

Université Mohamed Premier
Faculté des Sciences Juridiques,
Économiques et Sociales d'Oujda

STATISTIQUE DESCRIPTIVE

Chapitre 2 : Analyse d'une série de donnée bivariée

Première année sciences économiques et gestion
Semestre 1

Année Universitaire :
2021-2022

Sommaire

1. INTRODUCTION	3
2. ETUDE DES LIAISONS STATISTIQUES POUR DES DONNEES QUANTITATIVES	4
2.1. ANALYSE GRAPHIQUE	4
2.2. PARAMETRES DES DISTRIBUTIONS	6
2.2.1 <i>Paramètres des distributions marginales</i>	6
2.2.2 <i>Paramètres qui expriment la liaison</i>	8
a- <i>La covariance</i>	8
b- <i>Coefficient de corrélation linéaire de Pearson</i>	9
2.3. PREVISION DANS LE CAS D'UNE FORTE CORRELATION	10
2.3.1 <i>Coefficient de détermination</i>	10
2.3.2 <i>Ajustement linéaire par la méthode des moindres carrés</i>	12
2.3.3 <i>Ajustement non linéaire</i>	15
2.4. CAS DE DONNEES REGROUPEES EN CLASSES	16
2.4.1 <i>Distribution marginale</i>	18
2.4.2 <i>Distribution conditionnelle</i>	21
2.4.3 <i>Covariance et corrélation</i>	25

1. Introduction

But de l'étude de deux séries statistiques :

Chercher une *dépendance* (relation possible) et quel est le modèle qui explique la relation

Ou dire qu'il y a *indépendance*

Exemples

↘ Entre variables quantitatives :

- Existe-t-il un lien entre le taux de chômage et le taux d'inflation ?
- Peut-on expliquer le résultat d'un examen par le nombre d'heures de préparation ?
- Le prix d'un produit final dans nos magasins dépend-il des prix des matières premières le composant ?
- L'hypertension artérielle est-elle liée à l'âge ou au poids de l'individu ?

↘ Entre variables quantitatives et qualitatives :

- Existe-t-il un lien entre la consommation d'un produit et le sexe des consommateurs ?
- Peut-on lier le chiffre d'affaire réalisé par la vente d'un produit et les capacités de négociation des représentants ?
- Existe-t-il un lien entre le rendement scolaire mesuré sur une échelle de nombres réels et fumer le tabac ?

↘ Entre variables qualitatives :

- Pour la consommation d'un produit, on veut savoir si Existe-t-il un lien entre l'appréciation ou l'évaluation d'un produit et le sexe des clients ?
- Existe-t-il un lien entre la guérison d'un malade et le traitement médicale qu'il a obtenu ?
- Ou peut être y a-t-il une liaison entre la réponse à un traitement médicale et l'âge (enfant--adulte) ou le sexe du malade ?
- Existe-t-il un lien entre le rendement scolaire (mesuré sur des modalités qualitatives) et fumer le tabac ?

Mais... il faut faire attention :

La question de la liaison entre deux variables est différente de la question du sens de la causalité.

Exemple :

- Est-ce qu'on doit expliquer le poids d'un individu par sa taille ou dire que la taille dépend du poids ?
- Est-ce le prix qui détermine la demande ou la demande qui explique le niveau des prix?

Dans cet exposé on se limitera à l'observation de deux variables quantitatives.

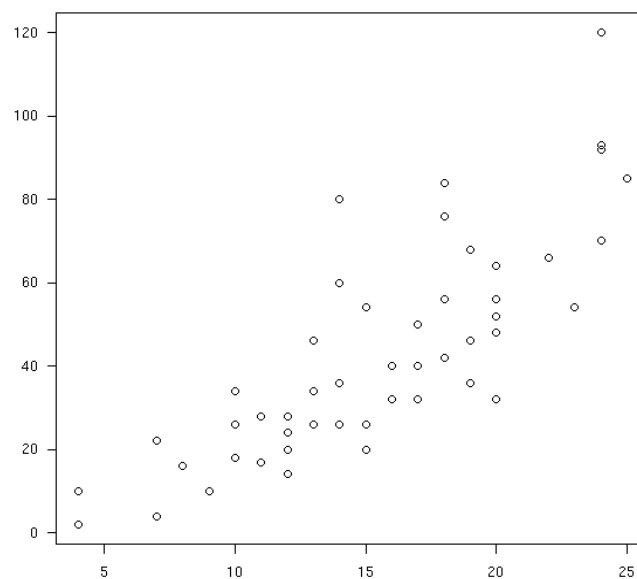
2. Etude des liaisons statistiques pour des données quantitatives

2.1. Analyse graphique

Tableau des données

x	x_1	x_2		x_i		x_n
y	y_1	y_2		y_i		y_n

Nuage de points



Existe-t-il une liaison entre les deux séries ?

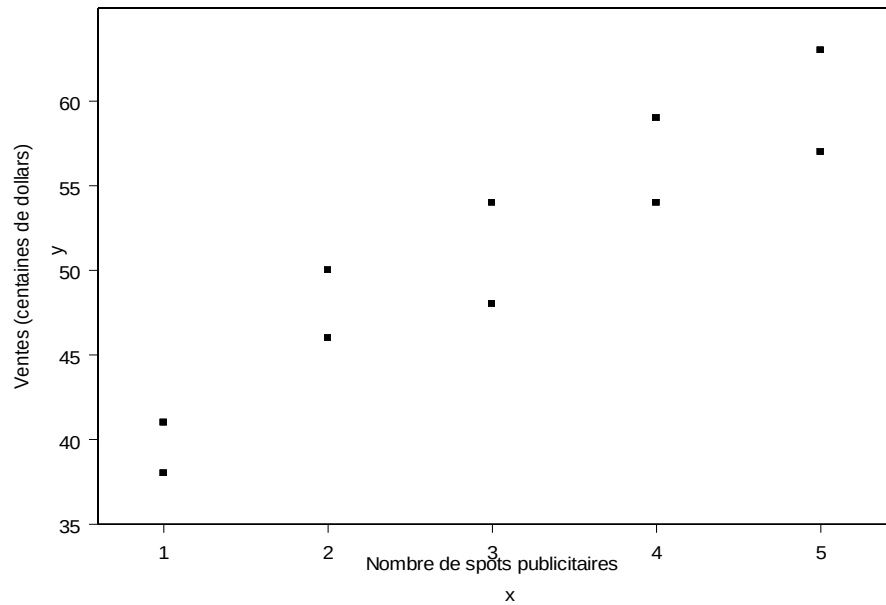
Un accroissement ou un décroissement d'une unité de la 1^{ère} série provoque-t-il un changement dans la 2^{ème} série ?

Exemple

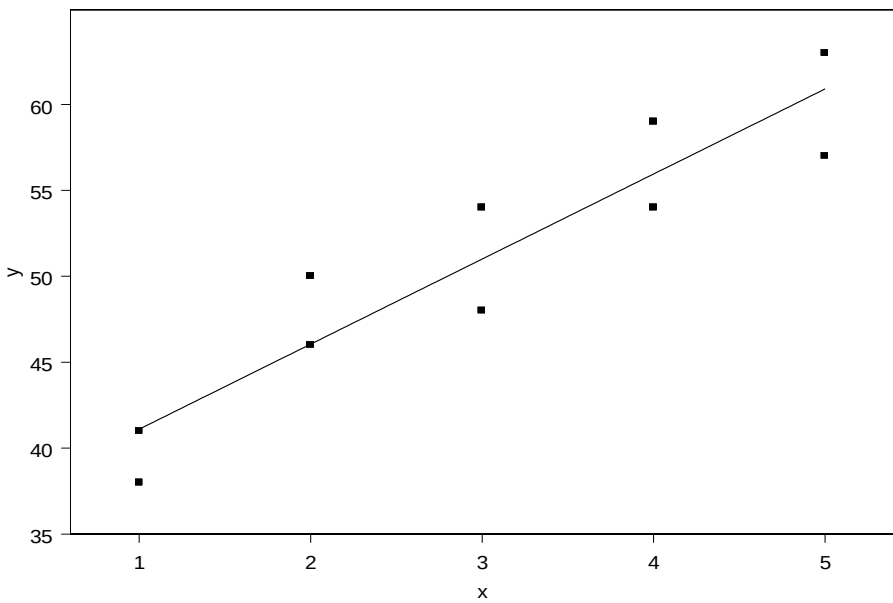
Tableau des données

Semaine	1	2	3	4	5	6	7	8	9	10
Nombre de spots publicitaires x	2	5	1	3	4	1	5	3	4	2
Volume des ventes (centaines de dollars) y	50	57	41	54	54	38	63	48	59	46

Analyse graphique : nuage de points



Ajustement linéaire de l'exemple

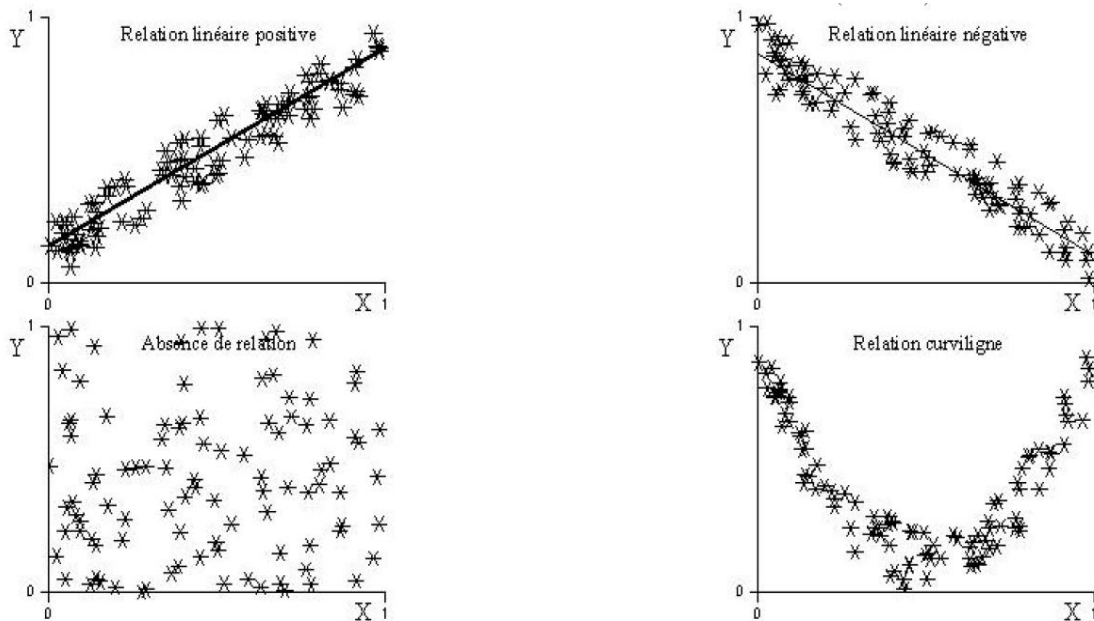


Une représentation graphique du nuage de points permet :

1. D'apprécier visuellement l'existence ou non d'une éventuelle liaison
2. De déterminer la forme de la liaison

Nuages de points : formes de liaison

Peut-on avoir une écriture $y = f(x)$



2.2. Paramètres des distributions

Les paramètres utilisés pour caractériser les distributions à deux variables sont de deux types,

- Les paramètres qui concernent une seule variable, ils servent à caractériser les diverses distributions marginales.
- Les paramètres qui servent à décrire les relations qui existent entre les deux séries d'observations considérées simultanément.

2.2.1 Paramètres des distributions marginales

Travailler avec chaque série séparément pour calculer les moyennes et les variances marginales

Exemple : x : nombre de spots publicitaires y : volume des ventes (centaines de dollars)

$$\bar{x} = \frac{30}{10} = 3$$

$$\bar{y} = \frac{510}{10} = 51$$

Semaine	Nombre de spots publicitaires x	Volume des ventes (centaines de dollars) y	x^2	y^2
1	2	50	4	2500
2	5	57	25	3249
3	1	41	1	1681
4	3	54	9	2916
5	4	54	16	2916
6	1	38	1	1444
7	5	63	25	3969
8	3	48	9	2304
9	4	59	16	3481
10	2	46	4	2116
Somme	30	510	110	26576
Moyenne	3	51	11	2657.6

Les variances sont

$$Var(x) = 11 - 9 = 2 \text{ et } Var(y) = 2657.6 - 51^2 = 56.6$$

Les écart-types sont

$$\sigma(x) = 1.41 \text{ et } \sigma(y) = 7.52$$

2.2.2 Paramètres qui expriment la liaison

Ces paramètres permettent de décrire les relations qui peuvent exister entre les deux séries :

- Covariance
- Coefficient de corrélation
- Coefficient de détermination
- Coefficient de Khi-deux (on va le voir pour les variables qualitatives)

a- La covariance

Pour avoir une idée sur la variation simultanée de x et y on peut utiliser la covariance de x et y :

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

Elle est aussi égal à :

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \quad (\text{formule de K oenig})$$

Le mot « covarier » d esigne varier ensemble.

Utilit  de la covariance

La covariance est positive si x et y ont tendance   varier dans le m me sens, et n gative si elles ont tendance   varier en sens contraire.

Plus la covariance s'approche de z ro plus on dit qu'il n'y a pas de liaison entre x et y

Exemple :

Semaine	Nombre de spots publicitaires x	Volume des ventes (centaines de dollars) y	xy
1	2	50	100
2	5	57	285
3	1	41	41
4	3	54	162
5	4	54	216
6	1	38	38
7	5	63	315
8	3	48	144
9	4	59	236
10	2	46	92
Somme	30	510	1629
Moyenne	3	51	162.9

On calcule la covariance dans l'exemple pour trouver

$$Cov(x, y) = 162.9 - 3 * 51 = 9.9$$

La covariance est positive, donc il y a une relation croissante entre x et y . Autrement dit les deux variables varient dans le même sens.

b- Coefficient de corrélation linéaire de Pearson

$$r(x, y) = corr(x, y) = \frac{cov(x, y)}{\sigma(x) * \sigma(y)}$$

avec $cov(x, y)$ la covariance entre x et y , $\sigma(x)$ est l'écart-type de x et $\sigma(y)$ celui de y .

On montre que :

$$|cov(x, y)| \leq \sigma(x) * \sigma(y)$$

- $r(x, y)$ est compris entre -1 et 1
- Plus il s'approche de 1 ou de -1 on dit qu'il y a une forte liaison linéaire entre les deux caractères
- Au voisinage de 1 l'accroissement est de même sens
- Au voisinage de -1 les deux séries évoluent dans deux sens opposés

2.3. Prédiction dans le cas d'une forte corrélation

L'intérêt d'une forte corrélation tient au fait que la connaissance de l'une des variables entraîne la connaissance de l'autre avec un faible risque d'erreur.

Dès lors, il convient de trouver la bonne relation mathématique qui existe entre les deux variables x et y :

$$y = f(x)$$

Prévision : quelle serait la valeur de y si on connaît déjà la valeur de x ?

Pour répondre à cette question

- Validation du modèle $y = f(x)$: commettre le minimum d'erreur
- Remplacer dans l'équation : si $x = x_0$, on cherchera la valeur $y = y_0$ inconnue en posant tout simplement $y_0 = f(x_0)$

2.3.1 Coefficient de détermination

On peut utiliser le carré de r , c'est le coefficient de détermination, r^2 est entre 0 et 1.

On estime qu'un coefficient de détermination supérieur à 75% donne une bonne régression des deux séries.

Le coefficient de corrélation devrait être supérieur à 0.87 ou inférieur à -0.87

Exemple

$$r(x, y) = \frac{9.9}{\sqrt{2} * \sqrt{56.6}} = 0.9305 > 0.87$$

$$r^2(x, y) = 0.8658 = 86.58\% > 75\%$$

La valeur du coefficient de détermination 0.8658 est nettement supérieure à 0.75

→ Plus le nombre de spots publicitaires augmente plus le volume des ventes augmente

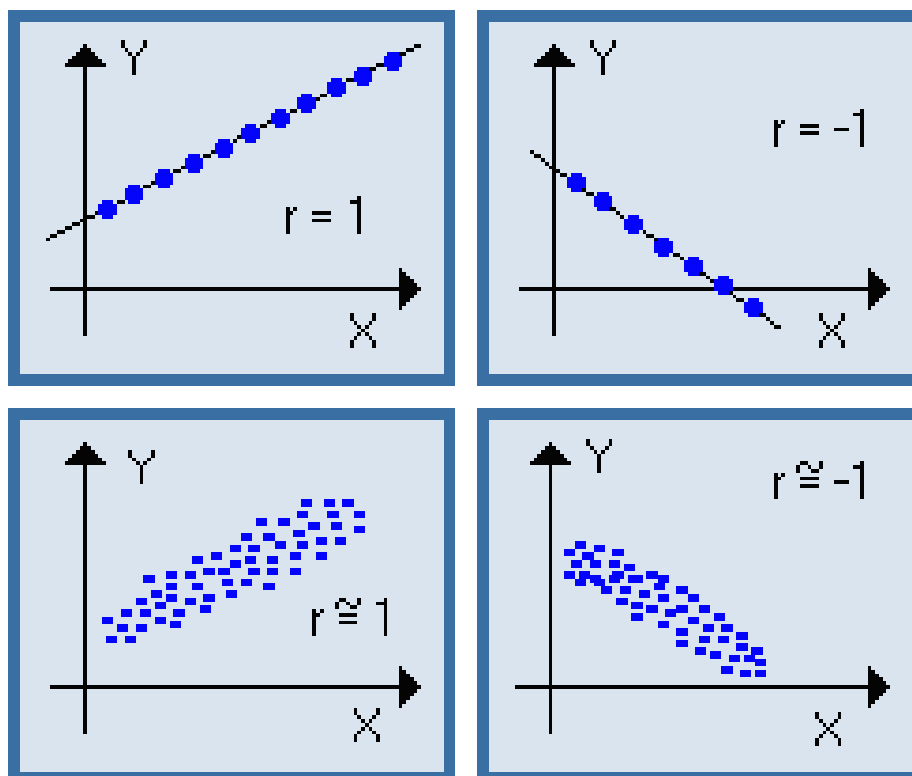
D'où on peut faire un ajustement linéaire de y « volume des ventes » en fonction de x « nombre de spots publicitaires »

Il est évident que y dépend de x . La variable y est la **variable à expliquer** et x la **variable explicative**. On n'a pas l'inverse.

On parle aussi de **variable endogène**, il s'agit de la variable **dépendante** qui doit être expliqué, c'est le cas ici de la variable y « volume des ventes ».

De même on dit d'une variable qu'elle est **exogène** si elle doit expliquer une autre variable, on dit aussi qu'elle est **indépendante**. Dans notre exemple il s'agit de la variable x « nombre de spots publicitaires ».

En général



Pour les deux premiers diagrammes ($r = 1$ et $r = -1$) la liaison linéaire est parfaite (en théorie), pour les deux autres diagrammes ($r \cong 1$ et $r \cong -1$) la liaison linéaire est de bonne qualité (c'est ce qu'on trouve en réalité).

Quelques propriétés

Pour tous nombres réels a , b , c et d , on a :

$$\text{Var}(a x + b) = \text{Var}(a x) = a^2 \text{Var}(x)$$

$$\text{Cov}(a x + b, c y + d) = \text{Cov}(a x, c y) = ac \text{Cov}(x, y)$$

$$r(a x + b, c y + d) = r(x, y) \text{ si } a \text{ et } c \text{ ont même signe}$$

$$r(a x + b, c y + d) = -r(x, y) \text{ si } a \text{ et } c \text{ ont des signes différents.}$$

Exemple :

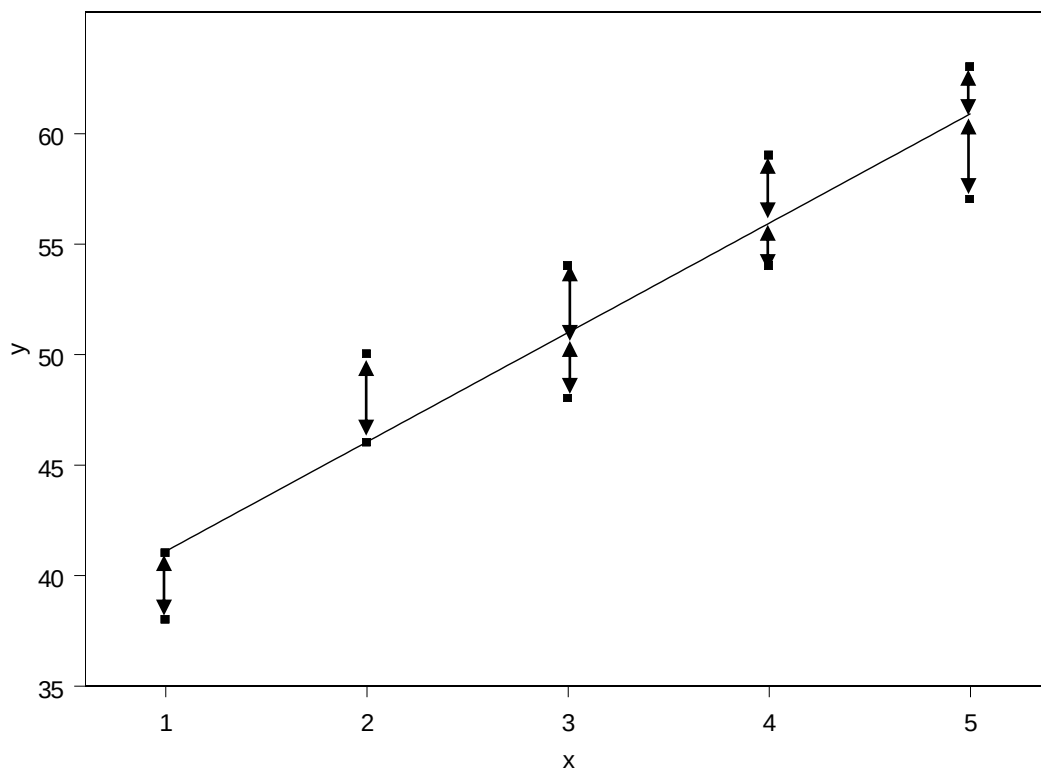
$$\text{Var}(2 x + 5) = \text{Var}(2 x) = 4 \text{Var}(x)$$

$$\text{Cov}(2 x + 5, 3 y + 1) = \text{Cov}(2 x, 3 y) = 6 \text{Cov}(x, y)$$

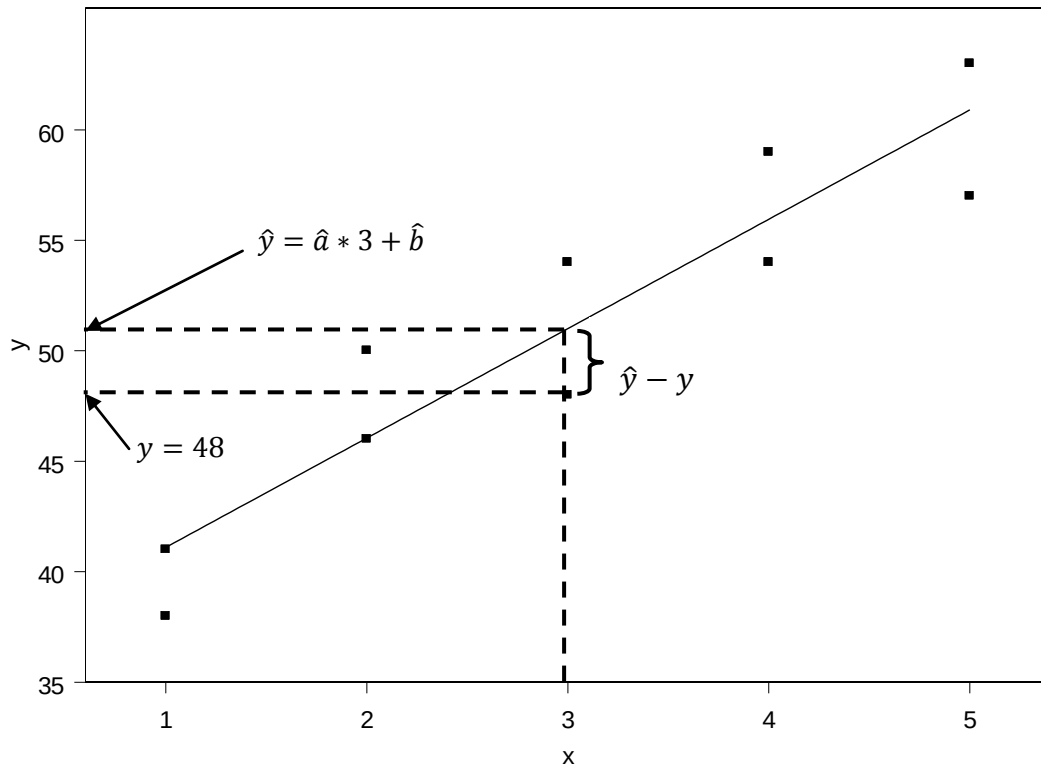
$$r(2 x + 5, 3 y + 1) = r(x, y)$$

$$r(2 x + 5, -2 y + 7) = -r(x, y)$$

2.3.2 Ajustement linéaire par la méthode des moindres carrés



Ajustement linéaire



L'erreur quadratique moyenne est

$$Q(f) = \frac{1}{n} \sum_{i=1}^n [y_i - f(x_i)]^2$$

Les différences $y_i - f(x_i)$ s'appellent les **résidus**. Si la somme des résidus est égale à 0 alors $Q(f)$ coïncide avec la variance des résidus.

Régression linéaire : on cherche une fonction f de la forme $f(x) = ax + b$

Posons

$$Q(a, b) = \frac{1}{n} \sum_{i=1}^n [y_i - ax_i - b]^2$$

Régression linéaire au sens des moindres carrées : on cherche le minimum de $Q(a, b)$; il est atteint pour les valeurs de

$$\hat{a} = \frac{\text{cov}(x, y)}{\text{var}(x)} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a} \bar{x}$$

Dans ce cas l'erreur quadratique coïncide avec la *variance des résidus*, elle est égale

$$Q(\hat{a}, \hat{b}) = \text{var}(y)[1 - r^2(x, y)] \quad \text{Variance résiduelle}$$

Plus le coefficient de détermination est proche de 1, plus l'erreur quadratique (variance résiduelle) due à l'ajustement est proche de 0.

Si on pose $y'_i = \hat{a}x_i + \hat{b}$ les valeurs estimées par la régression linéaire, on obtient :

$$\bar{y}' = \bar{y} \text{ et } \text{var}(y') = \frac{1}{n} \sum_{i=1}^n (y'_i - \bar{y}')^2 = r^2(x, y) * \text{var}(y)$$

$\text{var}(y')$ est la **variance expliquée** par la régression linéaire

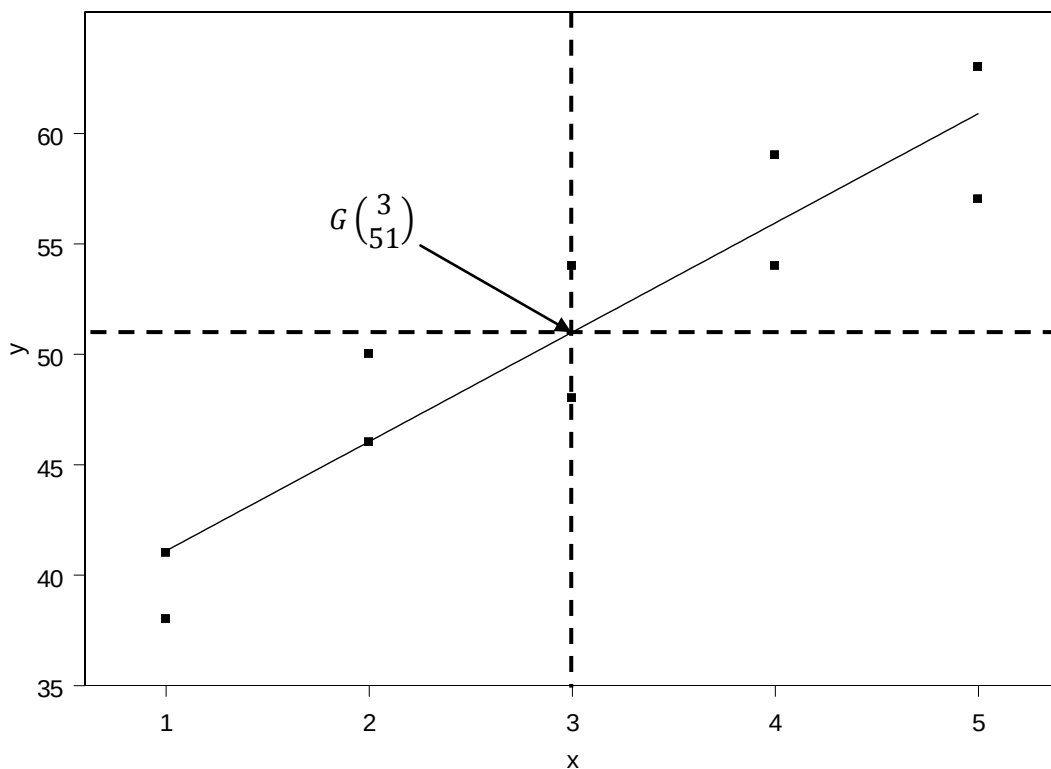
La **variance de y** s'écrit de la façon suivante

$$\text{var}(y) = r^2(x, y) * \text{var}(y) + [1 - r^2(x, y)] * \text{var}(y)$$

C'est la **somme de la variance expliquée plus la variance résiduelle**

Le coefficient de détermination est exprimé en pourcentage, exemple : dans la régression précédente, le pourcentage de la variance expliquée par la variable x est de 86.58%.

Remarque : La droite de régression linéaire $y = \hat{a}x + \hat{b}$ passe par la moyenne ou le centre de gravité $G = (\bar{x}, \bar{y})$ du nuage :



Exemple

$$\bar{x} = 3; \bar{y} = 51; Var(x) = 2; Var(y) = 56.6 \text{ et } Cov(x, y) = 9.9$$

D'où

$$\hat{a} = \frac{9.9}{2} = 4.95$$

$$\hat{b} = \bar{y} - \hat{a} \bar{x} = 36.15$$

L'erreur quadratique

$$Q(\hat{a}, \hat{b}) = 56.6 * (1 - 0.8658) = 7.595$$

La variance totale de y est $Var(y) = 56.6$

La variance expliquée par le modèle de régression linéaire est égale

$$56.6 * 86.58\% = 49.005$$

La droite de la régression est la suivante

$$y = 4.95 x + 36.15$$

Prévision : si on prévoit pour la semaine prochaine un nombre de spots publicitaires égal 3 alors le volume des ventes prévisionnelles sera approximativement :

$$y = 4.95 * 3 + 36.15 = 51$$

Soit 5100 \$

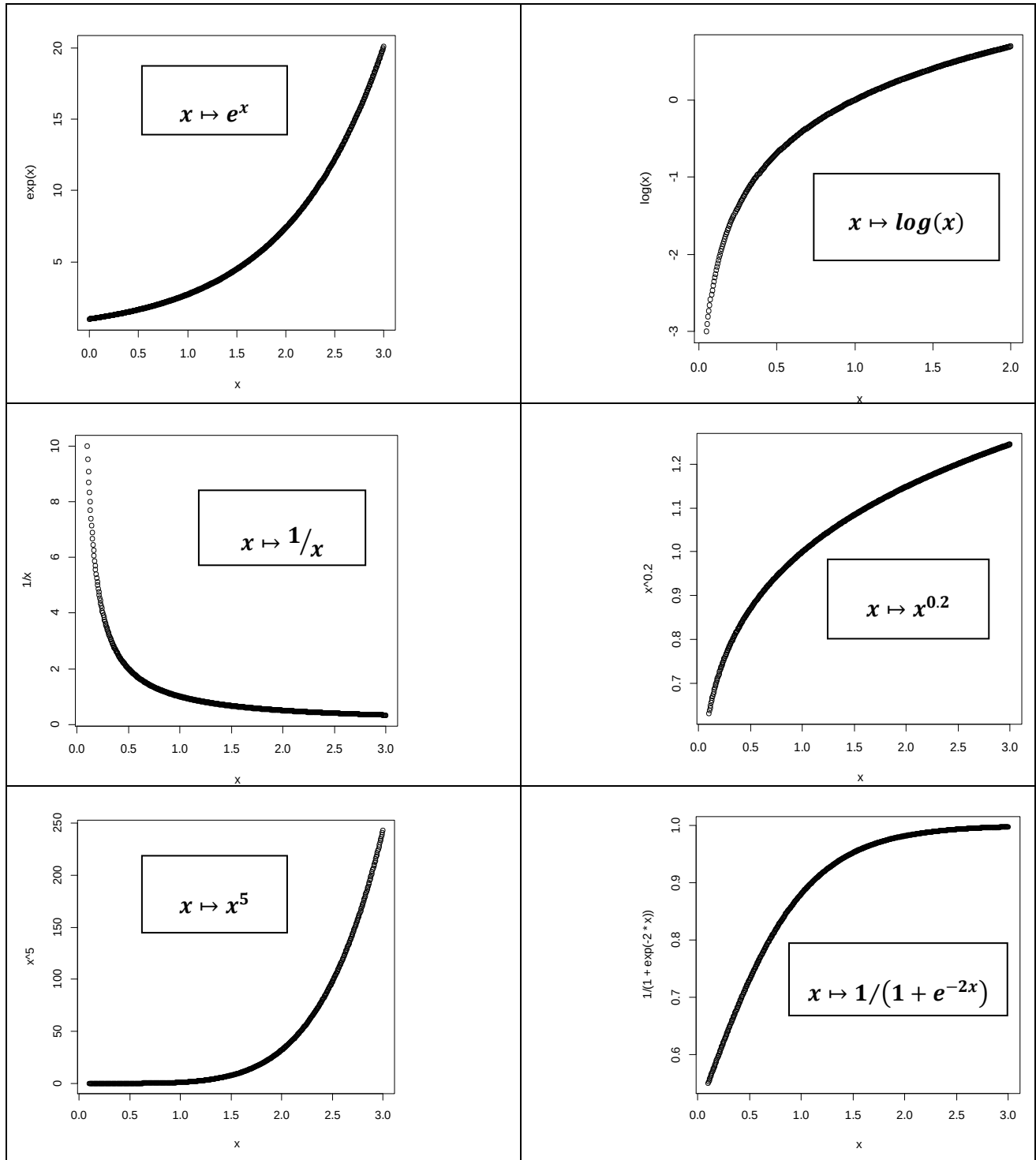
2.3.3 Ajustement non linéaire

Transformer le problème de manière à se ramener à une régression linéaire

Exemples de transformations

Famille	Fonctions	Transformation	Forme affine
exponentielle	$y = ae^{bx}$	$y' = \log(y)$	$y' = \log(a) + bx$
puissance	$y = ax^b$	$y' = \log(y) \text{ et } x' = \log(x)$	$y' = \log(a) + bx'$
inverse	$y = a + b/x$	$x' = 1/x$	$y = a + bx'$
logistique	$y = 1/(1 + e^{-(ax+b)})$	$y' = \log(y/(1-y))$	$y' = ax + b$

Exemples de formes :



2.4. Cas de données regroupées en classes

Tableau de contingence

On observe deux variables x et y sur une population P . On croise les données des deux variables x et y .

→ On obtient le tableau de contingence

On distingue deux cas :

- Le cas où les variables sont discrètes
 - On reprend tout ce que nous avons vu précédemment sans aucune perte d'informations
- Le cas où l'une au moins des variables est continue
 - Les classes sont constituées selon un procédé (plusieurs partitions possibles)
 - Chaque classe est représentée par son centre, cela va entraîner une perte d'informations

Remarque : on peut travailler directement avec la série de données sans regrouper les données en classes (moyens informatiques accessibles), mais dans certains cas les réponses sont recueillies sur des intervalles (par exemple un salaire qui se trouve entre deux valeurs [5000, 7000]).

Tableau de contingence

<i>Valeurs de y</i> <i>Valeurs de x</i>	y_1	y_2		y_j		y_L	<i>Total ou distribution marginale de x</i>
x_1	n_{11}	n_{12}		n_{1j}		n_{1L}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2L}	$n_{2\bullet}$
....							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{iL}	$n_{i\bullet}$
....							
x_K	n_{K1}	n_{K2}		n_{Kj}		n_{KL}	$n_{K\bullet}$
<i>Total ou distribution marginale de y</i>	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet j}$		$n_{\bullet L}$	n : nombre totale d'observations

2.4.1 Distribution marginale

C'est la distribution de x (resp. de y) indépendamment de y (resp. de x). Elle se lit sur la dernière colonne (resp. ligne) du tableau précédent.

Distribution marginale de x

<i>Valeurs de x</i>	<i>Effectifs</i>
x_1	$n_{1\bullet}$
x_2	$n_{2\bullet}$
....	
x_i	$n_{i\bullet}$
....	
x_K	$n_{K\bullet}$
<i>Total</i>	n

Les effectifs $n_{i\bullet}$ sont calculés par,

$$n_{i\bullet} = \sum_{j=1}^L n_{ij} , \quad i = 1, 2, \dots, K$$

Distribution marginale de y

<i>Valeurs de y</i>	<i>Effectifs</i>
y_1	$n_{\bullet 1}$
y_2	$n_{\bullet 2}$
....	
y_j	$n_{\bullet j}$
....	
y_L	$n_{\bullet L}$
<i>Total</i>	n

Les effectifs $n_{\bullet j}$ sont calculés par,

$$n_{\bullet j} = \sum_{i=1}^K n_{ij} , \quad j = 1, 2, \dots, L$$

On vérifie bien que

$$\sum_{i=1}^K n_{i\bullet} = \sum_{j=1}^L n_{\bullet j} = \sum_{i=1}^K \sum_{j=1}^L n_{ij} = n$$

- La moyenne et la variance de x sont :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} x_i \quad \text{et} \quad \text{var}(x) = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} x_i^2 - \bar{x}^2$$

- La moyenne et la variance de y sont :

$$\bar{y} = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} y_j \quad \text{et} \quad \text{var}(y) = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} (y_j - \bar{y})^2 = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} y_j^2 - \bar{y}^2$$

Pour le cas de variables continues les observations x_i et y_j seront remplacées simplement par les moyennes de classes m_i^x et m_j^y :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} m_i^x ; \quad \bar{y} = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} m_j^y ;$$

$$\text{var}(x) = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} (m_i^x - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^K n_{i\bullet} m_i^{x^2} - \bar{x}^2 ;$$

$$\text{var}(y) = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} (m_j^y - \bar{y})^2 = \frac{1}{n} \sum_{j=1}^L n_{\bullet j} m_j^{y^2} - \bar{y}^2$$

Exemple : Dans Le tableau de contingence suivant on donne la répartition conjointe du salaire mensuel y et de l'ancienneté en années x des salariés d'une entreprise.

$x \backslash y$ en milliers de DH	[5 ; 10[	[10 ; 15[	[15 ; 20[	Distribution marginale de x
[0 ; 8[	12	1	0	13
[8 ; 16[	4	8	0	12
[16 ; 24[	1	7	2	10
[24 ; 32[	0	1	14	15
Distribution marginale de y	17	17	16	$n = 50$

Les distributions marginales de x et y

Distribution marginale de l'ancienneté en années x

Ancienneté en années x	Moyenne de classe m_i^x	Effectif $n_{i\bullet}$
[0 ; 8[	4	13
[8 ; 16[	12	12
[16 ; 24[	20	10
[24 ; 32[	28	15
Total		50

Distribution marginale du salaire mensuel y

Salaire mensuel (exprimé en milliers) y	Moyenne de classe m_j^y	Effectif $n_{\bullet j}$
[5 ; 10[	7.5	17
[10 ; 15[	12.5	17
[15 ; 20[	17.5	16
Total		50

Les moyennes marginales sont

$$\bar{x} = \frac{1}{50} \sum_{i=1}^4 n_{i\bullet} m_i^x = \frac{1}{50} (13 * 4 + 12 * 12 + 10 * 20 + 15 * 28) = 16.32$$

$$\bar{x} = 16.32$$

$$\bar{y} = \frac{1}{50} \sum_{j=1}^3 n_{\bullet j} m_j^y = \frac{1}{50} (17 * 7.5 + 17 * 12.5 + 16 * 17.5) = 12.4$$

$$\bar{y} = 12.4$$

- Le salaire mensuel moyen est 12 400 DH
- L'ancienneté moyenne des salariés est 16.32 années (16 années et 4mois)

Les variances marginales sont

$$var(x) = \frac{1}{50} \sum_{i=1}^4 n_{i\bullet} m_i^{x^2} - \bar{x}^2 = \frac{1}{50} (13 * 4^2 + 12 * 12^2 + 10 * 20^2 + 15 * 28^2) - 16.32^2$$

$$var(x) = 87.578 \quad \text{et} \quad \sigma(x) = 9.358$$

$$var(y) = \frac{1}{50} \sum_{j=1}^3 n_{\bullet j} m_j^{y^2} - \bar{y}^2 = \frac{1}{50} (17 * 7.5^2 + 17 * 12.5^2 + 16 * 17.5^2) - 12.4^2$$

$$var(y) = 16.49 \quad \text{et} \quad \sigma(y) = 4.061$$

2.4.2 Distribution conditionnelle

- **La distribution de x connaissant y :**

La colonne de y_j dans le tableau de contingence entre x et y , qui décrit la distribution selon la variable x des $n_{\bullet j}$ individus présentant la valeur y_j de la variable y , constitue la distribution conditionnelle de x sachant $y = y_j$. On notera le caractère étudié par $x/y = y_j$ ou tout simplement x/y_j .

Distribution conditionnelle de x sachant y = y_j

<i>Valeurs de x</i>	<i>Effectif du caractère x/y_j</i>
x_1	n_{1j}
x_2	n_{2j}
....	
x_i	n_{ij}
....	
x_K	n_{Kj}
<i>Total</i>	$n_{\bullet j}$

La moyenne conditionnelle sera

$$\bar{x}/y = y_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^K n_{ij} x_i$$

On peut encore noter cette moyenne par $\bar{x}/y_j$

De même on définit la variance conditionnelle,

$$Var(x/y_j) = \frac{1}{n_{\bullet j}} \sum_{i=1}^K n_{ij} (x_i - \bar{x}/y_j)^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^K n_{ij} x_i^2 - (\bar{x}/y_j)^2$$

• ***La distribution de y connaissant x :***

Pour cette distribution on va considérer la ligne correspondant à x_i , pour décrire la distribution conditionnelle de y sachant $x = x_i$.

On note le caractère étudié par $y/x = x_i$ ou y/x_i .

Distribution conditionnelle de y sachant $x = x_i$

Valeurs de y	Effectif du caractère y/x_i
y_1	n_{i1}
y_2	n_{i2}
....	
y_j	n_{ij}
....	
y_L	n_{iL}
<i>Total</i>	$n_{i\bullet}$

La moyenne conditionnelle sera

$$\bar{y}/x_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^L n_{ij} y_j$$

et la variance conditionnelle,

$$Var(y/x_i) = \frac{1}{n_{i\bullet}} \sum_{j=1}^L n_{ij} (y_j - \bar{y}/x_i)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^L n_{ij} y_j^2 - (\bar{y}/x_i)^2$$

Exemple :

On reprend le dernier exemple,

$x \backslash y$ en milliers de DH	[5 ; 10[	[10 ; 15[	[15 ; 20[	Distribution marginale de x
[0 ; 8[	12	1	0	13
[8 ; 16[	4	8	0	12
[16 ; 24[	1	7	2	10
[24 ; 32[	0	1	14	15
Distribution marginale de y	17	17	16	$n = 50$

On veut déterminer la distribution de y/x_2 et y/x_4 ainsi que leurs moyennes conditionnelles.

La valeur de x_2 correspond à l'intervalle $[8 ; 16[$, on doit déterminer la distribution conditionnelle de y sachant $x = 12$.

Distribution conditionnelle de y sachant $x = 12$

<i>Valeurs de y</i>	Moyenne de classe m_j^y	<i>Effectif du caractère</i> y/x_2
$[5 ; 10[$	7.5	4
$[10 ; 15[$	12.5	8
$[15 ; 20[$	17.5	0
<i>Total</i>		12

La moyenne conditionnelle sera

$$\bar{y}/x_2 = \frac{1}{12} (4 * 7.5 + 8 * 12.5 + 0 * 17.5) = \frac{130}{12} = 10.833$$

$$\bar{y}/x_2 = 10\,833 \text{ DH}$$

Pour la valeur de x_4 , elle correspond à l'intervalle $[24 ; 32[$, il faut déterminer la distribution conditionnelle de y sachant $x = 28$.

Distribution conditionnelle de y sachant $x = 28$

<i>Valeurs de y</i>	Moyenne de classe m_j^y	<i>Effectif du caractère</i> y/x_4
$[5 ; 10[$	7.5	0
$[10 ; 15[$	12.5	1
$[15 ; 20[$	17.5	14
<i>Total</i>		15

La moyenne conditionnelle sera

$$\bar{y}/x_4 = \frac{1}{15} (0 * 7.5 + 1 * 12.5 + 14 * 17.5) = \frac{257.5}{15} = 17.167$$

$$\bar{y}/x_4 = 17\,167 \text{ DH}$$

On remarquera qu'en passant de l'ancienneté de 12 ans à une ancienneté de 28 ans, le salaire mensuel moyen passe de 10 833 DH à 17 167 DH avec une augmentation de 58.47%.

2.4.3 Covariance et corrélation

On calcule la covariance par la formule suivante adaptée au tableau de contingence :

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^L n_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^L n_{ij} x_i y_j - \bar{x} \bar{y}$$

Si les classes correspondent à des intervalles, on utilisera comme d'habitude, les moyennes de classes :

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^L n_{ij} (m_i^x - \bar{x})(m_j^y - \bar{y}) = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^L n_{ij} m_i^x m_j^y - \bar{x} \bar{y}$$

Le coefficient de corrélation se déduit

$$r(x, y) = \frac{cov(x, y)}{\sigma(x) * \sigma(y)}$$

Exemple :

Revenons à l'exemple précédent, on veut calculer la covariance entre x et y .

On déduira le coefficient de corrélation et le coefficient de détermination.

$$cov(x, y) = \frac{1}{50} \sum_{i=1}^3 \sum_{j=1}^4 n_{ij} m_i^x m_j^y - \bar{x} \bar{y} = \frac{1}{50} (11780) - 16.32 * 12.4 = 33.23$$

Le coefficient de corrélation sera

$$r(x, y) = \frac{cov(x, y)}{\sigma(x) * \sigma(y)} = \frac{33.23}{9.358 * 4.061} = 0.8745$$

$$r^2(x, y) = 0.8745^2 = 76.5\%$$

Conclusion : il y'a une forte liaison entre le salaire mensuel et l'ancienneté des salariés de l'entreprise, on peut proposer un ajustement linéaire des données par la méthode des moindres carrés. La droite de régression linéaire peut expliquée jusqu'à 76.5% de la variance de y .

Bibliographie

1. D.R. Anderson, D.J. Sweeney, T.A. Williams (2010), *Statistiques pour l'économie et la gestion*, De Boeck.
2. G. Baillargeon (2004), *Outils statistiques et analyse de données*, SMG.
3. B. GRAIS (1998), *Méthodes statistiques*, Techniques statistiques, Dunod.
4. P.-C. Pupion (2008) *Statistiques pour la gestion, applications avec Excel et SPSS*, Gestion Sup, Dunod
5. T.H. Wonnacott, R.J. Wonnacott (1991), *Statistique, économie-gestion-sciences-médecine*, ECONOMICA.