Chapitre 4

Étude d'une variable statistique à deux dimensions

Dans les chapitres précédents, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salaries en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit Ω une population et

$$\begin{aligned} Z\colon \Omega &\to \mathbb{R}^2, \\ w &\mapsto Z(w) = (X(w), Z(w)), \end{aligned}$$

ou directement

$$\begin{aligned} (X,Y)\colon \Omega &\to \mathbb{R}^2, \\ w &\mapsto (X(w), Z(w)). \end{aligned}$$

Dans ce cas, Z est dite variable statistique à deux dimensions avec $\mathrm{Card}(\Omega) = N$, avec N un entier fini. Le couple (X, Y) est appelé le couple de la variable statistique.

Exemple 20

- *On observe simultanément sur un échantillon de* 200 *foyers, le nombre d'enfants X et le nombre de chambre Y.*
- *On observe sur un échantillon de* 20 *foyers, le revenu mensuel X en Da et les dépenses mensuelles Y.*
- *Au près des étudiants pris au hasard parmi une section de L2 génie civil, on*

observe les notes de math3 X et de statistique Y.
- *Une entreprise mène une étude sur la liaison entre les dépenses mensuelles en publicité X et le volume des ventes Y qu'elle réalise.*

4.1 Représentation des séries statistiques à deux variables

Les séries statistiques à deux variables peuvent être présentées de deux façons.

Présentation 1

A chaque w_i, on associé (x_i, y_i), c'est à dire,

$$w_i \longrightarrow (x_i, y_i).$$

On rassemblera les données comme dans le tableau suivant

w_i	w_1	w_2	...	w_N
Variable X	$X(w_1)$	$X(w_2)$	...	$X(w_N)$
Variable Y	$Y(w_1)$	$Y(w_2)$	...	$Y(w_N)$

Cette représentation on la notera "présentation 1". Nous allons utiliser toujours les notations suivantes :

$$x_i := X(w_i)$$

et $y_i := Y(w_i)$.

Exemple 21
Soit Ω l'ensemble de 8 étudiants. Nous avons le tableau suivant

w_i	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$X(w)$	8	2	6	6	11	10	7	2
$Y(w)$	9	10	11	7	14	16	12	5

avec X représente le nombre d'heures passées à préparer l'examen de statistique par étudiant et Y représente la note sur 20 obtenue à l'examen par l'étudiant.

Lors de cette représentation, nous pouvons traduire le tableau associe dans une figure appelée "le nuage de points" ou "diagramme de dispersion" (voir Figure 4.1). Cette représentation est obtenue en mettant dans un repère cartésien chaque couple d'observation (x_i, y_j) par un point.

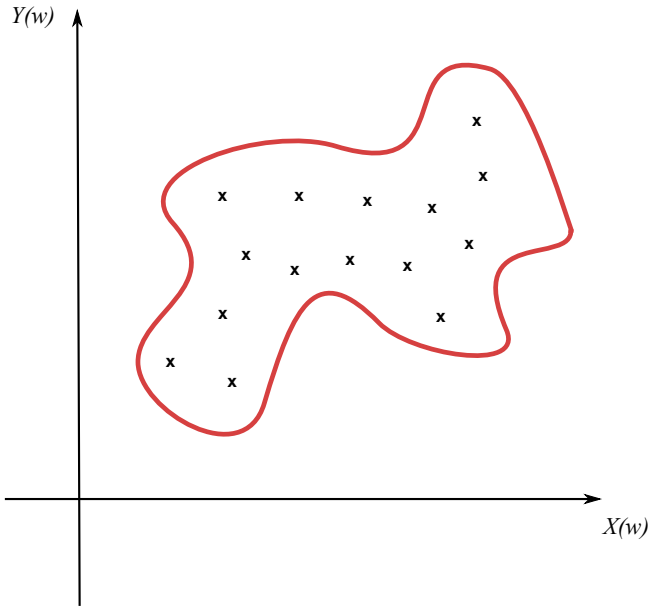


FIGURE 4.1: Représentation sous forme de nuage de points.

Présentation 2

Soit la variable statistique Z donnée par le couple (X, Y). Soient x_1, ..., x_k et y_1,...,y_l les valeurs prises respectivement par X et Y. Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à k et pour j allant de 1 à l,

$$z_{ij} := (x_i, y_j).$$

La variable statistique Z prend $k \times l$ valeurs. Lors de cette étude, nous avons le tableau à double entrée (ou tableau de contingence) suivant (discrète ou continue)

X \ **Y**	$C'_1 = [L'_1, L'_2[$ **ou** y_1	...	$C'_l = [L'_l, L'_{l+1}[$ **ou** y_l	**Marginale** % à **X**
$C_1 = [L_1, L_2[$ **ou** x_1	n_{11} **ou** f_{11}	...	n_{1l} **ou** f_{1l}	$n_{1\bullet}$ **ou** $f_{1\bullet}$
$C_2 = [L_2, L_3[$ **ou** x_2	n_{21} **ou** f_{21}	...	n_{2l} **ou** f_{2l}	$n_{2\bullet}$ **ou** $f_{2\bullet}$
$C_3 = [L_3, L_4[$ **ou** x_3	n_{31} **ou** f_{31}	...	n_{3l} **ou** f_{3l}	$n_{3\bullet}$ **ou** $f_{3\bullet}$
$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$
$C_k = [L_k, L_{k+1}[$ **ou** x_k	n_{k1} **ou** f_{k1}	...	n_{kl} **ou** f_{kl}	$n_{k\bullet}$ **ou** $f_{k\bullet}$
Marginale % à **Y**	$n_{\bullet 1}$ ou $f_{\bullet 1}$	...	$n_{\bullet l}$ ou $f_{\bullet l}$	N

Cette représentation on l'a notera "présentation 2". A chaque couple (x_i, y_i), on a n_{ij} est l'effectif qui représente le nombre d'individus qui prennent en même temps la valeur x_i et y_i, c'est à dire,

$$n_{ij} := \text{Card}\{w \in \Omega \ : Z(w) = z_{ij}\}.$$

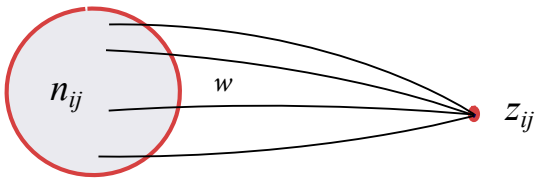


FIGURE 4.2: Le nombre d'individus qui prennent en même temps la valeur x_i et y_i.

Nous notons par f_{ij} la fréquence du coulpe (x_i, y_i). Cette fréquence est donnée par

$$f_{ij} := \frac{n_{ij}}{N},$$

avec

$$\begin{aligned} N &= \text{Card}(\Omega), \\ &= \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}, \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}. \end{aligned}$$

Le calcul ou le développement de cette double série est donné par

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} &= n_{11}+n_{12}+n_{13}+...+n_{1l}\\ &\quad +n_{21}+n_{22}+n_{23}+...+n_{2l}\\ &\quad \ddots \quad \ddots \quad \ddots \quad \ddots\\ &\quad +n_{k1}+n_{k2}+n_{k3}+...+n_{kl}.\end{aligned}$$

Remarque 16

Nous avons la propriété suivante,

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1.$$

Lois marginales

Sur la marge du tableau de contingence, on peut extraire les données seulement par rapport à X et seulement par rapport à Y (voir le tableau de contingence établi auparavant).

1. Effectifs et fréquences marginale par rapport à Y : nous avons, pour $j = 1...l$,

$$n_{\bullet j} := \sum_{i=1}^{k} n_{ij},$$

et

$$f_{\bullet j} := \frac{n_{\bullet j}}{N} = \sum_{i=1}^{k} f_{ij}.$$

2. Effectifs et fréquences marginale par rapport à X : nous avons, pour $i = 1...k$,

$$n_{i\bullet} := \sum_{j=1}^{l} n_{ij},$$

et

$$f_{i\bullet} := \frac{n_{i\bullet}}{N} = \sum_{j=1}^{l} f_{ij}.$$

Remarque 17
Nous avons les propriétés suivantes

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = N \qquad et \qquad \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1.$$

Exercice 23
Nous considérons 10 *salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes (pas encore traitées ou façonnées) sont données dans le tableau suivant,*

Salaire	6000	7400	7500	8200	8207	8900	9100	9900	9950	10750
Age	*15*	*26*	*20*	*43*	*47*	*37*	*52*	*34*	*50*	*44*

1. *Déterminer le tableau de contingence (X : âge, Y : salaire). Pour l'âge et pour le salaire, former respectivement des classes de pas de* 10 *ans et de* 1000 *Da.*
2. *Calculer* f_{21}, f_{12}, f_{45} *et* f_{33}.
3. *Déterminer les effectifs marginaux de* X *et de* Y. *Tracer le nuages de points.*
4. *Déterminer le tableau statistique des deux séries marginales* X *et* Y.

Solution : En utilisant les hypothèses, nous considérons les classes suivantes,

$$[15, 25[,\ [25, 35[,\ [35, 45[,\ [45, 55[,$$

pour l'âge et

$$[6, 7[,\ [7, 8[,\ [8, 9[,\ [9, 10[,\ [10, 11[,$$

pour le salaire ($\times 1000$). De plus, nous avons

$$Nombre\ de\ classe = \frac{e}{a_{\hat{a}ge}} = \frac{x_{\max} - x_{\min}}{a_{\hat{a}ge}} = \frac{52 - 15}{10} = 3.7 \simeq 4\ classes,$$

pour l'âge et

$$Nombre\ de\ classe = \frac{e}{a_{sal}} = \frac{y_{\max} - y_{\min}}{a_{sal}} = \frac{10750 - 6000}{1000} = 4.75 \simeq 5\ classes,$$

pour le salaire. Cette série statistique est représentée par le tableau suivant,

Age \ Salaire	$[6,7[$	$[7,8[$	$[8,9[$	$[9,10[$	$[10,11[$	$n_{i\bullet}$	$f_{i\bullet}$
$[15,25[$	1	1	0	0	0	0	0.2
$[25,35[$	0	1	0	1	0	2	0.2
$[35,45[$	0	0	2	0	1	3	0.3
$[45,55[$	0	0	1	2	0	3	0.3
$n_{\bullet j}$	1	2	3	3	1	10	1
$f_{\bullet j}$	0.1	0.2	0.3	0.3	0.1	1	$\varnothing$

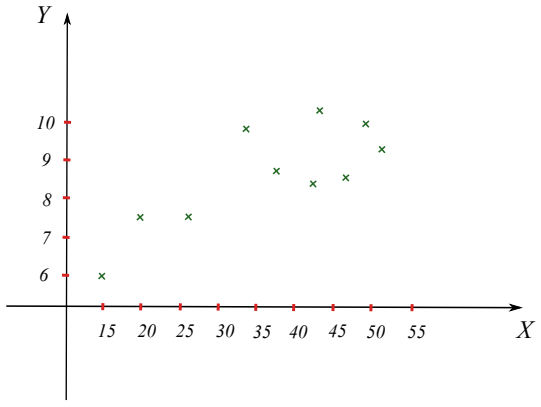
De ce fait, nous avons

$$f_{12} = \frac{n_{12}}{N} = \frac{1}{10} = 0.1, \qquad f_{21} = \frac{n_{21}}{N} = \frac{0}{10} = 0,$$

et

$$f_{45} = \frac{n_{45}}{N} = \frac{0}{10} = 0, \qquad f_{33} = \frac{n_{33}}{N} = \frac{2}{10} = 0.2.$$

Le nuage de points est tracé, à partir des données brutes, dans la figure suivante.



Enfin, les deux tableaux statistiques de X et de Y sont donnés, respectivement, par

X	$n_{i\bullet}$	$f_{i\bullet}$	x_i le centre
$[15, 25[$	2	0.2	20
$[25, 35[$	2	0.2	30
$[35, 45[$	3	0.3	40
$[45, 55[$	3	0.3	50

Y	$n_{i\bullet}$	$f_{i\bullet}$	y_i le centre
$[6, 7[$	1	0.1	6.5
$[7, 8[$	2	0.2	7.5
$[8, 9[$	3	0.3	8.5
$[9, 10[$	3	0.3	9.5
$[10, 11[$	1	0.1	10.5

4.2 Description numérique

4.2.1 Caractéristique des séries marginales

Dans le cas d'une variable statistique à deux dimensions X et Y, les moyennes sont données respectivement par

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i \quad \text{(moyenne } de\ X\text{)},$$

et

$$\overline{y} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} f_{\bullet j}y_j \quad \text{(moyenne } de\ Y\text{)}.$$

Remarque 18
Dans le cas continu, x_i et y_j représentent respectivement le centre des classes de X et Y, c'est à dire,

$$x_i = \frac{L_{i+1} + L_i}{2} \qquad et \qquad y_j = \frac{L_{j+1} + L_j}{2}.$$

Exemple 22
Nous calculons $\overline{x}$ et $\overline{y}$ pour l'exercice traité précédemment. Nous avons la moyenne d'âge

$$\overline{x} = \frac{1}{10}(40 + 60 + 120 + 150) = 37 \ ans.$$

et la moyenne du salaire

$$\overline{y} = \frac{1}{10}(6.5 + 15 + 25.5 + 28.5 + 10.5) \times 100 = 8600 \ Da.$$

Nous définissions maintenant la variance de X et la variance de Y comme suit,

$$Var(X) := \overline{x^2} - (\overline{x})^2, \qquad \text{avec} \quad \overline{x^2} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2,$$

et

$$Var(Y) := \overline{y^2} - (\overline{y})^2, \qquad \text{avec} \quad \overline{y^2} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j^2 = \sum_{j=1}^{l} f_{\bullet j}y_j^2.$$

Les écarts-type de X et de Y sont donnés, respectivement, par

$$\sigma_X := \sqrt{Var(X)} \qquad \text{et} \qquad \sigma_Y := \sqrt{Var(Y)}.$$

4.2.2 Série conditionnelle

La notion de série conditionnelle est essentielle pour comprendre l'analyse de la régression. Un tableau de contingence se compose en autant de séries conditionnelles suivant chaque ligne et chaque colonnes.

Série conditionnelle par rapport à X

Elle est notée par X/y_j (ou X_j) et on dit que c'est la série conditionnelle de X sachant que $Y = y_j$. Nous calculons dans ce cas la fréquence conditionnelle $f_{i/j}$ (f_i sachant j), pour $i = 1, ..., k$, par

$$f_{i/j} := \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

Nous avons aussi la moyenne conditionnelle $\overline{x}_j$, c'est à dire la moyenne des valeurs de X sous la condition y_j, elle est définie par

$$\overline{x}_j := \sum_{i=1}^{k} f_{i/j}x_i = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}x_i.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{X_j} := \sqrt{Var(X_j)}$ avec

$$Var(X_j) := \sum_{i=1}^{k} f_{i/j}(x_i - \overline{x}_j)^2 = \overline{x^2}_j - (\overline{x}_j)^2.$$

Série conditionnelle par rapport à Y

Elle est notée par Y/x_j (ou Y_j) et on dit que c'est la série conditionnelle de Y sachant que $X = x_i$. Nous calculons aussi dans ce cas la fréquence conditionnelle $f_{j/i}$ (f_j sachant i), pour $j = 1, ..., l$, par

$$f_{j/i} := \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

Nous avons aussi la moyenne conditionnelle $\overline{y}_i$, c'est à dire la moyenne des valeurs de Y sous la condition x_i, elle est définie par

$$\overline{y}_i := \sum_{j=1}^{l} f_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j \ .$$

Pour l'écart-type conditionnel, nous avons $\sigma_{Y_i} := \sqrt{Var(Y_i)}$ avec

$$Var(Y_i) := \sum_{j=1}^{l} f_{j/i} (y_j - \overline{y}_i)^2 = \overline{y^2}_i - (\overline{y}_i)^2.$$

4.2.3 Notion de covariance

Nous notons par $Cov(X, Y)$ la covariance entre les variables X et Y. La covariance est un paramètre qui donne la variabilité de X par rapport à Y (voir Figure 4.3).

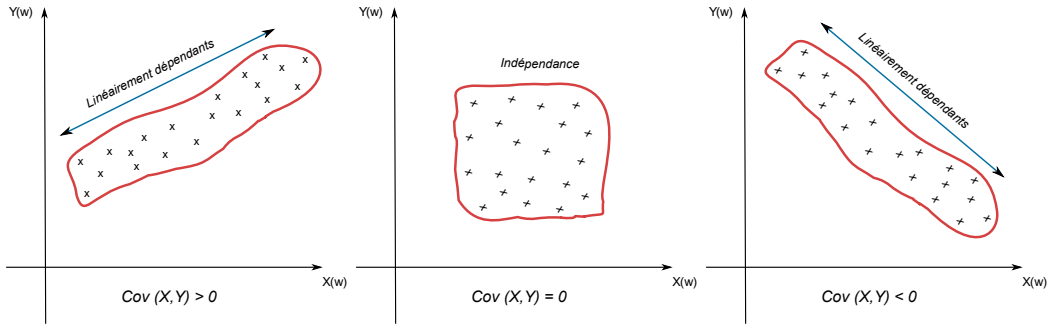


FIGURE 4.3: La covariance et la variabilité.

La covariance se calcule par l'expression suivante

$$Cov(X, Y) = \overline{xy} - \overline{x}\ \overline{y} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j - \overline{x}\ \overline{y}.$$

Nous avons aussi cette formule

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y}).$$

Remarque 19

Dans le cas où nous avons un tableau des données brutes "representation 1" (nous n'avons pas d'effectifs), nous avons les formules suivantes

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad et \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i.$$

De plus, nous avons

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i.$$

Remarque 20

La covariance est une notion qui généralise la variance, En effet,

$$Cov(X,X) = Var(X) \qquad et \qquad Cov(Y,Y) = Var(Y).$$

Cela provient de la définition, c'est à dire,

$$Cov(X,X) = \overline{xx} - \overline{x}\ \overline{x} = \overline{x^2} - \overline{x}^2 = Var(X).$$

Définition 25

On dit que deux variables statistiques X et Y sont indépendantes si et seulement si, pour tout i et j,

$$f_{ij} = f_{i\bullet} \times f_{\bullet j}.$$

Il suffit que cette égalité ne soit pas vérifiée dans une seule cellule pour que les deux variables ne soient pas indépendantes.. De manière équivalente, pour tout i et j,

$$N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}.$$

Dans ce cas, si X et Y sont indépendantes alors (réciproque est fausse) $Cov(X,Y) = 0$.

Cette définition donne une interprétation intéressante de d'indépendance ; elle signifie que dans ce cas, les effectifs des modalités conjointes peuvent se calculer uniquement à partir des distributions marginales, supposées « identiques » aux distributions de X et Y dans la population ; en d'autres termes, si X et Y sont indépendantes, les observations séparées de X et de Y donnent la même information qu'une observation conjointe.

4.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

4.3.1 Coefficient de corrélation

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires (voir ci-dessous). Il existe d'autres coefficients pour les relations non-linéaires et non-monotones, mais ils ne seront pas étudiés dans le cadre de ce cours.

Définition 26
La quantité

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

s'appelle le coefficient de corrélation.

Proposition 3
Le coefficient ρ_{XY} est compris entre $[-1,1]$, ou encore

$$| \rho_{XY} | \leq 1.$$

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y (voir Figure 4.4 et). Nous avons les deux caractéristiques suivantes (voir Figures 4.5 et 4.6) [1] :

1. Source : https ://en.wikipedia.org/wiki/Pearson_correlation_coefficient

– Plus le module de ρ_{XY} est proche de 1 plus X et Y sont liées linéairement.

– Plus le module de ρ_{XY} est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y.

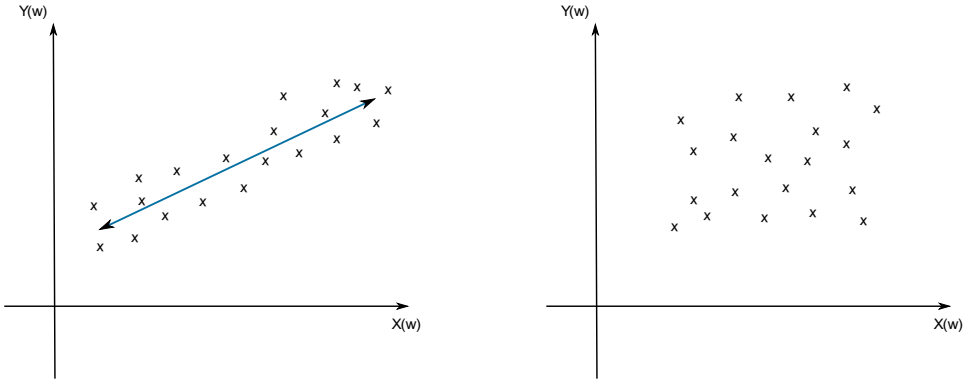


FIGURE 4.4: A gauche, le coefficient de corrélation est proche de 1. A droite, le coefficient de corrélation est proche de 0.

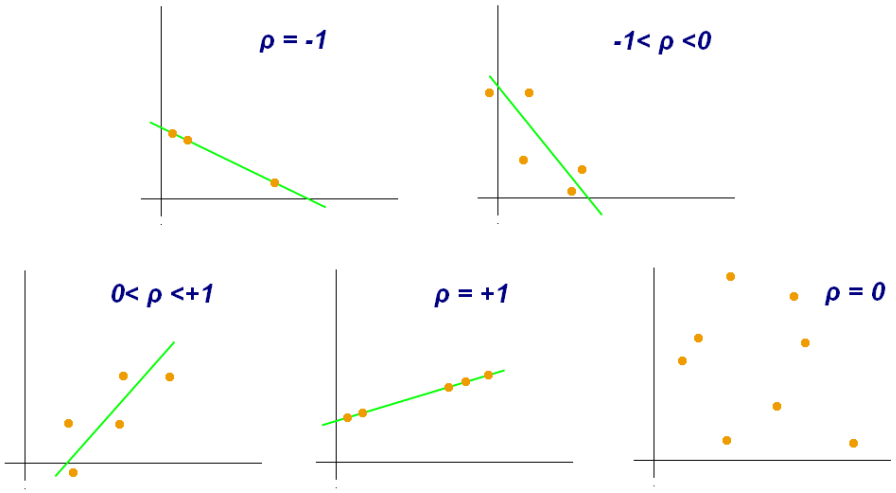


FIGURE 4.5: Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation .

Remarque 21
Par définition, si $\rho_{XY} = 0$, *alors* $Cov(X, Y) = 0$.

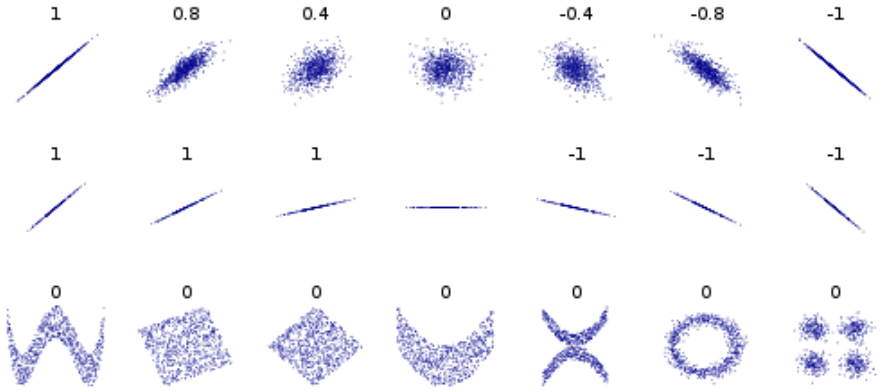


FIGURE 4.6: La corrélation reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (en bas). La figure au centre a une pente de 0, mais dans ce cas, le coefficient de corrélation est indéfini car la variance de Y est nulle. .

4.3.2 Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

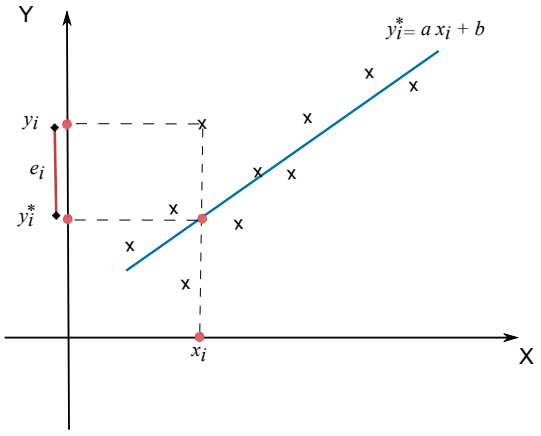


FIGURE 4.7: La droite la plus proche possible de chacun des points.

Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X, c'est à dire,

$$Y = aX + b,$$

telle que la distance entre le nuage de points et droite soit minimale. Cette distance matéria-

lise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédit par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des moindres carrés consiste à chercher la valeur des paramètres a et b qui minimise la somme des erreurs élevées au carré.

On pose

$$\sum_{i=1}^{n} e_i^2 = U(a,b),$$

avec e_i est l'erreur commise sur chaque observation, c'est à dire,

$$\mid e_i \mid = \mid y_i - y_i^\star \mid = \mid y_i - ax_i - b \mid .$$

La méthode des moindres carrées consiste donc à minimiser la fonction U (la somme des erreurs commises). Nous avons la condition de minimisation suivante,

$$\frac{\partial U}{\partial a} = \frac{\partial U}{\partial b} = 0,$$

avec

$$U(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

L'équation $\dfrac{\partial U}{\partial b} = 0$ donne

$$\sum_{i=1}^{n} -2(y_i - ax_i - b) = 0.$$

Ce qui implique que

$$\left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0 \right) \times \frac{1}{N}.$$

Par conséquent, nous obtenons

$$\overline{y} - a\overline{x} - b = 0,$$

c'est à dire,

$$\boxed{b = \overline{y} - a\overline{x}.}$$

De même, après calcule, $\dfrac{\partial U}{\partial a} = 0$ implique que

$$\boxed{a = \frac{Cov(X,Y)}{Var(X)}.}$$

Donc, la droite de régression, qui rend la distance entre elle et les points minimale, est

donnée par

$$D(Y/X): \ Y = aX + b,$$

avec

$$a = \frac{Cov(X,Y)}{Var(X)} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}.$$

Ou bien

$$D(X/Y): \ X = a'Y + b',$$

avec

$$a' = \frac{Cov(X,Y)}{Var(Y)} \qquad \text{et} \qquad b' = \overline{x} - a'\overline{y}.$$

Remarque 22

Le coefficient de corrélation ρ_{XY} permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir Figure 4.8),

- *Si $\mid \rho_{XY} \mid < 0.7$, alors l'ajustement linéaire est refusé (droite refusée).*
- *Si $\mid \rho_{XY} \mid \geq 0.7$, alors l'ajustement linéaire est accepté (droite acceptée).*

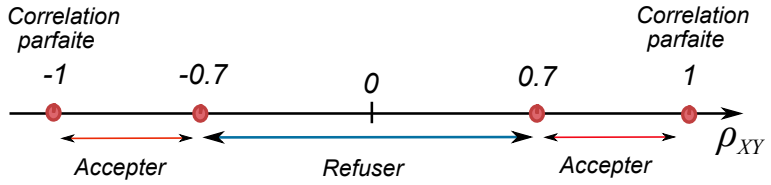


FIGURE 4.8: La zone d'acceptation ou de refus de l'ajustement linéaire.

4.4 Exercices corrigés

Exercice 24

Nous considérons 10 *joueurs et soient :*

- *Y la variable qui représente le nombre de jeux auquel un joueur joue.*
- *X la variable qui représente le gain ou perte ($+1$ s'il gagne* 10 *Da et -1 s'il perd* 10 *Da et* 0 *sinon).*

Nous avons le tableau de contingence suivant,

$\boldsymbol{X} \setminus \boldsymbol{Y}$	1	2	3	4	$n_{i\bullet}$
-1	*0*	*1*	*2*	*2*	
0	*1*	*1*	*0*	*1*	
1	*0*	*1*	*1*	*0*	
$n_{\bullet j}$					

1. *Compléter le tableau ci-dessus.*
2. *Calculer cov(X, Y).*

Solution Les lois marginales sont données dans ce tableau,

$\mathbf{X} \setminus \mathbf{Y}$	1	2	3	4	$n_{i\bullet}$
-1	0	1	2	2	**5**
0	1	1	0	1	**3**
1	0	1	1	0	**2**
$n_{\bullet j}$	**1**	**3**	**3**	**3**	**N=10**

La covariance est calculée à partir de

$$Cov(X, Y) = \overline{xy} - \overline{x}\ \overline{y}.$$

Nous avons

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{3} n_{i\bullet} x_i = -0.3,$$

et

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{4} n_{\bullet j} y_j = 2.8.$$

De plus, nous avons

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij} x_i y_j = -1.$$

Donc,

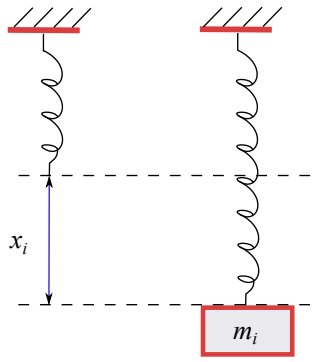
$$Cov(X, Y) = -0.16.$$

Exercice 25

Dans un TP de physique, on a les données suivantes :

x_i	0	0.5	1.1	1.5	1.9
m_i	0	10	20	30	40

La variable m_i représente les différentes masses appliquées comme dans le schéma ci-dessous et la variable x_i les hauteurs induits depuis l'état initial.



1. *Déterminer $D(m/x)$.*
2. *Déterminer $D(x/m)$.*
3. *Tracer le nuage de point et les deux droites. Représenter le point de coordonnée $(\overline{x}, \overline{y})$.*
4. *Peut-on déterminer x si $m = 51.75$ Kg ?*

Solution Nous déterminons facilement les moyennes $\overline{x} = 1$ et $\overline{m} = 1$. De plus, nous avons

$$Cov(x, m) = \overline{xm} - \overline{x}\ \overline{m} = 29.6 - 20 = 9.6.$$

Après calcul, nous avons aussi

$$Var(X) = \overline{x^2} - (\overline{x})^2 = 0.464 \qquad \text{et} \qquad Var(m) = 200.$$

Ce qui implique que

$$\sigma_x = 0.681 \qquad \text{et} \qquad \sigma_m = 14.14.$$

Dans ce cas, les coefficients de la droite sont donnés par

$$a = \frac{Cov(x, m)}{Var(X)} = 20.69 \qquad \text{et} \qquad b = \overline{m} - a\overline{x} = -0.69.$$

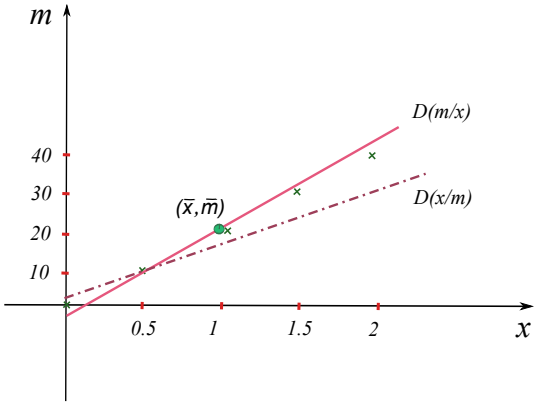
Par conséquent,

$$D(m/x) : \quad m = 20.69x - 0.69.$$

En renversant les axes, nous obtenons

$$D(x/m): \quad x = 0.048m + 0.04.$$

Nous traçons les deux droites dans la figure ci-dessous ($D(x/m)$ est la symétrie de $D(m/x)$ par rapport à la première bissectrice).



Si on calcule $\rho_{xm}= 0.997 > 0.7$, alors nous avons un ajustement linéaire accepté. Donc, on peut estimer x pour $m = 51.75$ et nous obtenons

$$x = 0.048 \times (51.75) + 0.04 = 2.52.$$

Exercice 26

- Le tableau de contingence suivant est entre le salaire mensuel X et l'ancienneté Y des ouvriers d'une entreprise.

$\boldsymbol{X}(\times 1000) \setminus \boldsymbol{Y}$	$[0,8[$	$[8,16[$	$[16,24[$	$[24,32[$	*Loi marginale*
$[20,30[$	*5*	*6*	*1*	*0*	
$[30,40[$	*2*	*4*	*3*	*3*	
$[40,50[$	*0*	*2*	*4*	*10*	
Loi marginale					

1. *Etudier les séries marginales.*
2. *Déterminer si les variables X et Y sont indépendantes.*
3. *Etudier les séries conditionnelles X/y_3 et Y/x_2 et présenter les résultats pour chaque groupe de séries conditionnelles.*

Solution *1- Nous complétons le tableau comme suit*

$\boldsymbol{X}(\times 1000) \setminus \boldsymbol{Y}$	$[0,8[$	$[8,16[$	$[16,24[$	$[24,32[$	$n_{i\bullet}$	$f_{i\bullet}$
$[20,30[$	*5*	*6*	*1*	*0*	***12***	***0.3***
$[30,40[$	*2*	*4*	*3*	*3*	***12***	***0.3***
$[40,50[$	*0*	*2*	*4*	*10*	***16***	***0.4***
$n_{\bullet j}$	***7***	***12***	***8***	***13***	***40***	***1***
$f_{\bullet j}$	***7/40***	***12/40***	***8/40***	***13/40***	***1***	$\varnothing$

Les moyennes après le calcule

$$\overline{x} = 36 \ (\times 1000),$$

et

$$\overline{y} = 17.4.$$

La variance et l'écart type de X

$$Var(X) = 69 \quad et \quad \sigma_X = 8.310.$$

La variance et l'écart type de Y

$$Var(Y) = 78.04 \quad et \quad \sigma_Y = 8.84.$$

2 - Si on choisit $i = 3$ *et* $j = 1$*, nous obtenons*

$$N \times n_{31} = 40 \times 0 = 0,$$

et

$$n_{3\bullet} \times n_{\bullet 1} = 16 \times 7 = 112,$$

qui sont bien évidement non égaux. Par conséquent, il existe i *et* j *tel que*

$$N \times n_{ij} \neq n_{i\bullet} \times n_{\bullet j}.$$

Donc, X *et* Y *ne sont pas indépendants.*

3 - La série X/y_3 *est la série conditionnelle de* X *sachant que* $Y = y_3$ *(*$j = 3$*). Sa moyenne*

est donnée par

$$\overline{x}_3 = 38.75.$$

La série Y/x_2 est la série conditionnelle de Y sachant que $X = x_2$ $(i = 2)$. Sa moyenne est donnée par

$$\overline{y}_2 = 16.67.$$

4.5 Exercices supplémentaires

Exercice 27

- Une usine produit des pièces d'une machine. Pour chaque pièce (individu), on dispose du coût de sa production (DA) et du temps nécessaire pour sa réalisation (en heures). Le tableau ci-après (série statistique) donne cette répartition :

Individu	*1*	*2*	*3*	*4*	*5*
Temps (X) mesuré en heures	*2*	*3*	*52*	*2*	*4*
Coût (Y) mesuré en Dinars	*10*	*16*	*23*	*12*	*18*

On donne

$$- \ Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right]. \qquad - \ Coefficient\ de\ corrélation = \frac{Cov(X,Y)}{\sigma(X)\ \sigma(Y)}.$$

$$- \ Droite\ de\ corrélation\ linéaire : \quad Y = \overline{y} - \frac{Cov(X,Y)}{Var(X)}\overline{x} + \frac{Cov(X,Y)}{Var(X)}X.$$

1. *Calculer la moyenne de la variable statistique X.*
2. *Calculer la moyenne de la variable statistique Y.*
3. *Calculer l'écart-type de la variable statistique X.*
4. *Calculer l'écart-type de la variable statistique Y.*
5. *Calculer la covariance des variable statistiques X et Y.*
6. *En supposant qu'il existe une corrélation linéaire entre X et Y, déterminer cette droite de corrélation.*
7. *Calculer le coefficient de corrélation. Conclusion ?*
8. *Une nouvelle pièce est réalisée en 6 heures. Estimer le coût de production de cette pièce en utilisant la droite de corrélation établie.*

Exercice 28

- Pour les données suivantes

X	1	2	7	4	6
Y	5	4	1	3	2

1. *Tracer le nuage de points.*
2. *Deviner le signe et la valeur du coefficient de corrélation.*
3. *Calculer le coefficient de corrélation, la pente et l'ordonnée à l'origine de la droite de régression.*

Exercice 29

- Soit X et Y deux variables statistiques mesurées sur un même individu. Par exemple, pour l'individu n°3, $X = 2$ et $Y = 8$.

Individu	1	2	3	4	5
X	3	4	2	5	3
Y	12	14	8	19	11

1. *Calculer la moyenne de la variable statistique X.*
2. *Calculer la moyenne de la variable statistique Y.*
3. *Calculer l'écart-type de la variable statistique X.*
4. *Calculer l'écart-type de la variable statistique Y.*
5. *Calculer la covariance des variable statistiques X et Y.*
6. *En supposant qu'il existe une corrélation linéaire entre X et Y, déterminer cette droite de corrélation.*
7. *Calculer le coefficient de corrélation. Conclusion ?*

Exercice 30

- On vous demande s'il existe une corrélation entre la population de chamois[2] *dans une commune et le nombre de permis de chasse enregistré par l'association de chasse locale.*

2. Le chamois est une sorte de chèvre des montagnes caractérisé par ses petites cornes en crochets.

Années	1999	2000	2001	2002	2003	2004
Chamois	3200	3650	3430	3890	4200	4350
Permis	202	231	240	225	245	263

Travail à faire :

- *Calculer le coefficient de corrélation entre ces deux séries.*
- *Tracer la droite d'ajustement.*

Exercice 31

- Une étude sur le chômage a été faite et qui s'intéresse à l'ancienneté du chômage (X) *moins de 24 mois, et l'âge* (Y) *entre* 20 *et* 35 *ans. Les résultats sont donnés par le tableau de contingence suivant :*

$\boldsymbol{X} \setminus \boldsymbol{Y}$	$[20, 25[$	$[25, 30[$	$[30, 35[$
$[0, 6[$	*10*	*8*	*5*
$[6, 12[$	*8*	*9*	*4*
$[12, 18[$	*15*	*11*	*9*
$[18, 24[$	*3*	*6*	*2*

1. *Quel est le nombre d'individus qui ont une ancienneté de chômage moins d'un an ?*
2. *Déterminer les deux distributions marginales.*
3. *Déterminer la distribution de* X *conditionnelle à* $Y = [25, 30]$*, c'est à dire,* $X/Y = [25, 30]$*.*
4. *Les variables* X *et* Y *sont elles indépendantes ? Justifier.*
5. *Donner la moyenne arithmétique.*
6. *Calculer le coefficient de corrélation linéaire. Commenter.*
7. *Donner l'équation de la droite de régression de* Y *en fonction de* X*.*
8. *Quel sera l'âge d'une personne ayant une ancienneté de chômage de* 15 *mois.*

Exercice 32

- On fait une étude statistique sur 10 sites de commerce électronique, ayant pour but de sonder sur une semaine le nombre de visiteurs et le nombre de commandes. On obtient le tableau suivant :

Le numéro du site (i)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Le nombre de connexion (x_i)	*80*	*100*	*115*	*110*	*70*	*125*	*105*	*90*	*110*	*95*
Le nombre de commandes (y_i)	*32*	*50*	*62*	*56*	*8*	*80*	*62*	*50*	*62*	*38*

1. *Calculer les moyennes arithmétiques de la variable statistique* X *et de la variable statistique* Y.
2. *Calculer les écarts-type de la variable statistique* X *et de la variable statistique* Y.
3. *Calculer la covariance entre* X *et* Y.
4. *Calculer le coefficient de corrélation linéaire entre* X *et* Y. *Commenter.*
5. *Déterminer la droite de corrélation* $Y = aX + b$.