Statistique Descriptive

6ème chapitre

Les séries statistiques à deux dimensions

1 • TABLEAUX ET GRAPHIQUES

A – Séries quantitatives connues individuellement

Exemple : on dispose des mesures de taille et de poids de 19 adolescents. Les données sont présentées par paires. Le premier élément de la paire correspond à la taille et le second au poids.

{{140 ; 38,2} ; {161 ; 44,3} ; {155 ; 46,1} ; {148 ; 38,2} ;{155 ; 50,5} ;{123 ; 22,4} ; {160 ; 40,4} ; {140 ; 34,7} ; {165 ; 50,5} ; {172 ; 50,5} ; {155 ; 38,1} ; {160 ; 57,3} ; {142 ; 39,3} ; {157 ; 46,1} ; {142;37,1} ; {148 ; 45,9} ; {180 ; 66,3} ; {167 ; 60} ; {165 ; 50,5}}

La présentation des données dans un tableau à deux dimensions est données ci-dessous, avec la représentation graphique la plus courante qui est celle dite du « **nuage de points** ».

Tableau 1

Taille	Poids
140	38,2
161	44,3
155	46,1
148	38,2
155	50,5
123	22,4
160	40,4
140	34,7
165	50,5
172	50,5
155	38,1
160	57,3
142	39,3
157	46,1
142	37,1
148	45,9
180	66,3
167	60
165	50,5

Figure 1

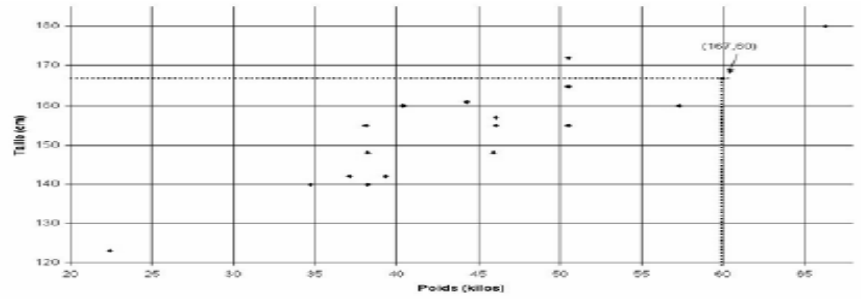
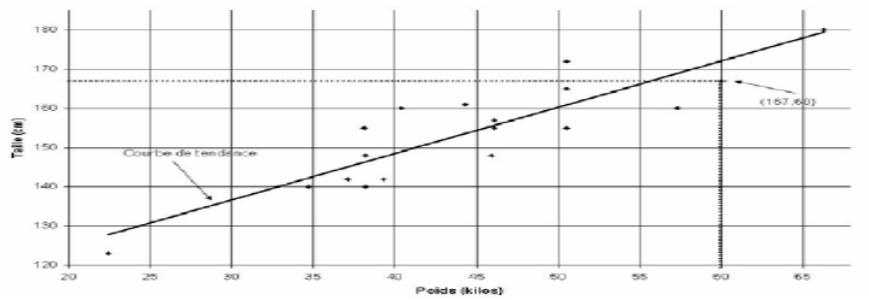


Figure 2



Ce graphique permet d'avoir un aperçu visuel de l'existence ou non d'une corrélation entre les deux variables, ici la taille et le poids. Ainsi, sur la figure 2, une droite « de tendance » a été ajoutée. Les coefficients de cette droite peuvent être calculés précisément (c'est l'objet du chapitre 6). On se contentera ici de noter que les points se regroupent assez bien autour de cette droite, ce qui semble confirmer que, toutes choses égales par ailleurs, il existe une relation positive entre la taille et le poids.

B – Séries quantitatives groupées

Exemple : Les données de l'exemple 1 concernant la taille et le poids de 19 adolescents ont été regroupées par classe dans le **tableau de contingence** ci-après.

Tableau 2

Taille \ Poids	[20 ;40[	[40 ;60[	[60 ;80]
[120 ;140[	1	0	0
[140 ;160[	6	4	0
[160 ;180]	0	6	2

C – Séries qualitatives

Exemple : supposons que l'on ait les données suivantes sur le sexe et le statut d'activité de 20 personnes. Les données sont présentées par paire. La première information concerne le sexe avec les deux modalités M et F. La seconde information concerne le statut d'activité, avec trois modalités (actif occupé [AO], chômeur [C], inactif [I]).

{{F ; AO} ; {M ; I} ; {F ; C} ; {F ; C} ; {M ; AO} ; {M ; AO} ; {M ; C} ; {F ; I} ; {F ;I } ; {F ;I } ; {M ; C} ;
{F ; AO} ; {F ; AO} ; {F ; AO} ; {M ; AO} ; {M ; C} ; {M ; AO} ; {F ; I} ; {F , C} ; {M , AO}}

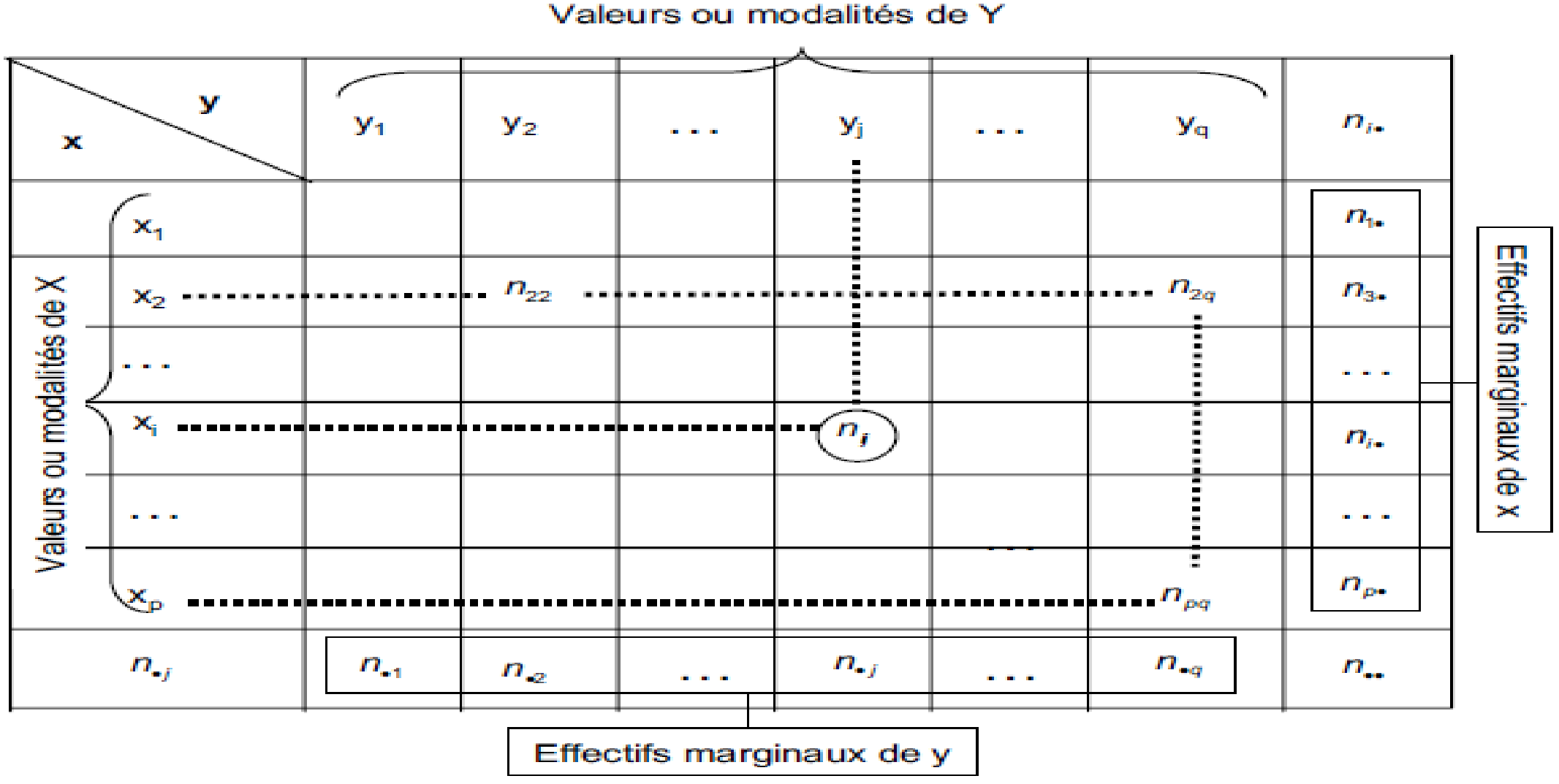
Regroupons ces données dans un tableau de contingence :

Tableau 3

Statut / Sexe	Actifs occupés	Chômeurs	Inactifs
Masculin	5	3	1
Féminin	4	3	4

Le tableau 4 représente un tableau de contingence sous forme symbolique.
À l'intersection de la modalité x_i et de la modalité y_i se trouve l'effectif correspondant.

Tableau 4



L'effectif n_{ij} représente le nombre d'individus qui ont à la fois la modalité/valeur x_i et la modalité/valeur y_j. On a ensuite les symboles suivants :

n_{22} : effectif des individus qui ont la modalité/valeur 2 de x et la modalité 2 de Y.

Par convention, on notre toujours la modalité/valeur de X (i) avant œlle de Y (j).

n_{2q} : effectif des individus qui ont la modalité/valeur 2 de x et la modalité q de Y.

n_{pq} : effectif des individus qui ont la modalité/valeur p de x et la modalité/valeur q de Y.

$n_{i\bullet}$: effectif des individus qui ont la modalité/valeur i (le « • » à la place du j signifie que l'on ne tient pas compte de Y). Exemple : $n_{1\bullet}$ désigne tout l'effectif des individus qui ont la modalité/valeur 1 de X.

Les distributions à deux caractères :

Introduction

Les cinq chapitres traitaient de la statistique à une seule dimension, c'est-à-dire de l'étude des séries statistiques selon un seul caractère. Ce caractère prenait diverses modalités (x_i) que l'on appelait valeurs de la variable dans le cas quantitatif (exemple : série du nombre de salariés en fonction de salaire, nombre d'individus selon diverses classes d'âges…etc).

Les tableaux des données que l'on obtenait étaient uniquement définis par deux colonnes : x_i et n_i.

On se propose maintenant d’étudier les populations, et les séries qui en découlent, suivant deux caractères quantitatifs prenant chacun diverses modalités (xi et yi). On étudiera, par exemple, un ensemble de salariés non plus seulement selon le salaire, mais encore selon l’âge, ou bien un ensemble d’individus selon leur poids et leur taille.

Les tableaux des données seront donc à deux dimensions, et pour pouvoir déterminer les caractéristiques des séries à deux caractères (moyennes, variances et autres moments), il faudra d’abord savoir compléter par les lignes et colonnes adéquates. C’est ce que nous verrons. Un problème nouveau surgira après : celui de la corrélation, où nous chercherons à déterminer si les variables sont liées, c’est-à-dire si l’une à une influence sur l’autre, ou bien si elles sont complètement indépendantes, eu égard à leurs variations. L’étude de la corrélation est celle des liaisons pouvant exister entre deux phénomènes (et entre « n » phénomènes pour la corrélation multiple)

- **Les tableaux de contingence :**

Les tableaux à double entrée représentant les séries statistiques à deux variables, prennent le nom de tableaux de contingence (appelés, parfois, également : Tableaux de corrélation).

- Construction des tableaux de contingence

Exemple 1 :

Le tableau de contingence relève d'un certain nombre de normes : soit la distribution de 17 500 jeunes salariés selon l'âge et le salaire net en milliers des dirhams.

Salaire (y_j) / Age (x_i)	[14 ; 15[	[15 ; 16[	[16 ; 17[	Total
[20 ; 22[	1 200	500	100	1 800
[22 ; 24[	2 500	3 500	600	6 600
[24 ; 26[	1 800	5 000	2 300	9 100
Total	5 500	9 000	3 000	17 500

La ligne et la colonne appelées ici « Total », prennent le nom de **« marges** ». Si l'on associe la marge du bas (ligne « Total » et la ligne du haut (yi), on obtient la distribution des 17 500 jeunes salariés selon leur salaire net (milliers des dirhams). Cette distribution est à une dimension. On l'appelle **distribution marginale du caractère y.** De même, la dernière colonne (Total) associé à la première, n'est autre que la distribution marginale des individus selon leur âge.

Si l’on s’en tenait uniquement à cette double représentation à une dimension, c’est à dire en utilisant seulement les distributions marginales, rien ne permettait de déterminer si l’âge a une influence sur le niveau de salaire, ou l’inverse. Le tableau de contingence, par contre, nous permet de voir comment se distribuent les effectifs de chaque modalité d’un caractère, suivant les modalités de l’autre. On a, en quelque sorte, « croisé » l’information.

*L*es séries statistiques à deux dimensions

II : Outils d'analyse

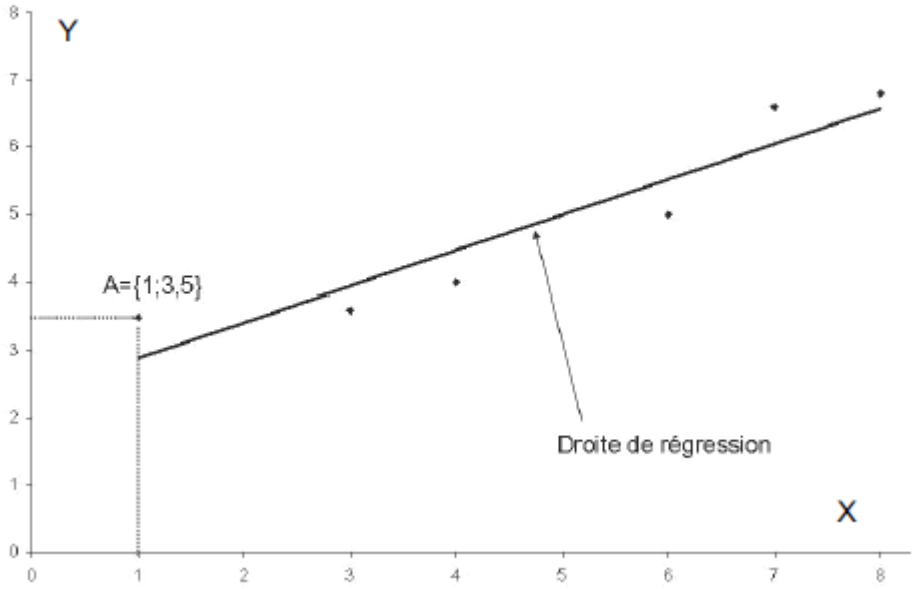
- Il est fréquemment nécessaire d'étudier les liens qui peuvent exister entre les deux (ou plus de deux) dimensions qui caractérisent une population statistique. Pour qualifier ces liens on parle de liaison statistique, de corrélation mais, c'est important de le préciser, il n'est jamais question de causalité, la statistique descriptive n'ayant pas pour objet de prouver des causalités.
- Ce chapitre se limite à l'étude des séries à deux dimensions, X et Y. Cela offre déjà un large éventail de possibilités si l'on se souvient que chacune de ces dimensions peut être quantitative, qualitative et que les données peuvent être groupées dans chaque cas par valeur ou groupes de valeurs. A ces différents cas, correspondent des outils d'analyse appropriés que nous allons évoquer successivement.

1 - SÉRIES QUANTITATIVES AVEC OBSERVATIONS CONNUES INDIVIDUELLEMENT

A –Liaison linéaire, liaison non linéaire, absence de liaison

- On s'intéresse à une statistique ayant deux dimensions que nous désignons par les variables X et Y. La notion de **courbe de régression** est un concept général qui va nous permettre de mettre en évidence au moyen d'un graphique s'il existe une relation entre ces deux variables 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 2 du chapitre 5) 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. Les éléments de chaque paire sont séparés par des points virgules afin de ne pas confondre la séparation des valeurs au sein de la paire, avec les décimales d'une valeur.
- 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 1) où, par convention, la valeur X se lit en abscisse et la valeur Y en ordonnée. Ainsi, la paire qui correspond au point A sur le nuage de points est la première paire de S.
- La valeur X = 1 se lit en abscisse et la valeur Y = 3,5 se lit en ordonnée. Il en va de même des cinq autres paires. Une main « experte » (celle du logiciel) a également tracé une droite entre les points : c'est la droite de régression ou droite de tendance

Figure 1

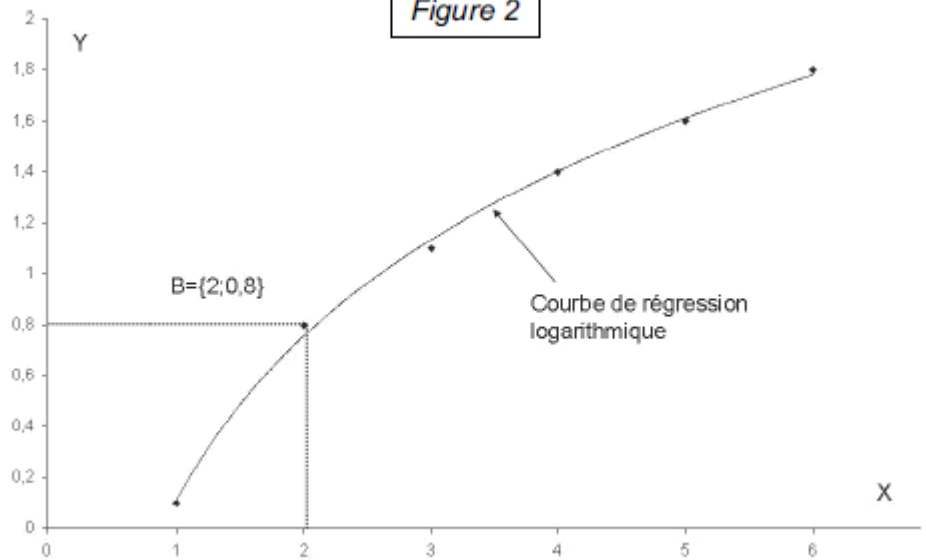


Nous verrons un peu plus loin comment le tracé de cette droite peut s'effectuer mathématiquement et quelles sont les propriétés de la droite de régression. Toutefois, il convient de noter dès maintenant que la relation ainsi établie entre X et Y n'est pas nécessairement linéaire. Pour le montrer, prenons un nouvel exemple

- **Exemple 2 :** 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 l'on a demandé à EXCEL de tracer une **courbe de régression** et que le logiciel a choisi un ajustement 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}}
- Quoique la très grande majorité des relations réelles entre variables ne soient pas linéaires, c'est néanmoins l'ajustement linéaire qui est retenu dans de nombreux cas, pour trois raisons :

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).

Figure 2

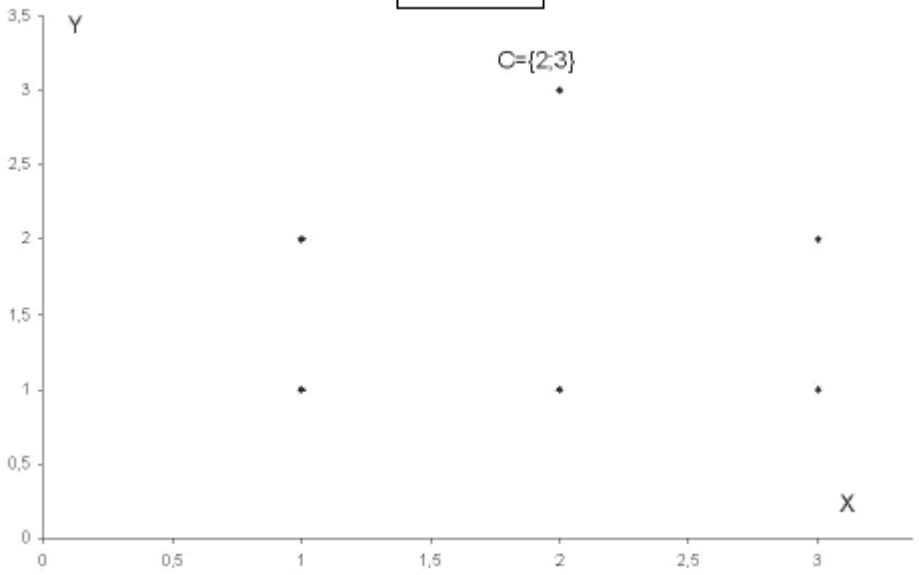

Y
X
B={2;0,8}
Courbe de régression
logarithmique
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
0
1
2
3
4
5
6

Pour finir, notons qu'il n'existe pas nécessairement de liaison entre deux variables, comme l'illustre l'exemple suivant d'**absence de relation**.

- **Exemple 3 :** Soit les données ci-dessous relatives aux deux variables X et Y. Cette fois le nuage de points évoque davantage un amas de points. On peut certes y voir une forme non linéaire (si on relie les points on obtient un dessin de maison), mais il resterait alors à interpréter cette relation.

  U = {{1 ; 1} ; {1 ; 2} ; {2 ; 3} ; {3 ; 2} ; {3 ; 1} ; {2 ; 1}}

Figure 3



B – La droite de régression linéaire

- ***1) Définition***

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

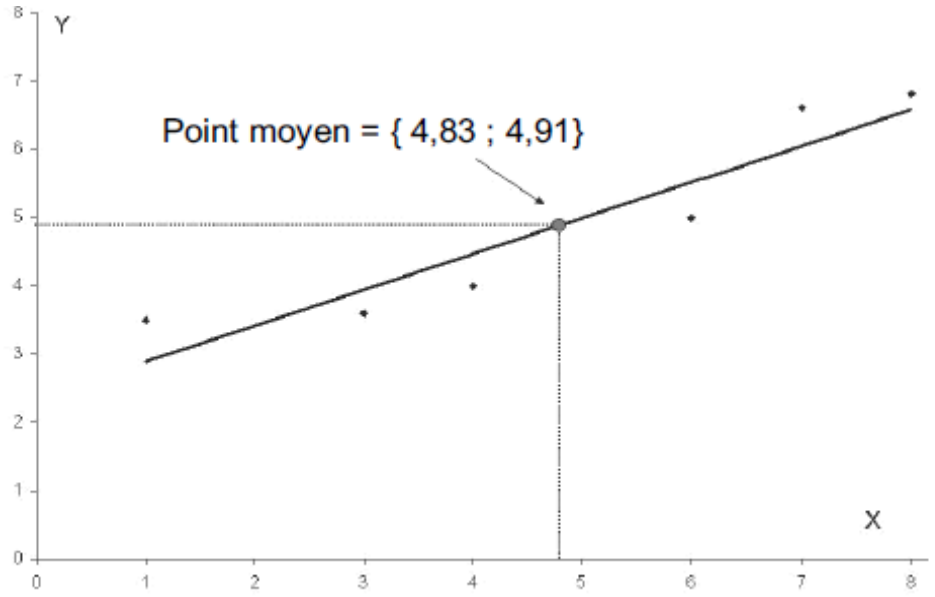
S = {{1 ; 3,5} ; {3 ; 3,6} ; {4 ; 4} ; {6 ; 5} ; {7 ; 6,6} ; {8 ; 6,8}}

La moyenne de X est donnée par (le « double barre » sur le X indique qu'il s'agit d'une moyenne marginale) :

$$\bar{\bar{x}} = \frac{1+3+4+6+7+8}{6} = \frac{29}{6} = 4,833$$

La moyenne marginale de Y est donnée par : $\bar{\bar{y}} = \frac{3,5+3,6+4+5+6,6+6,8}{6} = \frac{29,5}{6} = 4,91$

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



2) 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 suivante

$$a = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} \qquad b = \bar{y} - a\bar{x}$$

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

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

Par conséquent, la formule détaillée de *a* est :

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

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 :

Tableau 1

	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
Sommes ∑	29	29,5	160,9	175	156,01

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

$$\sum_{i=1}^{n} x_i = 29 \qquad \sum_{i=1}^{n} y_i = 29,5 \qquad \sum_{i=1}^{n} x_i y_i = 160,9 \qquad \sum_{i=1}^{n} x_i^2 = 175 \qquad \sum_{i=1}^{n} y_i^2 = 156$$

calculons a :

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{xy}}}{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2} = \frac{\frac{160,9}{6} - \frac{29}{6} \times \frac{29,5}{6}}{\frac{175}{6} - \left(\frac{29}{6}\right)^2} = 0,5258$$

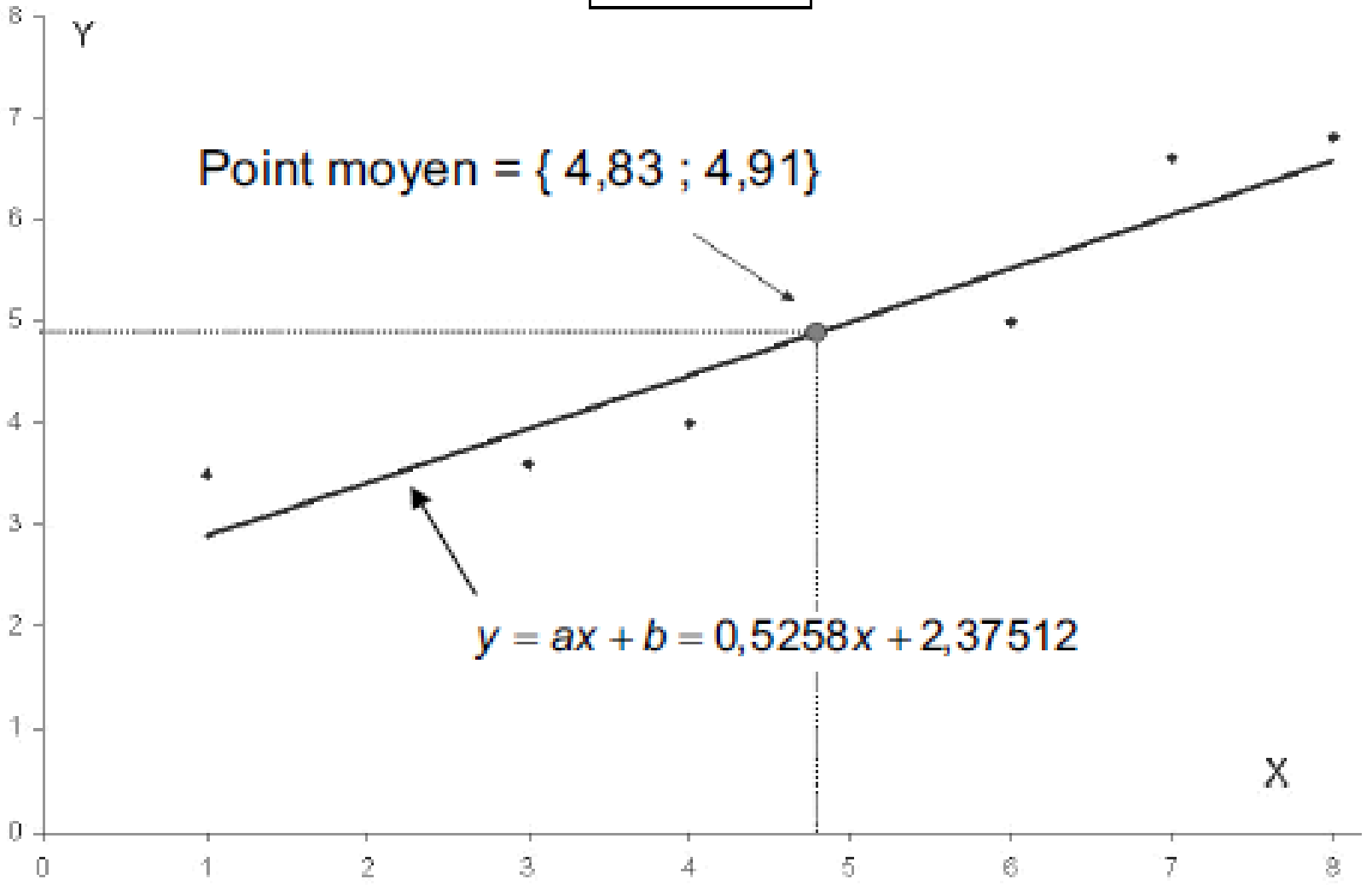
Une fois a connu, on en déduit b :

$$b = \overline{\overline{y}} - a\overline{\overline{x}} = \left(\frac{29,5}{6}\right) - 0,5258 \times \left(\frac{29}{6}\right) = 2,37512$$

L'équation de la droite de régression est donc :

$$y = ax + b = 0,5258x + 2,37512$$

Figure 5



3) Utilité de la droite de régression

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=0,5258 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 1 :** Soit la série S, déjà étudiée précédemment 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 dans l'équation de la droite pour obtenir Y :

$$\boldsymbol{Y = 0{,}5258x(12) + 2{,}37512 = 8{,}6847}$$

- **Exemple 2 :** Soit la série S, déjà étudiée précédemment et supposons que l'on veuille connaître la valeur X qui correspond à Y = 5 que l'on se donne. Dans ce cas, il suffit de remplacer Y par dans l'équation de la droite pour obtenir X :

$$5 = 0{,}5258x + 2{,}37512 \Longrightarrow x = 4{,}99212 \approx 5$$

C – Le coefficient de corrélation

1) Définition et calcul

- 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 :

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

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 :

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{\overline{x}}\,\overline{\overline{y}}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\overline{\overline{x}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y^2 - \left(\overline{\overline{y}}\right)^2}} = \frac{\frac{160,9}{6} - \frac{29}{6} \times \frac{29,5}{6}}{\sqrt{\frac{175}{6} - \left(\frac{29}{6}\right)^2}\sqrt{\frac{156}{6} - \left(\frac{29,5}{6}\right)^2}} = 0,9371$$

2) 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 $y = ax+b$) 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 *R2* est appelé **coefficient de détermination**. En pratique, il n'est pas nécessaire de passer par la formule . Il suffit en effet de calculer *r* et de l'élever au carré.

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

$$R^2 = r \times r = 0{,}9371^2 = 0{,}8781$$

Contrairement ou coefficient de corrélation, qui varie entre -1 et +1, le coefficient de corrélation 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.

Merci