Statistique de base ou descriptive

1.1 Introduction

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Les premières statistiques correctement élaborées ont été celles des recensements démographiques. C'est une méthode scientifique du traitement des données quantitatives. Elle s'applique à la plupart des disciplines : agronomie, biologie, sociologie, géologie, météorologie, sismologie… Le vocabulaire statistique est essentiellement celui de démographie.
Les ensembles étudiés sont appelés **population**. Les éléments de la population sont appelés **individus** ou **unités statistiques**. La population est étudiée selon un ou plusieurs **caractères.**

1.2 L'échantillonnage

Représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée. Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, **l'échantillon doit être représentatif** de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'**échantillonnage aléatoire** assure la représentativité de l'échantillon.

1.2.1 Echantillonnage aléatoire

Un échantillon est qualifié d'**aléatoire** lorsque chaque individu de la population a une **probabilité connue et non nulle** d'appartenir à l'échantillon. Le cas particulier le plus connu est celui qui affecte à chaque individu **la même probabilité** d'appartenir à l'échantillon.

1.2.2 Echantillonnage aléatoire simple

L'échantillonnage aléatoire simple est une méthode qui consiste à prélever **au hasard** et de **façon indépendante**, n individus ou unités d'échantillonnage d'une population à N individus. Chaque individu possède ainsi **la même probabilité** de faire partie d'un échantillon de n individus et chacun des échantillons possibles de taille n possède la même probabilité d'être constitué.
Les statistiques descriptives peuvent se résumer par le schéma suivant :

1.3 Vocabulaire statistique

Il faut savoir d'abord que l'étude statistique porte sur une population, donc on définit le vocabulaire ainsi les paramètres caractérisant une série statistique.

- Les ensembles sont appelés ***populations***. Comme un ensemble, une population statistique doit être clairement définie.
- Les éléments de la population sont appelés ***individus*** ou ***unités statistique***s, (que ce soient des hommes ou des automobiles…).
- La population est étudiée selon un ou plusieurs ***caractères ou variables.***

- **Un caractère** permet de déterminer une partition de la population selon diverses ***modalités***. Ainsi le sol est un *caractère* à deux ou plusieurs modalités : pulvérulent et cohérent, organique. Le sexe est un caractère à deux modalités : masculin ou féminin. Le caractère désigne une grandeur ou un attribut (propriété), observable sur un individu et susceptible de varier prenant ainsi différents états appelés **modalités**.
- On appelle **modalité** toute valeur

On appelle **modalité** toute valeur : $x_i \in X(P) = \{x_1, \quad x_2, \quad x_3, \cdots \quad x_i, \cdots \quad x_k\}$

Avec, k nombre de modalités différentes de X.

- **Variable (ou caractère) qualitative** : La variable est dite **qualitative** quand les modalités sont des **catégories** et ne sont pas mesurables (pour les automobiles la couleur est caractère ou variable qualitatif, profession, couleurs des yeux…).
- *Variable qualitative nominale* : La variable est dite qualitative ***nominale*** quand les modalités ne peuvent pas être ordonnées.
- *Variable qualitative ordinale* : La variable est dite qualitative ***ordinale*** quand les modalités peuvent être ordonnées.

Exemple :

Dans les catégories socioprofessionnelles, on admet d'ordonner les modalités : 'ouvriers', 'employés', 'cadres'. Si on ajoute les modalités 'sans profession', 'enseignant', 'artisan', l'ordre devient beaucoup plus discutable.

- **Variable (ou caractère) quantitative** : Une variable est dite **quantitative** si toute ses valeurs possibles sont **numériques**.
- *Variable quantitative discrète* : Une variable est dite **discrète**, si l'ensemble des valeurs possibles est dénombrable (si elle ne prend que des valeurs isolées, appartenant à un certain intervalle : nombre d'enfants dans une famille).
- *Variable quantitative continue* : Une variable est dite **continue**, si l'ensemble des valeurs possibles est continu.

1.4 La collecte de l'information

Le premier objet de la méthode statistique est de réunir les informations avant de les traiter. Les données sont recueillies soit par observation directe, soit indirectement.

- ❖ *Observation directe* : enquête menée par les statisticiens à l'aide de questionnaires qui sont ensuite dépouillés.
- ❖ *Observation indirecte :* statistiques d'une entreprise tirées de sa comptabilité, statistiques de naissances et des décès tirées de l'état civil ….

1.4.1 Différents types de collecte de l'information

- ❖ Les résultats statistiques peuvent être obtenus à partir d'une enquête exhaustive (approfondie) instantanée (dénombrement instantané ou recensement) ou d'un relevé continu (état civil).

- De même, l'enquête peut être ***exhaustive*** ou ***partielle***. L'enquête exhaustive porte sur toutes ***les unités*** de la population ; elle est utile mais souvent coûteuse. C'est pourquoi on a recours à des enquêtes partielles faites sur **un échantillon** de la population : il s'agit alors de **sondage** et il faut déterminer **un échantillon représentatif**, de manière que les résultats statistiques trouvés sur cet échantillon soient voisins de ceux que l'on aurait obtenus si on avait étudié la population entière.

1.5 Dépouillement des observations

Lorsque les observations sont obtenues, elles doivent être **classées** et **exploitées**. Pour chaque **caractère** à étudier, on définit un certain nombre **de classes** selon les **modalités**, puis on fait le tri des observations, c'est à dire une **répartition par classes**. Ces opérations peuvent être faites à la main ou à l'aide d'un ordinateur.

1.6 Tableaux statistiques

On peut représenter les données brutes d'une étude dans un tableau. Mais il est possible d'en déduire un tableau plus clair, en faisant **un regroupement par classes**. On choisit les classes pas trop nombreuses, mais suffisamment pour qu'il n'y ait pas de perte d'information. Il importe que les classes recouvrent tous les résultats et aient une intersection vide, d'où les formulations du type « de … à moins de …» ; la différence entre les deux extrémités est appelé **amplitude de la classe**.

On peut fixer le nombre de classes selon l'un des deux formules suivantes :

1) ***Règle de Sturge*** : $nombre.de\ classes, k = 1 + 3.22\ log_{10}\ n$
2) ***Règle de Yule :*** $nombre.de\ classes, k = 2.5\sqrt[4]{n}$

Avec n:effectif de l'échantillon

- **L'amplitude de la classe** (A) est donnée par l'expression :$A = \frac{Valeur\ max - valeur\ min}{nombre\ de\ classes}$
- **L'effectif d'une classe** est le nombre d'éléments de la population observés dans cette classe. On note $k$le nombre de valeurs distinctes ou modalités. Les valeurs distinctes sont notées $x_1,\ \ x_2,\ \ x_3, \cdots\ \ x_i, \cdots\ \ x_k$. On appelle ***effectif*** d'une modalité ou d'une valeur distincte, le nombre de fois que cette modalité (ou valeur distincte) apparaît. On note $\boldsymbol{n_i}$ l'effectif de la modalité $\boldsymbol{x_i}$
- **La fréquence de la modalité** $\boldsymbol{x_i}$ est le rapport de cet effectif à l'effectif total $\boldsymbol{N}$ de la population. La fréquence est exprimée en pourcentage. $f_i = \frac{n_i}{N}, ;\ \ i = 1,2 \ldots\ldots, k$

Application 1

On s'intéresse au variable ''état civil'' notée X relative à $20\ personnes$. La codification est

C	célibataire
M	Marié (e)
V	Veuf (ve)
D	divorcée

Le domaine de la variable X est $\{C, \quad M, \quad V, \quad D\}$. Considérons le résultat de dépouillement des observations suivant :

M	M	D	C	C	M	C	C	C	M
C	M	V	M	V	D	C	C	C	M

Ici $N = 20$

$$x_1 = M, x_2 = M, x_3 = D, x_4 = C, x_5 = C, \cdots\cdots, x_{20} = M$$

Le tableau statistique ci-dessous résume les résultats de dépouillement :

x_i	n_i	f_i
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10
Total	$N = 20$	$\sum f_i$ =1.00

Application 2

On s'intéresse à la charge de rupture d'un élément métallique en kilogrammes

711	862	851	912	922	791	825	935	895	758	8462
915	876	926	864	800	931	722	774	903	925	8633
853	700	885	857	844	907	917	786	820	930	8499
789	790	753	910	847	784	936	706	758	887	8160
941	909	784	882	859	903	925	704	792	888	8587
890	925	895	768	869	892	895	912	850	920	8816
763	805	795	759	916	853	789	942	712	764	8099
892	893	915	890	888	865	909	931	710	798	8691
914	794	931	701	772	935	887	880	933	905	8652
889	791	782	713	724	868	842	892	905	792	8198
										84797

On va regrouper les données en classes. Nous avons un effectif de $N = 100$ unités d'observations, ce qui nous donne d'après les règles de Sturge et Yule :

$$k = 1 + 3.22 \times \log_{10} 100 = 7.44$$

$$k = 2.5 \times \sqrt[4]{100} = 7.9$$

Dans cet exemple on prend 6 classes.

Charge (kg)	effectifs	fréquence
700 à moins de 750	10	0.1
750 à moins de 800	23	0.23
800 à moins de 840	4	0.04
840 à moins de 880	15	0.15
880 à moins de 920	32	0.32
920 et plus	16	0.16
Total	**100**	**1**

Représentation graphique des résultats

1. Cas de distributions quantitatives

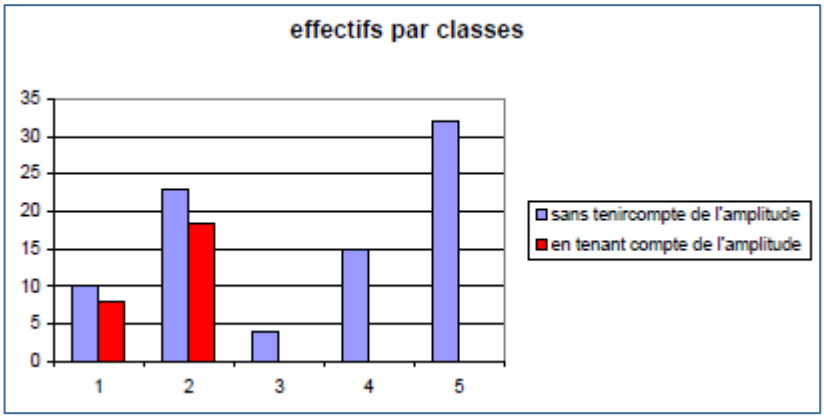
Les graphiques sont normalement réalisés en portant en abscisse la variable observée et en ordonnée l'effectif ou la fréquence.

- Dans le cas d'une variable ***discrète***, le graphique est un diagramme à bâtons, ainsi apparaît la discontinuité entre deux valeurs.
- Dans le cas d'une variable continue, le graphique est un histogramme. la surface limitée par l'histogramme doit être proportionnelle à l'effectif ou la fréquence. Il convient de prendre garde à l'amplitude des classes (on se ramène à la plus petite amplitude, amplitude élémentaire et on divise la hauteur du rectangle par la mesure de l'amplitude de la classe par rapport à cette amplitude élémentaire).

$$\boldsymbol{hauteur\ de\ rectangle, h = \frac{effectif\ ou\ fréquence \times amplitude\ élémentaire}{amplitude\ de\ la\ classe}}$$

De notre exemple on peut récolter les données suivantes :

Charge en kg (classe)	Effectifs	Amplitude	Hauteur de rectangle
700 à moins de 750	10	50	$h = (10 \times 40)/50 = 8$
750 à moins de 800	23	50	$h = (23 \times 40)/50 = 18.4$
800-840	4	40	$h = (4 \times 40)/40 = 4$
840-880	15	40	$h = (15 \times 40)/40 = 15$
880-920	32	40	$h = (32 \times 40)/40 = 32$



1.7 ANNEXE

Creating a histogram on EXCEL 2013

1. Start up Excel
2. Title the **A1** and **B2** column **Class Boundaries** and **Frequency** accordingly.

Histograms most often deal with intervals and frequency. On the horizontal *X*-axis will be the intervals data, which may also be called groups, segments, or bins. This is the grouped data. Frequency is on the vertical y-axis.

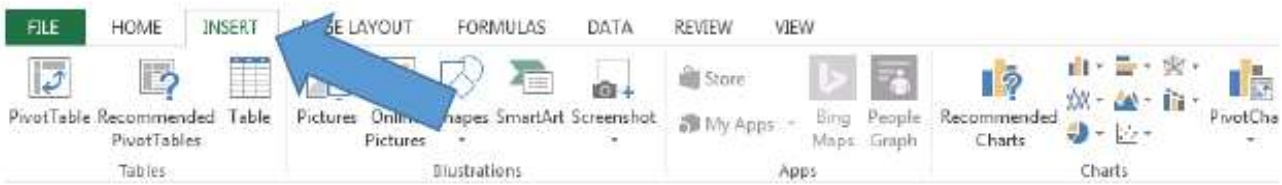
3. In the **Class Boundaries** and **Frequency** columns, input your data.

	A	B	C
1	Boundaries	Frequency	
2	99.5-104.5	2	
3	104.5-109.5	8	
4	109.5-114.5	18	
5	114.5-119.5	13	
6	119.5-124.5	7	
7	124.5-129.5	1	
8	129.5-134.5	1	
9			
10			
11			

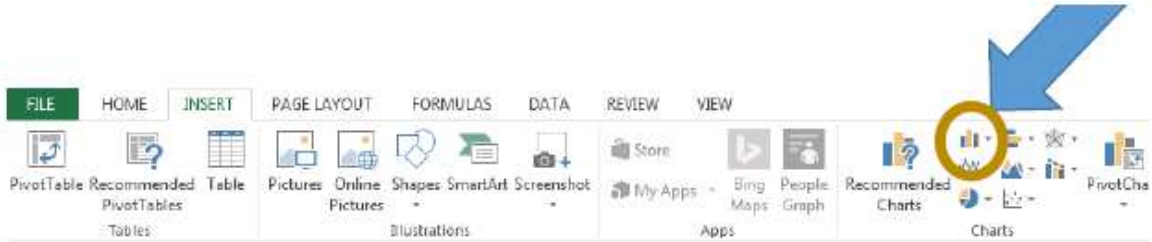
4. Once you have your raw data into Excel, select your data. In our example, We have selected cells A1 through B8.

	A	B	C
1	Boundaries	Frequency	
2	99.5-104.5	2	
3	104.5-109.5	8	
4	109.5-114.5	18	
5	114.5-119.5	13	
6	119.5-124.5	7	
7	124.5-129.5	1	
8	129.5-134.5	1	
9			
10			
11			
12			

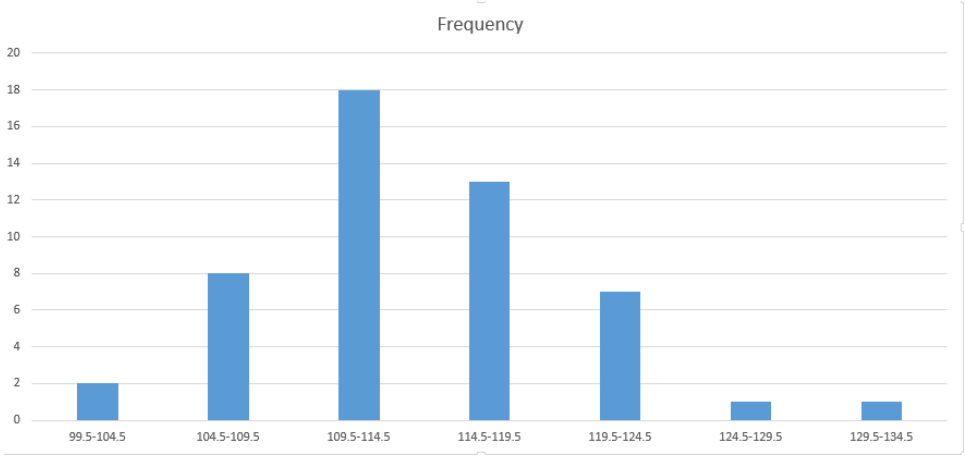
5. Select **INSERT** from the top toolbar.



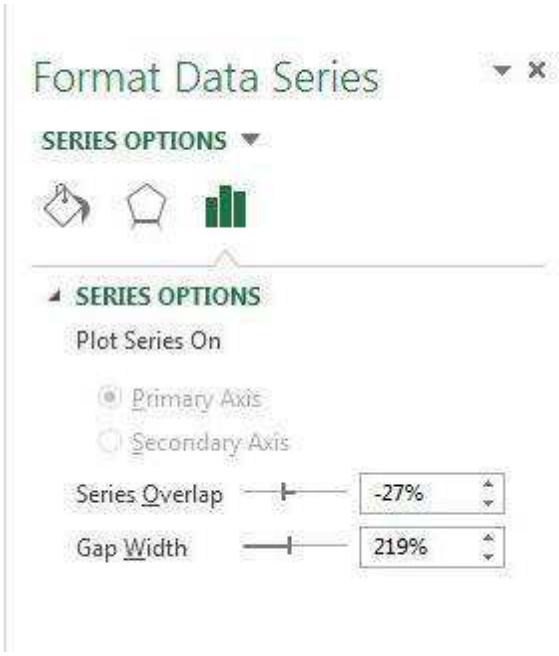
6. Click on **Insert Column Chart**, and select **Clustered Column**, from the **2-D Column Section**



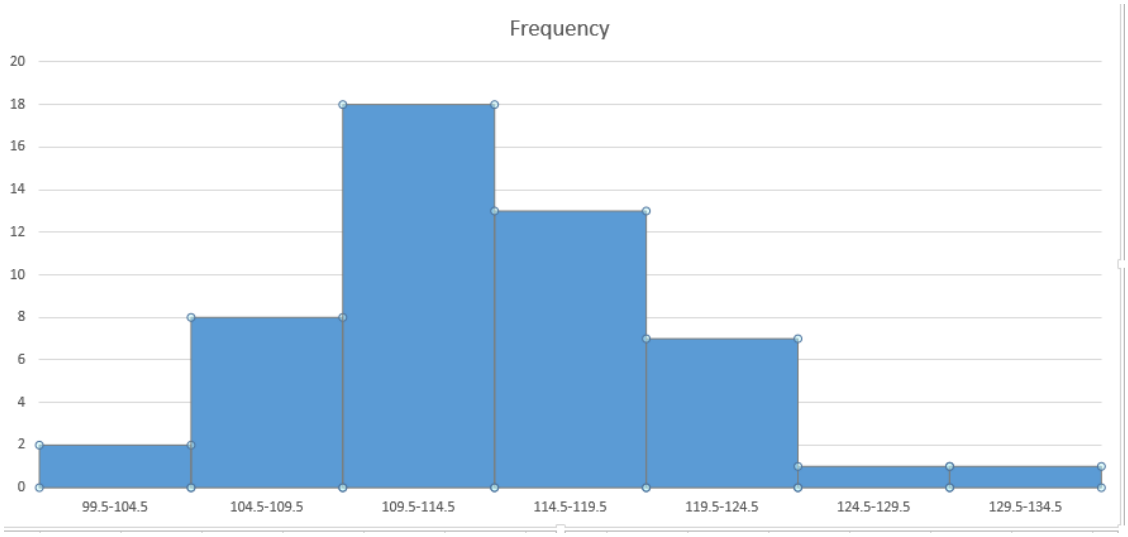
7. By now, you should have something that looks like this. **Right Click** on one of the **Bars** and select **Format Data Series**… from the drop down menu.



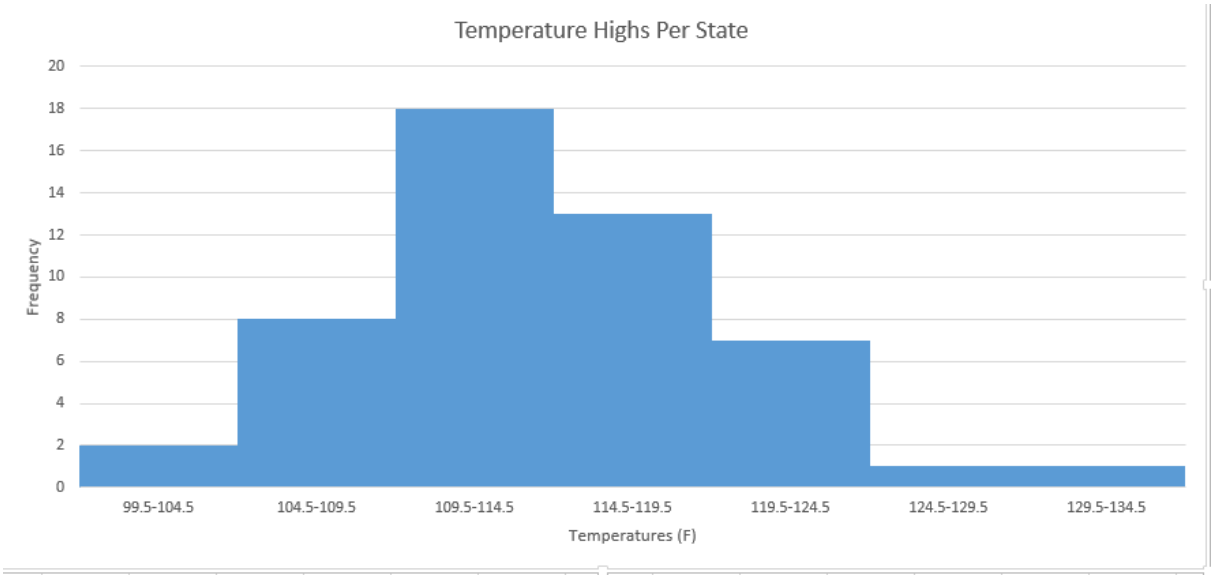
8. Set the **Gap Width** to **0%**



9. By now, you should be looking at something like this.



10. At the top toolbar select **Add Chart Elements** and add the axis labels and title for your graph.



Etude des series statistiques univariées (simples)

2.1 Introduction

Un tableau statistique ou un graphique sont parfois long à consulter, sans permettre d'avoir une idée suffisamment concise de la distribution statistique observée. Par définition, on appelle **série statistique** la donnée simultanée (dans un tableau) des **valeurs du caractère** étudié (noté $\boldsymbol{x_i}$), rangées dans l'**ordre croissan**t, et des effectifs (notés $\boldsymbol{n_i}$) de ces valeurs.

2.2 Série statistique associée à un caractère quantitatif discret

Supposons un échantillon composé de $\boldsymbol{n}$ ***éléments*** , numérotés de $\mathbf{1}$ à $\boldsymbol{n}$**.** Appelons $\boldsymbol{X}$ la valeur du caractère sur lequel porte l'étude. Soient $\boldsymbol{x_1, x_2, \cdots x_n}$, les valeurs de ce caractère ou **modalités** pour les éléments $\mathbf{1, 2, 3, \cdots, n}$ de la série. **L'étendue** de la série est l'écart qui sépare la plus grande et la plus petite valeur du caractère. Donc une série statistique associée à une variable quantitatif est l'ensemble des couples (paires) $\{(x_i, n_i)\}$, où x_i est la variable statistique et n_i désigne l'effectif associé à cette variable statistique.

L'effectif total de la série est nombre $\boldsymbol{n}$ d'éléments constituant l'échantillon étudié. Lorsque la valeur $\boldsymbol{x_i}$ du caractère apparue $\boldsymbol{n_i}$ fois dans la série statistique, on dit que $\boldsymbol{n_i}$ est la répétition de $\boldsymbol{x_i}$ ou l'effectif partiel relatif à $\boldsymbol{x_i}$ ou encore la fréquence absolue de $\boldsymbol{x_i}$**.** On la note comme suite ;

$$f_i = \frac{n_i}{N}$$

Donc la fréquence f_i d'une modalité i est tout simplement un rapport entre l'effectif correspondant n_i sur l'effectif total N

N: Effectif total de l'échantillon.

Exemple 2.1

Répartition de 150 éléments de structure suivant le degré de gravité de fissuration

Degré de gravité	0	1	2	3	4	5	6	
Nombre d'élément	11	22	45	40	19	11	2	150
Fréquence relative	0.07	0.14	0.3	0.266	0.126	0.07	0.01	$\cong 1.0$
Fréquence en% $f_i = \frac{n_i}{150} \times 100$	7.33	14.66	30	26.66	12.66	7.33	1.33	100

- **Effectif cumulé croissant** d'une valeur $\boldsymbol{x}$ est la somme des effectifs des valeurs $\boldsymbol{y}$ tels que $\boldsymbol{y \leq x}$
- **Effectif cumulé décroissant** d'une valeur $\boldsymbol{x}$ est la somme des effectifs des valeurs $\boldsymbol{y}$ tels que $\boldsymbol{y > x}$

Exemple 2.2

Les notes sur 20 obtenues lors d'un devoir dans une classe sont les suivantes

$$10, 8, 11, 9, 12, 10, 8, 10, 7, 9, 10, 11, 12, 10, 8, 9, 10, 9, 10, 11$$

- La **population** étudiée est la **classe** et les **individus** sont les **élèves**. L'effectif total $N = 20$ et **la note** obtenue en devoir est **le caractère discret** que l'on étudie.

- **La série statistique** définie par les effectifs est la suivante :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Effectifs (nombre d'élèves ayant la note) n_i	1	3	4	7	3	2

- **La série statistique** définie par les fréquences en pourcentage est la suivante :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Fréquences en % : $f_i = \frac{n_i}{20} \times 100\ n_i$	5%	15%	20%	35%	15%	10%

- **Les effectifs cumulés** sont les suivants :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Effectif cumulé croissant	1	4*	8	15	18	20
Effectif cumulé décroissant	19	16**	12	5	2	0

* : Nombre d'élèves ayant eu une note ≤ 8

** : Nombre d'élèves ayant eu une note > 8

2.3 Série statistique associée à un caractère quantitatif continu

Dans le cas d'un **caractère continu**, le nombre des valeurs discrètes est en principe infini. Pour éviter une répétition des fréquences très dispersée, on constitue des classes en divisant l'étendue de la série en un certain nombre d'intervalles partiels.

Définir une classe revient donc à fixer ses **limites**, le centre de la classe et **l'étendue** de la classe. Les **classes sont contiguës** et ne se chevauchent pas. Les classes sont d'**étendue égale** ou **inégale**.

Exemple 2.3

Poids des échantillons des sondages carottés, échelonnés entre $2.240kg\ et\ 4.490kg$

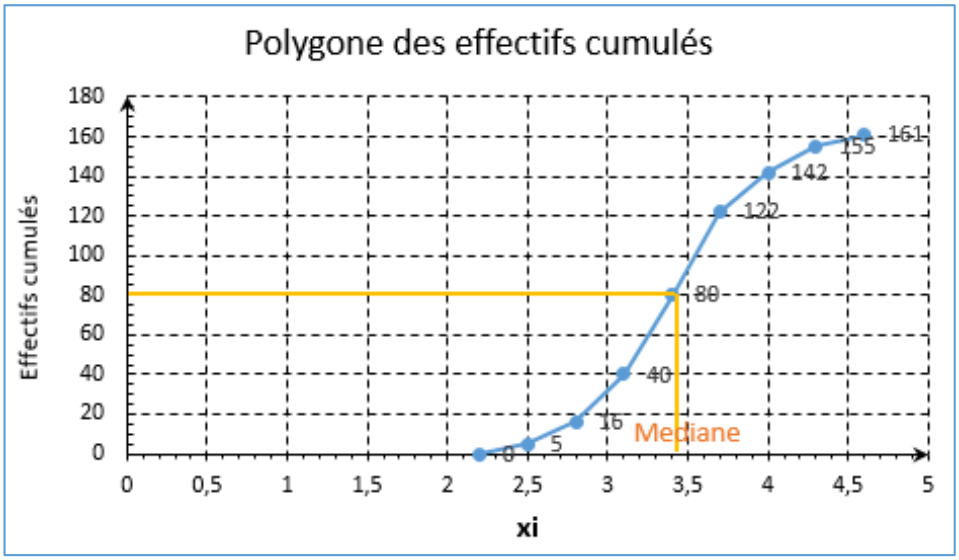
Classe	Limite de classe (kg)	Centre de classe	effectif	Fréquence relative f_i	(%)
1	2.2-2.5	2.350	5	0.031	3.1
2	2.5-2.8	2.650	11	0.068	6.8
3	2.8-3.1	2.950	24	0.148	14.8
4	3.1-3.4	3.230	40	0.248	24.8
5	3.4-3.7	3.550	42	0.259	25.9
6	3.7-4.0	3.850	20	0.124	12.4
7	4.0-4.3	4.150	13	0.080	8.0
8	4.3-4.6	4.450	6	0.037	3.70
		Total	=160	=1	=100

2.3.1 Polygone des effectifs cumulés

S'obtient en portant en ordonnée, au droit de chaque limite de classe figurant sur l'axe des abscisses, **la somme des effectifs** de toutes les classes inférieures. La ligne brisée joignant les points ainsi obtenue s'appelle le p**olygone des fréquences cumulées**. On obtiendra par la même manière le polygone des fréquences relatives cumulées si on porte en ordonnée la somme des fréquences relatives cumulées.

De l'exemple précédent on peut écrire :

Intervalle de classe	effectif	Effectif cumulé
2.2-2.5	5	5
2.5-2.8	11	16
2.8-3.1	24	40
3.1-3.4	40	80
3.4-3.7	42	122
3.7-4.0	20	142
4.0-4.3	13	155
4.3-4.6	6	161



2.4 Série statistique associée à un caractère qualitatif discret

Pour représenter les résultats d'une enquête relative à un caractère qualitatif. On regroupe les résultats en un nombre **de classe égal au nombre de modalités** du caractère étudié. A chaque classe est associé son effectif $\boldsymbol{n_i}$ ou sa **fréquence absolue**, ainsi que sa **fréquence relative :** $f_i = \frac{n_i}{N}$

Exemple 2.4 : L'analyse du sang de 100 individus a donné les résultats suivants :

Groupe sanguin	Fréquence absolue	Fréquence relative	Pourcentage
O	40	0.4	40%
A	43	0.43	43%
B	12	0.12	12%
AB	5	0.05	5%

2.4 Généralité sur les paramètres de position et de dispersion

Les paramètres de position et de dispersion sont un ensemble de valeurs caractéristiques qui permettent une représentation condensée de l'information contenue dans une série statistique. On distingue deux catégories de ces paramètres :

- les paramètres de position : la moyenne, la médiane, le mode et quantiles donnent l'ordre de grandeur de l'ensemble des mesures ;
- les paramètres de dispersion : variance, écart moyen, écart type, semi-interquartile, précisent le degré de dispersion des différentes valeurs d'une série autour d'une valeur centrale.

2.4.1 Paramètres de position

2.4.1.1 Moyenne arithmétique

Soit l'ensemble des mesures d'une même variable X : $(x_1, x_2, \cdots, x_n)$, la moyenne arithmétique notée $\boldsymbol{\mu}$ est définie par :

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemple 2.5

La moyenne arithmétique des valeurs 8, 5, 3, 6, 2 est :

$$\mu = \frac{8 + 5 + 3 + 6 + 2}{5} = \frac{24}{5} = 4.8$$

Lorsque les valeurs $x_1, x_2, \cdots, x_n$ se répètent respectivement $1, 2, 3 \cdots, n\ fois$ on obtient la moyenne arithmétique, en comptant chaque valeur x_i autant de fois qu'elle se présente. Ceci revient à pondérer la valeur x_i par l'effectif n_i qui lui correspond, on aura :

$$\mu = \frac{\sum_{i=1}^{k} n_i x_i}{N}$$

N: Effectif total de l'échantillon

Exemple 2.6

Si es valeurs 8, 5, 3, 6 et 2 se reproduisent respectivement 1, 4, 2, 2,1 fois, la moyenne arithmétique respectivement est :

$$\mu = \frac{1 \times 8 + 4 \times 5 + 2 \times 3 + 2 \times 6 + 1 \times 2}{10} = 4.8$$

$N = 1 + 4 + 2 + 2 + 1 = 10$

Exemple 2.7

Soit la distribution de la série statistique associée à un caractère continu :

Classe	8-10	10-12	12-14	14-16	16-18	18-20
n_i	1	2	4	6	5	2
$centre: x_i$	9	11	13	15	17	19
$n_i \times x_i$	9	22	52	90	85	38

$$\mu = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{9 + 22 + 52 + 90 + 85 + 38}{20}$$

2.4.1.3 Médiane

Est la valeur qui partage la série statistique (l'échantillon) en deux groupes de même effectif. Pour la calculer, il faut commencer par ordonner les valeurs en ordre de grandeurs croissantes ou décroissantes. **La médiane** se situé au centre de la série ainsi ordonnée, on la désigne par $m_{1/2}$

1.) Médiane d'une série statistique associée à un caractère discret

- Si la série possède un nombre **impair** de valeurs soit $2n+1$, **la médiane** sera la $(n+1)^{ième}$ valeur

$$\boldsymbol{M_{1/2} = x_{\left(\frac{n+1}{2}\right)}}$$

- Si la série compte un nombre pair de valeurs soit $2n$ valeurs, la médiane sera la moitié de la somme de la $(n)^{ième}$ et la $(n+1)^{ième}$ valeurs,

$$\boldsymbol{M_{1/2} = \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\}}$$

Exemple 2.8

Soit la série statistique d'un caractère discret et de nombre impair de valeurs 12, 3, 24, 1, 5, 8, 7

On l'ordonne en ordre croissant : 1, 3, 5, 7, 8, 12, 24.

7 et la médiane de la série.

Exemple 2.9

La médiane de la série suivante, comportant un nombre pair de valeurs : 4, 5, 8, 8, (9, 11), 12, 14, 17, 19. La médiane $M_{1/2} = \frac{9+11}{2} = \mathbf{10}$

On l'ordonne en ordre décroissant : 19, 17, 14, 12, (11, 9), 8, 8, 5, 4. La médiane

$$M_{1/2} = \frac{11+9}{2} = \mathbf{10}$$

Nota

La médiane n'a pas toujours de signification au point de vue statistique dans le cas des séries d'un caractère discret. C'est en particulier le cas lorsque plusieurs valeurs du caractère coïncident avec la valeur de la médiane.

Exemple 2.10

Dans le cas de la série suivante : 2, 2, 2, 3, 3, 3, 4, 4, 4, (5, 5), 5, 6, 6, 7, 7, 7, 8, 8, 9. La médiane est comprise entre la $10^{ième}$ et la $11^{ième}$ valeur, c.à.d. $M_{1/2} = \frac{5+5}{2} = \mathbf{5}$

La répétition de la valeur 5 ne permet d'autre conclusion que celle-ci : 9 valeurs sont inférieures à la médiane et 8 valeurs lui sont supérieures : **la médiane ne réalise pas d'équipartition effective de la série**.

Exemple 2.11

On trie la série statistique par ordre croissant des valeurs observées. Avec la série observée :

3, 2, 1, 0, 0, 1, 2

On obtient

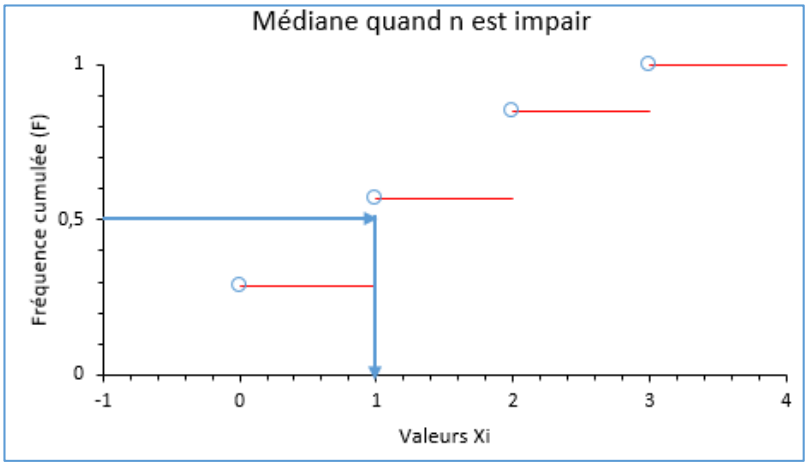
$$0, 0, 1, 1, 2, 2, 3$$

La médiane $m_{1/2}$ est la valeur qui se trouve au milieu de la série ordonnée :

$$0, 0, 1, \mathbf{1}, 2, 2, 3$$

La Figure ci-dessous montre la fonction de répartition de la série. La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur $1/2$.

$$\text{C.à.d. } F(M_{1/2}) = F(1) = \frac{1}{2} \Longrightarrow M_{1/2} = F^{-1}(1/2)$$



Cette figure montre la fonction de répartition de la série. La médiane peut être définie comme l'inverse de la fonction de répartition pour la valeur $1/2$.

$$M_{1/2} = F^{-1}(1/2)$$

Homework

Refaire le même travail dans le cas où n est pair et égal =8. représenter graphiquement la médiane sur la courbe discontinue de la fonction de répartion.

$$0, 0, 1, 1, 2, 2, 3, 4$$

2.) Médiane d'une série statistique associée à un caractère continu

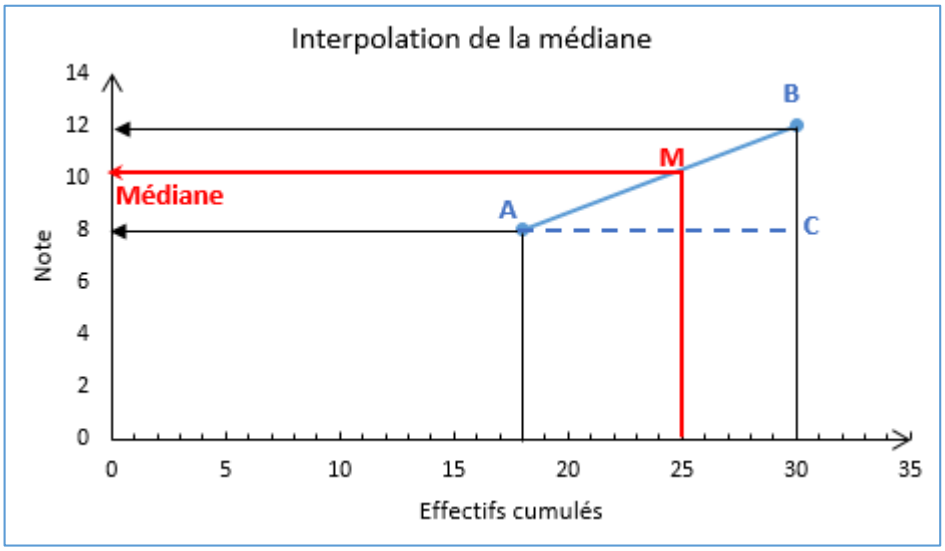
Dans ce cas **la médiane** s'obtient par **interpolation**. Si la variable est continue (regroupement par classe le calcul de la médiane se fait comme suit :

Utilisons la colonne des effectifs cumulés (Tableau ci-dessous) pour déterminer **la médiane** : Il y a 50 notes, 50% de l'effectif total c'est 25, la médiane ici est la note correspondante à l'effectif cumulé 25. D'après la ''effectif cumulé'' :

Note	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18*
[8 ; 12[	12	30*
[12 ; 15[	11	41
[15 ; 20[	9	50
	50	

- 18 personnes ont moins de 8
- 30 personne ont moins de 12

La médiane se trouve donc dans l'intervalle $[8;12[$ (**appelée classe médiane**), on va la déterminer par **interpolation linéaire**.



Les points A, M, B sont alignés ce qui traduit par les droites (AM) et (AB) ont même coefficient directeur ou on utilise le théorème de Thalès dans le ABC.

$$\frac{M_{1/2}-8}{25-18}=\frac{12-8}{30-18}\Longrightarrow M_{1/2}=7\times\frac{4}{12}+8=10.33$$

Exemple 2.12

On considère la série statistique suivante :

Classe	Fréquence absolue	Fréquence cumulée
$[38\,;40[$	11	11
$[40\,;42[$	28	39
$[42\,;44[$	16	55
$[44\,;46[$	25	80
$[46\,;48[$	15	95
$[46\,;48[$	5	100

Dans la classe $[42\,;44[$ se situent les observations comprises entre $39\ et\ 55$ donc en particulier l'observation de rang 50 qui correspond à la médiane $\boldsymbol{m_{1/2}}$

$$M_{1/2}=42+\frac{44-42}{16}\times(50-39)=\mathbf{43.375}$$

Homework

1. Calculer la moyenne et la médiane de la série statistique relative à la charge de rupture de l'élément métallique traité en chapitre I (application 2), page 4.
2. Représenter graphiquement (par histogrmme) la série statistique à variables continues et amplitudes égales suivante :

Classe	Effectifs n_i
[10 ; 20[	2
[20 ; 30[	5
[30 ; 40[	1
[40; 50[	3

3. Représenter graphiquement (par histogrmme) la série statistique à variables continues et amplitudes inégales suivante :

Classe	Effectifs n_i
[5 ; 10[	5
[10 ; 20[	10
[20 ; 25[	5
[25; 40[	15

2.4.1.3 Le Mode

Le mode d'une série statistique est la valeur la plus fréquente (dominante) de cette série. Dans la série 1, 2, 2, 4, 10. Le mode de cette série est, **2**.

- Pour une série statistique associée à un caractère continu (distribution par classes), on définit **la classe modale** : c'est **la classe** dont l'effectif est relativement **élevé** et on attribue au **mode** la **valeur centrale** de cette classe.

Exemple 2.13

Soit la série statistique à variables continues et amplitudes égales illustrée dans le Tableau ci-dessous :

Classes d'âges	Effectifs (n_i)
[10 ; 20[	7
[20 ; 30[	12
[30 ; 40[	3

La classe dont l'effectif le plus important est la classe [20 ; 30[, donc **la classe modale** de cette série est la classe **[20 ; 30[.**

Le mode de cette série est le centre de la classe modale soit :

$$Mo = \frac{20 + 30}{2} = 25$$

- *Cas d'une série statistique à variables continues et à amplitudes inégales*

Dans ce cas, pour calculer la médiane on procède par étape comme suite :

1. on rectifié les effectifs ;
2. on choisit la classe modale dont l'effectif rectifié le plus élevé (important) ;
3. le mode de la série est le centre de cette classe modale.

Soit la série statistique à variables continues et à amplitudes inégales suivante :

Classes d'âges	Amplitude	Effectifs (n_i)	Effectif rectifié
$[0\,;5[$	5	7	7
$[5\,;15[$	10	18	9
$[15\,;20[$	5	12	12
$[20\,;40[$	20	24	6

On prend par exemple l'amplitude des classes la plus répandue qui est 5 (on peut prendre les amplitudes 10 ou 20). On rectifié les effectifs des autres classes comme suit :

- l'amplitude de la classe $[5\,;15[$ est 10, c.à.d. deux fois l'amplitude de la classe $[0\,;5[$, donc on divise son effectif par 2, soit 9
- on fait la même chose pour la classe $[20\,;40[$

Donc la classe qui comporte **l'effectif rectifié** le **plus élevé** est la classe $[15\,;20[$, on dit que la classe $[15\,;20[$ est **la classe modale** de la série et son mode.

$$Mo = \frac{15+20}{2} = 17.50$$

2.4.1.4 Quantiles

La notion de **quantile** d'ordre $\boldsymbol{p}$ où $0 < p < 1$, généralise la médiane. Formellement un quantile est donné par **l'inverse de la fonction de répartition des effectifs F :**

$$Q_p = F^{-1}(p)$$

Si la fonction de répartition des effectifs, était continue et strictement croissante, la définition du quantile serait équivoque. La fonction de répartition est cependant discontinu et ''**par palier**''. Quand la fonction de répartition est par palier, il existe au moins neuf (9) manières différentes de définir les quantiles selon que l'on fasse ou non une interpolation de la fonction de répartition.

- Si $\boldsymbol{np}$ est un nombre entier, alors

$$Q_p = \frac{1}{2}\{x_{(np)} + x_{(np+1)}\}$$

- Si $\boldsymbol{np}$ n'est pas un nombre entier, alors

$$Q_p = x_{([np])}$$

Où $[np]$ représente le plus petite nombre entier supérieur ou égal à $\boldsymbol{np}$ **:** $[np] \geq np$

Remarques

- La médiane est le quantile d'ordre $p = \frac{1}{2}$;
- On utilise souvent :
  - $Q_{1/4}$ Le premier quartile,
  - $Q_{3/4}$ Le troisième quartile,
  - $Q_{1/10}$ Le premier décile,
  - $Q_{1/5}$ Le premier quintile,
  - $Q_{4/5}$ Le quatrième quintile,

$Q_{9/10}$ Le neuvième décile,

$Q_{0.05}$ Le cinquième percentile,

$Q_{0.95}$ Le nonante-cinquième percentile.

Si $F(x)$ est la fonction de répartition, alors $F(x_p) \geq p$.

Exemple 2.14

Soit la série statistique à variables discrètes : $12, 13, 15, 16, 18, 19, 22, 24, 25, 27, 28, 34$ contenant 12 observations $(n = 12)$

- **Le premier quartile** : comme $np = n \times p = 12 \times \frac{1}{4} = 3$ est **un nombre entier**, on a.

$$Q_{1/4} = \frac{1}{2}\{x_{(np)} + x_{(np+1)}\} = \frac{1}{2}\{x_{(3)} + x_{(4)}\} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{15 + 16}{2} = 15.5$$

- **La médiane** : Comme $np = 12 \times \frac{1}{2} = 6$ est nombre entier, on a.

$$Q_{1/2} = \frac{1}{2}\{x_{(np)} + x_{(np+1)}\} = \frac{1}{2}\{x_{(6)} + x_{(7)}\} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{19 + 22}{2} = 20.5$$

- **Le troisième quartile** : Comme $np = 12 \times \frac{3}{4} = 12 \times 0.75 = 9$ est un nombre entier on a.

$$Q_{3/4} = \frac{1}{2}\{x_{(np)} + x_{(np+1)}\} = \frac{1}{2}\{x_{(9)} + x_{(10)}\} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{25 + 27}{2} = 26$$

Exemple 2.15

Soit la série statistique à variables discrètes $12, 13, 15, 16, 18, 19, 22, 24, 25, 27$ contenant 10 observations $(n = 10)$.

- **Le premier quartile** : $np = n \times p = 10 \times \frac{1}{4} = 2.5$ **n'est pas un nombre entier**, on a.

$$Q_{1/4} = x_{([2.5])} = x_{(3)} = 15$$

- **La médiane** : Comme $np = 10 \times \frac{1}{2} = 5$ est nombre entier, on a.

$$Q_{1/2} = M_{1/2} = \frac{1}{2}\{x_{(np)} + x_{(np+1)}\} = \frac{1}{2}\{x_{(5)} + x_{(6)}\} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{18 + 19}{2} = 18.5$$

- **Le troisième quartile** : $np = 10 \times \frac{3}{4} = 10 \times 0.75 = 7.5$ n'est un nombre entier on a.

$$Q_{3/4} = x_{([7.5])} = x_{(8)} = 24$$

Exemple 2.16

Soit la série statistique à variables discrètes : $4, 5, 6, 11, 13, 14, 16$

5 est le ***premier*** *quartile*

11 est la médiane ou le ***deuxième*** *quartile*

14 est le ***troisième*** *quartile*

❖ Pour la série $4, 5, 6, 7, 11, 13, 14, 15, 16$

6 est le ***premier*** *quartile*

11 est la médiane ou le ***deuxième*** *quartile*

14 est le ***troisième*** *quartile*

Le nombre de termes inférieurs à $Q_{1/4}$ est donné par l'expression $np = 9 \times \frac{1}{4} = 2.25$ soit deux (2) termes et $Q_{3/4}$ est symétrique de $Q_{1/4}$ par rapport à la médiane $Q_{1/2}$ *ou* $M_{1/2}$.

2.4.2 Paramètres de dispersion

2.4.2.1 L'étendue

L'étendue est simplement la différence entre la plus grande et la plus petite valeur observée de la série statistique.

$$E = x_{(n)} - x_{(1)}$$

2.4.2.2 La distance interquartile

La distance interquartile est la différence entre le **troisième** et le **premier quartile**

$$IQ = Q_{3/4} - Q_{1/4}$$

2.4.2.3 La variance

La variance est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations (la moyenne des carrés moins le carré de la moyenne) :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Cette expression peut aussi s'écrire.

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Démonstration

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2 \times \frac{1}{n}\sum_{i=1}^{n}\mu x_i + \frac{1}{n}\sum_{i=1}^{n}\mu^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2 \times \frac{1}{n}\sum_{i=1}^{n}\mu x_i + \frac{1}{n}(n \times \mu^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu\frac{1}{n}\sum_{i=1}^{n} x_i + \mu^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu\mu + \mu^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu^2 + \mu^2$$

$$\boldsymbol{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2}$$

Lorsque les valeurs $x_1, x_2, \cdots, x_n$ se répètent respectivement $1, 2, 3 \cdots, n\ fois$ on obtient la variance, en comptant chaque valeur x_i autant de fois qu'elle se présente. Ceci revient à pondérer la valeur x_i par l'effectif n_i qui lui correspond, on aura :

$$\boldsymbol{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \mu^2}$$

$\boldsymbol{n}$: $Nombre\ total\ des\ effectif = \sum_{i=1}^{n} \boldsymbol{n_i}$

Quand on veut estimer la variance d'une variable $\boldsymbol{X}$ à partir d'un échantillon (une partie de la population sélectionnée au hasard) de taille $\boldsymbol{n}$, on utilise la variance corrigée $\boldsymbol{S_x^2}$ divisée par $\boldsymbol{n-1}$. c.à.d.

$$\boldsymbol{S_x^2} = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{n-1}\sigma^2$$

N.B :La plupart des logiciels statistiques calculent la variance corrigée $\boldsymbol{S_x^2}$ et non la variance $\boldsymbol{\sigma^2}$

Autre écriture de la variance

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} n_i (x_i - \mu)^2}{\sum_{i=1}^{n} n_i}$$

Pour alléger les calculs, on se sert du théorème de Koenig. Développons la somme S

$$S = \sum_{i=1}^{n} n_i (x_i - \mu)^2$$

$$= \sum_{i=1}^{n} n_i (x_i - \mu)^2 = \sum_{i=1}^{n} n_i x_i^2 - \sum_{i=1}^{n} 2 n_i x_i \mu + \sum_{i=1}^{n} n_i \mu^2$$

$$= \sum_{i=1}^{n} n_i x_i^2 - 2\mu \sum_{i=1}^{n} n_i x_i + n\mu^2 \qquad car\ \mu = \left(\frac{\sum_{i=1}^{n} n_i x_i}{n}\right)$$

$$O\grave{u}\ n = \sum_{i=1}^{n} \boldsymbol{n_i}$$

$$S = \sum_{i=1}^{n} n_i x_i^2 - 2\mu \times n\mu + n\mu^2$$

$$S = \sum_{i=1}^{n} n_i x_i^2 - n\mu^2$$

$$\sigma_x^2 = \left(\frac{\sum_{i=1}^{n} n_i x_i^2 - n\mu^2}{\sum_{i=1}^{n} n_i}\right)$$

Et en fin on obtient la même expression :

$$\boldsymbol{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \mu^2}$$

2.4.2.4 L'écart type

La caractéristique de dispersion la plus usuelle est en effet l'écart type. Puisque la moyenne arithmétique des écarts à la moyenne est nulle, on a recours à la moyenne quadratique de ces écarts. On définit :

L'écart type d'une série est la moyenne quadratique des écarts à la moyenne, autrement dit est la racine carrée de la variance.

$$\boldsymbol{\sigma_x = \sqrt{\sigma_x^2}}$$

Quand on veut estimer l'écart type d'une variable X à partir d'un échantillon de taille n, $\boldsymbol{S_x}$ on utilise la **variance corrigée** pour définir l'écart type.

$$\boldsymbol{S_x = \sqrt{S_x^2 =}\ \sigma_x\sqrt{\frac{n}{n-1}}}$$

N.B :La plupart des logiciels statistiques calculent l'écart type corrigé $\boldsymbol{S_x}$ et non la variance $\boldsymbol{\sigma_x}$

*Signification de l'écart type

Il existe une autre quantité représentante de la dispersion d'une série c'est l'étendue :

$$\boldsymbol{E = x_{(n)} - x_{(1)}}$$

$$\boldsymbol{Etendue = Valeur\ maximale - Valeur\ minimale}$$

Lorsque l'on compare deux séries de même nature, celle qui a l'écart type le plus élevé est la plus dispersée.

Exemple 2.17

Soit la série statistique 2, 3, 4, 4, 5, 6, 7, 9 de taille 8. On a :

$$\boldsymbol{\mu} = \frac{2+3+4+4+5+6+7+9}{8} = \frac{1\times 2+1\times 3+2\times 4+1\times 5+1\times 6+1\times 7+1\times 9}{8} = \mathbf{5}$$

$$\boldsymbol{\sigma_x^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{1}{8}[(2-5)^2+(3-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2+(9-5)^2]$$

$$= \frac{1}{8}[9+4+1+1+0+1+4+16] = \frac{36}{8} = \mathbf{4.5}$$

$$\boldsymbol{\sigma_x = \sqrt{\sigma_x^2} = \sqrt{4.5} = 2.121}$$

On peut également utiliser la formule de la variance $\boldsymbol{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \mu^2}$ ce qui est moins de calculs (surtout quand la moyenne n'est pas un nombre entier).

$$\boldsymbol{\sigma_x^2} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \mu^2$$

$$= \frac{1}{8}(2^2+3^2+4^2+4^2+5^2+6^2+7^2+9^2) - 5^2$$

$$= \frac{1}{8}(4+9+16+16+25+36+49+81) - 25$$

$$= \frac{236}{8} - 25$$

$$29.5 - 25 = 4.5$$

2.4.2.5 Coefficient de variation

L'étendue, la variance et l'écart type sont des paramètres de **dispersion absolue** qui mesurent la variation absolue des données. cependant, un écart type de $6\,mm$ n'a pas la même signification s'il se rapporte à des mesures de l'ordre de $160\,mm$ ou à des messure de l'ordre de $80\,mm$. Aussi dispose-t-on d'un indice de **dispersion relative** appelé **coefficient de variation** noté **COV**. Par définition, le coefficient de variation est égal à :

$$\boldsymbol{COV = \frac{\sigma}{\mu} \times 100}$$

Remarque : Ce coefficient cesse d'être éfficace (signifiant) pou $\boldsymbol{\mu}$ petit. Ce coefficient de variation a l'avantage d'être comparable pour toutes les séries statistiques.

De l'exemple de la charge de rupture (application. 2, page 4, chapitre 1), il est possible de calculer la moyenne, la variance, l'écart type et le coefficient de variation de cette série statistique associée à une variable continue.

$$\boldsymbol{\mu} = \frac{(10 \times 725 + 23 \times 775 + 4 \times 820 + 15 \times 860 + 32 \times 900 + 16 \times 940)}{100} = \frac{85095}{100} = 850.95$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} n_i x_i^2}{\sum_{i=1}^{n} n_i} - \mu^2$$

$$\sigma_x^2 = \frac{10 \times 725^2 + 23 \times 775^2 + 4 \times 820^2 + 15 \times 860^2 + 32 \times 900^2 + 16 \times 940^2}{100} - (850.95)^2$$

$$\boldsymbol{\sigma_x^2} = \frac{72911825}{100} - (850.95)^2 = 5002.35$$

$$\boldsymbol{\sigma} = \sqrt{5002{,}35} = 70.73\ kg$$

$$\boldsymbol{COV} = \frac{\sigma}{\mu} \times 100 = \frac{70.73}{850.95} \times 100 = 0.08$$

La série des charges apparait peu dispersée, car toutes les observations sont relativement voisines de la moyenne.

2.4.2.6 L'écart moyen absolu

L'écart moyen absolu est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre total des observations.

$$\boldsymbol{e_{moy} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|}$$

Exemple 2.18

On considère les deux séries de données suivantes $série1$: $95, 97, 100, 103, 105$

$série\ 2$: $50, 75, 100, 125, 150$

Elles ont la même moyenne arithmétique $(\mu = 100)$ et la même médiane $M_{1/2} = 100$. Cependant elles différent profondément. Cette différence est dite en statistique **la dispersion**. On constate que la deuxième série est beaucoup plus dispersée que la première. Il donc est important de caractériser une série statistique par les paramètres de tendance centrale (position), mais aussi par les caractéristiques de dispersion.

Série	μ	$M_{1/2}$	σ_x^2	σ_x	COV	e_{moy}
1	100	100	13.6	3.687	3.687	3.2
2	100	100	1250	35.355	35.355	30

Il est impossible de résumer ces écarts par leur moyenne arithmétique, puisque par définition de x :

$$\sum_{i=1}^{n}(x_i-\mu)=-n\mu+\sum_{i=1}^{n}x_i=-n\mu+n\mu \qquad car \qquad \sum_{i=1}^{n}x_i=\mu$$

Donc on aura alors recours à la moyenne des valeurs absolues des écarts, c'est **l'écart absolu moyen**.

2.4.2.7 L'écart moyen absolu

L'écart médian absolu est la somme des valeurs absolues des écarts à la médiane divisée par le nombre d'observations.

$$\boldsymbol{e_{med}=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-M_{1/2}\right|}$$

2.5 Moments

- On appelle à l'origine d'ordre $\boldsymbol{r}$ le paramètre :

$$\boldsymbol{m'_r=\frac{1}{n}\sum_{i=1}^{n}x_i^r}$$

- On appelle moment centré d'ordre $\boldsymbol{r}$ le paramètre :

$$\boldsymbol{m_r=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^r}$$

Les moments généralisent la plupart des paramètres. On a en particulier.

$$\boldsymbol{m'_1=\mu}$$
$$\boldsymbol{m_1=0}$$
$$\boldsymbol{m'_2=\frac{1}{n}\sum_{i}x_i^2=\sigma_x^2+\mu^2}$$
$$\boldsymbol{m_2=\sigma_x^2}$$

Nous verrons par la suite que les moments d'ordres supérieurs $(r=3,4)$ sont utilisés pour mesurer **la symétrie** et **l'aplatissement.**

2.6 Paramètres de forme

2.6.1 Coefficient d'asymétrie de Fisher (Skewness)

Le moment centré d'ordre trois est défini par

$$\boldsymbol{m_3=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^3}$$

Cette quantité peut prendre des valeurs positives, négatives ou nulles. L'asymétrie se mesure au moyen du **coefficient d'asymétrie de Fisher**.

$$g_1 = \frac{m_3}{\sigma_x^3}$$

2.6.2 Coefficient d'asymétrie de Pearson

Le coefficient d'asymétrie de Pearson est basé sur une comparaison de la moyenne arithmétique et le mode, standardisée par l'écart type :

$$A_P = \frac{\mu - Mo}{\sigma_x}$$

Tous les coefficients d'asymétrie ont les mêmes propriétés. Ils sont nuls si la distribution est symétrique, négatifs si la distribution est allongée à gauche (letft asymmetry) et positifs si la distribution est allongée à droite (right asymmetry) (Fig).

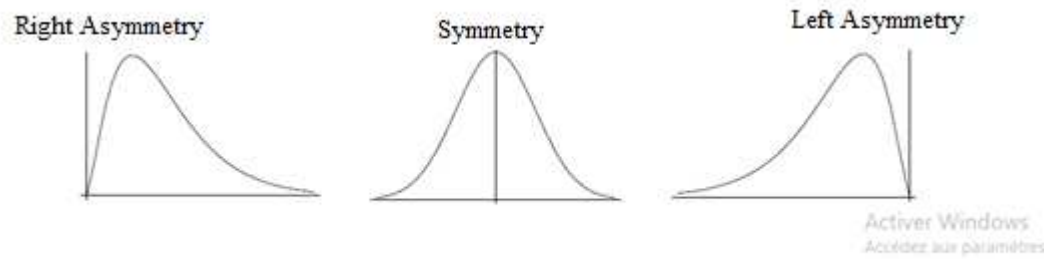


Fig. Asymétrie d'une distribution

2.7 Paramètre d'aplatissement (Kurtosis)

L'aplatissement est mesuré par le coefficient d'aplatissement de **Pearson**.

$$\beta_2 = \frac{m_4}{\sigma_x^4}$$

Ou par le coefficient de **Fisher**

$$g_2 = \beta_2 - 3 = \frac{m_4}{\sigma_x^4} - 3$$

Où m_4 est le moment centré d'ordre $\mathbf{4}$ et σ_x^4 est le carré de la variance.

- Une courbe mésokurtique si $g_2 \approx 0$
- Une courbe leptokurtique si $g_2 > 0$. La courbe de distribution est plus pointue et possède des queues plus longues.
- Une courbe platykurtique si $g_2 < 0$. La courbe de distribution est plus arrondie et possède des queues plus courtes.

Dans la figure ci-dessous, on présente un exemple de deux distributions de même moyenne et de même variance. La distribution plus pointue est leptokurtique, l'autre est plus mésokurtique. La distribution leptokurtique a une queue plus épaisse.

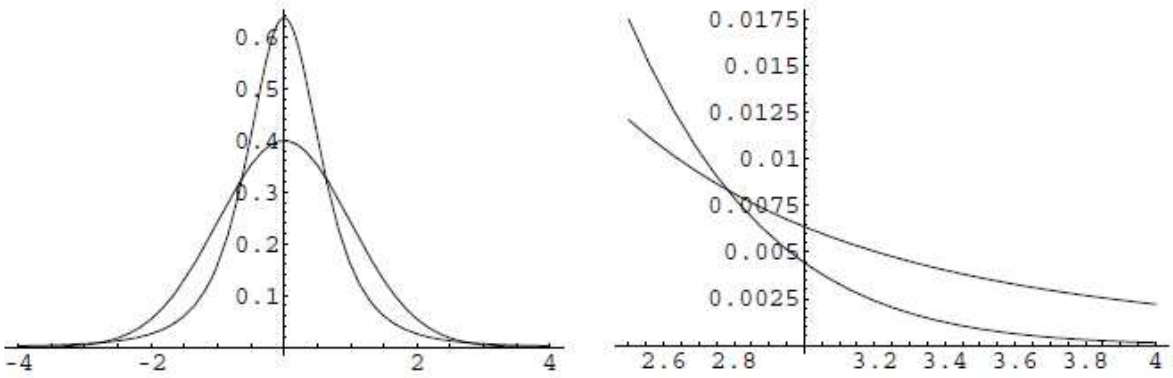


Fig. Distribution mésokurtique et leptokurtique

Etude des series statistiques doubles

3.1 Introduction

Dans le chapitre précédent, on a étudié une population selon un seul caractère. Cependant il est souvent utile de considérer à la fois plusieurs caractères de la même population : taille, âge, poids, qualification,…Nous nous limitons dans le présent chapitre à l'étude simultanée de deux caractères ou variables x et y.

3.2 Série statistique double

Une série statistique double peut être donnée comme l'énumération d'un certain nombre de résultats. On s'intéresse à deux variables $\boldsymbol{x}$ $\boldsymbol{et}$ $\boldsymbol{y}$. Ces deux variables sont mesurées sur les $\boldsymbol{n}$ unités d'observation. Pour chaque unité on obtient donc deux mesures. La série statistique est alors suite de $\boldsymbol{n}$ couples des valeurs prises par les deux variables sur chaque individu.

$$(\boldsymbol{x_1}, \boldsymbol{y_1}), \dots, (\boldsymbol{x_i}, \boldsymbol{y_i}), \dots, (\boldsymbol{x_n}, \boldsymbol{y_n})$$

Chacune des deux variables peut être, soit quantitative, soit qualitative.

3.3 Série statistique associée à deux variables quantitatives

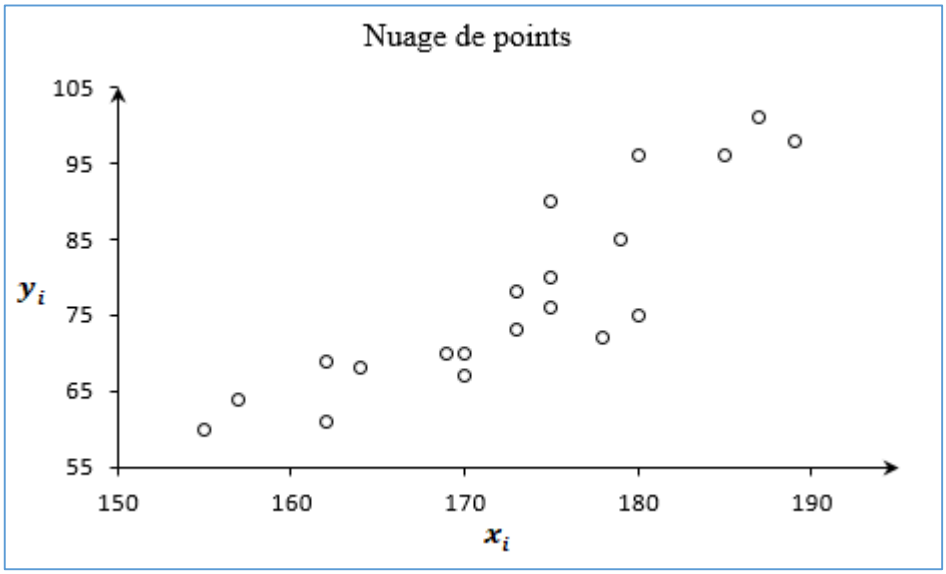
Dans ce cas chaque couple est composée de deux valeurs numériques. Un couple de nombres (entiers ou réels) peut toujours être représenté comme un point dans un plan.

$$(\boldsymbol{x_1}, \boldsymbol{y_1}), \dots, (\boldsymbol{x_i}, \boldsymbol{y_i}), \dots, (\boldsymbol{x_n}, \boldsymbol{y_n})$$

Exemple 3.1

Soit les résultats de mesure de deux caractères $(x, y) de\ vingt\ (20)$ individus.

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
155	60	162	69	180	75	175	90
162	61	169	70	175	76	180	96
157	64	170	70	173	78	185	96
170	67	178	72	175	80	189	98
164	68	173	73	179	85	187	101



3.3.1 Analyse des variables

Les variables $x\ et\ y$ peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2$$

Ces paramètres sont appelés paramètres marginaux : moyennes marginales, variances marginales, écart types marginaux, quantiles marginaux,…

3.3.2 Covariance

La covariance est définie comme suite.

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

- **La covariance** peut prendre des valeurs positives, négatives ou nulles,
- Quand $x_i = y_i$ pour tout $i = 1, \dots, n$, **la covariance** est égale à **la variance.**
- **La covariance** peut également s'écrire

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_x \mu_y$$

Démonstaration

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - y_i \mu_x - x_i \mu_y + \mu_x \mu_y)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \mu_x - \frac{1}{n}\sum_{i=1}^{n} x_i \mu_y + \frac{1}{n}\sum_{i=1}^{n} \mu_x \mu_y$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y$$

$$\Longrightarrow \boldsymbol{cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_x \mu_y}$$

3.3.3 Corrélation

Est une relation réciproque reliant deux variables aléatoires, dont l'un appelle logiquement l'autre.

- Le **coefficient de corrélation** indique le degré de liaison entre deux variables aléatoires.

$$r_{xy} = \frac{\boldsymbol{cov(x,y)}}{\sigma_x \sigma_y} = \frac{\boldsymbol{cov(x,y)}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

- Le **coefficient de détermination** $(\boldsymbol{R^2})$ est le carré du coefficient de corrélation

$$\boldsymbol{R_{xy}^2 = \frac{[cov(x,y)]^2}{\sigma_x^2 \sigma_y^2}}$$

NOTA :

1. Le coefficient de corrélation est toujours compris entre -1 *et* 1 *c.* à. *d.* $-1 \leq r_{xy} \leq 1$,
2. Le coefficient de détermination est compris entre 0 *et* 1, *c.* à. *d.* $0 \leq R_{xy}^2 \leq 1$,
3. Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante,
4. Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante,
5. Si le coefficient de corrélation est nul ou proche de zéro, il n'y pas de dépendance linéaire. On peut cependant avoir une dépendance non linéaire avec un coefficient de corrélation nul (voir Fig).

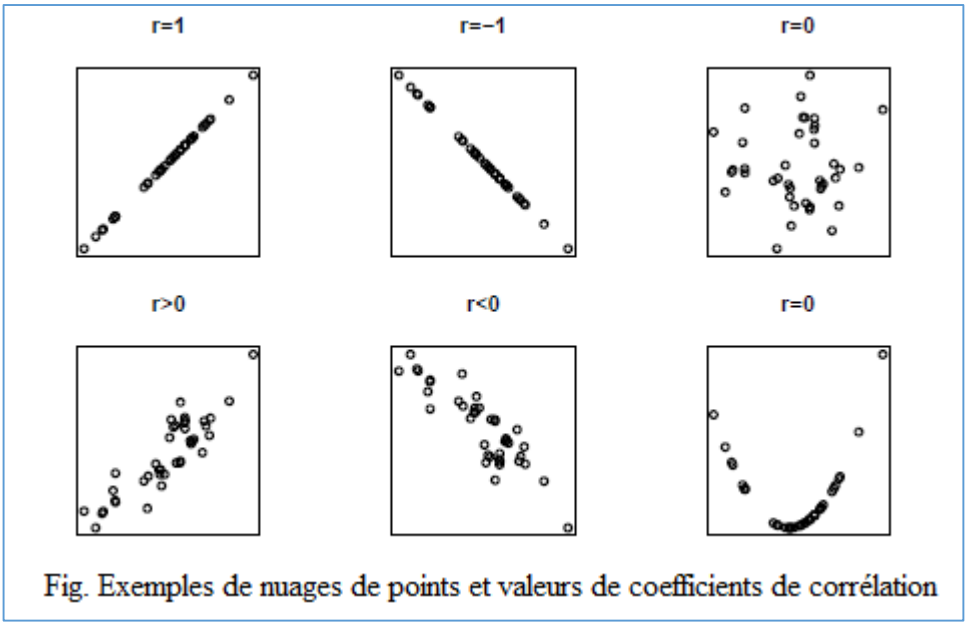


Fig. Exemples de nuages de points et valeurs de coefficients de corrélation

3.3.4 Droite de régression

L'ajustement linéaire consiste à remplacer le nuage de points par une droite, cette droite est dite droite de régression. Cette droite ajuste au mieux un nuage de points au sens des moindres carrées.

On se propose de déterminer une droite telle que les valeurs de $\boldsymbol{y}$ estimées le long de cette droite pour les différentes valeurs de $\boldsymbol{x_i}$ diffèrent peu des valeurs de $\boldsymbol{y_i}$.

Si $\boldsymbol{y = ax + b}$, est l'équation d'une telle droite, la méthode ***des moindres carrés*** consiste à déterminer un couple unique des valeurs $\boldsymbol{a\ et\ b}$ telles que la somme des carrés des écarts entre les valeurs estimées sur la droite, soit minimum. C'est-à-dire

$$\sum_{i=1}^{n} (ax_i + b - y_i)^2 \quad minimum.$$

La droite ainsi obtenue est appelée droite de régression de $\boldsymbol{y\ par\ rapport}$ à $\boldsymbol{x}$, notée $Dy(x)$.

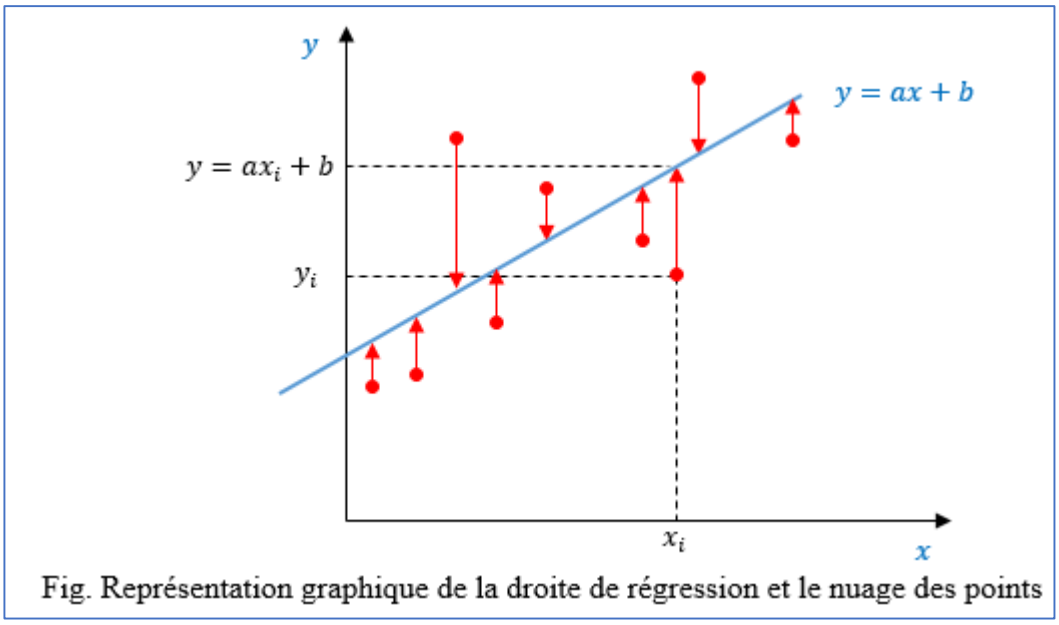


Fig. Représentation graphique de la droite de régression et le nuage des points

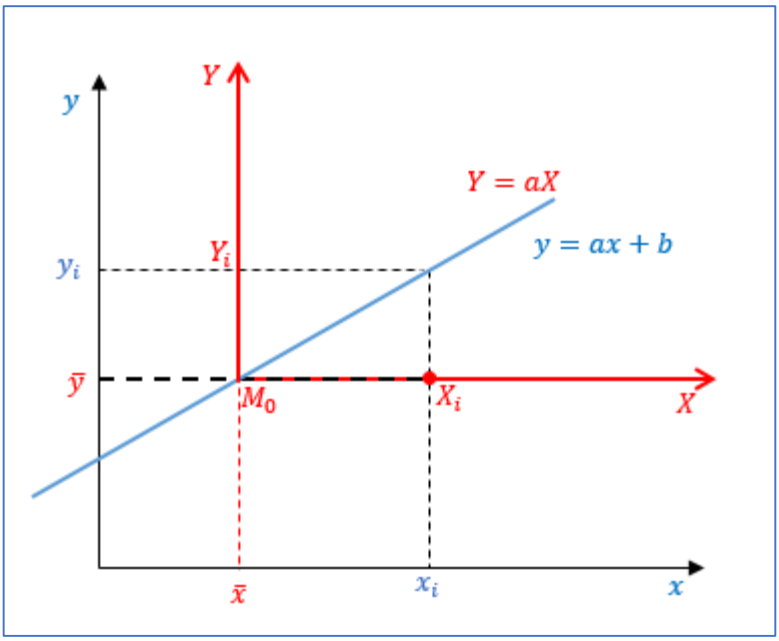
Nous admettons sans le démontrer, que cette droite passe par le point moyen M_0, appelé centre de gravité du nuage et ayant pour coordonnées (Voir Fig. ci-dessous) :

$$\mu_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Où $\boldsymbol{n}$ est le nombre d'observations.

Par rapport aux nouveaux axes de coordonnées M_0X, M_0Y, la droite de régression aura pour équation : $Y = aX$.



Le **coefficient de régression** $\boldsymbol{a}$ se détermine en minimisant donc la quantité.

$$\varphi(a) = \sum_{i=1}^{n}(Y_i - aX_i)^2 = \sum_{i=1}^{n} Y_i^2 + a^2 \sum_{i=1}^{n} X_i^2 - 2a \sum_{i=1}^{n} X_i Y_i$$

Cette quantité est un trinôme de second degré par rapport à $\boldsymbol{a}$ où X_i *et* Y_i sont des constantes

Cette quantité sera minimum pour la valeur de $\boldsymbol{a}$ qui annule la dérivée de cette quantité par rapport à $\boldsymbol{a}$, c.à.d. $\boldsymbol{\varphi'(a)}$

$$\varphi'(a) = -2\sum_{i=1}^{n} X_i Y_i + 2a \sum_{i=1}^{n} X_i^2$$

$$\Longrightarrow a = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

De la Figure précédente, il apparait que

$$X_i = x_i - \bar{x} \quad et \quad Y_i = y_i - \bar{y}$$

Donc on peut écrire l'expression de $\boldsymbol{a}$ sous la forme.

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

Ou encore.

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

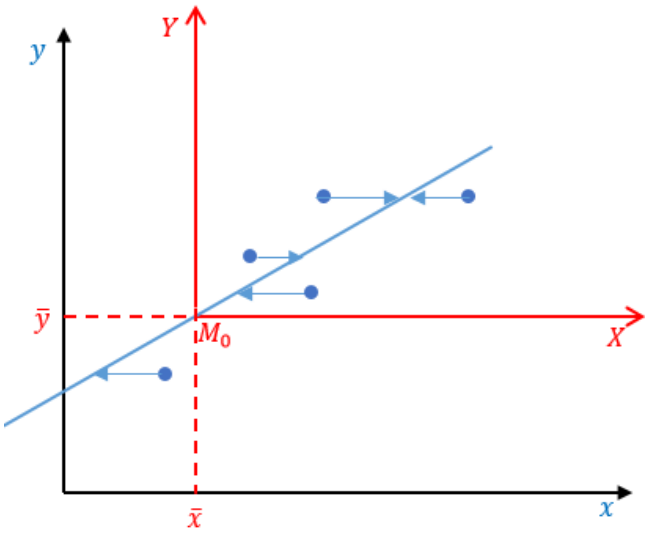
Le numérateur est la **covariance** de x et y tandis que, le dénominateur n'étant autre que la **variance**.

On peut écrire l'expression de $\boldsymbol{a}$ sous la forme

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

- On peut définir de même la droite de régression de $\boldsymbol{x}$ par rapport à $\boldsymbol{y}$, notée $Dx(y)$.

On détermine une droite $\boldsymbol{X = a'Y}$ en minimisant la somme des carrés des distances parallèles à l'axe des abscisses (Voir Fig. ci-dessous).



La symétrie des calculs conduit à :

$$a' = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}$$

- On constate que le carré du coefficient de corrélation est :

$$R^2 = aa'$$

3.3.5 Ajustements

Est de tracer la droite de régression $Dy(x)$ qui passe au plus près lieux des points de nuage et d'en trouver son équation de type $y = ax + b$

1. *Ajustement par la méthode de Mayer*

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage de points.

- Pour une série statistique à deux variables, X *et* Y, dont les valeurs sont des couples (x_i, y_i). Les coordonnées du **point moyen G** de la série sont :

$$x_C = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
$$y_C = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Exemple 3.2

Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de rugby de 2001 à 2006.

Année	2001	2002	2003	2004	2005	2006
Rang x_i	1	2	3	4	5	6
Nombre d'adhérents y_i	70	90	115	140	170	220

Le but est d'étudier cette série à deux variables (rang et nombre d'adhérents) afin de prévoir l'évolution du nombre d'adhérents pour les années suivantes.

✓ Déterminer les coordonnées des points moyens suivants :

1. G_1 des années allant de 2001 à 2003 ;
2. G_2 des années allant de 2004 à 2006 ;
3. G, point moyen du nuage tout entier.

✓ Déterminer l'équation de la droite d'ajustement D_1 qui passe par deux points moyens du nuage de points moyennant la méthode de **Mayer**.

Solution

Coordonnées de G_1: $\begin{cases} x_{G_1} = \frac{1+2+3}{3} = 2 \\ y_{G_1} = \frac{70+90+115}{3} = 91.7 \end{cases} \Rightarrow G_1(2,\ 91.7)$

Coordonnées de G_2: $\begin{cases} x_{G_2} = \frac{4+5+6}{3} = 5 \\ y_{G_2} = \frac{140+170+220}{3} = 176.7 \end{cases} \Rightarrow G_2(5,\ 176.7)$

Coordonnées de G: $\begin{cases} x_G = \frac{1+2+3+4+5+6}{6} = 3.5 \\ y_G = \frac{70+90+115+140+170+220}{6} = 134.2 \end{cases} \Rightarrow G(3.5,\ 134.2)$

✓ La droite D_1n'est pas parallèle à l'axe des ordonnées, elle a donc pour équation $y = ax + b$

$$a = \frac{y_{G_2} - y_{G_1}}{x_{G_2} - x_{G_1}} = \frac{176.7 - 91.7}{5 - 2} 28.3$$

De plus, elle passe par le point $G_1(2,\ 91.7)$

$$\Rightarrow y_{G_1} = ax_{G_1} + b \Leftrightarrow 91.7 = 28.3 \times 2 + b \Rightarrow b = 35.1$$

L'équation de la droite D_1: $y = 28.3\ x + 35.1$

2. Ajustement par la méthode des moindres carrées

Il s'agit d'obtenir une droite équidistante des points situés de part et d'autre d'elle-même. Pour ce faire, on cherche à minimiser la somme des distances des points à la droite au carré. On considère une série statistique à deux variables, représentée par un nuage justifiant un ajustement affine.

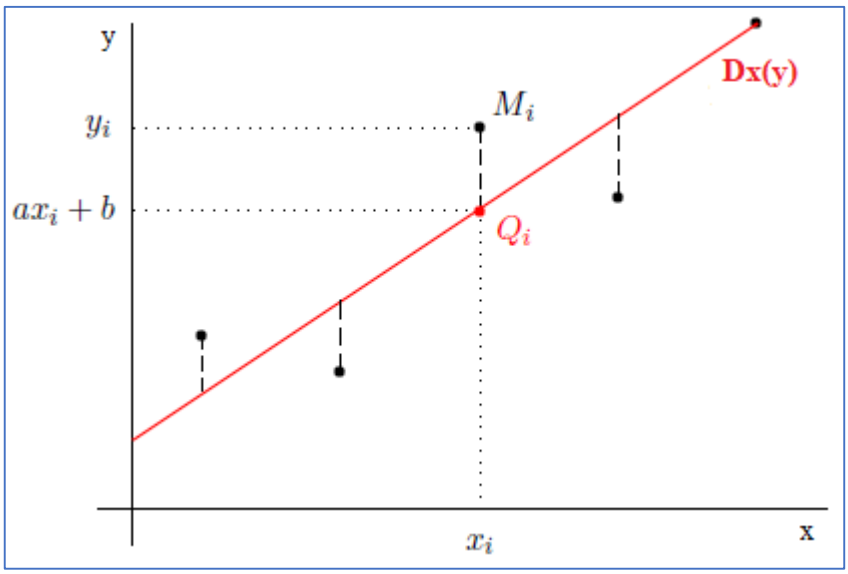
$$\sum_{i=1}^{n} (M_i Q_i)^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

Remarque

Il serait judicieux de s'intéresser à la droite Dx(y) qui minimise la quantité

$$\sum_{i=1}^{n} [x_i - (a' y_i + b')]^2$$

Cette droite est appelée droite de régression de x *en* y.



Propriété 1

La droite de régression $Dy(x)$ a pour équation $y = ax + b$ où

$$\begin{cases} a = \dfrac{cov(x,y)}{\sigma_x^2} \\ b\ vérifie\ \bar{y} = a\bar{x} + b \end{cases}$$

Propriété 2

Le point moyen G du nuage appartient toujours à la droite de régression de $y\ en\ x$.

Question 3

-Déterminer l'équation de la droite d'ajustement D_2 de $y\ en\ x$ obtenue par la méthode des moindres carrées de l'exemple 3.2, et la tracer.

Les coefficients de régression de la droite D_2 sont $a = 29\ et\ b = 32.7 \implies y = 29x + 32.7$

- pour tracer la droite D_2, il choisir au moins deux points sur cette droite. Par exemple : $(0,32.7)\ et\ (8,264.7)$

3. Ajustement exponentiel

On remarque qu'un ajustement affine (moindres carrées) ne semble pas très approprié pour ce nuage de points à partir de l'année 2006 (exemple 3.2). On se propose de déterminer un ajustement plus juste.

On pose $z = \ln y$ ou $y\ prend\ les\ valeurs\ (70,\ 90, \dots, 220)$

Donc, il suffit de calculer les valeurs $\ln y_i$ pour chaque valeur de $\boldsymbol{i}$

x_i	1	2	3	4	5	6
z_i	4.248	4.50	4.745	4.942	5.136	5.394

Question 4

- Déterminer l'équation de la droite d'ajustement D_3 de $z\ en\ x$ obtenue par la méthode des moindres carrées.

$$z = ax + b \implies a = 0.224\ et\ b = 4.045 \implies z = 0.224x + 4.045$$

On déduit la relation entre $y\ et\ x$ puis on trace la courbe représentative de la fonction $y = f(x)$.

$$\text{On a } \begin{cases} z = 0.22x + 4.045 \\ z = \ln y \end{cases} \implies \ln y = 0.22x + 4.045$$

On compose par la fonction exponentielle :

$$e^{\ln y} = e^{0.224x + 4.045}$$
$$= (e^{0.224})^x \times e^{4.045}$$
$$= (1.251)^x \times 57.111$$
$$\implies \boldsymbol{y = 57.111 + 1.251^x}$$

Pour tracer cette courbe donner des valeurs à x pour obtenir les ordonnées y

Comparaison des résultats des méthodes d'ajustement

On supposant que les ajustements conduit dans cet exemple restent valables pour les années suivantes. On calcule $y\ lorsque\ x\ de\ l'année$ 2007 c.à.d. au rang 7

- Méthode de Mayer : $y = 28.3 \times 7 + 35.1 = 233.2$ environ 233 $adhérents$
- Ajustement affine $y = 29 \times 7 + 32.7 = 235.7\ soit\ 236\ adhérents$

- Ajustement exponentiel : $y = 57.112 + 1.024^{7} = 273.9\ soit\ 274\ adhérents$

3.3.6 Corrélogramme

Pour mieux comprendre un Corrélogramme décrivant une opération de mesures statistiques. Examinons l'exemple de mesure suivant relatif au pourcentage (la teneur) des gros grains du sable ($\%\ cg$) que contienne les échantillons des sédiments prélevés sur une traversée d'un cours d'eau. Les échantillons sont pris tous les cinq mètres ($5\ m$) d'intervalle (voir Fig. ci-dessous).

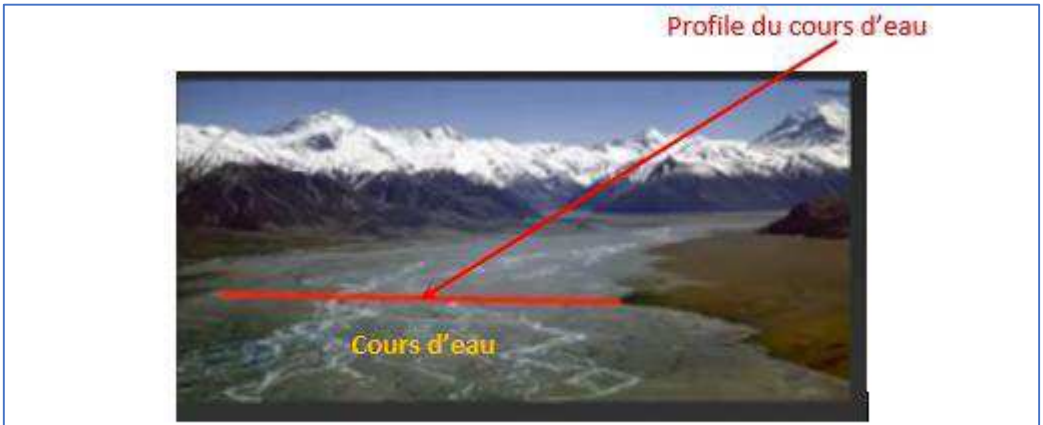
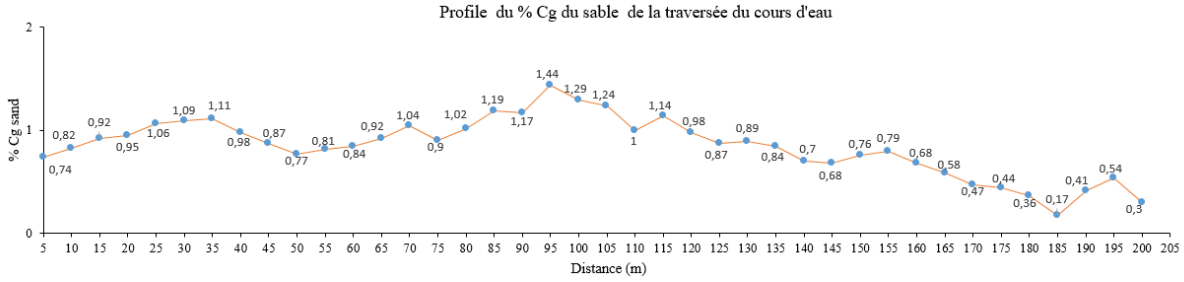


Fig. Profile du cours d'eau sur lequel les échantillons de sédiments ont été pris

Les résultats d'échantillonnage ont permis de tracer le graphique ci-dessous représentant le pourcentage %cg en fonction de la distance de profile.



Quelques paramètres statistiques du profile

En utilisant le Tableur Excel on peut aisément déterminer ces paramètres.

Paramètres de position :

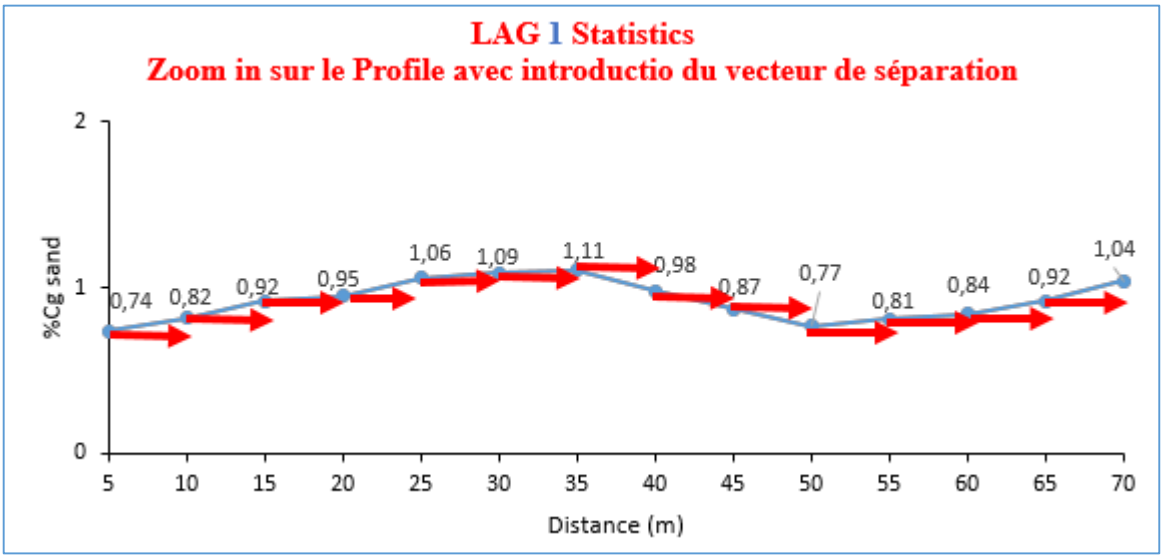
- Moyenne arithmétique : $\boldsymbol{\mu = 0.84425}$
- Médiane :$\boldsymbol{M_{1/2} = Q_{1/2} = 0.87}$
- Valeur minimale : $\mathbf{0.17}$
- Valeur maximale : $\mathbf{1.44}$
- Mode : $\boldsymbol{Mo = 0.92}$

Paramètres de dispersion :

- Ecart moyen absolu : $\boldsymbol{e_{moy} = 0.2170373}$
- Variance standard : $\boldsymbol{\sigma_x^2 = 0.07956865}$

- Ecart type standard : $\boldsymbol{\sigma_x = 0.28207916}$
- Covariance standard : $\boldsymbol{COV = 33.4118048\%}$

Pour mieux expliquer la description d'un Corrélogramme, on intervient sur une partie du profile en faisant introduire la notion du vecteur de séparation (*separation vector or* ***LAG***) or **distance de décalage** dans une orientation définie au préalable.



Chaque vecteur de séparation est caractérisé par deux points : la queue (Tail) et une tête (Head). Le Tableau ci-dessous regroupe les ordonnées de ces caractéristiques de vecteurs de séparation.

Queue	Tête	Queue	Tête	Queue	Tête	Queue	Tête
0.74	0.82	0.81	0.84	1.24	1.0	0.79	0.68
0.82	0.92	0.84	0.92	1.0	1.14	0.68	0.58
0.92	0.95	0.92	1.04	1.14	0.98	0.58	0.47
0.95	1.06	1.04	0.90	0.98	0.87	0.47	0.44
1.06	1.09	0.90	1.02	0.87	0.89	0.44	0.36
1.09	1.11	1.02	1.19	0.89	0.84	0.36	0.17
1.11	0.98	1.19	1.17	0.84	0.70	0.17	0.41
0.98	0.87	1.17	1.44	0.70	0.68	0.41	0.54
0.87	0.77	1.44	1.29	0.68	0.76	0.54	0.30
0.77	0.81	1.29	1.24	0.76	0.79	-	-

$$\begin{cases} \boldsymbol{\mu_{Queue} = 0.8582051} \\ \boldsymbol{\sigma_{x_{Queue}} = 0.27141803} \\ \boldsymbol{\mu_{Tête} = 0.84692308} \\ \boldsymbol{\sigma_{x_{Head}} = 0.28525292} \end{cases}$$

*Calcul de coefficient de corrélation

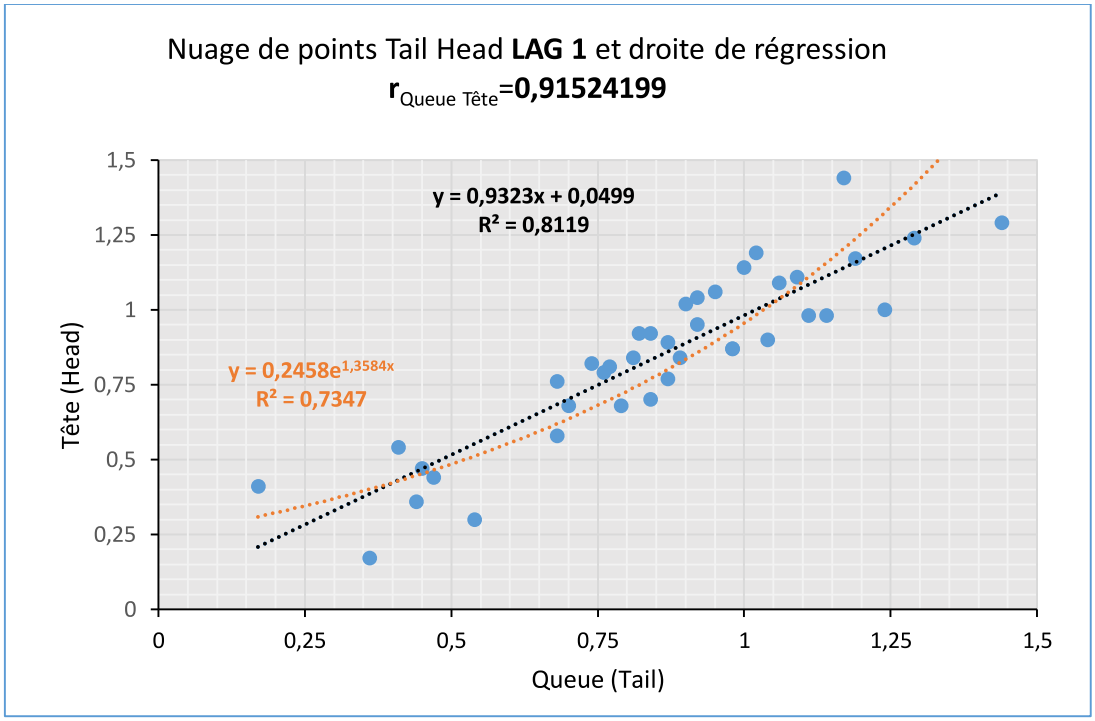
$$r_{xy} = \frac{\boldsymbol{cov(x,y)}}{\sigma_x \sigma_y}$$

$$cov(Queue, Tête) = 0.07086012$$

$$\sigma_{x_{Queue}} * \sigma_{x_{Head}} = 0.077422278$$

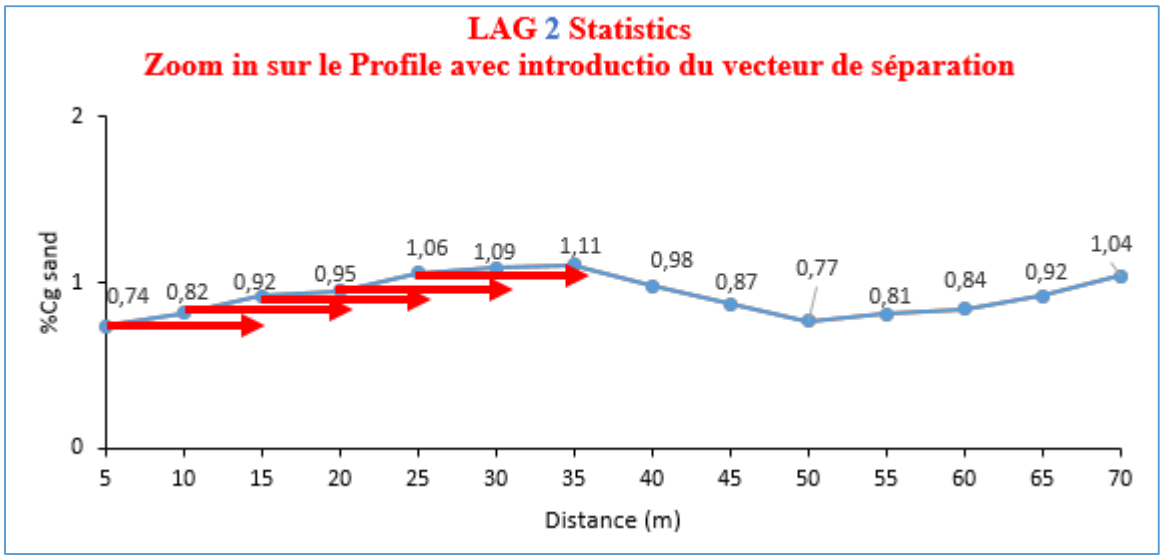
$$r_{Queue\ Tête} = \frac{cov(Queue, Tête)}{\sigma_{Queue}\sigma_{Tête}} = 0.91524199$$

***Nuage de points correspondant au LAG 1**



***Détermination de la distribution spatiale pour LAG 2**

Cette fois ci, la distance de séparation (pas de distance) est fixée à deux pas (LAG 2). La Figure sous-indiquée illustre l'orientation des vecteurs de séparation pour un pas double de la distance de décalage (LAG 2).



Le Tableau ci-dessous regroupe les ordonnées de caractéristiques relatives aux vecteurs de séparation pour ce cas.

Queue	Tête	Queue	Tête	Queue	Tête	Queue	Tête
0.74	0.92	0.81	0.92	1.24	1.14	0.79	0.58
0.82	0.95	0.84	1.04	1.0	0.98	0.68	0.47
0.92	1.06	0.92	0.9	1.14	0.87	0.58	0.44
0.95	1.09	1.04	1.02	0.98	0.89	0.47	0.36
1.06	1.11	0.9	1.19	0.87	0.84	0.44	0.17
1.09	0.98	1.02	1.17	0.89	0.70	0.36	0.41
1.11	0.87	1.19	1.44	0.84	0.68	0.17	0.54
0.98	0.77	1.17	1.29	0.70	0.76	-	-
0.87	0.81	1.44	1.24	0.68	0.79	-	-
0.77	0.84	1.29	1.0	0.76	0.68	-	-

$$\begin{cases} \mu_{Queue} = 0.87891892 \\ \sigma_{x_{Queue}} = 0.26253871 \\ \mu_{Tête} = 0.86243243 \\ \sigma_{x_{Head}} = 0.27805281 \end{cases}$$

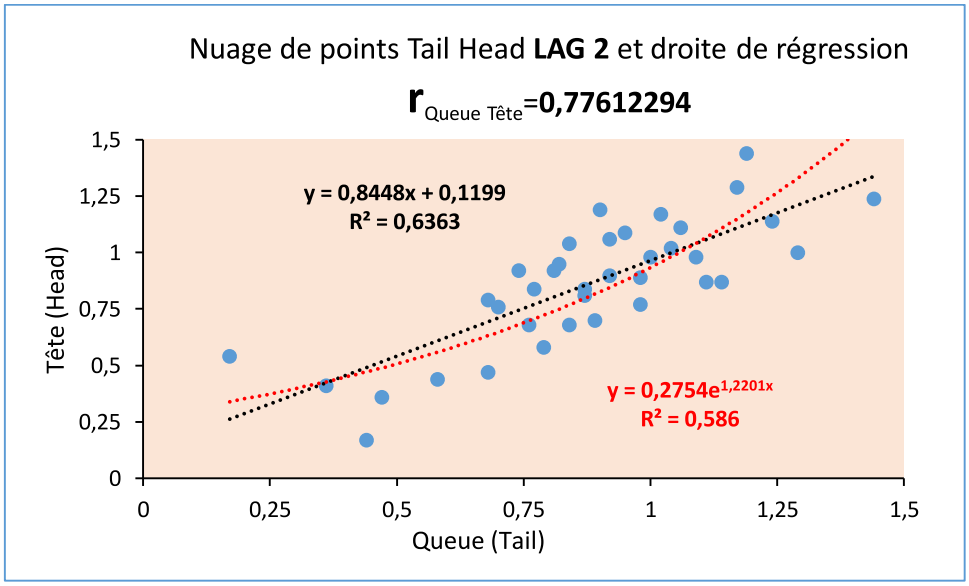
***Calcul de coefficient de corrélation**

$$cov(Queue, Tête) = 0.05665668$$

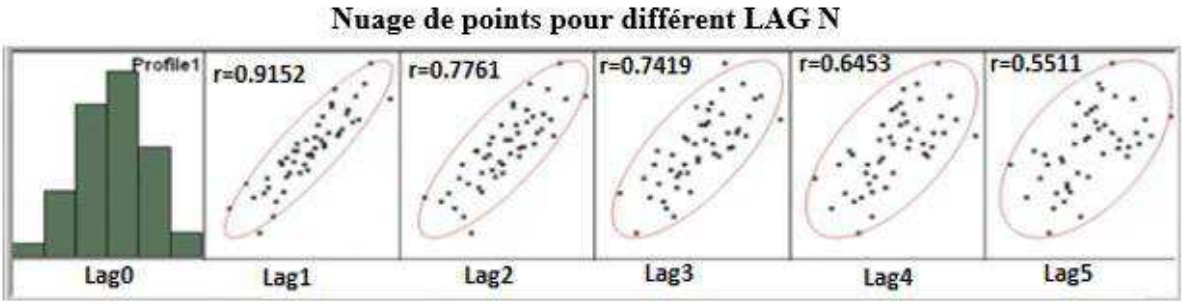
$$\sigma_{x_{Queue}} * \sigma_{x_{Head}} = 0.07299962$$

$$r_{Queue\ Tête} = \frac{cov(Queue, Tête)}{\sigma_{Queue}\sigma_{Tête}} = 0.77612294$$

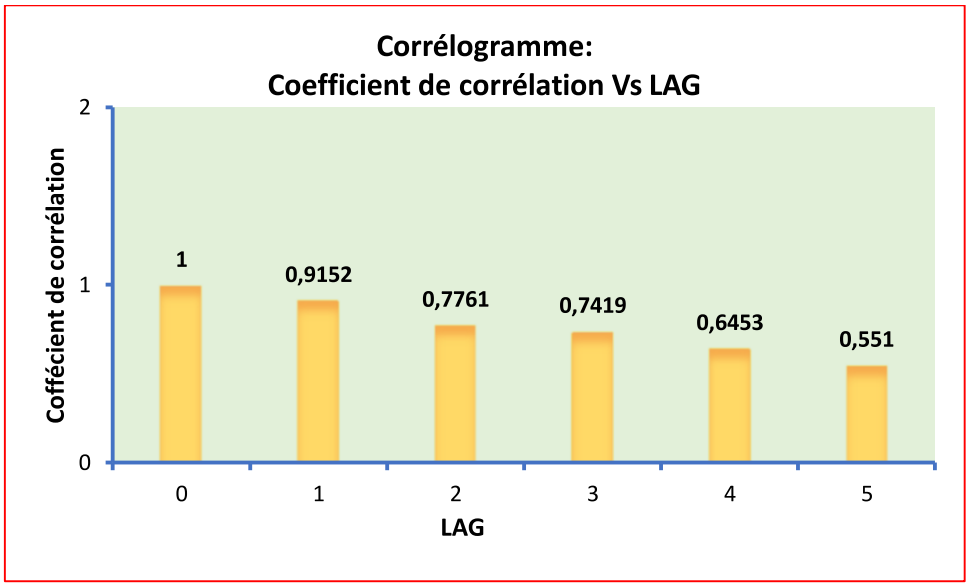
***Nuage de points correspondant au LAG 2**



- En augmentant le pas de décalage des vecteurs de séparation pour $N = 3, 4\ et\ 5$, on a pu constater que le coefficient de corrélation ne cesse à diminuer et le nuage de points devient de plus en plus très dispersé. La Figure sous-indiquée illustre le degré de dispersion du nuage de points en fonction du pas de séparation et le coefficient de corrélation liant la tête et la queue des vecteurs de séparation.



- Le Corrélogramme est le graphique représentant la relation entre le coefficient de corrélation et celle de la distance de séparation ou LAG.



Homework

Soit le profil montré ci-dessous, représentant le pourcentage des grains grossiers des sédiments de sable (**% cg**) contenus dans les échantillons pris le long de ce profil à travers un cours d'eau.

- -Présenter graphiquement les différentes dispersions spatiales (nuage de points) en fonction des **LAG N** ($N = 0, 1, 2, 3\ et\ 4$) et les coefficients de corrélations y afférentes.
- -Tracer le Corrélogramme correspondant.