Faculté des sciences exactes
Département des Mathématiques et Informatique
Matière : Analyse de données

<u>Correction de la Fiche de TD n° 01</u>

<u>Solution exercice n° 01 :</u>

- Variable 1 : qualitative, non discret et non continue
- Variable 2 : qualitative, non discret et non continue
- Variable 3 : non qualitative, discret et continue
- Variable 4 : qualitative, non discret et non continue
- Variable 5 : non qualitative, discret et non continue
- Variable 6 : non qualitative, non discret et continue
- Variable 7 : non qualitative, non discret et continue
- Variable 8 : non qualitative, non discret et continue

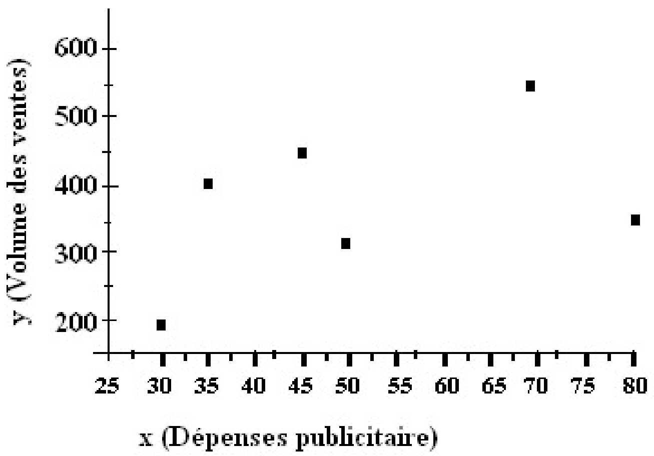
<u>Solution exercice n° 02 :</u>

- Variable 1 : non ordinale, non nominale et dichotomique
- Variable 2 : ordinale, non nominale et dichotomique
- Variable 3 : non ordinale, nominale et non dichotomique
- Variable 4 : non ordinale, non nominale et dichotomique
- Variable 5 : non ordinale, nominale et non dichotomique

<u>Solution exercice n° 03 :</u>

<u>Question (1) :</u>

Le graphique du nuage de points entre les dépenses en publicité et le volume des ventes que l'entreprise réalise :



A de cette représentation, nous soupçonnons l'existence d'une relation entre X et Y, mais la corrélation n'est pas très forte. (Nous la vérifions avec la question 3)

Question (2) : Ajustement de la droite de régression.

La réponse à cette question permet d'exprimer la relation à l'aide d'une équation mathématique : Y=ax+b

Xi	Yi	XiYi	X_i^2	Y_i^2
70	**580**	**40600**	**4900**	**336400**
80	**380**	**30400**	**6400**	**14400**
30	**200**	**6000**	**900**	**4000**
50	**310**	**15500**	**2500**	**96100**
35	**400**	**14000**	**1225**	**160000**
45	**450**	**20250**	**2025**	**202500**
310	**2320**	**126750**	**17950**	**979400**

$$\overline{Y} = \frac{\sum Y_i}{N} = \frac{2320}{6} = 386.66 \approx 387$$

$$\overline{X} = \frac{\sum X_i}{N} = \frac{310}{6} = 51.6 \approx 52$$

$$Var(X) = S^2(X) = \frac{\sum X_i^2}{N} - \overline{X^2} = \frac{17950}{6} - (52)^2 = 288$$

$$S\ (X) = \sqrt{288} = 16.97 \approx 17$$

$$Var(Y) = S^2(Y) = \frac{\sum Y_i^2}{N} - \overline{Y^2} = \frac{979400}{6} - (387)^2 = 13464$$

$$S\ (Y) = \sqrt{13464} = 116.03 \approx 116$$

$$Cov(X,Y) = \frac{\sum XY}{N} - \overline{X}\overline{Y} = \frac{126750}{6} - (52)(387) = 1001$$

Nous obtenons alors les valeurs a et b de la droite : Y=ax+b

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{1001}{288} = 3.48$$

$$b = \overline{Y} - a\overline{X} = 387 - (3.48)(52) = 387 - 180.96 \approx 387 - 181 = 206$$

d'où : **Y=3.48x+206**

Question 3 : Calcul du coefficient de corrélation

$$r = \frac{Cov(X,Y)}{S_{ex}S_{ey}} = \frac{1001}{17 * 116} \approx 0.507$$

Question 4 : Interprétation du coefficient de corrélation

Pour cette série, nous pouvons conclure qu'il y a une corrélation positive mais très forte entre le volume des ventes et les dépenses en publicité de cette entreprise. En effet, seulement : 26 % ($r^2 = 0.51^2 = 0.26$) **de la fluctuation totale de Y se trouve expliquée par le lien entre X et Y.**

Solution exercice n° 04

Les formules de calcul des moyennes et les variances de *T* et *V*, leurs covariances et le coefficient de corrélation linéaire :

Pour T : $\overline{T} = \frac{\sum T_i}{N}$, $Var(T) = \frac{\sum T_i^2}{N} - \overline{T^2}$

Pour V : $\overline{V} = \frac{\sum V_i}{N}$, $Var(V) = \frac{\sum V_i^2}{N} - \overline{V^2}$

Covariance : $Cov(T,V) = \frac{\sum TV}{N} - \overline{T}\,\overline{V}$

Coefficient de corrélation : $r = \frac{Cov(T,V)}{\sqrt{Var(T).Var(V)}}$

Les coefficients de la droite d'ajustement se calcul par les formules :

$a = \frac{Cov(T,V)}{Var(T)}$ **,** $b = \overline{V} - a\overline{T}$

Afin d'arriver à répondre à toutes ces questions, on a besoin d'un tableau de cinq colonnes :

T_i	V_i	T_i^2	V_i^2	T_iV_i
10	**8**	**100**	**64**	**80**
15	**15**	**225**	**225**	**225**
20	**23**	**400**	**529**	**460**
25	**31**	**625**	**961**	**775**
30	**38**	**900**	**1444**	**1140**
35	**46**	**1225**	**2116**	**1610**
$\sum T_i = 135$	$\sum V_i = 161$	$\sum T_i^2 = 3475$	$\sum V_i^2 = 5339$	$\sum T_iV_i = 4290$
$\overline{T} = 22.5$	$\overline{V} = 26.83$	**579.17**	**889.83**	**715**

1- **Moyenne arithmétiques :** $\overline{T} = 22.5$ **et** $\overline{V} = 26.83$

2- **Variances :** $Var(T) = \frac{\sum T_i^2}{N} - \overline{T^2} = 579.17 - (22.5)^2 = 72.92$

$$Var(V) = \frac{\sum V_i^2}{N} - \overline{V^2} = 889.83 - (26.83)^2 = 169.9811$$

3- **Covariance :** $Cov(T,V) = \frac{\sum TV}{N} - \overline{T}\,\overline{V} = 715 - (22.5)(26.83) = 111.325$

Coefficient de corrélation : $r = \frac{Cov(T,V)}{\sqrt{Var(T).Var(V)}} = \frac{111.325}{\sqrt{72.92.169.9811}} \approx 0.9999 = 1$

4- **Droite d'ajustement :**

$$a = \frac{Cov(T,V)}{Var(T)} = \frac{111.325}{72.92} \approx 1.527$$

$$b = \overline{V} - a\overline{T} = 26.83 - 1.527.22.5 = -7.5275$$

Ainsi la droite d'ajustement a pour équation : **V= 1.527.T – 7.53**

Solution exercice n° 05

Le tableau concerne deux variables x et y mesurées sur un échantillon de taille 10 individus (i=1..10) permettant de calculer les paramètres nécessaires :

Individus	x_i	y_i	x_i^2	y_i^2	$x_i\, y_i$
1	0.5	0	0.25	0	0
2	-0.1	1.2	0.01	1.44	-0.12
3	-0.5	0.5	0.25	0.25	-0.25
4	-0.3	0.1	0.09	0.01	-0.03
5	0	2.5	0	6.25	0
6	1.6	-0.7	2.56	0.49	-1.12
7	2	2	4	4	4
8	2.4	1.2	5.76	1.44	2.88
9	0.5	3.5	0.25	12.25	1.75
10	2.7	-0.9	7.29	0.81	-2.43
Total	$\sum x_i = 8.8$	$\sum y_i = 9.4$	$\sum x_i^2 = 20.46$	$\sum y_i^2 = 26.94$	$\sum x_i y_i = 4.68$
	$\overline{X} = \frac{\sum x_i}{N} = \frac{8.8}{10} = 0.88$	$\overline{Y} = \frac{\sum y_i}{N} = \frac{9.4}{10} = 0.94$			

A. **Avec la méthode des moindres carrées**

$\overline{X} = 0.88$, $\overline{Y} = 0.94$

$$Var(X) = \frac{\sum X_i^2}{N} - \overline{X^2} = \frac{20.46}{10} - (0.88) = 1.27$$

$$Var(Y) = \frac{\sum y_i^2}{N} - \overline{Y^2} = \frac{26.94}{10} - (0.94) = 1.81$$

$$Cov(X,Y) = \frac{\sum XY}{N} - \overline{X}\overline{Y} = \frac{4.68}{10} - 0.88 * 0.94 = -0.36$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{-0.36}{\sqrt{1.27 * 1.81}} = -0.24$$

1- Calculer la droite de régression observée de y par rapport à x : Dy/x

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{-0.36}{1.27} = -0.28, \quad b = \overline{Y} - a\overline{X} = 0.94 + 0.28 * 0.88 = 1.19$$

2- Calculer la droite de régression observée de x par rapport à y : Dx/y

L'équation de la droite de régression de x par rapport à y (Dx/y) aura ainsi équation sous la forme symétrie précédente (voir cours) :

(Dx/y) : $\boldsymbol{x = \alpha y + \beta}$ avec $\alpha = \frac{COV(X,Y)}{Var(Y)}$ ou $\alpha = r\frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}}$ et

$$\boldsymbol{\beta = \overline{x} - \alpha\overline{y}}$$

Mise sous la forme y=a'x+b' avec a'=$\mathbf{1}/\alpha$ et b'=$-\beta/\alpha$.

α=-0.36/1.81=-0,19889503, β=0,88- α *0,94= 1,06696133

a'=1/ α=-5.02, b'=- β/ α= -5,36444444

(les résultats finaux des calculs sont données exactement, les valeurs intermédiaires prises ici sont approximatives)

Les deux droites :

$$\boxed{\begin{array}{l}\mathbf{Dy/x : y = -0.28\ x + 1.19}\\ \mathbf{Dx/y : y = -5.02\ x + 5.36}\end{array}}$$

3- Interprétation des résultats :

Le coefficient de corrélation linéaire r = -0.24, qui présume une très mauvaise approximation linéaire de ce nuage, sa tendance décroissante est confirmée par la présence du signe négatif.

L'angle entre ces deux droites est important, ce qui ne fait que confirmer la très mauvaise approximation linéaire du nuage (**Voir Figure ci-dessous**).

B. <u>Avec la méthode des moindres rectangles</u>

On cherche à approximer ce nuage par une droite **D** de vecteur ***u'=[α1, α2].***

Pour atteindre cet objectif, on applique la formulation de la méthode des moindres rectangles :

- Calcul de la matrice V des variances-Covariance :

$$V = \begin{bmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & Var(x_2) \end{bmatrix} = \begin{bmatrix} 1{,}27 & -0{,}36 \\ -0{,}36 & 1{,}81 \end{bmatrix}$$

La dérivée du Lagrangien s'annule quand **Vu-λu** = 0 ➔ **(V-λI).u=0** où I est la matrice identité 2X2 et u le vecteur propre associé à la valeur propre λ (seulement la plus grande valeur propre λ_1).

Cherchons les valeurs propres de la matrice V en résolvant l'équation du second degré :

$$P(\lambda)=\text{Dét}(V-\lambda I)=\begin{vmatrix} \mathbf{1{,}27-\lambda} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{1{,}81-\lambda} \end{vmatrix}=\lambda^2-3.08\lambda+2.17$$

$$\Rightarrow \boxed{\mathbf{\lambda_1=1.99 \text{ et } \lambda_2=1.09 \ (\max \lambda=\lambda_1)}}$$

Pour trouver les vecteurs propres associés à ces valeurs propres, nous devons résoudre :

(V-λI).u=0

$$(V-\lambda_1 I).u_1=\begin{vmatrix} \mathbf{1{,}27-1.99} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{1{,}81-1.99} \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} \mathbf{-0.72} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{-0.18} \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} \mathbf{0} \\ \mathbf{0} \end{vmatrix}$$

Ce qui conduit à résoudre le système :

$$-0.72x_1-0.36x_2=0 \quad \text{eq.1}$$
$$-0.36x_1-0.18x_2=0 \quad \text{eq.2}$$

Comme l'équation (2) est la double de l'équation (1) on aura $2x_1+x_2=0$ c'est-à-dire :

$\boldsymbol{u_1' = (x_1, -2x_1)}$

Il reste à nommer ce vecteur à 1 : $\|\boldsymbol{u_1}\|^2 = \mathbf{1} = (\boldsymbol{x_1^2} + \mathbf{4}\boldsymbol{x_1^2}) = \mathbf{5}\boldsymbol{x_1^2}$, on en déduit que : $\boldsymbol{x_1} = \mathbf{1}/\sqrt{\mathbf{5}}$ ainsi que :

$$\boxed{\boldsymbol{u_1}' = (\mathbf{0.4472}, -\mathbf{0.8944})}$$

Le vecteur propre u_2 associé à la valeur propre λ_2.

$$(V-\lambda_2 I).u1=\begin{vmatrix} 1{,}27-1.09 & -0{,}36 \\ -0{,}36 & 1{,}81-1.09 \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix}=\begin{vmatrix} 0.18 & -0{,}36 \\ -0{,}36 & 0.72 \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix}=\begin{vmatrix} 0 \\ 0 \end{vmatrix}$$

ce qui conduit à résoudre le système :

$0.18x_1 - 0.36x_2=0$ éq.1
$-0.36x_1 + 0.72x_2=0$ éq.2

Comme l'équation 2 est la double de l'équation 1 on aura $x_1-2x_2=0$, c'est-à-dire
$u_2^{'}=(x_1,-x_1/2)$

Il reste à remplacer à nommer ce vecteur à 1. $\|u_2\|^2=1=(x_1^2+x_1^2/4)=5x_1^2/4$, on en déduit que $x_1=2/\sqrt{5}$ ainsi que :

$$\boxed{u_2=(0.8944,-0.4472)'}$$

On obtient ainsi la droite de régression orthogonale, portée par le vecteur :
$u_1'=(0.4472,-0.8944)$ correspondant à la plus grande valeur propre $\lambda_1=1.99$ en résolvant :

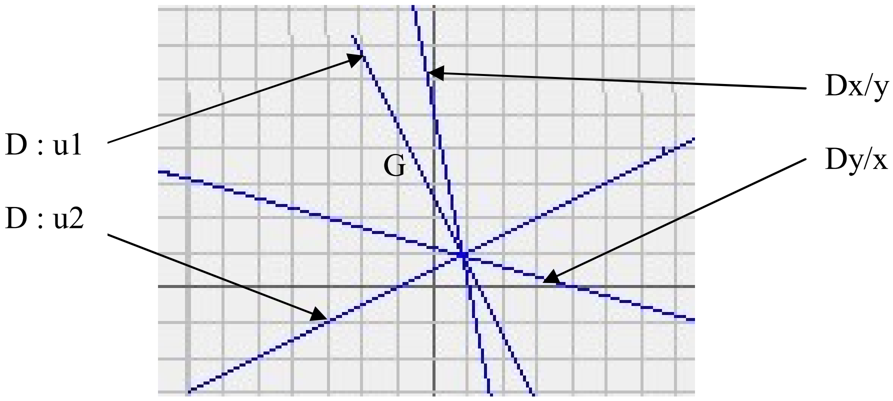
$$x_2=\frac{-0.8944}{0.4472}x_1+0.94-\frac{-0.8944}{0.4472}0.88$$

⇨ $\boxed{\text{Soit } x_2=-2x_1+2.7}$ **D : u1**

L'équation de la deuxième droite de régression, associée à la deuxième valeur propre, orthogonale à la précédente :

$$x_2=\frac{-0.4472}{0.8944}x_1+0.94-\frac{-0.4472}{0.8944}0.88$$

⇨ $\boxed{\text{Soit } x_2=0{,}5x_1+0.5}$ **D : u2**



On peut montrer que la pente de cette droite de régression orthogonale est comprise entre les deux droites de régression **$D_{y/x}$ et $D_{x/y}$,** c'est-à-dire qu'elle passe par **G** et se situe entre ces deux droites.