I.4 - Dispersion et concentration

Dans cette partie s'intéresse à la variabilité des données au sein d'une série. Ainsi, une fois la moyenne connue, on peut compléter la connaissance d'une série pour apprécier dans quelle mesure les données sont dispersées ou au contraire concentrées autour de la moyenne. Les caractéristiques de dispersion et/ou de concentration sont nombreuses (l'intervalle de variation, la variance, l'écart-type, le coefficient de variation, les intervalles interquartiles et interdéciles et l'écart médiale-médiane…).

I.4.1- L'intervalle de variation

L'intervalle, ou « spread » c'est la différence entre la valeur maximale et la valeur Minimale.

Exemple : soit deux élèves dont les notes dans quatre matières ont été les suivantes :

Élève A : {8, 9, 10, 11, 12} Élève B : {2, 4, 16, 18}

L'étendue des notes de A est 12-8 = 4, tandis que l'étendue des notes de B est 18 – 2 = 16. On notera pourtant que la moyenne des deux élèves est de 10. Mais B a des notes beaucoup plus dispersées que A. En fait, si on fait le rapport 16/4, on voit que les notes de B sont 4 fois plus dispersées que celles de A. Il est a noté que l'intervalle de variation est trop sensible aux valeurs extrêmes.

Soit la série suivante {1016, 774, 1008, 8, 1001, 999, 1100}

Apres avoir classer les chiffres par ordre croissant : {8, 774, 999, 1001, 1008, 1016, 1100}

L'intervalle de variation est donc donné par E = 1100 – 8 = 1092. On constate que la valeur de l'intervalle de variation est exagérément augmentée par la présence du chiffre 8.

I.4.1- L'intervalle interquartile

L'intervalle interquartile est défini comme étant l'étendue des 50% de valeurs situées au milieu d'une série de données classées. C'est une de la variation qui n'est pas influencée par les valeurs extrêmes, contrairement à l'intervalle de variation.

L'intervalle interquartile se calcule en procédant aux quatre étapes suivantes :

a. Classement des données de la série par ordre croissant.
b. Trouver la médiane de la série pour séparer celle-ci en deux séries : la première série contient les données inférieures à la médiane et la seconde les données supérieures à la médiane.
c. Déterminer la médiane des deux nouvelles séries, sans inclure dans aucune d'elle la médiane de la série initiale.
- La médiane de la première série est appelée « 1^{er} quartile » et noté Q_1.
- La médiane de la seconde série est appelée « 2^{eme} quartile » et noté Q_3.

d. Calculer IQ, l'intervalle interquartile par la formule :

$$IQ = Q_3 - Q_1$$

Exemple : cas de série avec nombre de valeurs impair.

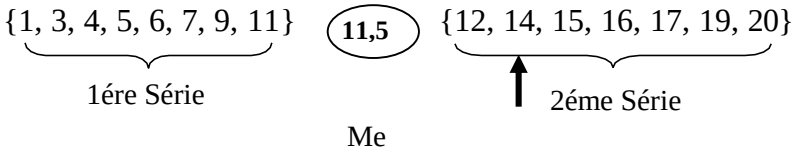
{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6,5}

Pour déterminer l'intervalle interquartile dans ce cas on doit d'abord :

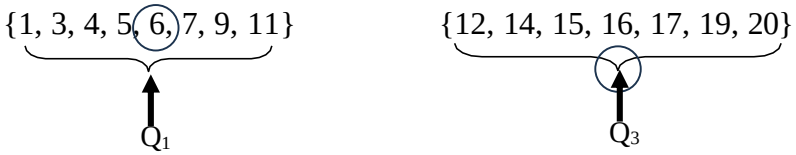
1) Classer les données en ordre croissant

{1, 3, 4, 5, 6, 7, 9, 11, 12, 14, 15, 16, 17, 19, 20}

2) puis nous déterminons la médiane et séparons la série en deux « sous-séries » : pour cela nous avons, (n+1)/2=(14+1)/2=7,5. L'intervalle médiane est donc constitué par la 7ème et la 8ème valeur, c'est-à-dire [11-12]. Et la médiane est égale a (11+12)/2=11,5.



3) Déterminons ensuite la médiane de chacune de ces deux nouvelles séries.



4) Il reste plus qu'à calculer l'intervalle interquartile :

$$IQ = Q_3 - Q_1 = 16 - 6 = 10$$

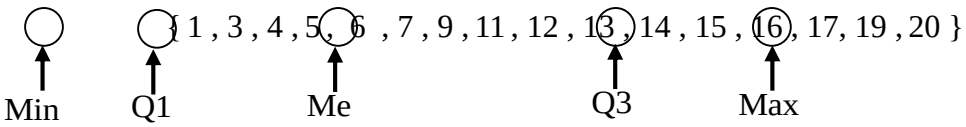
I.4.1- La boîte à moustache

La boîte à moustache, appelée en anglais « Box Plot », est un graphique qui résume la dispersion d'une série à partir de 5 valeurs : la valeur minimale et la valeur maximale (ce sont les « moustaches »), l'intervalle interquartile (désigné par ses deux valeurs Q1 et Q3) et la médiane (ces trois dernières valeurs constituant la « boîte »).

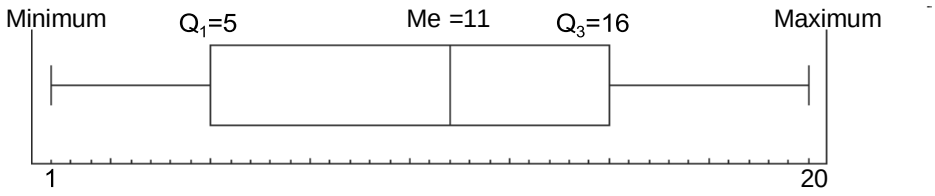
Exemple : soit la série de chiffres suivante, où aucune valeur n'est répétée. Le nombre de chiffres est impair.

{4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6, 5}

Nous savons que Me = 11, Q_1 = 5 et Q_3 = 16 pour les avoir calculés à l'exemple cité précédemment. Quant aux valeurs minimale et maximale, elles sont respectivement égales à 1 et 20. Classons la série par ordre croissant pour mieux faire apparaître les différentes valeurs impliquées dans la boîte à moustache.



Le graphique dit de la « boîte à moustache » correspondant est donc :



A – Utilité de la boîte à moustache pour comparer des séries

La boîte à moustache permet de comparer des séries du point de vue de leur dispersion mais aussi de leur caractéristique de tendance centrale (puisque la médiane est repérée).

Exemple : soient les notes sur 20 de 4 groupes d'étudiants :

Groupe A {1, 2, 2, 12, 5, 5, 9, 5, 7, 11, 7, 8, 2}

Groupe B {16, 13, 15, 13, 11, 13, 16, 3, 18, 11}

Groupe C {8, 8, 8, 7, 4, 16, 13, 16, 18, 11}

Groupe D {12, 10, 6, 8, 5, 16, 12, 15, 10, 15, 12, 10}

La comparaison des graphiques boîtes à moustaches de chaque groupe permet d'avoir une bonne idée de la dispersion des notes, tout en visualisant la note médiane (qui est souvent jugée préférable à la note moyenne).

Alors il suffit de tracer le graphique de boite a moustache des différentes séries :

Pour ce faire il faut déterminer les paramètres de dispersion (la valeur minimale, maximale, médiane le premier quartile et le deuxième quartile) de la série de données considérée.

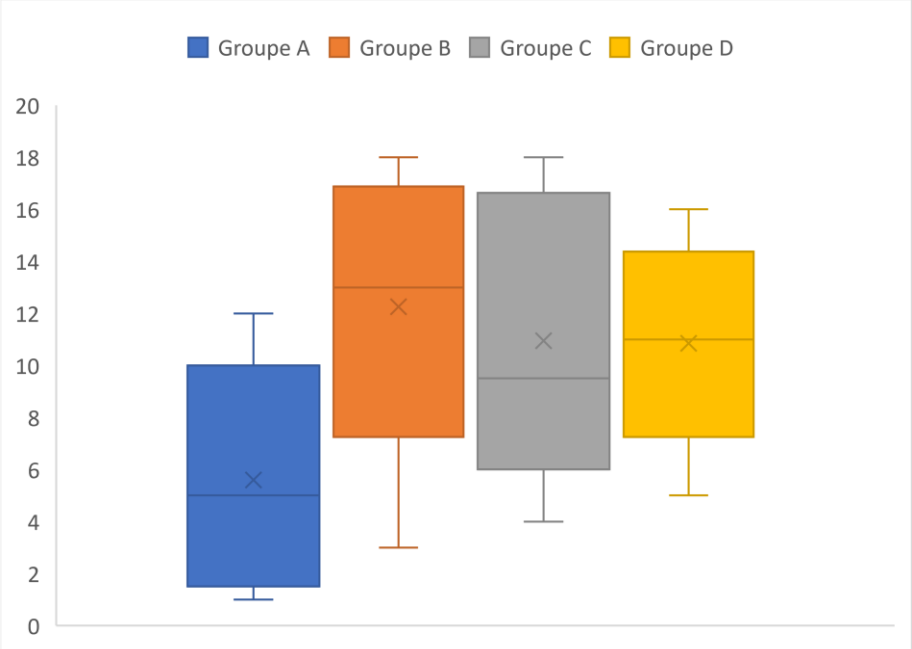
Groupe A	Groupe B	Groupe C	Groupe D
1	3	4	5
2	11	7	6
2	11	8	8
2	13	8	10
5	13	8	10
5	13	11	10
5	15	13	12
7	16	16	12
7	16	16	12
8	18	18	15
9			15
11			16
12			

Apres avoir classer les données on obtient les résultats suivants:

Paramètres	**Groupe A**	**Groupe B**	**Groupe C**	**Groupe D**
Min	**1**	**3**	**4**	**5**
Max	**12**	**18**	**18**	**16**
Médiane	**5**	**13**	**9.5**	**11**

Quartile 1	**2**	**11.5**	**8**	**9.5**
Quartile 3	**8**	**15.75**	**15.25**	**12.75**

Ensuite en trace le graphique en boite a moustaches des différentes séries :



Suivant la position de la médiane au sein de la boîte, on peut en déduire des informations sur la forme de la distribution d'une série.

2) Si la médiane est proche du centre de la boîte, c'est que la distribution est symétrique.
3) Si la médiane est à gauche du centre de la boîte, c'est que la distribution est étalée à droite.
4) Si la médiane est à droite du centre de la boîte, c'est que la distribution est étalée à gauche.

De même, en comparant la longueur respective de chaque moustache, on peut en déduire des informations sur la forme de la distribution.

1) Si les moustaches sont à peu près de la même longueur, c'est que la distribution est symétrique.
2) Si la moustache de droite est plus longue que la moustache de gauche, c'est que la distribution est étalée à droite.
3) Si la moustache de gauche est plus longue que la moustache de droite, c'est que la distribution est étalée à gauche.

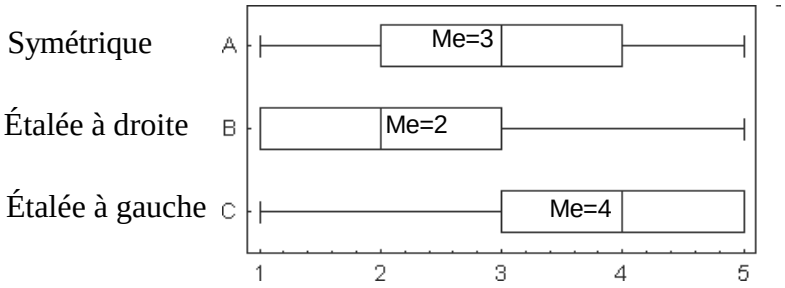
Exemple : Soit les trois séries, dont les distributions (voir les diagrammes en boites a moustaches) sont respectivement symétrique (Me=3), étalée à droite (Me = 2) et étalée à gauche (Me = 4) :

A = {1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5}

B = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5}

C = {1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5}

Les boîtes à moustaches correspondantes ont bien les caractéristiques précitées :



I.4.3- Variance, écart-type et coefficient de variation

La variance, l'écart-type et le coefficient de variation sont les indicateurs les plus fréquemment utilisés pour mesurer la dispersion d'une série. Ces indicateurs renseignent sur la **dispersion des données autour de la moyenne**.

Plus les données sont concentrées autour de la moyenne, plus les valeurs de ces trois indicateurs sont faibles. Inversement, plus les données sont dispersées autour de la moyenne, plus ces trois indicateurs sont élevés.

I.4.2.1- La variance

Soit une série de valeurs d'une variable X : $\{x_1, x_2, ….., x_k\}$. Soit les effectifs associés : $\{n_1, n_2, ….., n_k\}$. La variance de cette série s'écrit :

Si l'effectif considéré est celui d'une population on écrit :

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

Comme on peut appliquer la formule développée :

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

Si l'effectif considéré est celui d'un échantillon on écrit :

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

Ou la formule développée :

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

Si $\{n_1, n_2, \ldots..,n_k\} = \{1, 1,\ldots.,1\}$ et que k = n c'est-à-dire pour des données connues individuellement ou non répétées, la variance d(e la série s'écrira :

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

Ou la formule développée :

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

Lorsque les données sont groupées par classe, c'est le centre de classe c_i, qui remplace x_i et la formule de la variance devient :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.(c_i - \bar{x})^2$$

Ou la formule développée :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.(c_i)^2 - (\bar{x})^2$$

1) Mode de calcul de la formule

Pour calculer la variance, on applique successivement les étapes suivantes:

a) Calcul de la moyenne ;
b) Calcul des écarts à la moyenne ;
c) Calcul des carrés des écarts à la moyenne ;
d) Somme des carrés des écarts à la moyenne ;
e) Division par n.

L'exemple ci-après illustre cette méthode.

Exemple : soit la série {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

Les étapes a), b), c) et d) sont facilitées par la disposition en tableau :

x	$(x_i - x)$ (Étape b)	$(x_i - x)^2$ (Étape c et d)
2	-6,2	38,44
5	-3,2	10,24
7	-1,2	1,44
1	-7,2	51,84
9	0,8	0,64
13	4,8	23,04
6	-2,2	4,84
15	6,8	46,24
8	-0,2	0,04
16	7,8	60,84
82		**237,6**

La moyenne (**étape a**) :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}(x_i) = \frac{82}{10} = 8{,}2$$

La variance (**étape e**) :

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

$$\sigma^2 = \frac{237{,}6}{10} = 23{,}76$$

B – L'écart-type et le coefficient de variation

1) *L'écart-type*

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

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

Généralement, si aucune valeur n'est répétée ou si les données ne sont pas regroupées par valeur, on aura :

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

Exemple 1 : Soit la série {2, 5, 7, 1, 9, 13, 6, 15, 8, 16}

La variance de cette série a été calculée dans l'exemple précèdent, elle est égale à :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.(x_i - \bar{x})^2 = \frac{237{,}6}{10} = 23{,}76$$

D'où l'écart type est égale :

$$\sigma = \sqrt{23{,}76} = 4{,}87$$

2) *Le coefficient de variation*

Le coefficient de variation est donné par la formule suivante :

$$C_v = \frac{\sigma}{\bar{x}}.100$$

On utilisant les données de l'exemple précèdent, le coefficient de variation dans ce cas est égale a :

$$C_v = \frac{\sigma}{\bar{x}}.100 = \frac{4{,}87}{8{,}2}.100 = 59{,}4\%$$

I.4 - Les séries statistiques à deux dimensions

Il est fréquemment nécessaire d'étudier les liens qui peuvent exister entre les deux (ou plus de deux) dimensions qui caractérisent une population statistique. Pour qualifier ces liens on parle de liaison statistique, de corrélation mais, c'est important de le préciser, il n'est jamais question de causalité, la statistique descriptive n'ayant pas pour objet de prouver des causalités.

A – Régression linéaire

La notion de **courbe de régression** est un concept général permettant de mettre en évidence au moyen d'un graphique s'il existe une relation entre deux variables X et Y et quelle est la nature de cette relation.

La courbe de régression est en fait un tracé que l'on fait passer entre les observations d'un nuage de points. Le plus souvent, on essaie de tracer une droite (voir la figure de l'exemple ci-dessous) que l'on désigne alors par **droite de régression** ou, plus simplement par l'expression **droite de tendance**.

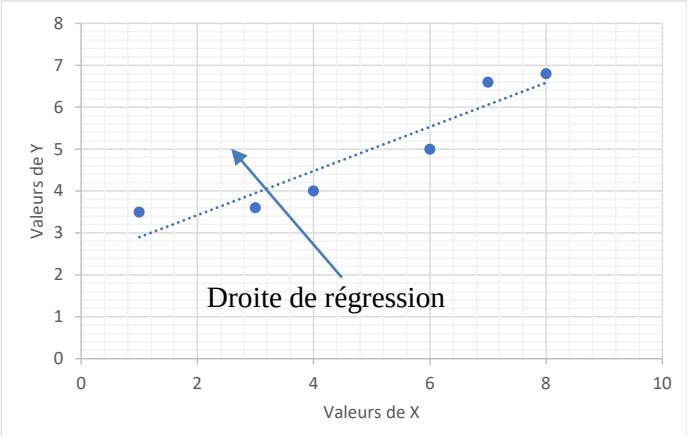
Exemple 1 : Soit S la série de données ci-dessous relatives aux deux variables X et Y, présentées par paires. Le premier élément de la paire correspond à la valeur de X et le second à la valeur de Y.

S = {{1 ; 3,5} ; {3 ; 3,6} ; {4 ; 4} ; {6 ; 5} ; {7 ; 6,6} ; {8 ; 6,8}}

Représentons ces données à l'aide d'un **nuage de points** (figure ci-dessous) où, par convention,

la valeur X se lit en abscisse et la valeur Y en ordonnée.

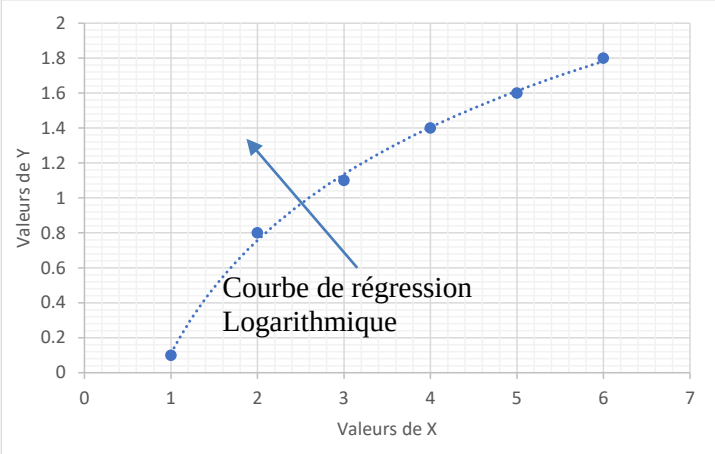
La droite qui passe par le point moyen $(\bar{\bar{x}}, \bar{\bar{y}})$ de l'ensemble des points X,Y, cette droite s'appelle droite de régression ou droite de tendance.



Toutefois, il convient de noter que la relation ainsi établie entre X et Y n'est pas nécessairement linéaire. Pour le montrer, prenons un nouvel exemple.

Exemple : Soit les données ci-dessous relatives aux deux variables X et Y. Cette fois le nuage de points évoque davantage une courbe logarithmique qu'une droite linéaire. C'est pourquoi le nuage de points s'ajuste bien par une **courbe de régression logarithmique**, donc **non linéaire**.

T = {{1 ; 0,1} ; {2 ; 0,8} ; {3 ; 1,1} ; {4 ; 1,4} ; {5 ; 1,6} ; {6 ; 1,8}}



La grande majorité des relations réelles entre variables ne sont pas linéaires, Mais c'est l'ajustement linéaire qui est retenu dans de nombreux cas, pour les raisons suivantes :

1) L'ajustement linéaire est beaucoup plus simple à traiter mathématiquement.
2) Beaucoup de relations sont approximativement linéaires si l'on prend un intervalle de variation suffisamment petit.
3) Certaines relations peuvent être rendues linéaires par un changement de variable approprié (généralement une transformation logarithmique).

B – La droite de régression linéaire

Le **point moyen** est le point qui a pour coordonnées la moyenne de X et la moyenne de Y. On l'appelle aussi le **centre de gravité**.

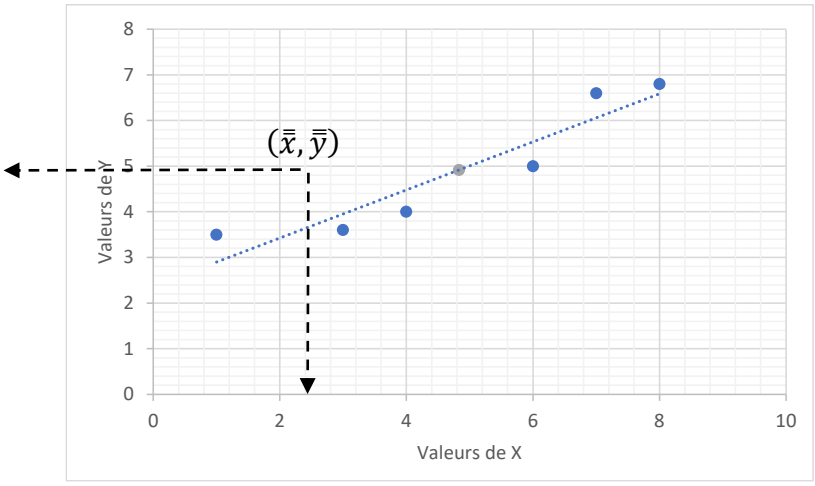
La **droite de régression** est une droite qui passe par le **point moyen**. C'est aussi la droite qui **minimise la somme des carrés des écarts des observations**. Une fois connue, l'équation de cette droite permet de résumer la série et de faire des prévisions.

Exemple : Soit la série S déjà étudiée au paragraphe A

S = {{1 ; 3,5} ; {3 ; 3,6} ; {4 ; 4} ; {6 ; 5} ; {7 ; 6,6} ; {8 ; 6,8}}

On doit déterminer d'abord le point moyen c'est-à-dire le point dont les coordonnées sont $(\bar{\bar{x}}, \bar{\bar{y}})$.

Le graphique de la figure 4, illustre le point moyen :



1) Calcul des coefficients

L'équation de la droite de régression se calcule ainsi. Soit la droite d'équation:

$$y = ax + b$$

Si nous voulons que cette droite soit ajustée à un nuage de points dans le plan {X,Y}, il faut calculer les coefficients a et b en appliquant les formules suivantes :

$$a = \frac{Cov(x,y)}{\sigma_x^2}$$

Et

$$b = \bar{\bar{y}} - a.\bar{\bar{x}}$$

où cov(x, y) représente la covariance de (x, y) et se calcule ainsi :

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

Exemple : calculons *a* et *b* dans le cas de la série S :

S = {{1 ; 3,5}, {3 ; 3,6}, {4 ; 4}, {6 ; 5}, {7 ; 6,6}, {8 ; 6,8}}

Pour faciliter les calculs, adoptons la disposition en tableau suivante :

X	Y	XY	X^2	Y^2
1	3,5	3,5	1	12,25
3	3,6	10,8	9	12,96
4	4	16	16	16
6	5	30	36	25
7	6,6	46,2	49	43,56
8	6,8	54,4	64	46,24
29	**29,5**	**160,9**	**175**	**156**

Ensuite, calculons les sommes dont nous avons besoin dans la formule de *a*.

Puis on calcul les différents paramètres moyenne écart type et covariance entre x et y.

Moyenne de la série X :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad = \frac{29}{6} = 4{,}83$$

Moyenne de la série y :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(y_i) \quad = \frac{29{,}5}{6} = 4.91$$

Ecart type de la série X :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i)^2 - (\bar{x})^2 = \frac{1}{6} \times 175 - (4{,}83)^2 = 5{,}84$$

Ecart type de la série Y :

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{k}(y_i)^2 - (\bar{y})^2 = \frac{1}{6} \times 156 - (4{,}91)^2 = 1{,}79$$

Covariance entre X et Y :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{k}(x_i.y_i) - \bar{\bar{x}}.\bar{\bar{y}} = \frac{1}{6} \times 160{,}9 - 4{,}83 \times 4{,}91 = 3{,}05$$

Calculons maintenant *a* :

$$a = \frac{Cov(x,y)}{\sigma_x^2} = \frac{3.05}{5{,}84} = 0.52$$

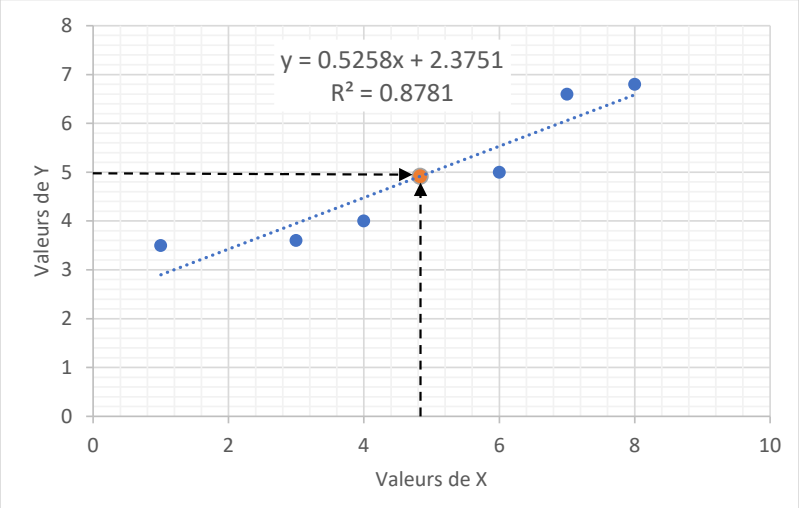
Une fois a connu, on en déduit b :

$$b = \bar{\bar{y}} - a.\bar{\bar{x}} = 4{,}91 - 0{,}52 \times 4{,}83 = 2{,}38$$

L'équation de la droite de régression est donc :

$$y = 0.52.x + 2.38$$

La figure 5 ci-dessous illustre l'équation de cette droite. Nous vérifions à nouveau que cette droite passe par le point moyen.



La droite de régression sert d'abord à **vérifier l'existence d'une relation linéaire** et la nature de celle-ci. Ainsi, dans notre exemple, le coefficient directeur de la droite a=0,52 est positif ce qui dénote une relation positive : x et y varient dans le même sens.

La droite de régression sert ensuite à **faire des prévisions**. Ainsi, nous pouvons utiliser l'équation de la droite de régression pour calculer des valeurs de Y associées à une valeur de X que l'on se donne.

Exemple : Soit la série S, étudiée dans l'exemple précédent et supposons que l'on veuille connaître la valeur Y qui correspond à X = 12 que l'on se donne et qui ne figure pas dans S. Dans ce cas, il suffit de remplacer X par sa valeur dans l'équation de la droite pour obtenir Y :

$$y = 0{,}52 \times 12 + 2{,}38 = 8{,}68$$

C – Le coefficient de corrélation

Le coefficient de corrélation mesure la plus ou moins grande dépendance entre les deux caractères X et Y. On le désigne par la lettre "r" et il varie entre -1 et +1 :

$$r = \frac{Cov(x,y)}{\sigma_x.\sigma_y}$$

Plus r est proche de +1 ou de -1, plus les deux caractères sont dépendants. Plus il est proche de 0, plus les deux caractères sont indépendants.

Exemple : Calculons le coefficient de corrélation de la série S :

$$r = \frac{Cov(x,y)}{\sigma_x . \sigma_y} = \frac{3{,}05}{\sqrt{5{,}84} \times \sqrt{1{,}79}} = 0{,}94$$

1) *Coefficient de corrélation et coefficient de détermination*

Il existe un lien entre le coefficient de corrélation et la droite de régression. Ce lien est donné par la formule :

$$R^2 = a \times a'$$

où ***a*** est le coefficient de la droite de régression de y en x (c'est-à-dire la droite de régression de la forme y = a.x+b) et où a' est le coefficient de la droite de régression de x en y (c'est-à-dire le coefficient de la droite de régression de x en y).

Le terme R^2 est appelé **coefficient de détermination**. En pratique, il n'est pas nécessaire de passer par la formule
$R^2 = a \times a'$. Il suffit en effet de calculer r et de l'élever au carré.

Exemple : Calculons le coefficient de détermination de la série S :

$$R^2 = r \times r \;\Rightarrow\; r = \sqrt{R^2} = \sqrt{0.8781} = 0{,}9371$$

Contrairement au coefficient de corrélation, qui varie entre -1 et +1, le coefficient de détermination varie entre 0 et 1. Il sert aussi à mesurer la corrélation des deux variables, mais ne donne aucune indication sur le sens (positif ou négatif) de la corrélation. Plus il est proche de 0, plus la corrélation est faible. Plus il est proche de 1, plus la corrélation est élevée.