ECE 2 - Année 2017-2018
Lycée français de Vienne
Mathématiques - F. Gaunard
http://frederic.gaunard.com

T.P. n°4

Simulations de couples de v.a. discrètes
Statistiques descriptives bivariées

1 Introduction

Dans une population donnée, il peut arriver que l'on souhaite étudier simultanément deux caractères X et Y. On peut alors s'intéresser aux propriétés de chacun de ces deux caractères séparément (statistiques univariées) mais aussi, on peut s'intéresser au(x) lien(s) entre ces deux caractères (statistiques bivariées). On étudie alors le couple de caractères $Z = (X, Y)$.

☞ Les notions introduites dans ce TP permettent de donner un sens pratique aux notions probabilistes introduites dans le Chapitre 9 du cours (*Couples de v.a. discrètes*).

Notamment, on peut penser que l'une des deux variables, X par exemple, est une cause l'autre (Y dans ce cas). On dit alors que X est la **variable explicative** et que Y est la **variable à expliquer**.

Exemple. En interrogeant un échantillon de la population mondiale Ω, on peut étudier le lien entre l'âge et l'acuité visuelle. *A priori*, la variable explicative est alors l'âge (caractère X) et la variable à expliquer est l'acuité visuelle (caractère Y).

Pour le $k-$ième individu de l'échantillon interrogé, on note $M_k = (x_k, y_k)$ le couple de résultats obtenus et l'idée naturelle est alors de tracer le **nuage de points** comme expliqué ci-après.

2 Modèle de régression

2.1 Nuage de points et point moyen

On appelle **nuage de points** associé à la série statistique (X, Y) l'ensemble des points M_k de coordonnées (x_k, y_k) tracés dans un repère orthonormé du plan (où $X = (x_k)$ et $Y = (y_k)$).

L'examen du nuage de points permet de faire des constatations qualitatives:

- est-il concentré ou dispersé?
- relève-t-on une tendance?
- y a-t-il des valeurs *a priori* aberrantes?

Le **point moyen** du nuage est le point de coordonnées $(\overline{x}, \overline{y})$, où $\overline{x}$ désigne la moyenne des x_k et $\overline{y}$ celle des y_k.

☞ On utilise la fonction `plot2d()` pour représenter le nuage de points ainsi que le point moyen (que l'on différencie).

Exercice 1. Nous allons comparer deux sondage sur le lien entre la taille X et le poids Y d'un individu au Japon et aux États-Unis.

(1) *A priori*, quelle est la variable explicative et la variable à expliquer?
(2) On représente les résultats du sondage au Japon dans le tableau ci-dessous

individu i	1	2	3	4	5	6	7	8
taille x_i (cm)	161	170	152	181	163	145	168	175
poids y_i (kg)	58	66	52	73	60	45	65	68

(a) Tracer le nuage de points de cet exemple et placer le point moyen avec un symbole différent.
(b) Modifier les axes: on souhaite que l'axe des abscisses aille de 140 à 200 (cm) et que l'axe des ordonnées aille de 0 à 120 (kg).

(3) Mêmes questions avec le tableau des résultats du sondage aux États-Unis

individu i	1	2	3	4	5	6	7	8
taille x_i (cm)	169	195	177	182	166	155	189	174
poids y_i (kg)	90	95	115	90	70	60	80	70

2.2 Fonction de régression et erreur d'ajustement

Si l'on considère X comme variable explicative et Y comme variable à expliquer, chercher un **modèle de régression** consiste à savoir si Y est (à peu près) fonction de X (*i.e.* $Y = f(X)$) ou plus précisément à mettre Y sous la forme

$$Y = f(X) + \varepsilon,$$

où ε est une variable aléatoire appelée **erreur d'ajustement**. On considère alors que la régression est correcte si ε est d'espérance nulle et que le nuage de points des **résidus** (X, ε) ne se répartit pas autour de 0 selon une direction précise.

Si une telle formule existe, on dit que les variables sont **corrélées**. Sinon, on dit qu'elles sont non corrélées.

Exercice 2. Dans chacun des deux cas de l'Exercice 1, dire si les variables X et Y sont corrélées ou non.

Exercice 3.

(1) Recopier et compléter les instructions suivantes pour créer une série statistique `X` dont les composantes vérifient $x_i = i + \mathcal{U}_i$, où les $\mathcal{U}_i$ sont des v.a. indépendantes de même loi uniforme sur $[-1/2; 1/2]$ et créer une série statistique Z dont les composantes sont des réalisations de variables aléatoires indépendantes suivant une loi normale $\mathcal{N}(0, 0.25)$.

```
K=1:20;
X= K+ grand(1, 20, ......., ........, ......)
Z=grand(1, 20, 'nor', 0, .........)
```

(2) On définit $Y = X + Z$.
(a) Représenter le nuage de points associé au couple (X, Y) et son point moyen.
(b) Intepréter le résultat obtenu à l'ajout de la commande

```
plot2d(K, K, style=5)
```

2.3 Covariance empirique

En complète analogie avec la covariance (théorique) définie dans le cours, on peut définir la **covariance empirique** d'un échantillon du couple (X, Y) par la formule

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}),$$

où, comme précédemment, $\overline{x}$ et $\overline{y}$ désignent les moyennes des séries X et Y. De cette définition, on peut tirer une formule permettant de calculer celle-ci avec SciLab

```
cov=mean((x-mean(x)).*(y-mean(y)))
```

☞ Utiliser la formule de König-Huygens pour proposer une autre instruction permettant le calcul de la covariance

```
cov=.....................................................
```

☞ La commande `corr(x,y,1)` renvoie également la covariance du couple `(x,y)`. En particulier, `corr(x,x,1)` permet d'obtenir la variance de `x`.

2.4 Tableau de contingence

Parfois, les données statistiques apparaissent sous la forme d'un **tableau de contingence**. Plus précisément, si $x_1, ..., x_\ell$ sont les valeurs prises par X et $y_1, ..., y_m$ celles prises par Y, on note $f_{i,j}$ la fréquence d'apparition du couple (x_i, y_j).

$X \setminus Y$	y_1	y_2	$\cdots$	y_m	
x_1	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,m}$	$f_{1,\bullet}$
x_2	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,m}$	$f_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_ℓ	$f_{\ell,1}$	$f_{\ell,2}$	$\cdots$	$f_{\ell,m}$	$f_{\ell,\bullet}$
	$f_{\bullet,1}$	$f_{\bullet,2}$	$\cdots$	$f_{\bullet,m}$	

☞ $f_{i,\bullet} = \sum_{j=1}^{m} f_{i,j}$ est la **fréquence marginale** d'apparition de x_i;

☞ $f_{\bullet,j} = \sum_{i=1}^{\ell} f_{i,j}$ est la **fréquence marginale** d'apparition de y_j;

On peut alors former la matrice de contingence $F = (f_{i,j})$. Dans ce cas, l'instruction `covar(X, Y, F)` renvoie $\mathrm{cov}(X, Y)$.

Exercice 4. On considère une variable aléatoire X suivant une loi de Poisson de paramètre 2 et Y une variable aléatoire dont la loi conditionnelle sachant $X = k$ est une loi de Poisson de paramètre k.

(1) Recopier et compléter la suite d'instructions ci-dessous pour effectuer 1000 réalisations indépendantes du couple (X, Y) et conserver les résultats dans deux vecteurs ligne `X` et `Y`.

```
X=grand(1, 1000, ............ , 2);
Y=zeros(1, 1000);
for k=1:1000
    Y(k)=grand(1, 1, ............, ............);
end
```

(2) Déterminer le maximum `l` des valeurs prises par `X` et le maximum `m` de celles prises par `Y`. Construire la matrice de contingence F. (Attention: le couple (X, Y) est à valeurs dans $[\![0; \mathtt{l}]\!] \times [\![0; \mathtt{m}]\!]$.)

(3) Déterminer les fréquences marginales de `X` et `Y`.

(4) Déterminer le coefficient de corrélation de `X` et `Y`.

3 Régression linéaire : Méthode des moindres carrés

On se place dans la situation où l'on souhaite savoir si la la "courbe" de Y en fonction de X peut être approximée par une droite. On cherche donc deux constante a et b telles que

$$Y = aX + b + \varepsilon.$$

On utilise alors la méthode des moindres carrés qui nous donne l'équation de la droite la plus proche des points en terme de distance, c'est à dire l'unique droite D d'équation $y = ax + b$ qui rend minimale la somme des carrés des erreurs d'ajustement

$$d^2(a,b) = \frac{1}{n}\sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

Le résultat suivant donne la valeur de a et b et est admis. Sa démonstration repose sur des outils d'*algèbre bilinéaire* non accessibles avec le programme d'ECE (la série statistique $aX + b\mathbf{1}$ est le *projeté orthogonal* de Y sur le plan de $\mathbb{R}^n$ engendré par X et $\mathbf{1}$ pour le produit scalaire $\langle X, Y\rangle = E(XY)$).

Théorème 1. *La droite la plus proche du nuage de points associé au couple (x, y) est la droite d'équation $y = ax + b$ avec*

$$a = \frac{\mathrm{cov}(x,y)}{V(x)}, \quad \textit{et} \quad b = \overline{y} - a/\overline{x}.$$

En particulier, cette droite passe par le point moyen $(\overline{x}, \overline{y})$.

Exercice 5. Déterminer et représenter (avec le nuage de points) la droite de régression pour le premier cas (Japon) de l'Exercice 1.

Tout comme dans le cours, on définit le coefficient de corrélation linéaire (empirique) du couple (x, y) le nombre ρ par la formule

$$\rho = \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)},$$

où $\sigma(x)$ et $\sigma(y)$ désignent les écarts-type de x et de y.

Théorème 2. *Soit ρ le coefficient de corrélation linéaire du couple (x, y). Alors*

(i) $\rho \in [-1; 1]$*;*

(ii) $\rho = \pm 1$ *si et seulement si la régression $Y = aX + b$ est exacte.*

☞ Par conséquent, si $|\rho|$ est proche de 1 **et qu'on a visualisé une relation linéaire entre les données**, on peut confirmer qu'il y a bien corrélation linéaire entre X et Y.

☞ En sciences humaines et en sciences économiques, une valeur de $|\rho|$ de l'ordre de $0,85$ est souvent considérée comme bonne.

Exercice 6. Un chercheur en sociologie veut analyser, s'il existe une relation linéaire entre la densité de population dans les villes et le taux de criminalité correspondant. Le taux de criminalité Y est indiqué en nombre de crimes par 10000 habitants et la densité de population X est mesurée en milliers d'habitants par km^2.

Région i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	7.1	5.8	11.5	2.1	3.7	3.6	7.5	4.2	3.8	10.3	8.6	7.2
y_i	12	9	15	4	4	2	10	3	5	11	10	11

(1) Si la région 3 a une superficie de 20 km^2, quel est le nombre de crimes dans cette région?
(2) Tracer le nuage de points de ces observations.
(3) Calculer les coefficients de la droite de régression.
(4) Le taux de criminalité et la densité de population sont-ils corrélés?

(5) À quelle augmentation du taux de criminalité pouvons-nous nous attendre pour une variation de 1000 habitants par km^2 de la densité de population?
(6) Estimer le taux de criminalité le plus plausible pour une densité de population de 7500 habitants par km^2.

4 Extraits d'annales de concours

Exercice 7. (D'après **ECRICOME 2016**)

Dans tout l'exercice, X et Y sont deux variables aléatoires définies sur le même espace probabilisé et à valeurs dans $\mathbb{N}$. On dit que les deux variables X et Y sont **échangeables** si :

$$\forall (i,j) \in \mathbb{N}^2, \quad P([X=i] \cap [Y=j]) = P([X=j] \cap [Y=i])$$

Résultats préliminaires

(1) On suppose que X et Y sont deux variables indépendantes et de même loi.
Montrer que X et Y sont échangeables.
(2) On suppose que X et Y sont échangeables. Montrer, à l'aide de la formule des probabilités totales, que :

$$\forall i \in \mathbb{N}, \quad P(X=i) = P(Y=i)$$

Étude d'un exemple

Soient n, b et c trois entiers strictement positifs.
Une urne contient initialement n boules noires et b boules blanches. On effectue l'expérience suivante, en distinguant trois variantes.

- On pioche une boule dans l'urne. On définit X la variable aléatoire qui vaut 1 si cette boule est noire et 2 si elle est blanche.
- On replace la boule dans l'urne et :
  - ⋆ Variante 1 : on ajoute dans l'urne c boules de la même couleur que la boule qui vient d'être piochée.
  - ⋆ Variante 2 : on ajoute dans l'urne c boules de la couleur opposée à celle de la boule qui vient d'être piochée.
  - ⋆ Variante 3 : on n'ajoute pas de boule supplémentaire dans l'urne.
- On pioche à nouveau une boule dans l'urne. On définit Y la variable aléatoire qui vaut 1 si cette seconde boule piochée est noire et 2 si elle est blanche.

(3) (a) Compléter la fonction Scilab suivante, qui simule le tirage d'une boule dans une urne contenant b boules blanches et n boules noires et qui retourne 1 si la boule tirée est noire, et 2 si la boule tirée est blanche.

```
function res = tirage( b , n )
        r = rand()
        if ..... then
                res = 2
        else
                res = 1
        end
endfunction
```

(b) Compléter la fonction suivante, qui effectue l'expérience étudiée avec une urne contenant initialement b boules blanches, n boules noires et qui ajoute éventuellement c boules après le premier tirage, selon le choix de la variante dont le numéro est `variante`.
Les paramètres de sortie sont :

- `x` : une simulation de la variable aléatoire X
- `y` : une simulation de la variable aléatoire Y

```
function [ x , y ] = experience ( b , n , c , variante )
        x = tirage ( b , n )
        if variante == 1 then
                if x == 1 then
                        ..........
                else
                        ..........
                end
        else if variante == 2 then
                ..........
                ..........
                ..........
                ..........
                ..........
        end
        y = tirage ( b , n )
endfunction
```

(c) Compléter la fonction suivante, qui simule l'expérience N fois (avec $N \in \mathbb{N}^*$), et qui estime la loi de X, la loi de Y et la loi du couple (X, Y).
Les paramètres de sortie sont :

- `loiX` : un tableau unidimensionnel à deux éléments qui estime [P(X=1), P(X=2)]
- `loiY` : un tableau unidimensionnel à deux éléments qui estime [P(Y=1), P(Y=2)]
- `loiXY` : un tableau bidimensionnel à deux lignes et deux colonnes qui estime :

$$\begin{bmatrix} P([X=1]\cap[Y=1]) & P([X=1]\cap[Y=2]) \\ P([X=2]\cap[Y=2]) & P([X=1]\cap[Y=2]) \end{bmatrix}$$

```
function [ loiX, loiY , loiXY ] = estimation(b,n,c,variante,N)
        loiX = [ 0 , 0 ]
        loiY = [ 0 , 0 ]
        loiXY = [ 0 , 0 ; 0 , 0 ]
        for k = 1 : N
                [x , y] = experience( b , n , c , variante )
                loiX(x) = loiX(x) + 1
                ..........
                ..........
        end
        loiX = loiX / N
        loiY = loiY / N
        loiXY = loiXY / N
endfunction
```

(d) On exécute notre fonction précédente avec $b = 1$, $n = 2$, $c = 1$, $N = 10000$ et dans chacune des variantes. On obtient :

```
-->[loiX,loiY,loiXY] = estimation(1,2,1,1,10000)
 LoiXY =
        0.49837 0.16785
        0.16697 0.16681
 LoiY =
        0.66534 0.33466
```

```
 LoiX =
        0.66622 0.33378

-->[loiX,loiY,loiXY] = estimation(1,2,1,2,10000)
 LoiXY =
        0.33258 0.33286
        0.25031 0.08425
 LoiY =
        0.58289 0.41711
 LoiX =
        0.66544 0.33456

-->[loiX,loiY,loiXY] = estimation(1,2,1,3,10000)
 LoiXY =
        0.44466 0.22098
        0.22312 0.11124
 LoiY =
        0.66778 0.33222
 LoiX =
        0.66564 0.33436
```

En étudiant ces résultats, émettre des conjectures quant à l'indépendance et l'échangeabilité de X et Y dans chacune des variantes.
On donne les valeurs numériques approchées suivantes :

$$0.33 \times 0.33 \simeq 0.11$$
$$0.33 \times 0.41 \simeq 0.14$$
$$0.33 \times 0.58 \simeq 0.19$$
$$0.33 \times 0.66 \simeq 0.22$$
$$0.41 \times 0.66 \simeq 0.27$$
$$0.58 \times 0.66 \simeq 0.38$$
$$0.66 \times 0.66 \simeq 0.44$$

(4) On se place dans cette question dans le cadre de la variante 1.
   (a) Donner la loi de X.
   (b) Déterminer la loi du couple (X, Y).
   (c) Déterminer la loi de Y.
   (d) Montrer que X et Y sont échangeables mais ne sont pas indépendantes.

Exercice 8. (D'après **HEC 2016**)

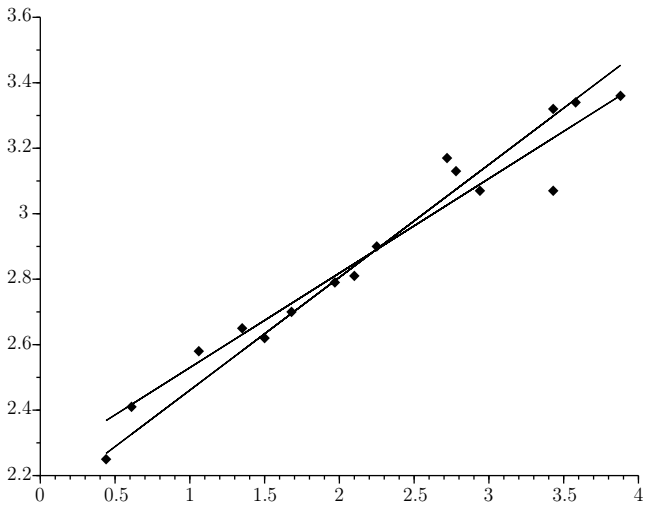
On rappelle qu'en `Scilab`, les commandes `variance` et `corr` permettent de calculer respectivement la variance d'une série statistique et la covariance d'une série statistique double.
Si $(v_i)_{1\leq i\leq n}$ et $(w_i)_{1\leq i\leq n}$ sont deux séries statistiques, alors la variance de $(v_i)_{1\leq i\leq n}$ est calculable par `variance(v)` et la covariance de $(v_i, w_i)_{1\leq i\leq n}$ est calculable par `corr(v,w,1)`.

(1) On a relevé pour $n = 16$ entreprises qui produisent le bien considéré à l'époque donnée, les deux séries statistiques $(u_i)_{1\leq i\leq n}$ et $(t_i)_{1\leq i\leq n}$ reproduites dans les lignes (1) et (2) du code `SciLab` suivant dont la ligne (5) est incomplète :

```
(1) u=[1.06,0.44,2.25,3.88,0.61,1.97,3.43,2.10,1.50,1.68,2.72,1.35,2.94,2.78,3.43
,3.58]
(2) t=[2.58,2.25,2.90,3.36,2.41,2.79,3.32,2.81,2.62,2.70,3.17,2.65,3.07,3.13,3.07
,3.34]
```

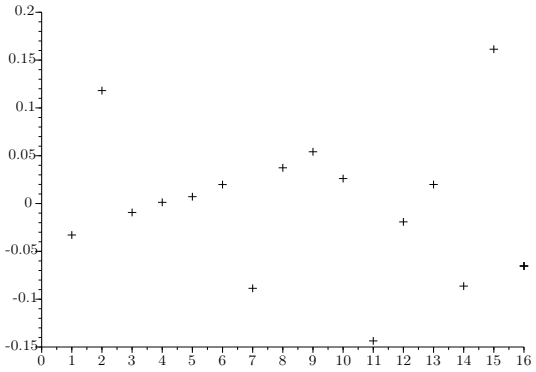
```
(3) plot2d(u,t,-4) // -4 signifie que les points sont représentés par des
losanges.
(4) plot2d(u,corr(u,t,1)/variance(u)*u+mean(t)-corr(u,t,1)/variance(u)*mean(u))
// équation de la droite de régression de t en u.
(5) plot2d(u,.....)
// équation de la  droite de régression de u en t.
```

Le code précédent complété par la ligne (5) donne alors la figure suivante :



(a) Compléter la ligne (5) du code permettant d'obtenir la figure précédente.
(b) Interpréter le point d'intersection des deux droites de régression.
(c) Estimer graphiquement les moyennes empiriques $\overline{u}$ et $\overline{t}$.
(d) Le coefficient de corrélation empirique de la série statistique double $(u_i, t_i)_{1\leq i\leq 16}$ est-il plus proche de -1, de 1 ou de 0 ?
(e) On reprend les lignes (1) et (2) du code précédent que l'on complète par les instructions (6) à (11) qui suivent et on obtient le graphique ci-dessous :

```
(6) a0=corr(u,t,1)/variance(u)
(7) b0=mean(t)-corr(u,t,1)/variance
(u)*mean(u)
(8) t0=a0*u+b0
(9) e=t0-t
(10) p=1:16
(11) plot2d(p,e,-1)
```



Que représente ce graphique ? Quelle valeur peut-on conjecturer pour la moyenne des ordonnées des 16 points obtenus sur le graphique ? Déterminer mathématiquement la valeur de cette moyenne.