, négligeant le fait que ces valeurs extrêmes ne sont pas nécessairement représentatives de la dispersion de l'ensemble des chiffres de la série.

La variance et l'écart-type sont influencées par les valeurs extrêmes d'une série, mais cette influence est diluée dans le calcul d'ensemble. Par ailleurs, pour avoir une idée de la dispersion au sein d'une série de chiffre en excluant les valeurs extrêmes atypiques, on dispose de l'intervalle interquartile (et ses cousins proches que sont l'intervalle interdécile et les centiles).

On pourrait donc légitimement se contenter d'étudier ici la variance et l'écart-type d'une part, l'intervalle interquartile d'autre part. Mais on ajoutera cependant à ces deux indicateurs, la définition et le calcul des extrémités de séries.

A - Minimum, maximum, intervalle de variation et rapport de variation

1) Minimum et maximum d'une série

Soit la série suivante : {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}

En classant les valeurs de la plus petite à la plus grande, on remarque que la valeur minimum est 0 et que la valeur maximum est 5.

{0,1,1,1, 2, 2,2,2,2, 3,3,3,3,3, 4,4,4,4, 5,5}

2) Intervalle de variation (ou « étendue »)

L'intervalle de variation (IV) ou l'étendue de la série est simplement une façon de résumer le minimum et le maximum de la série en un seul chiffre. On l'obtient ainsi :

Intervalle de variation de la série = valeur maximale – Valeur minimale

La série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5} a pour intervalle de variation $5-0 = 5$.

3) Rapport de variation

Le **rapport de variation** est simplement le rapport de la valeur maximale à la valeur minimale. Par exemple, si on calcule le rapport de variation de la série précédente on obtient 5/0. Ce calcul est impossible On en déduit qu'il n'est pas toujours possible de calculer le rapport de variation !!!

4) Milieu arithmétique

Les valeurs minimale et maximale ne reflètent pas ou très mal la dispersion des valeurs au sein d'une série de chiffres car il peut s'agir (et il s'agit souvent) de valeurs atypiques, extrêmes, exceptionnelles, le reste des valeurs de la série pouvant être quant à lui dispersé ou resserré autour du mode, de la médiane, de la moyenne, voir du milieu arithmétique de l'intervalle de variation.

C'est la raison pour laquelle l'intervalle de variation n'est pas utilisé comme indicateur de la dispersion. Il faut néanmoins le repérer, ne serait-ce que pour calculer le milieu arithmétique :

$$\text{Milieu arithmétique} = \text{intervalle de variation} : 2.$$

La connaissance de l'intervalle de variation ainsi que du milieu arithmétique est utile pour le tracé de divers graphiques (ne serait-ce que pour déterminer le minimum et le maximum de l'axe des abscisses) et en particulier pour l'étude de la forme de la distribution sous-jacente à une série de chiffres à partir du graphique de TUKEY, même si la longueur du graphique de TUKEY n'est pas nécessairement égale à l'intervalle de variation (c'est seulement lorsqu'il n'y a pas de valeur atypique que les deux se confondent, voir le chapitre 5).

B - Intervalles interquartile et interdécile

L'intervalle de variation ne donne qu'une idée imprécise et parfois fautive de la dispersion de la série, car les valeurs extrêmes peuvent être exceptionnelles et le reste de la population statistique être concentré sur un intervalle beaucoup plus restreint. D'où l'idée de calculer **l'intervalle interquartile** qui donne une idée plus précise de la dispersion des valeurs d'une série (ou d'une distribution). On peut aussi calculer **l'intervalle interdécile** (notamment pour la répartition des richesses, des salaires, etc.). Avant de définir les intervalles interquartile et interdécile, il convient cependant de définir les quartiles et les déciles.

1) Définition et calcul des quartiles

Les **quartiles** sont les **trois valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **quatre sous populations** de même taille. On les désigne respectivement par Q_1 , Q_2 et Q_3 .

On notera que $Q_2 = Me$. Autrement dit, le deuxième quartile n'est autre que la médiane que nous avons déjà étudiée. **Il est important de noter qu'il n'existe pas d'algorithme universellement accepté pour déterminer les quartiles Q_1 et Q_3 .** Cependant, cette diversité de méthodes n'est pas réellement un problème car plus la série contient de chiffres, plus les méthodes de calcul des quartiles donnent des résultats identiques. Pour les très grandes séries, les différences entre méthodes deviennent négligeables. Dans ce qui suit, nous utiliserons la formule employée par le logiciel EXCEL de Microsoft⁶.

⁶ Voir la notice technique sur le site de Microsoft : <http://support.microsoft.com/kb/103493/fr> . La notice présente la formule différemment mais aboutit au même résultat. Les principales autres méthodes de calcul des quartiles sont résumées sur le site [Mathworld](http://mathworld.wolfram.com/Quartile.html). L'avantage de la méthode Microsoft est qu'il n'est pas nécessaire d'appliquer une formule différente suivant que n est pair ou impair. En outre, le même principe s'applique au calcul des déciles et des centiles (voir plus loin).

Le logiciel **Mathematica** détermine quant à lui les quartiles de la façon suivante :

- `Quartiles[list]` is equivalent to `Quantile[list, {1/4, 1/2, 3/4}, {{1/2, 0}, {0, 1}}]`. »
- The second quartile is equivalent to `Median[list]`. »
- For even `Length[list]`, the first quartile is equivalent to the median of the $\frac{n}{2}$ smallest elements in `list`.
- For odd `Length[list]`, the first quartile is equivalent to the average of the median of the $\frac{n-1}{2}$ smallest elements and the median of the $\frac{n+1}{2}$ smallest elements in `list`.
- The third quartile is defined as for the first, but with the largest rather than smallest elements.

S'agissant de la série précédente, on trouve ainsi $Q_3 = 11,175$ avec Mathematica :

```
In[4]:= Quartiles[{8.4, 8.5, 8.9, 9, 9.1, 9.5, 9.5, 9.6, 10.5, 11.1, 11.1, 11.2, 11.5, 11.6, 11.7}]
Out[4]:= {9.025, 9.6, 11.175}
```

Un autre algorithme de calcul est proposé dans [l'annexe 1 à ce chapitre](#).

Prenons l'exemple de la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a $n=20$

Le calcul de Q_i ($i=1,2$ ou 3) s'effectuera dès lors au moyen la formule suivante :

$$Q_i = a(k) + [f \times [a(k+1) - a(k)]]$$

S'agissant du premier quartile on aura donc :

$$Q_1 = a(k) + [f \times [a(k+1) - a(k)]]$$

Pour comprendre cette formule, il faut d'abord définir g :

$$g = [(1/4)(n-1)] + 1.$$

Une fois que l'on a défini g , on peut en déduire k et f :

- k représente le rang de la valeur dans le classement du tableau ci-dessous et est égal à la partie entière de g
- f est la partie décimale de g .
- $a(k)$ est la valeur associée au rang k et $a(k+1)$ est la valeur associée au rang $k+1$ Si l'on applique cette formule, on voit que :

$$g = [(1/4)(20-1)] + 1 = 5,75$$

Par conséquent $k = 5$ et $f = 0,75$. On a donc $a(k) = 2$ et $a(k+1) = 2$. Dès lors :

$$Q_1 = 2 + [0,75 \times (2-2)] = 2$$

On peut suivre la formule sur le tableau ci-après :

	Série classée	Rang
	0	1
	1	2
	1	3
	1	4
$Q_1 = 2$	$a(k) \rightarrow 2$	5 $\leftarrow k$
	$a(k+1) \rightarrow 2$	6 $\leftarrow k+1$
	2	7
	2	8
	2	9
	3	10
	3	11
	3	12
	3	13
	3	14
	4	15
	4	16
	4	17
	4	18
	5	19
	5	20

Le calcul de Q_3 s'effectue au moyen de la même formule que pour Q_1 , soit :

$$Q_3 = a(k) + [f \times [a(k+1) - a(k)]]$$

Mais avec un changement dans la définition de g. Désormais on a :

$$g = \lfloor \frac{3}{4} (n-1) \rfloor + 1$$

Prenons toujours l'exemple de la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a n=20. Dès lors :

$$g = \lfloor \frac{3}{4} (n-1) \rfloor + 1 = \lfloor \frac{3}{4} (20-1) \rfloor + 1 = 15,25$$

Par conséquent $k = 15$ et $f = 0,25$. On a donc $a(k) = 4$ et $a(k+1) = 4$. Dès lors :

$$Q_3 = 4 + \lfloor 0,25 \times (4-4) \rfloor = 4$$

On peut suivre la formule sur le tableau ci-après :

	Série classée	Rang
	0	1
	1	2
	1	3
	1	4
	2	5
	2	6
	2	7
	2	8
	2	9
	3	10
	3	11
	3	12
	3	13
	3	14
$Q_3 = 4$	$a(k) \rightarrow 4$	15 $\leftarrow k$
	$a(k+1) \rightarrow 4$	16 $\leftarrow k+1$
	4	17
	4	18
	5	19
	5	20

2) Définition, calcul et interprétation de l'intervalle interquartile

L'intervalle interquartile (IIQ) est la différence entre le troisième quartile et le premier quartile. Il s'écrit :

$$IIQ = Q_3 - Q_1$$

L'intervalle interquartile sert à apprécier la dispersion de la série, de façon absolue, ou bien par comparaison avec une autre série (à condition que les valeurs de l'autre série soient exprimées dans la même unité). En effet, les valeurs Q_1 et Q_3 délimitent une plage au sein de laquelle **environ**⁷ 50% des valeurs de la série sont concentrées. **Plus cet intervalle est large, plus la série est dispersée.** Dans l'exemple que nous avons utilisé, l'IIQ est égal à $4 - 2 = 2$.

⁷ C'est pour cette raison que les algorithmes de calcul diffèrent.

3) Définition et calcul des déciles

Les **déciles** sont les **neuf valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **dix sous populations** de même taille. On les désigne respectivement par $D_1, D_2, \dots, D_9$.

Il est logique d'utiliser le même algorithme de calcul pour les déciles que pour les quartiles. Par conséquent, nous utiliserons ici l'algorithme employé par les concepteurs du logiciel EXCEL de Microsoft.

On a ainsi :

$$D_i = a(k) + [f \times [a(k+1) - a(k)]]$$

Pour déterminer k et f, on prendra :

- pour $D_1 \rightarrow g = (1/10)(n-1) + 1$
- pour $D_2 \rightarrow g = (2/10)(n-1) + 1$

Et ainsi de suite jusqu'à D_9 :

- pour $D_9 \rightarrow g = (9/10)(n-1) + 1$

Très souvent (mais pas toujours), on ne s'intéresse qu'aux déciles extrêmes, c'est-à-dire D_1 et D_9 .

Exemple : Soit la série {21, 17, 18, 21, 40, 9, 21, 9, 18, 18, 19, 2, 2, 5, 20, 21, 14, 15, 21, 9, 14, 21, 24, 5, 9, 9, 1, 1, 2, 5, 13, 40, 40, 26, 28, 30, 30, 35, 39, 16, 17}

Classons-la par ordre croissant et calculons la valeur de g associée à D_1 , puis la valeur de g associée à D_9 et déduisons-en D_1 et D_9 . On obtient :

				1	1	$g=(1/10)(41-1)+1=5$
				1	2	
				2	3	On en déduit $k=5$ et $f=0$
				2	4	
			$a(k)$	2	5	$\leftarrow k=5$
			$a(k+1)$	5	6	
				5	7	
				5	8	
				9	9	
				9	10	
				9	11	
				9	12	
				9	13	
				13	14	
				14	15	
				14	16	
				15	17	
				16	18	
				17	19	
				17	20	
				18	21	
				18	22	
				18	23	
				19	24	
				20	25	
				21	26	
				21	27	
				21	28	
				21	29	
				21	30	
				21	31	
				24	32	
				26	33	
				28	34	$g=(9/10)(41-1)+1=37$
				30	35	On en déduit $k=37$ et $f=0$
				30	36	
			$a(k)$	35	37	$\leftarrow k=37$
			$a(k+1)$	39	38	
				40	39	
				40	40	
				40	41	

[Fichier EXCEL](#)

On voit ainsi que $D_1 = 2$ et que $D_9 = 35$.

4) Intervalle interdécile

Bien qu'il puisse aussi servir à apprécier la dispersion d'une série, l'intervalle interdécile sert surtout à mettre en évidence l'inégalité au sein des distributions de revenus ou de richesse⁸. En effet :

- les valeurs D_1 et D_9 délimitent une plage au sein de laquelle **environ**⁹ 80% des valeurs de la série sont concentrées.
- les valeurs D_1 et D_9 délimitent aussi les 20% de valeurs qui sont à l'extérieur de l'intervalle interdécile (10% de valeurs les plus faibles de la série et 10% de valeurs les plus élevées de la série).

Plus l'intervalle interdécile est large, plus la série est dispersée. Dans notre exemple, puisque $D_1 = 2$ et que $D_9 = 35$, l'IID est égal à $35 - 2 = 33$.

5) Note sur les centiles

Les **centiles** sont les **99 valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **cent sous populations** de même taille. On les désigne respectivement par $C_1, C_2, \dots, C_{99}$.

Il est logique d'utiliser le même algorithme de calcul pour les centiles que pour les déciles. Par conséquent, nous utiliserons ici l'algorithme employé par les concepteurs du logiciel EXCEL de Microsoft.

On a ainsi :

$$C_i = a(k) + [f \times [a(k+1) - a(k)]]$$

Pour déterminer k et f, on prendra :

- pour $C_1 \rightarrow g = (1/100)(n-1) + 1$
- pour $D_2 \rightarrow g = (2/100)(n-1) + 1$

Et ainsi de suite jusqu'à D_9 :

- pour $D_9 \rightarrow g = (99/100)(n-1) + 1$

Très souvent (mais pas toujours), on ne s'intéresse qu'aux centiles extrêmes, c'est-à-dire C_1 et C_{99} .

Contrairement aux quartiles et aux déciles, on calcule rarement un intervalle intercentile (ce n'est pas interdit cependant). Pour le calculer, il suffit de faire l'opération $C_{99} - C_1$.

⁸ Par exemple, si l'on classe les 27 pays de l'UE par ordre de PIB par habitant croissant pour l'année 2008, on voit alors que les pays qui sont avant le premier décile sont la Roumanie, la Bulgarie et la Pologne, tandis que les pays qui sont après le 9^{ème} décile sont Les Pays-Bas, l'Irlande et le Luxembourg. [Voir le fichier EXCEL](#).

⁹ C'est pour cette raison que les algorithmes de calcul sont différents.

C - Variance, écart-type et coefficient de variation

Ces trois statistiques sont liées entre elles. Elles sont toutes les trois des indicateurs de la dispersion d'une série par rapport à sa valeur moyenne. Le plus simple est de commencer par l'étude de la variance.

1) La variance

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. De même que la moyenne, elle se résume à un seul chiffre qui s'obtient par un calcul que nous allons décomposer ci-après.

a) Définition

La définition de la variance d'une série de chiffres est donnée par la formule¹⁰ :

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Où :

- V désigne la variance des n valeurs associées aux n unités statistiques de la population. Chacune de ces valeurs est désignée par x_i , le i étant un indice qui varie de 1 à n ($i = 1, \dots, n$).
- $\bar{x}$ est la **moyenne arithmétique simple** des n valeurs associées aux unités statistiques x_i ($i = 1, \dots, n$).

b) Exemple

Soit la série $S = \{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5\}$ ou $n=20$. Pour calculer la variance de cette série, on procède ainsi :

- **Toujours se ramener à une série** : par exemple, si au lieu d'avoir une série on avait la distribution suivante il faudrait d'abord la transformer en série:

x_i	n_i
0	1
1	3
2	5
3	5
4	4
5	2

¹⁰ Nous donnons ici la formule de la variance des valeurs associées aux unités statistiques d'une population et non la variance d'un **échantillon** de cette population, dont la définition est légèrement différente. En effet, pour calculer la variance d'un échantillon, on divise par $n-1$ au lieu de diviser par n . Mais dans ce cas le « n » de l'échantillon est évidemment beaucoup plus petit que le « n » de la population et l'on différencie alors les deux en désignant par N le nombre d'unités statistiques de la population et par n le nombre d'unités statistiques de l'échantillon. De plus, si l'on veut extraire plusieurs échantillons de la population, on est amené à rajouter un indice aux n pour les distinguer (on prendra alors l'indice j puisque l'indice i est déjà utilisé pour désigner les unités statistiques elles-mêmes).

- **Calculer la moyenne arithmétique simple :**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{4+0+1+1+2+2+2+3+3+4+2+3+4+5+2+1+3+3+4+5}{20} = \frac{54}{20} = 2,7$$

- **Retrancher ensuite cette moyenne de chacune des 20 valeurs de la série** (colonne 3 du tableau ci-après).
- On obtient ainsi une série qui comprend des valeurs négatives (car certaines valeurs de la variable sont inférieures à la moyenne et donc si on retranche la moyenne elles deviennent négatives) et des valeurs positives (car certaines valeurs de la variable sont supérieures à la moyenne et donc si on retranche la moyenne elles restent positives).
- Afin de tenir compte à la fois des distances positives et négatives, on ne peut pas additionner immédiatement les valeurs de la colonne 3. Il faut d'abord élever au carré chacune de ces valeurs, de façon à obtenir une série de valeurs positives (colonne 4).
- Cette série de valeurs positives reflète déjà en elle-même la dispersion par rapport à la moyenne. Mais il faut encore additionner ces valeurs pour avoir un chiffre unique (dernière valeur en caractère gras dans la colonne 4)
- Diviser ensuite ce total par n , pour avoir en fait une sorte de moyenne. C'est pour cela que l'on dit que la variance n'est finalement que « la moyenne du carré des écarts à la moyenne ». Et l'on obtient la variance de notre série de chiffres, soit ici :

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{36,2}{20} = 1,81$$

Tableau pour la disposition du calcul de la variance

[illegible]

c) Utilité de la variance

La variance n'est pas d'une grande utilité en soi. On peut seulement dire que plus elle est élevée, plus la dispersion autour de la moyenne est élevée. Mais comme les écarts à la moyenne ont été élevés au carré, le chiffre obtenu, quoiqu'exprimé dans l'unité de la variable, est généralement assez élevé et « encombrant ». C'est pourquoi, on utilise surtout la variance comme **calcul intermédiaire** pour obtenir **l'écart-type** et le **coefficient de variation**.

2) L'écart-type

a) Définition

La définition de l'**écart-type** d'une série de chiffres est donnée par la formule :

$$\sigma_x = \sqrt{V(x)}$$

En d'autres termes, l'écart-type est égal à la racine carrée de la variance.

b) Exemple

Soit la série $S = \{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5\}$ ou $n=20$.

Nous voulons calculer l'écart-type. Nous avons déjà calculé la variance :

$$V(x)=1,81$$

Il suffit donc de prendre la racine carrée de la variance et l'on obtient :

$$\sigma_x = \sqrt{V(x)} = \sqrt{1,81} = 1,345$$

c) Utilité de l'écart-type

De façon générale :

- **si l'écart-type est faible**, cela signifie que les valeurs sont assez concentrées autour de la moyenne.
- **si l'écart-type est élevé**, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

Exemple : Dans une usine, le fait d'avoir un écart-type aussi bas que possible peut constituer un **objectif de contrôle de qualité**. Soit une entreprise qui fabrique un certain composant et qu'un des éléments du contrôle de la qualité consiste à mesurer le diamètre du composant. Chaque composant aura donc son diamètre mesuré. On calculera ensuite le diamètre moyen, puis l'écart-type. Si l'écart-type est faible, cela signifie que les pièces ont dans l'ensemble un diamètre proche de la moyenne, donc que leur diamètre se ressemble. *À la limite, un écart-type nul signifie que toutes les pièces ont le même diamètre.* Inversement, plus l'écart-type est élevé, plus il y a de pièces dont le diamètre s'écarte de la moyenne et qui risque de ne pas cadrer avec le système auxquelles elles sont destinées. Une version plus élaborée de cet exemple constitue la base de la méthode de contrôle de qualité dite **six-sigma**, ou « 6 fois l'écart-type », sigma étant le symbole utilisé pour représenter l'écart-type d'une population.

Une autre utilité de l'écart-type est de servir de bornes. On regardera par exemple combien de valeurs de la série sont comprises dans l'intervalle défini par :

$$[\bar{x} - \sigma_x; \bar{x} + \sigma_x]$$

Dans l'exemple de $S = \{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5\}$, dont nous avons déjà calculé la moyenne simple (2,7), nous obtenons l'intervalle suivant :

$$[2,7 - 1,345 ; 2,7 + 1,345] = [1,355 ; 4,045]$$

Pour savoir combien d'unités ont des valeurs comprises dans cet intervalle, il faut classer la série par ordre croissant des valeurs de la variable et ensuite encadrer les valeurs comprises dans l'intervalle. On voit ainsi que 14 unités sur 20 ont des valeurs comprises dans l'intervalle, ce qui représente 70% de la population. Si ces valeurs se rapportaient aux caractéristiques des pièces d'un processus industriel on pourrait juger que la qualité est acceptable ou bien au contraire se fixer comme objectif d'augmenter le nombre de valeurs qui sont dans cet intervalle. On voit que cet intervalle est avec l'intervalle interquartile une autre façon de mesurer la dispersion d'une série.

Nombres d'unités statistiques dont la valeur est comprise dans l'intervalle

$$[\bar{x} - \sigma_x; \bar{x} + \sigma_x]$$

Colonne 1	Colonne 2	
	Valeurs de la variable X (désignées par x_i)	
x_2	0	
x_3	1	
x_4	1	
x_{16}	1	
x_5	2	1,355 x_5 x_6 x_7 x_{11} x_{15} x_8 x_9 x_{12} x_{17} x_{18} x_1 x_{10} x_{13} x_{19} 4,045 14 valeurs sur 20 (70%) sont comprise dans l'intervalle $[\bar{x} - \sigma_x; \bar{x} + \sigma_x]$
x_6	2	
x_7	2	
x_{11}	2	
x_{15}	2	
x_8	3	
x_9	3	
x_{12}	3	
x_{17}	3	
x_{18}	3	
x_1	4	
x_{10}	4	
x_{13}	4	
x_{19}	4	
x_{14}	5	
x_{20}	5	

3) Le coefficient de variation

L'écart-type est un outil intéressant pour mesurer la dispersion d'une série, ainsi que nous venons de le voir. Mais il possède une limite : **Il est exprimé dans l'unité de la variable dont il mesure la dispersion des valeurs.**

Ainsi, par exemple, si on veut comparer les dispersions de deux séries qui sont exprimées dans des unités différentes, on ne peut pas.

D'où l'idée de calculer le **coefficient de variation** qui est égal à l'écart-type divisé par la moyenne, c'est-à-dire :

$$C_v = \frac{\sigma_x}{\bar{x}}$$

4) Décomposition de la variance (deux dimensions)

Lorsqu'on souhaite connaître la variance au sein d'une population étudiée en fonction de deux dimensions, il est possible d'effectuer une décomposition de la variance totale en deux éléments :

- La variance intra population
- La variance inter population

Exemple : soit la distribution des notes obtenues par 50 étudiants dans une matière (les notes sont sur 20). Ces étudiants sont classés par groupe de TD. On a donc deux dimensions (première dimension : la note sur 20 ; deuxième dimension : le groupe). Les données sont présentées dans le tableau ci-après ([voir fichier EXCEL](#)) :

		Deuxième dimension : Le groupe, qui est un caractère à 4 modalités (A,B,C et D)			
		A	B	C	D
Première dimension : la notation qui est une variable qui peut prendre des valeurs entre 0 et 20 (y compris des valeurs décimales ici limitées à 0,5)	2,5	2,5	5	4,5	
	3	6,5	5,5	4,5	
	5,5	8	5,5	7	
	8	8	9,5	7,5	
	9	8,5	10	9,5	
	9,5	12	10,5	10	
	10,5	13	10,5	11	
	12	14,5	13	11	
	14	15	13,5	12	
	16	15	16,5	12,5	
	16,5	17	17	13	
	18		17	15	
			18,5		
			19		
			20		

On peut naturellement calculer la variance totale des 50 notes, de même que la moyenne totale. On obtient ([voir fichier EXCEL](#)) :

Variance des 50 notes	Moyenne des 50 notes
21,0264	11,06

On peut ensuite poser la question suivante : Quelle part de la variance totale des notes est attribuable à la variation des notes au sein de chaque groupe (**variance intra population**) et quelle part de la variance totale des notes est attribuable à la variance des notes entre les groupes (**variance inter population**) ?

Pour répondre à cette question, il faut appliquer la formule suivante¹¹ :

Variance totale = Composante intra population + Composante inter population

La composante intra population est la *moyenne pondérée des variances de chaque groupe* et se calcule comme suit :

Composante intra : $(n_A/n) V(A) + (n_B/n) V(B) + (n_C/n) V(C) + (n_D/n) V(D)$

Où $n_A + n_B + n_C + n_D = n$ et où $V(A)$, $V(B)$, $V(C)$ et $V(D)$ sont les variances des notes dans chaque groupe.

Dans notre exemple, cela donne ([voir fichier EXCEL](#)) :

$$(12/50)V(A) + (11/50)V(B) + (15/50)V(C) + (12/50)V(D) = 19,68268181818180$$

La composante inter population est la *variance pondérée des moyennes de chaque groupe* et se calcule comme suit :

Composante inter : $(n_A/n) (\text{moyenne des notes du groupe A} - \text{moyenne totale})^2 +$
 $(n_B/n) (\text{moyenne du groupe B} - \text{moyenne totale})^2 +$
 $(n_C/n) (\text{moyenne du groupe C} - \text{moyenne totale})^2 +$
 $(n_D/n) (\text{moyenne du groupe D} - \text{moyenne totale})^2 +$

Dans notre exemple, cela donne ([voir fichier EXCEL](#)) :

$$(12/50)(10,375-11,06)^2 + (11/50)(10,91-11,06)^2 + (15/50)(12,73-11,06)^2 + (12/50)(9,79-11,06)^2 + = 1,343718181818180$$

$$\text{On vérifie que : } 19,68268181818180 + 1,343718181818180 = 21,0264$$

Cette décomposition permet de voir que c'est la variabilité des notes au sein de chaque groupe qui explique l'essentiel de la variance totale $(19,68/21,0264)*100 =$

¹¹ La démonstration s'appuie sur le théorème de KOËNIG-HUYGENS. Le lecteur intéressé peut se reporter à l'ouvrage de Gérard CALOT, [Statistique descriptive](#), Chapitre 3, propriétés de la moyenne, pages 51 et suivantes.

Exemple de décomposition de la variance (voir fichier EXCEL)

31

3 – Résumé

On distingue **deux types de statistiques résumées** : Les statistiques de **tendance « centrale »** (**mode, moyenne, médiane**) et les **statistiques de dispersion** (**intervalle interquartile** essentiellement). Ces deux types peuvent être combinés pour donner des statistiques résumées plus élaborées telles que l'**écart-type**, la **variance** et le **coefficient de variation**.

Le **mode** d'une série est la valeur qui revient le plus fréquemment dans la série.

La **moyenne arithmétique simple** d'une série est égale à la somme des valeurs de la série, divisée par le nombre total de valeurs de la série.

La **médiane** est la valeur de la variable (et dans certains cas, la modalité du caractère) qui partage la population, dont les unités statistiques ont préalablement été classées par ordre croissant des valeurs (de la variable considérée), en deux sous populations égales.

L'**intervalle de variation** d'une série est égal à la valeur maximale de la série moins la valeur minimale de la série.

Le **rapport de variation** est simplement le rapport de la valeur maximale à la valeur minimale. On ne peut pas le calculer quand la valeur minimale est 0.

Les **quartiles** sont les **trois valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs, en **quatre sous populations** de même taille. On les désigne respectivement par Q_1 , Q_2 et Q_3 . Personne n'est d'accord sur la façon de calculer les quartiles, mais tout le monde est d'accord pour utiliser le logiciel EXCEL. C'est pourquoi nous retenons ici l'algorithme de calcul proposé par les concepteurs de ce logiciel. Cependant, il ne faut pas s'étonner du fait qu'il existe de nombreuses formules alternatives (dont une est proposée en annexe à ce chapitre¹²).

L'**intervalle interquartile** (IIQ) est la différence entre le troisième quartile et le premier quartile.

Les **déciles** sont les **neuf valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **dix sous populations** de même taille. On les désigne respectivement par D_1 , D_2 , ..., D_9 . Le mode de calcul des déciles se base sur le même algorithme que celui des quartiles.

L'**intervalle interdécile** est la différence entre le neuvième décile et le premier décile.

Les **centiles** sont les **99 valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **cent sous populations** de même taille. On les désigne

¹² Voir aussi, plus haut dans ce chapitre, la note de bas de page relative à la méthode de calcul proposée par le logiciel Mathematica, très performant, mais moins utilisé.

respectivement par $C_1, C_2, \dots, C_{99}$. Le mode de calcul des déciles se base sur le même algorithme que celui des quartiles et des déciles.

La **variance** est égale à la moyenne des carrés des écarts à la moyenne arithmétique de la série.

Lorsqu'on souhaite connaître la variance au sein d'une population étudiée en fonction de deux dimensions, il est possible d'effectuer une décomposition de la variance totale en deux éléments suivant la formule :

Variance totale = Composante intra population + Composante inter population

La composante intra population est la *moyenne pondérée des variances de chaque groupe*. **La composante inter population** est la *variance pondérée des moyennes de chaque groupe*

L'**écart-type** est égal à la racine carrée de la variance.

Le **coefficient de variation** est égal à l'écart-type divisé par la moyenne

Lorsque l'on étudie simultanément deux dimensions d'une population, il est parfois utilisé de décomposer la variance totale en deux :

- La composante **intra population** (ou moyenne pondérée des variances).
- La composante **inter population** (ou variance pondérée des moyennes).

Annexe 1 : Méthode alternative pour le calcul des quartiles

Cette méthode ne correspond pas à celle employée par EXCEL, ni par les autres logiciels de calcul. **Elle n'est donnée ici qu'à titre d'information parce qu'elle est la plus logique.** C'est aussi la méthode qui est [proposée dans Wikipedia](#) :

i) **Toujours se ramener à une série** : Si les données se présentent sous forme d'une distribution par valeurs, convertir la distribution en série.

ii) **Classer la série** : Une fois la série constituée, l'ordonner en classant les chiffres par ordre croissant.

iii) **Déterminer le quotient et le reste de la division de n par 4** : Soit n le nombre d'éléments de la série et p le quotient de la division de n par 4. Quatre cas peuvent se présenter, suivant les quatre valeurs possibles du reste de la division de n par 4. On peut en effet avoir $n=4p$ (pas de reste) ; $n=4p+1$ (reste 1); $n=4p+2$ (reste 2); $n=4p+3$ (reste 3). Envisageons successivement ces quatre cas.

a) Cas où $n = 4p$

C'est le cas où, quand on divise n par 4, on trouve p et que le reste est nul. Dans ce cas, on a :

Q_1 = moyenne entre la p^{e} et $(p+1)^{\text{e}}$ valeur.

$Q_2 = Me$ = moyenne entre la $(2p)^{\text{e}}$ valeur et la $(2p+1)^{\text{e}}$ valeur.

Q_3 = moyenne entre la $(3p)^{\text{e}}$ valeur et la $(3p+1)^{\text{e}}$ valeur

Exemple : si l'on prend la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a $n = 4p = 20 \Leftrightarrow p=5$.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
	1	4	
(p) ^{ème} valeur	2	5	p
(p+1) ^{ème} valeur	2	6	p+1
	2	7	
	2	8	
	2	9	
(2p) ^{ème} valeur	3	10	2p
(2p+1) ^{ème} valeur	3	11	2p+1
	3	12	
	3	13	
	3	14	
(3p) ^{ème} valeur	4	15	3p
(3p+1) ^{ème} valeur	4	16	3p+1
	4	17	
	4	18	
	5	19	
	5	20	

[Fichier EXCEL](#)

Par conséquent, on a :

$$Q_1 = \text{moyenne entre la } p^{\text{e}} \text{ et la } (p+1)^{\text{e}} \text{ valeur} = (2+2)/2=2$$

$$Q_2 = \text{Me} = \text{moyenne entre la } (2p)^{\text{e}} \text{ valeur et la } (2p+1)^{\text{e}} \text{ valeur} = (3+3)/2=3$$

$$Q_3 = \text{moyenne entre la } (3p)^{\text{e}} \text{ valeur et la } (3p+1)^{\text{e}} \text{ valeur} = (4+4)/2=4$$

Les 4 groupes de valeurs sont : $\{0, 1, 1, 1, 2\}$, $\{2, 2, 2, 2, 3\}$, $\{3, 3, 3, 3, 4\}$, $\{4, 4, 4, 5, 5\}$

b) Cas où $n = 4p + 1$

Dans ce cas, le reste de la division par 4 est 1 et l'on a :

$$Q_1 = \text{moyenne entre la } p^{\text{e}} \text{ et la } (p+1)^{\text{e}} \text{ valeur.}$$

$$Q_2 = (2p+1)^{\text{e}} \text{ valeur.}$$

$$Q_3 = \text{moyenne entre la } (3p+1)^{\text{e}} \text{ valeur et la } (3p+2)^{\text{e}} \text{ valeur.}$$

Exemple : si l'on prend la série $\{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3\}$ on a $n=17$ et $n = 4p + 1$, avec $p = 4$.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
(p) ^{ème} valeur →	1	4	← p
(p+1) ^{ème} valeur →	2	5	← p+1
	2	6	
	2	7	
(2p) ^{ème} valeur →	2	8	← 2p
(2p+1) ^{ème} valeur →	2	9	← 2p+1
	3	10	
	3	11	
(3p) ^{ème} valeur →	3	12	← 3p
(3p+1) ^{ème} valeur →	3	13	← 3p+1
(3p+2) ^{ème} valeur →	4	14	← 3p+2
	4	15	
	4	16	
	5	17	

[Fichier EXCEL](#)

Par conséquent, on a :

$$Q_1 = \text{moyenne entre la } p^{\text{e}} \text{ et la } (p+1)^{\text{e}} \text{ valeur} = (1+2)/2=1,5$$

$$Q_2 = (2p+1)^{\text{e}} \text{ valeur} = 2$$

$$Q_3 = \text{moyenne entre la } (3p+1)^{\text{e}} \text{ valeur et la } (3p+2)^{\text{e}} \text{ valeur} = (3+4)/2=3,5$$

Les 4 groupes de valeurs sont : $\{0, 1, 1, 1\}$, $\{2, 2, 2, 2\}$, $\underline{2}$, $\{3, 3, 3, 3\}$, $\{4, 4, 4, 5\}$

On a exclu $Q_2=Me$ pour obtenir 4 groupes égaux.

c) Cas où $n = 4p + 2$

Dans ce cas, le reste de la division par 4 est 2 et l'on a :

$$Q_1 = (p+1)^{\text{e}} \text{ valeur.}$$

$$Q_2 = \text{moyenne entre la } (2p+1)^{\text{e}} \text{ valeur et la } (2p+2)^{\text{e}} \text{ valeur.}$$

$$Q_3 = (3p+2)^{\text{e}} \text{ valeur}$$

Exemple : si l'on prend la série $\{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3\}$, on a $n=18$ et $n = 4p+2$, avec $p = 4$.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
(p)ème valeur →	1	4	← p
(p+1)ème valeur →	2	5	← p+1
	2	6	
	2	7	
(2p)ème valeur →	2	8	← 2p
(2p+1)ème valeur →	2	9	← 2p+1
(2p+2)ème valeur →	3	10	← 2p+2
	3	11	
(3p)ème valeur →	3	12	← 3p
(3p+1)ème valeur →	3	13	← 3p+1
(3p+2)ème valeur →	3	14	← 3p+2
	4	15	
	4	16	
	4	17	
	5	18	

[Fichier EXCEL](#)

Par conséquent, on a :

$$Q_1 = (p+1)^{\text{e}} \text{ valeur} = 2$$

$$Q_2 = \text{moyenne entre la } (2p+1)^{\text{e}} \text{ valeur et la } (2p+2)^{\text{e}} \text{ valeur} = (2+3)/2=2,5$$

$$Q_3 = (3p+2)^{\text{e}} \text{ valeur} = 3$$

Les 4 groupes de valeurs sont : $\{0, 1, 1, 1\}$, $\underline{2}$, $\{2, 2, 2, 2\}$, $\{3, 3, 3, 3\}$, $\underline{3}$, $\{4, 4, 4, 5\}$

On a exclu Q_1 et Q_3 pour obtenir 4 groupes égaux.

d) Cas où $n = 4p + 3$

Dans ce cas, le reste de la division par 4 est 3 et l'on a :

$$Q_1 = (p+1)^{\text{e}} \text{ valeur.}$$

$$Q_2 = (2p+2)^{\text{e}} \text{ valeur.}$$

$$Q_3 = (3p+3)^{\text{e}} \text{ valeur.}$$

Exemple : si l'on prend la série $\{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4\}$ on a $n=19$ et $n = 4p+3$, avec $p= 4$.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
(p)ème valeur →	1	4	← p
(p+1)ème valeur →	2	5	← p+1
	2	6	
	2	7	
(2p)ème valeur →	2	8	← 2p
(2p+1)ème valeur →	2	9	← 2p+1
(2p+2)ème valeur →	3	10	← 2p+2
	3	11	
(3p)ème valeur →	3	12	← 3p
(3p+1)ème valeur →	3	13	← 3p+1
(3p+2)ème valeur →	3	14	← 3p+2
(3p+3)ème valeur →	4	15	← 3p+3
	4	16	
	4	17	
	4	18	
	5	19	

[Fichier EXCEL](#)

Par conséquent, on a :

$$Q_1 = (p+1)^{\text{e}} \text{ valeur} = 2$$

$$Q_2 = (2p+2)^{\text{e}} \text{ valeur} = 3$$

$$Q_3 = (3p+3)^{\text{e}} \text{ valeur} = 4$$

Les 4 groupes de valeurs sont : $\{0, 1, 1, 1\}$, 2, $\{2, 2, 2, 2\}$, 3, $\{3, 3, 3, 3\}$, 4, $\{4, 4, 4, 5\}$

On a exclu Q_1 , Q_2 et Q_3 pour obtenir 4 groupes égaux.

Formules de détermination des 3 quartiles (Q₁, Q₂ et Q₃) d'une série statistique

n= nombre d'unités statistiques de la série

p = quotient de la division de n par 4

$\left[\begin{array}{l} Q_1 = \frac{p^{\text{ème}} \text{ valeur} + (p+1)^{\text{ème}} \text{ valeur}}{2} \\ n = 4p \left\{ \begin{array}{l} Q_2 = \frac{(2p)^{\text{ème}} \text{ valeur} + (2p+1)^{\text{ème}} \text{ valeur}}{2} \\ Q_3 = \frac{(3p)^{\text{ème}} \text{ valeur} + (3p+1)^{\text{ème}} \text{ valeur}}{2} \end{array} \right. \end{array} \right.$	$\left. \begin{array}{l} Q_1 = \frac{p^{\text{ème}} \text{ valeur} + (p+1)^{\text{ème}} \text{ valeur}}{2} \\ Q_2 = (2p+1)^{\text{ème}} \text{ valeur} \\ Q_3 = \frac{(3p+1)^{\text{ème}} \text{ valeur} + (3p+2)^{\text{ème}} \text{ valeur}}{2} \end{array} \right\} n = 4p + 1$
$\begin{array}{c} \circ \\ \circ \\ Q_1 = (p+1)^{\text{ème}} \text{ valeur} \\ \circ \\ n = 4p + 2 \left\{ \begin{array}{l} Q_2 = \frac{(2p+1)^{\text{ème}} \text{ valeur} + (2p+2)^{\text{ème}} \text{ valeur}}{2} \\ Q_3 = (3p+2)^{\text{ème}} \text{ valeur} \end{array} \right. \\ \circ \\ \circ \end{array}$	$\left. \begin{array}{l} Q_1 = (p+1)^{\text{ème}} \text{ valeur} \\ Q_2 = (2p+2)^{\text{ème}} \text{ valeur} \\ Q_3 = (3p+3)^{\text{ème}} \text{ valeur} \end{array} \right\} n = 4p + 3$

Tableau récapitulatif

[Fichier EXCEL](#)

Annexe 2 : Formule de décomposition de la moyenne

Soit une population totale de n unités statistiques, composée de k groupes. Pour simplifier, les groupes sont désignés par des lettres : A, B, C, etc. À partir du 27^{ème} groupe, il faut deux lettres, puis trois, etc., mais peu importe en pratique, car en général le nombre de groupes est limité à quelques uns, comme dans l'exemple de ce chapitre où il n'y a que 4 groupes A, B, C et D.

La population totale est donc égale à la somme des populations des groupes, ce qui peut s'écrire :

$$n = n_A + n_B + \dots + n_k$$

La moyenne arithmétique d'une variable X s'écrit pour le groupe A :

$$\bar{x}_m = \frac{1}{n_m} (x_1^m + x_2^m + \dots + x_{n_m}^m)$$

Et la moyenne générale de la variable X dans la population s'écrit (en remplaçant A par B et C respectivement) :

$$\begin{aligned} \bar{x} = \frac{1}{n} & \left[(x_1^A + x_2^A + \dots + x_{n_A}^A) + (x_1^B + x_2^B + \dots + x_{n_B}^B) + \dots \right. \\ & \left. + (x_1^m + x_2^m + \dots + x_{n_m}^m) + \dots + (x_1^k + x_2^k + \dots + x_{n_k}^k) \right] \end{aligned}$$

Ce qui peut se réécrire :

$$\begin{aligned} \bar{x} = \frac{n_A}{n} \frac{(x_1^A + x_2^A + \dots + x_{n_A}^A)}{n_A} & + \frac{n_B}{n} \frac{(x_1^B + x_2^B + \dots + x_{n_B}^B)}{n_B} + \dots \\ & + \frac{n_m}{n} \frac{(x_1^m + x_2^m + \dots + x_{n_m}^m)}{n_m} \dots + \frac{n_k}{n} \frac{(x_1^k + x_2^k + \dots + x_{n_k}^k)}{n_k} \end{aligned}$$

D'où la formule :

$$\bar{x} = \frac{n_A}{n} \bar{x}_A + \frac{n_B}{n} \bar{x}_B + \dots + \frac{n_m}{n} \bar{x}_m + \dots + \frac{n_k}{n} \bar{x}_k$$

Exercices corrigés :

TD3 2008 : [Sujets](#) - [Corrigés](#)

TD3 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#)