

Série d'exercice n°3

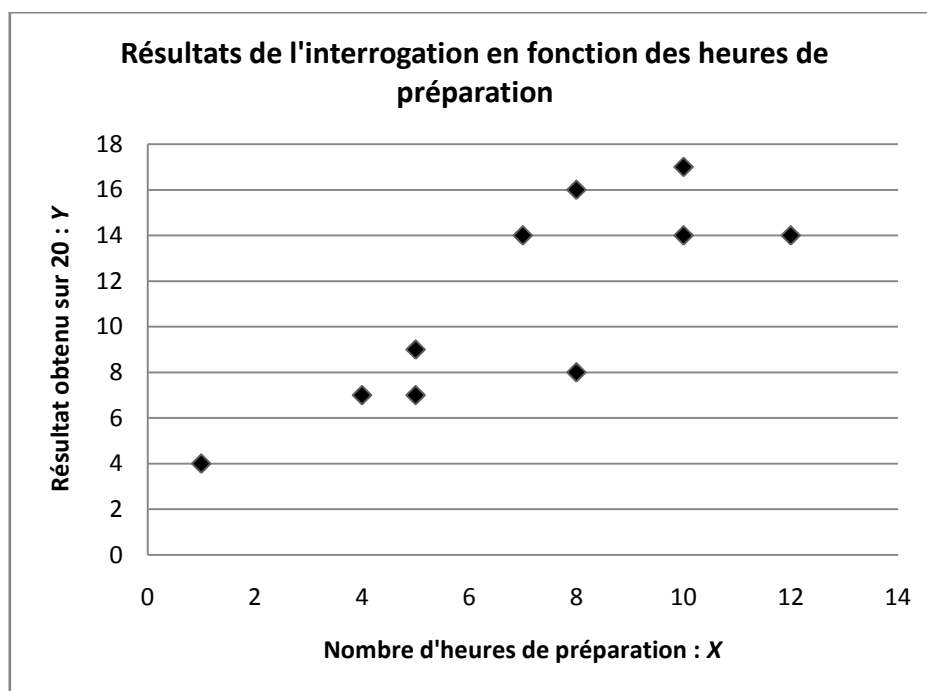
Exercice 1 : On demande à dix étudiants le nombre d'heures X passées à préparer une interrogation. Par ailleurs, on relève pour chacun d'eux le résultat obtenu Y sur un total de 20 :

X	5	1	4	12	10	8	7	8	5	10
Y	7	4	7	14	14	8	14	16	9	17

1. Représenter graphiquement la série statistique.
2. Déterminez le coefficient de corrélation de Pearson de cette série. Peut-on définir un ajustement linéaire du nuage de points observé ?
3. Dans l'affirmative, faites l'ajustement et interprétez les résultats obtenus.
4. Peut-on améliorer les résultats de l'ajustement en considérant le modèle $y = ax^b$ (faites une transformation logarithmique).

Solution de l'exercice 1 :

1. Représentation graphique :



2. Calculons le coefficient de corrélation ; pour cela on doit calculer tout d'abord le coefficient de covariance et les écart-types des deux caractères X et Y .

	X	Y	X^2	Y^2	XY
	5	7	25	49	35
	1	4	1	16	4
	4	7	16	49	28
	12	14	144	196	168
	10	14	100	196	140
	8	8	64	64	64
	7	14	49	196	98
	8	16	64	256	128
	5	9	25	81	45
	10	17	100	289	170
Total	70	110	588	1392	880

Les mesures caractéristiques dont on a besoin sont données dans le tableau suivant :

n	$\bar{x}_n$	$\bar{y}_n$	S_X^2	S_Y^2	S_X	S_Y	$Cov(X, Y)$	Corrélation : r	Coef.Detrm. r -deux
10	7	11	9,8	18,2	3,1305	4,2661	11	0,8237	67,84%

On a un coefficient de corrélation proche du seuil 0,87 ; on a un assez bon ajustement linéaire du nuage représenté dans le graphe.

3. La droite de régression linéaire sera

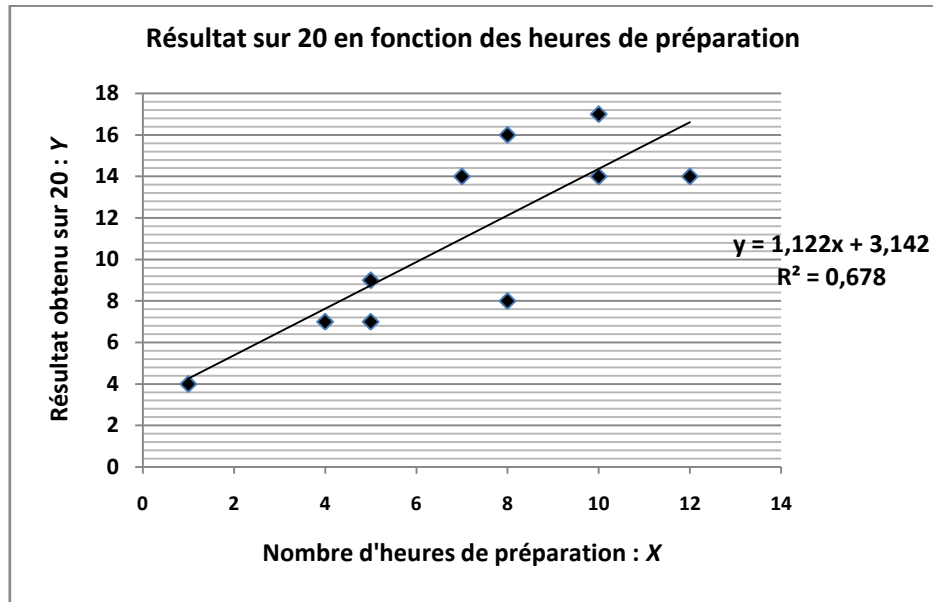
$$Y = 1,122 X + 3,142$$

Cette droite peut expliquer jusque 67,84% de la variance des notes Y. Ou d'une autre manière on peut dire que

« 67,84% de la variabilité des notes est expliquée par la variabilité des durées de préparation des interrogations »

Le modèle linéaire suggère qu'une augmentation d'une heure de préparation va produire 1,12 point de plus dans l'interrogation.

$$(\hat{a} = 11/9,8 = 1,12245 \text{ et } \hat{b} = 11 - 1,12245 * 7 = 3,1428)$$



4. On remarque une légère forme parabolique (non linéaire) sur le graphe ci-dessus qui peut se voir en se rapprochant des environs de la note 20.

Si on fait une transformation logarithmique des données càd en posant :

$$\begin{cases} y' = \log(y) \\ x' = \log(x) \end{cases}$$

Les données (x', y') seront :

$x' = \log(x)$	0,7	0	0,6	1,08	1	0,9	0,85	0,9	0,7
$y' = \log(y)$	0,85	0,6	0,85	1,15	1,15	0,9	1,15	1,2	0,95

Le coefficient de corrélation calculé sera égale à

$$r = 0,8778$$

Une nette amélioration est prévue avec le modèle puissance $y = ax^b$ et si on calcule $\log(y)$ on obtient le modèle de régression linéaire suivant :

$$y' = \log(a) + bx'$$

qui peut expliquer jusqu'à 77,06% de la variance de y' .

Exercice 2 : Soit un échantillon de 450 ménages classés en fonction de deux caractères ;

X : le revenu

Y : la consommation d'un bien

Y \ X	10	12	16	20
20	38	29	16	0
25	46	54	19	1
30	13	64	41	7
40	0	0	27	48
50	0	0	31	16

1. Déterminer les distributions marginales de X et Y .
2. Déterminer les moyennes marginales et les variances marginales de X et Y .
3. Soit les variables conditionnelles $Y/X = 20$ et $Y/X = 50$, calculer leurs moyennes ; que peut on déduire ?
4. Déterminez le coefficient de corrélation de Pearson de cette série. Peut-on envisager un ajustement linéaire ?
5. Dans l'affirmative, faite l'ajustement et interprétez les résultats obtenus.

Solution de l'exercice 2 :

1. On va compléter le tableau des effectifs en sommant les lignes et les colonnes :

		Y				Total : $n_{i.}$
		10	12	16	20	
X	20	38	29	16	0	83
	25	46	54	19	1	120
	30	13	64	41	7	125
	40	0	0	27	48	75
	50	0	0	31	16	47
Total : $n_{.j}$		97	147	134	72	450

La distribution marginale de X correspond à la dernière colonne tandis que celle de Y correspond à la dernière ligne.

2. On a les tableaux de distributions de X et de Y :

Distribution marginale de X

X	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
20	83	1660	33200
25	120	3000	75000
30	125	3750	112500
40	75	3000	120000
50	47	2350	117500
<i>Total</i>	450	13760	458200

$\bar{x}_n$	$\bar{y}_n$	S_X^2	S_Y^2	S_X	S_Y
30,58	14,04	1018,22-935,14 = 83,22	11,71	9,12	3,42

Distribution marginale de Y

Y	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
10	97	970	9700
12	147	1764	21168
16	134	2144	34304
20	72	1440	28800
Total	450	6318	93972

Les résultats sont les suivants :

$\bar{x}_n$	$\bar{y}_n$	S_X^2	S_Y^2	S_X	S_Y
30,58	14,04	83,22	11,71	9,12	3,42

3. Les distributions des variables $Y/X = 20$ et $Y/X = 50$ sont les suivantes :

$Y/X = 20$	n_{1j}	$n_{1j} y_j$
10	38	380
12	29	348
16	16	256
20	0	0
<i>Total</i>	83	984
<i>Moyenne</i>	11,86	

$Y/X = 50$	n_{5j}	$n_{5j} y_j$
10	0	0
12	0	0
16	31	496
20	16	320
<i>Total</i>	47	816
<i>Moyenne</i>	17,36	

Plus on augmente le revenu plus la consommation moyenne augmente.

4. Déterminons le coefficient de corrélation linéaire de Pearson :

Avant on doit calculer le coefficient de covariance

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^5 \sum_{j=1}^4 n_{ij} x_i y_j - \bar{x}_n \bar{y}_n$$

Le tableau ci-dessous donne les valeurs des quantités $n_{ij} x_i y_j$

		Y			
		10	12	16	20
X	20	7600	6960	5120	0
	25	11500	16200	7600	500
	30	3900	23040	19680	4200
	40	0	0	17280	38400
	50	0	0	24800	16000
Total		23000	46200	74480	59100

$$Cov(x, y) = \frac{23000 + 46200 + 74480 + 59100}{450} - 30,58 * 14,04 = 21,31.$$

Le coefficient de corrélation est égal à $r(x, y) = 0,68$. Si on calcule le coefficient de détermination on trouve la valeur 46,62% ; cette valeur est faible ; donc un ajustement linéaire est à proscrire.

Exercice 3 : Ce tableau présente l'évolution du chiffre d'affaire en million d'Euros (CA) d'une filiale de la firme américaine VIXON entre 2002 et 2006.

Années	2002	2003	2004	2005	2006
CA	8	12	35	40	70

1. Etablir par la méthode des moindres carrés la relation linéaire liant le chiffre d'affaires Y au temps t en considérant $t = 1$ en 2002.
2. Vérifier la qualité de l'ajustement et déterminer les variances expliquée V_E et résiduelle V_R de Y .
3. Faire un ajustement logarithmique ($Y = a \log(t) + b$) et comparer la qualité de l'ajustement avec celui de la question 1.

Solution de l'exercice 3 :

Années	1	2	3	4	5
CA	8	12	35	40	70

1. On a besoin de calculer certaines quantités caractéristiques :

$\bar{t}_n$	$\bar{y}_n$	S_t^2	S_y^2	S_t	S_y	$Cov(t, y)$	$r(t, y)$
3	33	2	497,6	1,414	22,307	30,4	0,964

Le modèle de régression linéaire est le suivant :

$$y = 15,2 t - 12,6$$

2. Si on calcule le coefficient de détermination r-deux : $r^2(t, y) = 92,86\%$; on peut dire qu'on a un très bon ajustement linéaire, c.à.d. que le modèle de régression linéaire s'ajuste bien aux nuages des points qui représentent les données et explique 92,86% de la variance du CA.; la variance expliquée V_E et la variance résiduelle V_R de Y sont données dans le tableau suivant :

Variance expliquée V_E	Variance résiduelle V_R
462,08	35,52

3. Pour l'ajustement logarithmique $Y = a \log(t) + b$ on trouve un coefficient de détermination de 80,53% ; cet ajustement est de qualité inférieure à l'ajustement précédent.

Remarque : On peut tout de même chercher à améliorer légèrement le premier ajustement en considérant un ajustement logarithmique de la forme $\log(Y) = a t + b$; le coefficient de détermination est de 95,06%.

Exercice4 : (examen rattrapage avril 2017) On observe entre 1955 et 2014, le pourcentage de jeunes canadiens dans l'ensemble de la population canadienne. On résume dans le tableau suivant le pourcentage moyen des périodes de dix ans.

On veut expliquer le pourcentage par les périodes qui sont en nombre de six, pour prévoir le pourcentage moyen de la période entre 2015 et 2024.

Années	[1955;1964]	[1965;1974]	[1975;1984]	[1985;1994]	[1995;2004]	[2005;2014]
Périodes i	1	2	3	4	5	6
Pourcentage P_i (%)	33	30	23	20	19	16

1. Calculer le coefficient de corrélation et interpréter le résultat.
2. Donner l'équation de la régression linéaire en vérifiant la qualité de l'ajustement.
3. Donner le pourcentage moyen des jeunes canadiens dans toute la population pour la période $i = 7$ (cette période correspond aux années entre 2015 et 2024)

Solution de l'exercice 4 :

1. On complète le tableau suivant :

Années	[1955; 64]	[65; 74]	[75; 84]	[85; 94]	[95;2004]	[2005;2014]	Total
Périodes i	1	2	3	4	5	6	21
Pourcentage P_i (%)	33	30	23	20	19	16	141
i^2	1	4	9	16	25	36	91
P_i^2	1089	900	529	400	361	256	3535
$i * P_i$	33	60	69	80	95	96	433

Le coefficient de corrélation entre i et P est égale à

$$r = \frac{cov(P, i)}{\sigma_P * \sigma_i} = \frac{\frac{433}{6} - \frac{141}{6} * \frac{21}{6}}{\sqrt{\left(\frac{3535}{6} - \left(\frac{141}{6}\right)^2\right) * \left(\frac{91}{6} - \left(\frac{21}{6}\right)^2\right)}} = \frac{-10,083}{\sqrt{36,9167 * 2,9167}} = -0,9717$$

Interprétation : Il y a une forte liaison entre les deux variables P et i , et puisque $r < 0$ les deux variables changent dans deux sens opposés c'est-à-dire qu'avec le temps qui progresse on voit bien une diminution du pourcentage des jeunes canadiens.

2. L'équation d'ajustement de P en fonction de i s'écrit : $P = \hat{a} * i + \hat{b}$

$$\begin{cases} \hat{a} = \frac{cov(P, i)}{\sigma_i^2} \\ \hat{b} = \bar{P} - \hat{a} * \bar{i} \end{cases}$$

Après calcul on trouve :

$$\begin{cases} \hat{a} = \frac{-10,083}{2,9167} = -3,457 \\ \hat{b} = 23,5 + 3,457 * 3,5 = 35,6 \end{cases}$$

L'équation de l'ajustement sera la suivante $P = -3,457 i + 35,6$

Si on veut vérifier la qualité de l'ajustement on doit calculer le coefficient de détermination.

On trouve $r^2 = 94,42\%$, ainsi le modèle de régression linéaire explique 94,42% de la variance de P , ce pourcentage est très satisfaisant.

3. Pour calculer le pourcentage moyen des jeunes canadiens dans toute la population pour la période égale à 7 , on utilise l'équation de la régression :

$$P = -3,457 i + 35,6$$

Le pourcentage moyen des jeunes canadiens dans toute la population pour la période entre 2015 et 2024 est estimé à

$$-3,457 * 7 + 35,6 = \mathbf{11,4\%}$$

Années	[1955; 1964]	[1965; 1974]	[1975; 1984]	[1985; 1994]	[1995; 2004]	[2005; 2014]	[2015; 2024] Prévision
Périodes i	1	2	3	4	5	6	7
Pourcentage P_i (%)	33	30	23	20	19	16	11,4