Université 08Mai 1945 Guelma
1 Licence Sciences Alimentaires (Procédés agroalimentaires) (1 COFFEE)
STATISTIQUES APPLIQUEES

Corrigé-type de Série N°5 de Statistiques Appliques
(Etude de corrélation et Régression 1)

Exercice 1

Un laboratoire fabrique depuis 10 ans un vaccin destiné à immuniser contre une certaine maladie, la production en millions de doses étalée sur dix ans est fournie par le tableau suivant :

x_i = rang de l'année	0	1	2	3	4	5	6	7	8	9
y_i = production en 10^6 doses	48	45	39	33	29	26	23	20	18	15

1) Ajuster graphiquement cette distribution à deux variables. Quels sont les inconvénients de cette méthode d'ajustement ?
2) Utiliser la calculatrice pour trouver les droites de régression Dy(x) et Dx(y). Laquelle de ces deux droites vous paraît présenter le plus d'intérêt ?
3) Quelle serait la production de la $12^{ème}$ année ?

Solution

1) Ajustement aux jugés de cette distribution : C'est une droite de la forme $y = a * x + b$ qui passe par deux points par exemple le point A (1 ; 45) et le point B (8 ; 18).
Calculons donc les deux coefficients a et b à partir des coordonnées des points A et B.

$$\begin{cases} y_1 = a\,x_1 + b \\ y_2 = a\,x_2 + b \end{cases} \Rightarrow \begin{cases} 45 = a + b \\ 18 = a * 8 + b \end{cases} \Rightarrow \begin{cases} a = -3.857142857 \\ = 48.857142857 \end{cases}$$

D'où la droite d'ajustement aux jugés : $y = -3.857\,x + 48.857$
On pourrait prendre deux autres points A'(2 ; 39) et B'(7 ; 20). On aurait pu avoir une autre droite d'ajustement aux jugés…L'inconvénient de cette méthode d'ajustement est que le résultat obtenu n'est pas unique.
2) En utilisant la calculatrice : Dy(x) = 46.3455 – 3.721 x ; Dx(y) = 12.2509 – 0.262 y
Remarque : ces 2 droites passent par le point moyen G ($\bar{x} = 4.5$; $\bar{y} = 29.6$).
3) La production de la $12^{ème}$ année : $\hat{y}(11) = 5.412$

Exercice 2

Pour étudier les mécanismes hormonaux de la puberté on a mesuré les concentrations de deux hormones : l'œstradiol et l'œstrone pour un groupe de 8 adolescentes. Les résultats sont :

x_i = concentration œstradiol pg/ml	7.5	16.5	22	30	39	54	69	
y_i = concentration œstrone pg/ml	9	18.5	21.5	27	32.5	48.5	57	

On note par H le point moyen des quatre premiers points du nuage et par K le point moyen des quatre autres points.
1) Calculer les coordonnées des points H et K et déterminer la droite d'ajustement Y.
2) Utiliser la méthode des moindres carrés ordinaires pour déterminer Y(X).
3) Calculer la covariance et le coefficient de corrélation linéaire.

Solution

1) Calcul des coordonnées des points H et K : l'abscisse de H est la moyenne de x_1 à x_4 et son ordonnée est la moyenne de y_1 à y_4 d'où H (19 ; 19). l'abscisse de K est la moyenne de x_5 à x_8 et son ordonnée est la moyenne de y_5 à y_8 d'où K (59.75 ; 49).

$$\begin{cases} y_1 = a\,x_1 + b \\ y_2 = a\,x_2 + b \end{cases} \Rightarrow \begin{cases} 19 = a * 19 + b \\ 49 = a * 59.75 + b \end{cases} \Rightarrow \begin{cases} a = 0.736196319 \\ b = 5.012269939 \end{cases}$$

D'où la droite d'ajustement de Mayer : y = 0.736 x + 5.012

Soit le point moyen G ($\bar{x}$ = 39.375 ; $\bar{y}$ = 34), la droite de Mayer passe par le point G.

2) La droite des moindres carrés est y = 0.728 x + 5.3211
3) Cov (x ; y) = $\frac{13941.25}{8}$ – 39.375 * 34 = 403.90625 ; r = $\frac{403.90625}{23.5488 * 17.27721}$ = 0.99274858

Exercice 3

1) Si le coefficient de corrélation entre x et y est égal à 0.66, σ_x^2 = 34.18 et σ_y^2 = 121.36 ; $\bar{x}$ = 30 et $\bar{y}$ = 57.55 Trouver les deux droites de régression.
2) Les droites de régression relatives à un ensemble donné sont : Dy(x) → 45.414 = 2.993x – y et Dx(y) → x – 0.228y = 31.576 Calculer le coefficient de corrélation r.

Solution

1) La pente a = $\frac{0.66 * \sigma_y}{\sigma_x}$ = 1.243642887 b = y – a x = 20.24071339

Dy(x) = 1.243642887 x + 20.24071339 Dx(y) = a'y + b' avec aa' = 0.66² et b'= $\bar{x}$ – a'$\bar{y}$

Dx(y) = 0.35026132 y + 9.842460997 Ces 2 droites passent par G ($\bar{x}$ = 30 ; $\bar{y}$ = 57.55)

2) On a : a a' = r² d'où r = $\sqrt{2.993 * 0.228}$ = 0.826077478 ≈ 82.61 %

Exercice 4

Le tableau suivant concerne les âges auxquels 100 couples se sont mariés :

Classes	Femmes Y	[17 ; 22[	[22 ; 27[	[27 ; 32[	[32 ; 37[	Σ
Maris X	Centres					
[20 ; 25[		14	9	1	0	
[25 ; 30[		18	7	2	1	
[30 ; 35[		4	13	3	1	
[35 ; 40[		1	9	10	2	
[40 ; 45[		0	1	2	2	
Σ						

1) Compléter le tableau. Calculer le tableau de contingence des fréquences.
2) Calculer les distributions, les moyennes et les variances marginales de X et de Y.
3) Calculer la covariance entre X et Y ainsi que le coefficient de corrélation linéaire.
4) Trouver par la méthode des moindres carrés, les deux droites de régression Dy(x) et Dx(y).

Solution

1) Les centres et les fréquences sont figurés dans le tableau suivant :

Classes	Femmes Y	[17 ; 22[	[22 ; 27[	[27 ; 32[	[32 ; 37[	Σ
Maris X	Centres	19.5	24.5	29.5	34.5	
[20 ; 25[	22.5	0.14	0.09	0.01	0	0.24
[25 ; 30[	27.5	0.18	0.07	0.02	0.01	0.28
[30 ; 35[	32.5	0.04	0.13	0.03	0.01	0.21
[35 ; 40[	37.5	0.01	0.09	0.10	0.02	0.22
[40 ; 45[	42.5	0	0.01	0.02	0.02	0.05
Σ		0.37	0.39	0.18	0.06	1

2) La distribution marginale de x :

x_i	22.5	27.5	32.5	37.5	42.5
n_i	24	28	21	22	5

La distribution marginale de y :

y_j	19.5	24.5	29.5	34.5
n_j	37	39	18	6

Les moyennes et variances marginales sont :
$\bar{x} = 30.3$ $\bar{y} = 24.15$ $\sigma^2_x = 36.66$ $\sigma^2_y = 19.6275$

3) La covariance COV(x ; y) = 15.48
Le coefficient de corrélation r = 0.577088251

4) Dy(x) = 11.35556465 + 0.422258592 x a a' = r² d'où : a' = 0.788689338

$\bar{x} = a' \bar{y} + b'$ d'où : b' = 11.25315246 Dx(y) = 11.25315246 + 0.788689338 y

Ces 2 droites passent par le point moyen G ($\bar{x} = 30.3$; $\bar{y} = 24.15$).

Exercice 5

Une étude a été menée pour déterminer si les céréales, au son d'avoine, contribuent à abaisser le taux de cholestérol sérique. Quatorze hommes choisis au hasard ont été placés sur un régime comprenant du son d'avoine (ou des flocons de maïs) : après deux semaines, leurs taux de cholestérol des lipoprotéines de basse densité (LDL) ont été enregistrés. Chaque homme est ensuite passé à un régime alternatif. Après une deuxième période de deux semaines, le taux de cholestérol LDL de chaque homme a été enregistré à nouveau. Les données sont présentées dans le tableau suivant :

Sujet	LDL (mmol/L)	
	Son d'avoine	Flocons d'avoine
1	4.61	3.84
2	6.42	5.57
3	5.40	5.85
4	4.54	4.8
5	3.98	3.68
6	3.82	2.96
7	5.01	4.41
8	4.34	3.72
9	3.8	3.49
10	4.56	3.84
11	5.35	5.25
12	3.89	3.73
13	2.25	1.84
14	4.24	4.14

Calculer l'intervalle de confiance à 95% pour la moyenne des différences des taux de cholestérol LDL (mmol/L : son d'avoine – flocons d'avoine).

On suppose dans cet exercice que cette différence suit la loi normale.

Solution

Le tableau suivant donne les différences des taux de cholestérol LDL : (mmol/L : différence d = son d'avoine – flocons d'avoine).

1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.77	0.85	– 0.45	– 0.26	0.3	0.86	0.6	0.62	0.31	0.72	0.1	0.16	0.41	0.1

Calcul de $\bar{x}$; et de s = écart-type de l'échantillon.

La calculatrice donne : $\bar{x} = 0.363571428$; $s = 0.405455245$

On a à traiter un petit échantillon de taille $n = 14 < 30$.

Donc c'est la loi de Student qui va être utilisée avec les degrés de liberté

ddl = n – 1 = 13.

L'intervalle de confiance est donné par la formule :

$$IC = \left[\bar{x} \pm t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right] avec\ \ t_{\alpha,n-1} = 2.16$$

$$\text{I.C.} = [0.363571428 \pm 2.16\ \frac{0.405455245}{\sqrt{14}}] = [0.363571428 \pm 0.23406294]$$

$$= [0.129508487 ; 0.597634368] \approx [0.130 ; 0.598].$$

Exercice 6

Un échantillon de 80 stimulateurs cardiaques étudié a donné les résultats suivants : La moyenne est : $\bar{x} = 0.31$ et l'écart-type est s = 0.015

1) Donner un intervalle de confiance à 0.95 % pour la moyenne des stimulateurs cardiaques.
2) Quelle sera la taille n de l'échantillon si l'erreur absolue est inférieure à 0.001 ?

Solution

1) On est en présence d'un grand échantillon de taille $n = 80 > 30$. La loi normale sera utilisée.
La formule donnant l'I.C. $= [\bar{x} - u_\alpha \frac{s}{\sqrt{n}} ; \bar{x} + u_\alpha \frac{s}{\sqrt{n}}] \Rightarrow$
$P [\bar{X} - u_\alpha \frac{S}{\sqrt{n}} < \mu < \bar{X} + u_\alpha \frac{S}{\sqrt{n}}] = 1 - \alpha = 0.95$. Avec les données de l'hypothèse :

$\bar{x} = 0.31$ et $s = 0.015$ sont des observations de $\bar{X}$ et S sur un échantillon de n = 80, et de la table n° 1 on tire $u_\alpha = 1.96$ alors l'intervalle de confiance demandé sera :

$$\text{I.C.} = [0.31 - 1.96\frac{0.015}{\sqrt{80}} ; 0.31 + 1.96\frac{0.015}{\sqrt{80}}] = [0.31 \pm 0.003287] =$$

$$[0.306713 ; 0.313287] \approx [0.307 ; 0.313]$$

2) L'erreur absolue $u_\alpha \frac{s}{\sqrt{n}}$ doit être $< 0.001 \Rightarrow 1.96\frac{0.015}{\sqrt{n}} < 0.001 \Rightarrow n > 29.4^2 = 864.36$
On prend un échantillon de taille n = 865 stimulateurs pour avoir 1 erreur absolue qui ne dépasserait pas 0.001