Mathématiques pour la biologie (math39)
Licence Biologie - Aix Marseille Université
Feuille 2 : liens entre variables

Exercice 1/
Une étude de Moore et McCabe (1989) a mis en évidence l'effet de la cigarette sur le cancer du poumon dans la population anglaise. Sur 25 catégories socio-professionnelles, les auteurs de l'étude ont étudié les deux indices suivants :

- Le premier indice Y représente le nombre journalier moyen de cigarettes fumées par un individu d'une catégorie, divisé par le nombre journalier moyen de cigarettes fumées par un individu de la population totale (ici la population anglaise).
- Le deuxième indice, X, est le taux de décès par cancer du poumon dans une catégorie sur le taux de décès par cancer du poumon dans la population totale.

Les 25 catégories de l'étude sont les suivantes : 1 = Farmers, foresters, and fisherman ; 2 = Miners and quarrymen ; 3 = Gas, coke and chemical makers ; 4 = Glass and ceramics makers ; 5 = Furnace, forge, foundry, and rolling mill workers ; 6 = Electrical and electronics workers ; 7 = Engineering and allied trades ; 8 = Woodworkers ; 9 = Leather workers ; 10 = Textile workers ; 11 = Clothing workers ; 12 = Food, drink, and tobacco workers ; 13 = Paper and printing workers ; 14 = Makers of other products ; 15 = Construction workers ; 16 = Painters and decorators ; 17 = Drivers of stationary engines, cranes ; 18 = Laborers not included elsewhere ; 19 = Transport and communications workers ; 20 = Warehousemen, storekeepers, packers, and bottlers ; 21 = Clerical workers ; 22 = Sales workers ; 23 = Service, sport, and recreation workers ; 24 = Administrators and managers ; 25 = Professionals, technical workers, and artists ;

On trouve les valeurs suivantes pour X et Y :

Catégorie	Y	X
1	77	84
2	137	116
3	117	123
4	94	128
5	116	155
6	102	101
7	111	118
8	93	113
9	88	104
10	102	88
11	91	104
12	104	129
13	107	86
14	112	96
15	113	144
16	110	139
17	125	113
18	133	146
19	115	128
20	105	115
21	87	79
22	91	85
23	100	120
24	76	60
25	66	51

Le nuage de points associé est donné dans la figure 1.

1. Quelle est la population statistique ? Quelles sont les variables observées ?

2. Quel lien peut-il exister entre les variables ?

3. On donne les valeurs des moyennes et des variances de X et de Y : $m_X = 102.88$, $m_Y = 109$, $Var(X) = 283.9$, $Var(Y) = 654.6$. On donne également la covariance entre X et Y : $COV(X, Y) = 308.8$. Rappeler la définition de la covariance et de la corrélation.

4. Calculer le coefficient de corrélation entre X et Y.

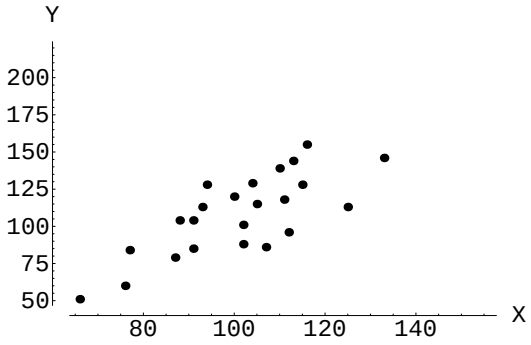

Figure 1: Nuages de points formé par X et Y

5. En déduire la relation qui existe entre ces deux variables.
6. Comment l'interpréter ?
7. Comment pourrait-on prendre en compte la variable catégorie socio-professionnelle ?

Exercice 2/
Commenter les nuages de points de la figure 2.

Exercice 3/
Pour savoir si la couleur des yeux et la couleur des cheveux sont des caractères indépendants on a dressé le tableau suivant à partir d'une population de 6800 individus (cf Tassi) :

⋱ Cheveux Yeux ⋱	blond	brun	noir	roux	total
bleu	1768	807	189	47	2811
gris/vert	946	1387	746	53	3132
brun	115	438	288	16	857
total	2829	2632	1223	116	6800

1. Quelle est la population statistique, les variables étudiées ?

2. Quel lien peut-on mesurer entre la couleur des yeux et celle des cheveux ?

3. Si ces deux variables étaient indépendantes on devrait avoir en théorie les effectifs qui sont donnés dans le tableau suivant

⋱ Cheveux Yeux ⋱	blond	brun	noir	roux	total
bleu	1169	1088	506	48	$n_1 = 2811$
gris/vert	1303	1212	563	53	$n_2 = 3132$
brun	357	332	154	15	$n_3 = 857$
total	$q_1 = 2829$	$q_2 = 2632$	$q_3 = 1223$	$q_4 = 116$	$n = 6800$

   Rappeler la méthode de calcul de ces effectifs.

4. Calculer la valeur de D, la distance du Khi-deux obtenue à partir de ces deux tableaux.

5. Quelle est la valeur maximum possible pour D.

6. Que penser vous du lien entre les deux variables ?

7. Quelle est la part de ce lien qui s'explique par les individus blonds aux yeux bleus ?

Exercice 4/
On demande à deux médecins urgentistes de classer 8 symptômes du plus grave, au moins grave. On classe leurs réponses dans le tableau suivant :

Symptôme	Rang médecin 1	Rang médecin 2
1	1	3
2	3	2
3	5	5
4	6	4
5	7	7.5
6	4	7.5
7	8	6
8	2	1

1. Quelle est la population statistique et les variables étudiées ?
2. Comment peut-on étudier le lien entre les variables ?
3. D'où vient la valeur 7.5 dans le tableau ?
4. Calculer les moyennes et les variances des deux séries de rangs.
5. On donne la covariance entre les deux séries de rangs : $COV(R_1, R_2) = 3.56$. Calculer le coefficient de corrélation de Spearman entre les deux jugements des médecins.
6. Comment peut-on l'interpréter ?

Exercice 5/
On désire mesurer l'efficacité d'un traitement, noté traitement A, sur des patients atteints de leucémie myéloïde chronique. Après un an de traitement on mesure le taux (en pourcentages) de chromosome de Philadelphie chez 17 personnes ayant suivi le traitement et chez 16 personnes sans traitement atteintes du même cancer. Le tableau suivant contient ces pourcentages (un pourcentage nul signifie la disparition du chromosome responsable de la maladie)

Sans traitement	Avec traitement
35	44
64	81
87	18
65	16
20	25
95	32
50	0
55	0
25	71
64	11
90	38
91	21
54	81
0	12
16	10
85	0
	10

1. Quelle est la population statistique et les variables étudiées ?
2. Comment peut-on mesurer le lien entre les variables ?
3. On donne les moyennes des individus (A=traitement, ST = sans traitement) : $m_A = 27.65$ et $m_{ST} = 56$. On donne également les variances par groupe d'individus : $VAR(A) = 687.3$ et $VAR(ST) = 836.75$. Puis la variance totale $V = 960.5$. Calculer la variance expliquée par le facteur traitement.
4. Calculer la variance résiduelle.
5. Retrouver la variance totale.
6. Calculer le rapport de corrélation.
7. Interpréter ce résultat.

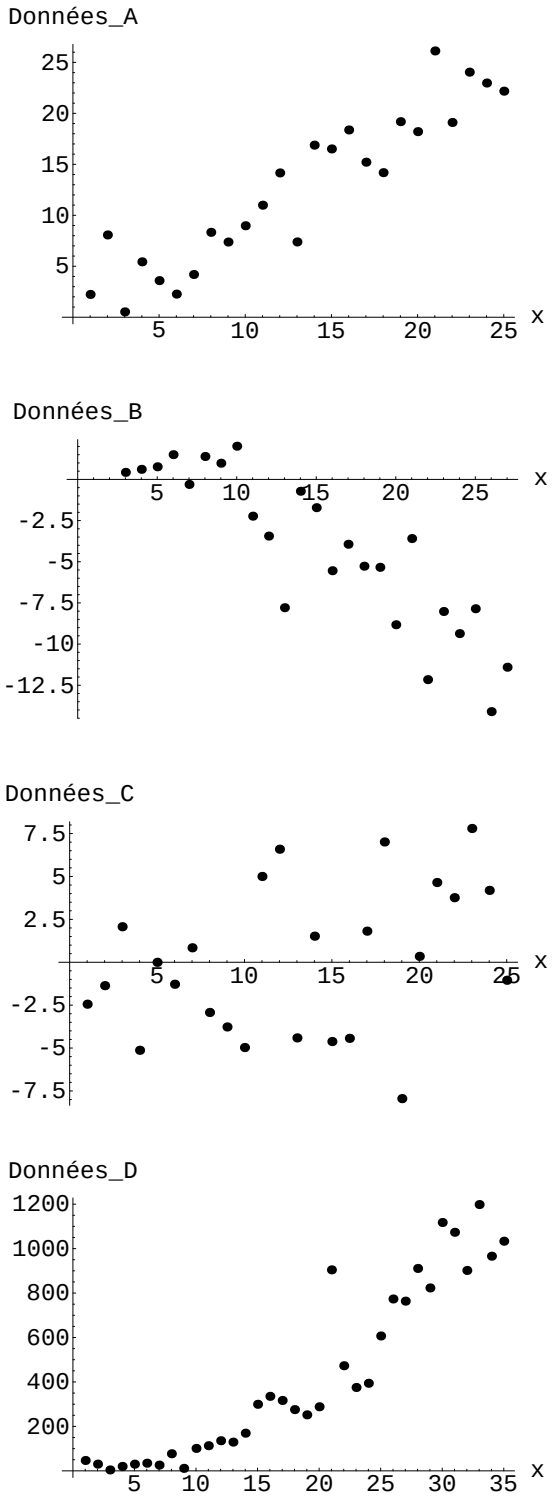

Figure 2: Quatre nuages de points