4 Séries statistiques à deux variables

4.1- Séries doubles

S'il faut étudier, pour une même population de n individus, simultanément, deux caractères, il s'agit d'attribuer à chaque individu une valeur numérique liée à l'un des caractères (ce qui définit une variable statistique x) et une valeur numérique liée à l'autre caractère (ce qui définit une variable statistique y).

Ainsi, les couples de valeurs prises par les variables statistiques, x et y, constituent une *série statistique double*, représentée par l'ensemble de ses éléments : $\{(x_1, y_1),(x_2, y_2),...,(x_n, y_n)\}$. x_i est ici la mesure du 1e caractère pour l'individu i et y_i la mesure du 2e caractère pour le même individu.

Par exemple, le relevé du poids et de la taille de chaque étudiant de 3e automobile fournit un ensemble de couples (poids, taille) qui forme une série double.

Autre exemple : pour un certain nombre de véhicules, la TVA payée à l'achat du véhicule neuf dépend du prix payé à l'achat et l'ensemble des couples (montant TVA, prix payé) forme une série double.

Il faut remarquer que $S_x = \{x_1, x_2,..., x_n\}$ et $S_y = \{y_1, y_2,..., y_n\}$ sont deux séries simples. Si les valeurs distinctes de la variable x sont suffisamment nombreuses, les éléments de S_x peuvent être groupés en J classes. De même, si les valeurs distinctes de la variable y sont en assez grand nombre, les éléments de S_y peuvent être groupés en K classes. Il est possible alors d'adopter une représentation par un tableau à deux entrées tel que celui du *tableau* 4.1, où n_{ij} donne le nombre d'individus (effectifs) présentant une valeur du 1e caractère dans la $i^{ème}$ classe du 1e caractère et une valeur du 2e caractère dans la $j^{ème}$ classe du 2e caractère.

2e caractère / 1e caractère	Classe 1	Classe 2	…	Classe K
Classe 1	n_{11}	n_{12}	…	n_{1K}
Classe 2	n_{21}	n_{22}	…	n_{2K}
:	:	:	…	:
:	:	:	…	:
Classe J	n_{J1}	n_{J2}	…	n_{JK}

Tableau 4.1

Si les valeurs distinctes des variables ne sont pas trop nombreuses, les deux séries S_x et S_y sont représentées dans un tableau du type de ceux utilisés auparavant, en colonnes, comme dans le *tableau* 4.2, où il est alors évident que chaque couple (x_i, y_i) est celui des valeurs des deux caractères correspondant au $i^{ème}$ individu. A chaque couple (x_i, y_i) est bien sûr attaché l'*effectif* 1.

x_i	y_i
x_1	y_1
x_2	y_2
x_3	y_3
:	:
:	:
x_n	y_n

Tableau 4.2

4.2- Représentation graphique d'une série double

Puisqu'une série statistique double est un ensemble de couples, il est pratique de représenter une telle série, dans un plan cartésien, par les points dont les coordonnées sont les couples (x_i, y_i) de la série.

L'ensemble de tous ces points forme le ***nuage de points*** de la série double. La *figure* 4.1 illustre cet ensemble.

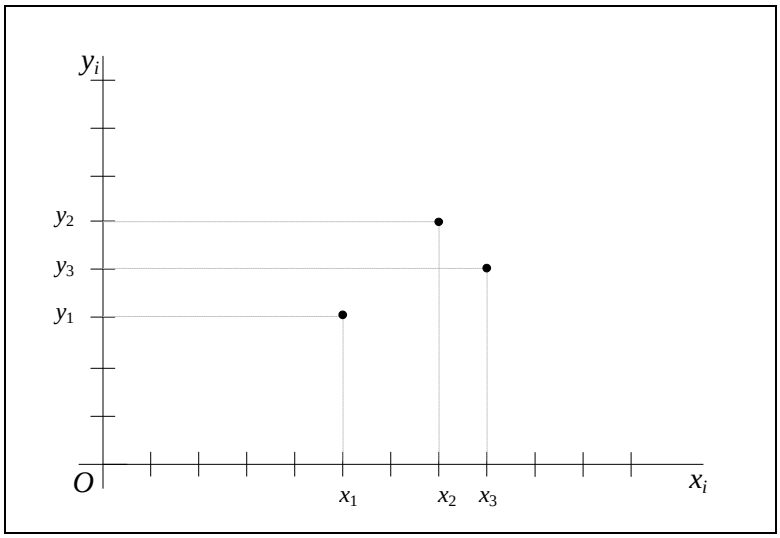


Figure 4.1

Il faut noter que, si la série est recensée, il peut se présenter plusieurs couples pour un même point en coordonnées cartésiennes. Dans ce cas, il est alors possible d'utiliser une représentation pondérée des différents couples (par exemple, chaque case de la représentation est hachurée avec une densité proportionnelle à l'effectif de cette case).

4.3- Notion de corrélation

Une série statistique double présente donc les valeurs de deux caractères relevés sur n individus. Se pose alors la question de l'existence d'un lien entre ces deux caractères.

Cette liaison s'exprime parfois simplement par une loi mathématique. Par exemple, la TVA payée à l'achat d'un véhicule neuf dépend du prix payé à l'achat : $T = 0{,}25P$. Cette relation est une ***liaison fonctionnelle*** entre les deux variables.

A l'opposé de ce cas où le lien entre les deux variables est fonctionnel, il existe des séries qui ne présentent aucun lien entre les variables : les variables sont indépendantes. Par exemple, si x_i note le nombre de chômeurs à Liège chacun des jours ouvrables du mois de septembre et y_i, le débit de la Meuse à Liège, ces mêmes jours ouvrables, il n'y a vraisemblablement aucune relation entre les valeurs x_i et y_i.

Entre le cas de la liaison fonctionnelle et le cas de l'indépendance entre les valeurs statistiques, il y a de nombreux cas où il existe, entre les valeurs prises par des variables statistiques, une certaine liaison qu'il est impossible de formuler mathématiquement de façon précise. Ainsi, par exemple, il existe un lien entre le prix d'un véhicule neuf et la puissance du moteur mais il y a des véhicules de petite cylindrée plus chers que certains véhicules de cylindrée moyenne.

Dans ces cas, il y a ***dépendance statistique*** ou ***corrélation*** entre ces variables statistiques (il y a corrélation entre le prix et la cylindrée d'un véhicule neuf).

4.4- Visualisation d'un lien entre les données

La représentation graphique d'une série double par son nuage de points permet de prévoir le type de liaison entre les variables en partant de l'allure générale du nuage.

Si le nuage a l'allure présentée sur la *figure* 4.2, par exemple, la corrélation peut être supposée de type ***linéaire*** et le nuage peut être ajusté par une droite d'équation $y = ax + b$.

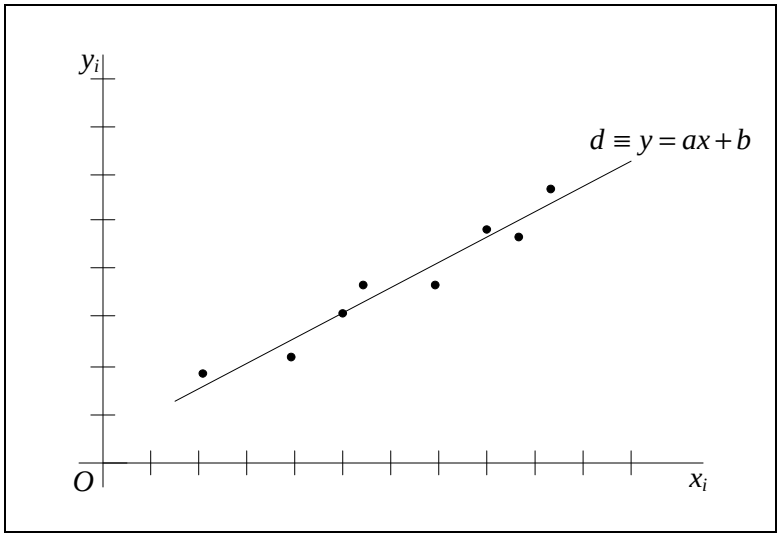


Figure 4.2

Un nuage du type de celui de la *figure* 4.3 laisse supposer une corrélation de type parabolique et le nuage peut être ajusté par une parabole d'équation : $y = ax^2 + bx + c$.

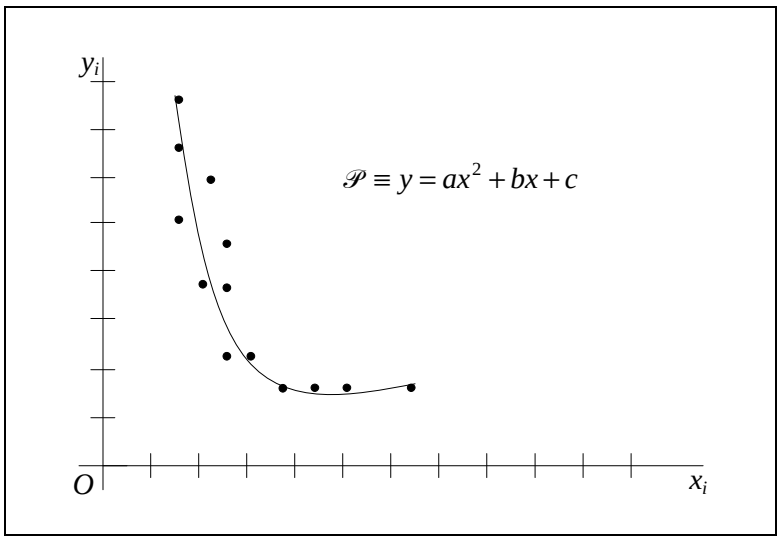


Figure 4.3

Un nuage comme celui de la *figure* 4.4 peut laisser supposer un lien logarithmique et le nuage de points peut être ajusté par une courbe d'équation $y = a \ln x + b$.

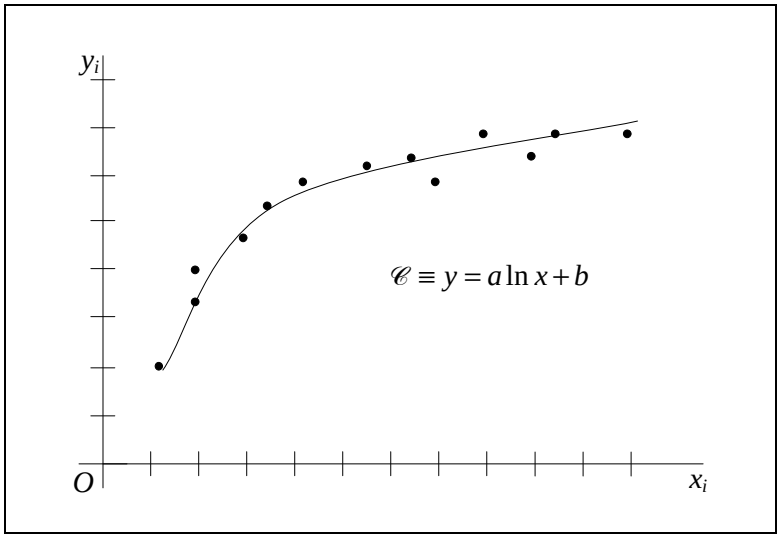


Figure 4.4

Le nuage de la *figure* 4.5 fait penser, quant à lui, à une corrélation nulle, c'est à dire à l'indépendance des variables.

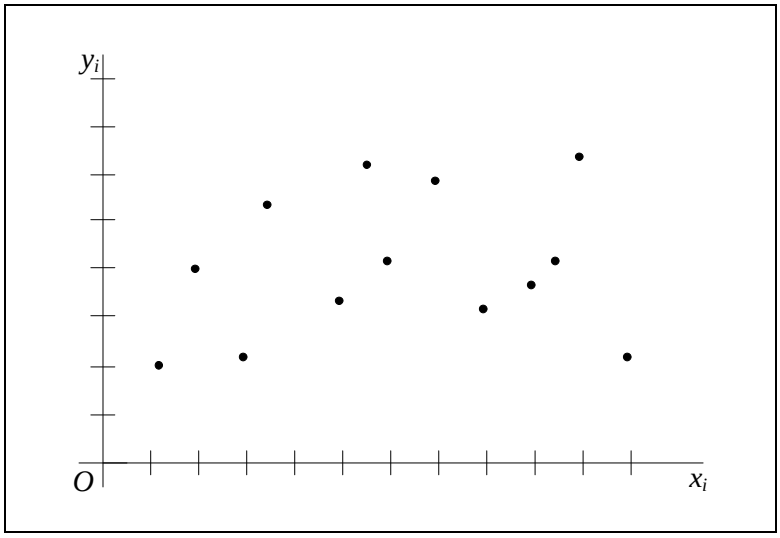


Figure 4.5

En supposant que l'information est correcte, c'est à dire que les données sont suffisamment nombreuses et précises, le problème qui va être abordé est de découvrir s'il existe un lien entre les variables et ensuite de mesurer ce lien en calculant ce qui est appelé le ***coefficient de corrélation***.

Seul le cas de la corrélation linéaire sera étudié en détail, les corrélations de type puissance, exponentielle ou logarithmique pouvant être linéarisées[1].

4.5- Ajustement linéaire par la méthode des moindres carrés

Il s'agit d'exprimer par une fonction linéaire, les valeurs de la variable considérée comme variable dépendante en fonction des valeurs de la variable considérée comme variable indépendante.

Il faut donc trouver :

- une fonction $y = f(x)$ permettant d'ajuster la variable dépendante y à partir de la variable indépendante x,
- ou une fonction $x = g(y)$ permettant d'ajuster la variable dépendante x à partir de la variable indépendante y.

Pour ce faire, la méthode des moindres carrés réduit le plus possible les écarts parallèles à l'axe vertical δ_i entre les valeurs observées et les valeurs données par la loi linéaire et rend minimale la somme des carrés de ces écarts δ_i. La *figure* 4.6 présente cette façon de faire.

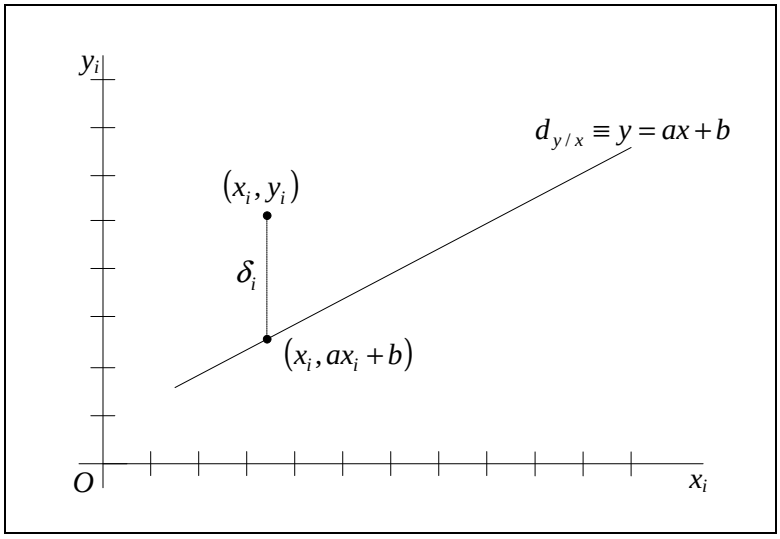


Figure 4.6

$$\delta_i = |y_i - (ax_i + b)| \quad \text{et} \quad \sum_{i=1}^{n} \delta_i^2 \text{ doit être minimum.}$$

Il est possible également de réduire le plus possible les écarts δ_i' parallèles à l'axe horizontal entre les valeurs observées et les valeurs données par la linéarité recherchée. Il faut alors rendre minimum la somme des carrés de ces écarts δ_i' comme sur la *figure* 4.7.

[1] Voir le cours d'***Analyse numérique*** de 2e année de la section « Automobile ».

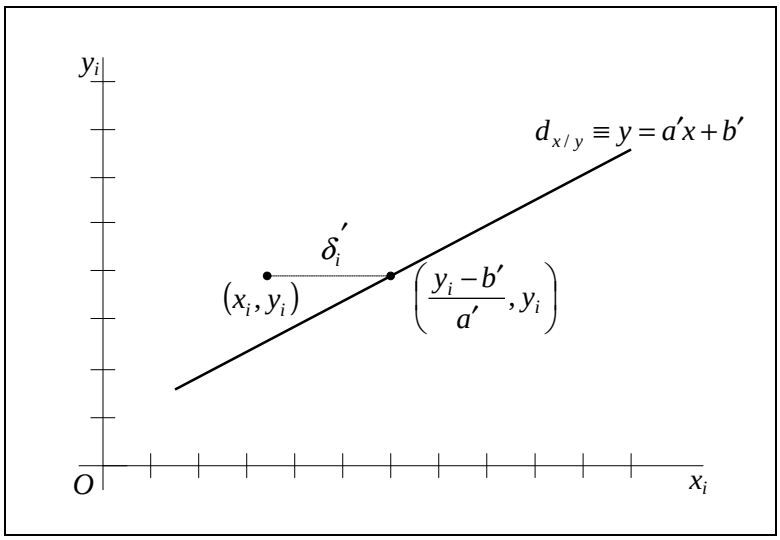


Figure 4.7

$$\delta_i' = \left| x_i - \left(\frac{y_i - b'}{a'} \right) \right| \quad \text{et} \quad \sum_{i=1}^{n} \delta_i'^2 \quad \text{doit être minimum.}$$

La droite des moindres carrés $d_{y/x} \equiv y = ax + b$ obtenue à partir des écarts δ_i parallèles à l'axe vertical a pour coefficient angulaire

$$a = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}} \quad \text{ou} \quad a = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n\sigma_x^2}$$

et pour ordonnée à l'origine

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

Elle passe par le point $C(\bar{x}, \bar{y})$[1].

L'équation de $d_{y/x}$ peut aussi s'écrire $d_{y/x} \equiv y - \bar{y} = a(x - \bar{x})$.

La droite des moindres carrés $d_{x/y} \equiv y = a'x + b'$ obtenue à partir des écarts δ_i' parallèles à l'axe horizontal a pour coefficient angulaire

$$a' = \frac{1}{a^*} \quad \text{si} \quad a^* = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i y_i^2 - n\bar{y}^2} \quad \text{ou} \quad a^* = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n\sigma_y^2}$$

et pour ordonnée à l'origine

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

Elle passe également par le point $C(\bar{x}, \bar{y})$ et son équation peut s'écrire $d_{x/y} \equiv y - \bar{y} = a'(x - \bar{x})$.

Exemple (théorique)

Soit une série double dont les valeurs des variables x_i et y_i sont données dans le *tableau* 4.3. Il s'agit de rechercher les deux droites de régression relatives à cette série double et de les représenter.

[1] Voir le cours d'***Analyse numérique*** de 2e année de la section « Automobile ».

x_i	y_i
2	7
3	10
4	10
5	14
6	14

Tableau 4.3

Vu la forme des coefficients a et a' (ou a^*) : il est pratique de compléter le *tableau* 4.4 de données par une colonne reprenant les valeurs des produits $x_i y_i$, une colonne les x_i^2 et une colonne les y_i^2.

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	2	7	14	4	49
	3	10	30	9	100
	4	10	40	16	100
	5	14	70	25	196
	6	14	84	36	196
Totaux	20	55	238	90	641

Tableau 4.4

$$n = 5 \qquad \bar{x} = \frac{20}{5} = 4 \qquad \bar{y} = \frac{55}{5} = 11$$

$$\sigma_x^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = 2 \qquad \sigma_y^2 = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 = 7{,}2$$

Ainsi :

$$a = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n\sigma_n^2} = \frac{238 - 5.4.11}{5.2} = 1{,}8$$

et

$$b = \bar{y} - a\bar{x} = 11 - 1{,}8.4 = 3{,}8$$

$d_{y/x} \equiv y = 1{,}8x + 3{,}8$

D'autre part :

$$a^* = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n\sigma_y^2} = \frac{238 - 5.4.11}{5.7{,}2} = 0{,}5$$

d'où

$$a' = \frac{1}{a^*} = 2$$

et

$$\begin{aligned} b' &= \bar{y} - a'\bar{x} \\ &= 11 - 2.4 \\ &= 3 \end{aligned}$$

$d_{x/y} \equiv y = 2x + 3$

La *figure* 4.8 représente le nuage de points et les deux droites de régression.

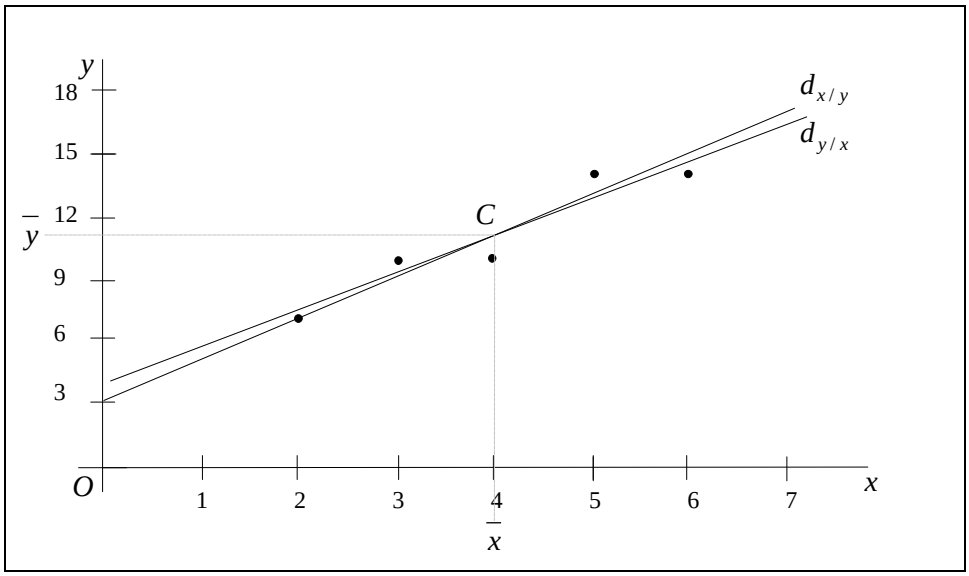


Figure 4.8

4.6- Coefficient de corrélation

Le coefficient de corrélation « mesure » la qualité d'un quelconque ajustement, il est une mesure de l'intensité de la corrélation. Le développement suivant consiste à calculer la valeur minimale de $\sum_i \delta_i^2$ et permet de définir le coefficient de corrélation linéaire.

Le coefficient angulaire et l'ordonnée à l'origine de la droite des moindres carrés $d_{y/x}$ sont

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n\sigma_x^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

où $$n\sigma_n^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

et rendent minimale l'expression

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

La valeur minimale de cette expression peut alors être calculée.

$$
\begin{aligned}
\sum_{i=1}^{n}\delta_i^2 &= \sum_{i=1}^{n}\left[(y_i - ax_i) - b\right]^2 \\
&= \sum_i (y_i - ax_i)^2 + \sum_i b^2 - 2\sum_i (y_i - ax_i).b \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i + \sum_i b^2 - 2b\sum_i y_i + 2ab\sum_i x_i \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i + nb^2 - 2nb\bar{y} + 2nab\bar{x} \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i + nb^2 - 2nb\left(\bar{y} - a\bar{x}\right) \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i + nb^2 - 2nb^2 \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i - nb^2 \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i - n\left(\bar{y} - a\bar{x}\right)^2 \\
&= \sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_i y_i - n\bar{y}^2 - na^2\bar{x}^2 + 2na\bar{x}\bar{y} \\
&= \underbrace{\sum_i y_i^2 - n\bar{y}^2}_{n\sigma_y^2} + a^2\underbrace{\left(\sum_i x_i^2 - n\bar{x}^2\right)}_{n\sigma_x^2} - 2a\underbrace{\left(\sum_i x_i y_i - n\bar{x}\bar{y}\right)}_{n\sigma_x^2 a} \\
&= n\sigma_y^2 - na^2\sigma_x^2 \\
&= n\sigma_y^2\left(1 - a^2\frac{\sigma_x^2}{\sigma_y^2}\right)
\end{aligned}
$$

En posant :

$r = a\dfrac{\sigma_x}{\sigma_y}$, le coefficient de corrélation[1],

il vient

$$\sum_i \delta_i^2 = n\sigma_y^2\left(1 - r^2\right)$$

La somme des carrés des écarts étant nécessairement supérieure ou égale à 0, il faut que le coefficient r soit tel que :

$$r \in [-1,1]$$

La valeur de $\sum_i \delta_i^2$ est maximal si $r = 0$ et nulle si $r = -1$ ou $r = +1$.

Ainsi :

- si $r = -1$ ou $r = +1$, la corrélation est parfaite ($r = -1$ ou $r = +1$ correspond à $\sum_i \delta_i^2 = 0$ ou encore $\delta_i = 0$, $\forall i$ et chaque point (x_i, y_i) est sur la droite des moindres carrés $d \equiv y = ax + b$) ;
- si $r = 0$, la corrélation est nulle (ou les variables sont indépendantes, $\sum_i \delta_i^2$ est maximale) ;
- sinon, $-1 < r < 1$.

[1] Il faut remarquer, à partir de ce point des notes, que *r* et *a* sont de même signe puisque σ_x et σ_y sont positifs.

Il est possible de donner une interprétation rapide du nombre r. Dire que $r = +0{,}9$ signifie, d'une part que la variation de y et x se fait dans le même sens (puisque $r > 0$ et compte tenu de la remarque énoncée *page* 40[1]) et que, d'autre part, $r^2 = 0{,}81$ donc $81\,\%$ des points sont dus à l'existence d'une liaison linéaire, d'une dépendance linéaire des y par rapport aux x, tandis que $19\,\%$ des points du nuage sont indépendants de cette liaison.

Cette première interprétation du nombre r ne tient pas compte de l'effectif total. Or, une valeur de r est plus crédible pour un effectif de 1000 que pour un effectif de 10. Des tables de valeurs significatives de r sont alors utilisées. Ces tables sont basées sur le fait que la population étudiée suit la loi normale de ***Gauss*** et que la corrélation à envisager est linéaire. Elles sont établies par le calcul des probabilités. Si la population ne suit pas la loi normale, ces tables restent valables dans une certaine mesure.

Ces tables présentées dans le *tableau* 4.5 relient les valeurs de r à un paramètre L qui représente le degré de liberté. Pour la corrélation linéaire :

$$L = n - 2$$

Coefficient de sécurité					Coefficient de sécurité				
	0,9	**0,95**	**0,98**	**0,99**		**0,9**	**0,95**	**0,98**	**0,99**
L					**L**				
					20	0,36	0,42	0,49	0,54
					25	0,32	0,38	0,45	0,49
					30	0,30	0,35	0,41	0,45
3	0,81	0,88	0,93	0,96	**35**	0,27	0,32	0,38	0,42
4	0,73	0,81	0,88	0,92	**40**	0,26	0,30	0,36	0,39
5	0,67	0,75	0,83	0,87	**45**	0,24	0,29	0,34	0,37
6	0,62	0,71	0,79	0,83	**50**	0,23	0,27	0,32	0,35
7	0,58	0,67	0,75	0,80					
8	0,55	0,63	0,72	0,76	**60**	0,21	0,25	0,29	0,32
9	0,52	0,60	0,69	0,73	**70**	0,20	0,23	0,27	0,30
10	0,50	0,58	0,66	0,71	**80**	0,18	0,22	0,26	0,28
11	0,48	0,55	0,63	0,68	**90**	0,17	0,21	0,24	0,27
12	0,46	0,53	0,61	0,66	**100**	0,16	0,19	0,23	0,25
13	0,44	0,51	0,59	0,64					
14	0,43	0,50	0,57	0,62	**150**	0,14	0,16	0,19	0,21
15	0,41	0,48	0,56	0,61	**200**	0,12	0,14	0,17	0,18
16	0,40	0,47	0,54	0,59	**300**	0,10	0,11	0,13	0,15
17	0,39	0,46	0,53	0,58	**400**	0,08	0,10	0,12	0,13
18	0,38	0,44	0,52	0,56	**500**	0,07	0,09	0,10	0,11
19	0,37	0,43	0,50	0,55					
					1000	0,05	0,06	0,07	0,08

Tableau 4.5

L'exemple suivant montre comment se servir d'une telle table dans différents cas : trois calculs de corrélation ont abouti à un coefficient r égal à $0{,}5$ avec des effectifs respectifs de 8, de 16 et de 61 éléments.

- $r = 0{,}5$; $n = 8$ et donc $L = 6$.
  Sur la ligne $L = 6$, la valeur $r = 0{,}5$ ne se trouve pas. La plus petite valeur est $0{,}62$ qui correspond à un coefficient de sécurité de $90\,\%$. Comme r est inférieur à $0{,}62$; il est possible d'affirmer que le risque de non-corrélation linéaire est supérieur à $10\,\%$.

- $r = 0{,}5$; $n = 16$ et donc $L = 14$.

[1] Il faut remarquer, à partir de ce point des notes, que *r* et *a* sont de même signe puisque σ_x et σ_y sont positifs.

Le coefficient $r = 0{,}5$ correspond à un coefficient de sécurité de $95\,\%$. Il y a donc $5\,\%$ de risque que la corrélation ne soit pas linéaire.

- $r = 0{,}5$; $n = 61$ et donc $L = 59$.
  Sur la ligne $L = 60$ ne se trouve pas la valeur de $r = 0{,}5$. Le plus grand nombre est $0{,}32$ qui correspond à un coefficient de sécurité de $99\,\%$. Il y a donc moins de $1\,\%$ de risque que la corrélation étudiée ne soit pas linéaire.

Remarques

- Corrélation n'implique pas Causalité. Une forte corrélation peut être due à une évolution temporelle des variables x et y dans le même sens, à une même influence externe ou à une simple coïncidence.
- Le type de corrélation concerné par r est linéaire. Pour des séries dont le nuage de points se présente comme ceux de la *figure* 4.9, r est voisin de 0. Pourtant, il existe une corrélation entre les variables x et y.

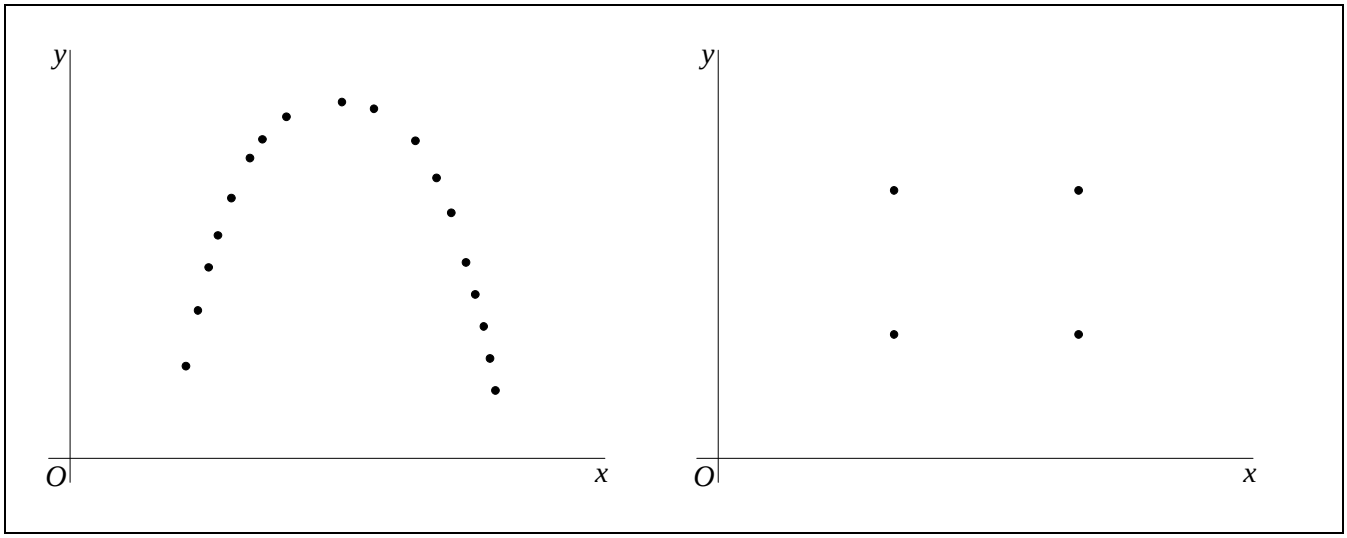


Figure 4.9

Les *figures* 4.10 donnent quelques formes typiques de nuages de points en relation avec les valeurs du coefficient de corrélation linéaire r.

4.7- Propriétés des coefficients a , a^* et r

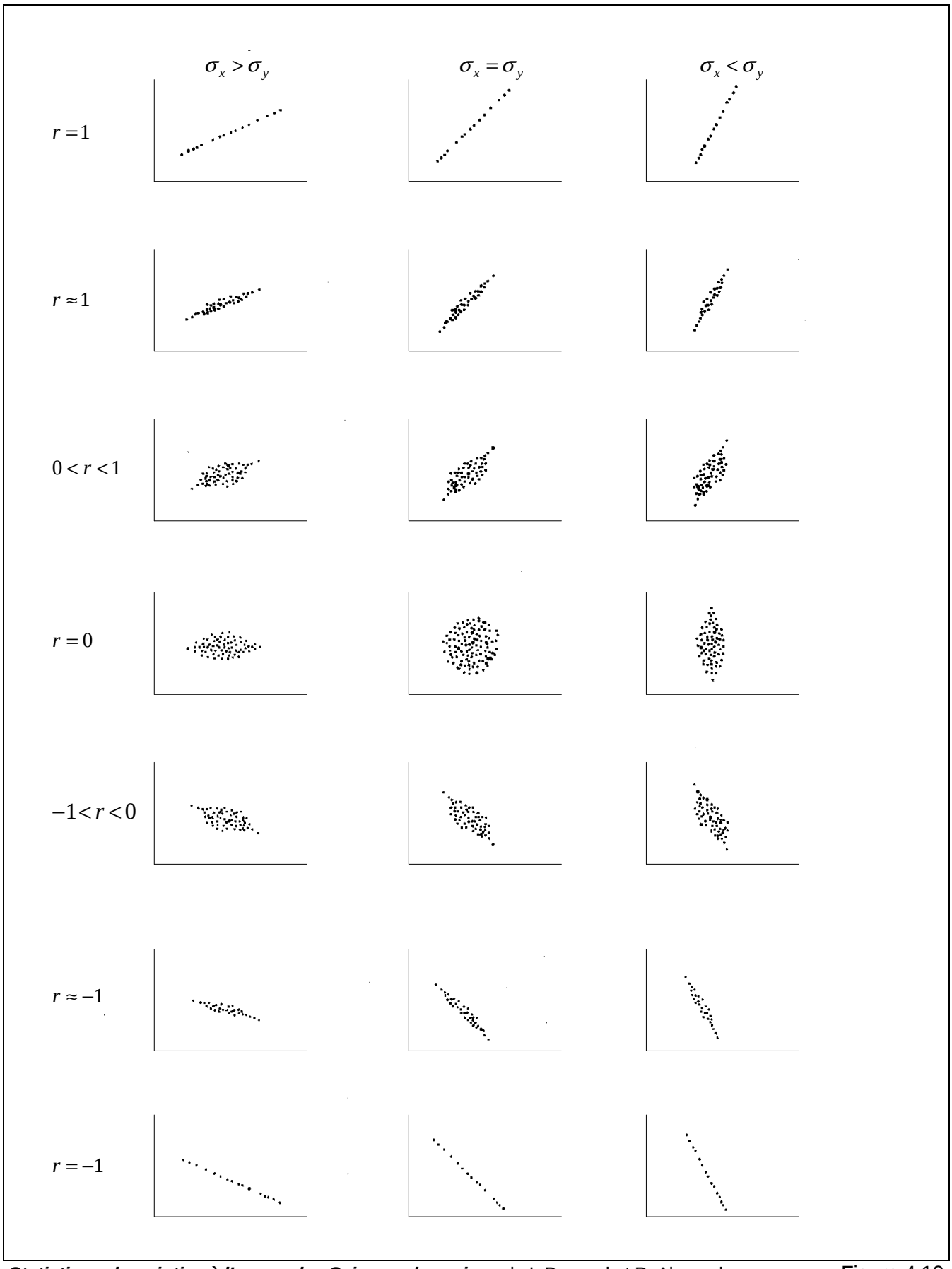
$$r^2 = a.a^*$$

En effet :

$$r = a\frac{\sigma_x}{\sigma_y}$$

$$a = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{n\sigma_y^2}$$

$$a^* = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{n\sigma_y^2}$$



Statistique descriptive à l'usage des Sciences humaines de L.Bragard et P. Alexandre Figure 4.10

et

$$r^2 = a.a^* \Leftrightarrow \quad a^2 \frac{\sigma_x^2}{\sigma_y^2} = a.a^*$$

$$\Leftrightarrow \quad a \frac{\sigma_x^2}{\sigma_y^2} = a^*$$

$$\Leftrightarrow \quad \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n\sigma_x^2} \cdot \frac{\sigma_x^2}{\sigma_y^2} = \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n\sigma_y^2}$$

$$\Leftrightarrow \quad \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n\sigma_y^2} = \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n\sigma_y^2}$$

En conséquence

1- a et a^* sont de même signe.

2- r , a et a^* sont de même signe puisque, par définition, r et a sont de même signe.

3- Le lien entre r , a et a^* permet de simplifier un peu le calcul des coefficients des droites de régression, en suivant le schéma suivant :

$$\left.\begin{array}{c} \overline{x}, \overline{y} \\ \sigma_x, \sigma_y \\ \sum_i x_i.y_i \\ n \end{array}\right\} \rightarrow a$$

$$\rightarrow d_{y/x} \equiv y - \overline{y} = a\left(x - \overline{x}\right)$$

$$\rightarrow r = a\frac{\sigma_x}{\sigma_y}$$

$$\rightarrow a^* = \frac{r^2}{a}$$

$$\rightarrow a' = \frac{1}{a^*}$$

$$\rightarrow d_{x/y} \equiv y - \overline{y} = a'\left(x - \overline{x}\right)$$

4.8- Un autre coefficient : le coefficient d'amélioration

Le coefficient d'amélioration A s'exprime en $\%$ et est défini par :

$$A = 100\left(1 - \sqrt{1 - r^2}\right).$$

Exemple

Si $r = 1$, alors $A = 100\,\%$ et la corrélation est parfaite ;

si $r = 0$, $A = 0\,\%$ et la corrélation est nulle ;

si $r = 0{,}9$; $A = 56\,\%$.