Table des matières

Chapitre 1

Variable Statistique à deux dimensions

Introduction

Dans la statistique descriptive on s'intéresse à étudier un caractère ou une variable pour une population donnée. On peut être mené à étudier deux caractères simultanément, on fait appel dans ce cas à la statistique à deux variables.

1.1 Définition

Soient X, Y ceux variables définies sur le même ensemble Ω, $card(\Omega) = n$ le couple $Z = (X, Y)$ est appelé variable statistique à deux dimensions.
L'ensemble des valeurs prises par la variable statistique Z=(X,Y) est donné par le tableau :

ω_i	ω_1	ω_2	...	ω_n
$X(\omega_i)$	$X(\omega_1)$	$X(\omega_2)$ ...		$X(\omega_n)$
$Y(\omega_i)$	$Y(\omega_1)$	$Y(\omega_2)$	...	$Y(\omega_n)$

Soient $x_1,\ x_2, ..., x_n$ les valeurs prises par la V.S X et $y_1,\ y_2, ..., y_n$ les valeurs prises par la V.S Y

Exemple 1 :
Dans le tableau suivant on donne pour chaque ville, le nombre moyen d'heures d'ensoleillement par an ainsi que la température moyenne.

Ville	V1	V2	V3	V4	V5	V6
Heures d'ensoleillement	2790	2072	2767	1729	1574	1833
Température	14.7	11.4	14.2	10.8	9.7	11.2

La population : les 6 villes.
la variable 1 : Nombre d'heures d'ensoleillement par an.
la variable 2 : La température moyenne.

Exemple 2 :

Le tableau ci dessous permet de suivre l'évolution de l'espérance de vie des femmes en France de 1990 à 1999.

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Espérance de vie	80.9	81.1	81.4	81.8	81.9	82	82.3	82.4	82.4	

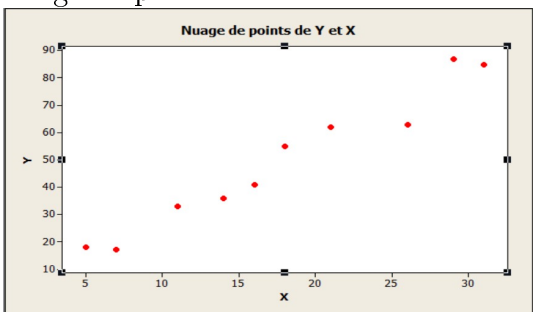
La population : les femmes de France.
La variable 1 : l'année.
La variable 2 : l'espérance de vie.

Remarque :

Dans une série statistique à deux variables si l'une des variable est le temps (année par exemple), on l'appelle série chronologique.

1.2 Présentation des données

*La première présentation d'une série statistique double est celle qui est de la forme $z_i = (x_i, y_i)$, où les données sont sous forme de couples.
Les données sont représentées par des points dans un repère orthonormés, nuage de points.



*La deuxième présentation est celle du tableau de contingence.

1.2.1 Tableau de contingence

***Tableau de contingence des effectifs :**

L'effectif de la valeur (x_i, y_j), noté n_{ij}.
$n_{ij} = \ card\,\{\omega \in \Omega,\ tq(X(\omega), Y(\omega)) = (x_i, y_j),\ \text{où } 1 \leq i \leq l,\ 1 \leq j \leq k\}$

$x_i \vert y_j$	y_1	y_2	...	y_k	Marge de X
x_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
...	...	...	...	...	...
x_l	n_{l1}	n_{l2}	...	n_{lk}	$n_{l.}$
Marge de Y	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	N

L'effectif marginal de la V.S X noté $n_{i.} = \sum_{j=1}^{k} n_{ij}$.
L'effectif marginal de la V.S Y noté $n_{.j} = \sum_{i=1}^{l} n_{ij}$.
$\sum_{j=1}^{k} \sum_{i=1}^{l} n_{ij} = \sum_{i=1}^{l} \sum_{j=1}^{k} n_{ij} = N$.
$card(\Omega) = N$.

***Tableau de contingence des fréquences :**

La fréquence de la valeur (x_i, y_j) noté $f_{ij} = \frac{n_{ij}}{N}$, $1 \leq i \leq l,\ 1 \leq j \leq k$.
La loi de la variable statistique $Z = (X, Y)$ résume l'ensemble des valeurs $x_i,\ y_j,\ f_{ij}$, il se présente sous la forme suivante ;

$x_i \vert y_j$	y_1	y_2	...	y_k	Marge de X
x_1	f_{11}	f_{12}	...	n_{1k}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2k}	$f_{2.}$
...	...	...	...	...	...
x_l	f_{l1}	f_{l2}	...	f_{lk}	$f_{l.}$
Marge de Y	$f_{.1}$	$f_{.2}$	...	$f_{.k}$	1

La fréquence marginale de la valeur x_i notée $f_{i.} = \sum_{j=1}^{k} f_{ij}$.
La fréquence marginale de la valeur y_j notée $f_{.j} = \sum_{i=1}^{l} f_{ij}$.
***Remarque :**
$\sum_{i=1}^{l} \sum_{j=1}^{k} f_{ij} = \sum_{j=1}^{k} \sum_{i=1}^{l} f_{ij} = 1$.
Exemple :
Une expérience a été réalisée sur 234 personnes pour étudier la relation qui existe entre l'âge X en années et le temps de sommeil Y en heures, on a obtenu le tableau suivant ;

X|Y	$[5, 7[$	$[7, 9[$	$[9, 11[$	$[11, 15]$	Marge de X
$[1, 3[$	0	0	2	36	38
$[3, 11[$	0	3	12	26	41
$[11, 19[$	2	8	35	16	61
$[19, 31[$	0	26	22	3	51
$[31, 59]$	22	15	6	0	43
Marge de Y	24	52	77	81	234

le tableau de contingence des fréquences sera alors

X|Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$	Marge de X
$[1,3[$	0	0	$2/234$	$36/234$	$38/234$
$[3,11[$	0	$3/234$	$12/234$	$26/234$	$41/234$
$[11,19[$	$2/234$	$8/234$	$35/234$	$16/234$	$61/234$
$[19,31[$	0	$26/234$	$22/234$	$3/234$	$51/234$
$[31,59]$	$22/234$	$15/234$	$6/234$	0	$43/234$
Marge de Y	$24/234$	$52/234$	$77/234$	$81/234$	1

1.2.2 Fréquence marginale conditionnelle de y_j sachant que $X = x_i$

La fréquence marginale conditionnelle de y_j sachant que $X = x_i$ notée $f_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$.
La loi marginale de Y sachant que $X = x_i$ est représentée par le tableau suivant ;

Y	y_1	...	y_k
$f_{j\|i}$	$\frac{f_{i1}}{f_{i.}}$	...	$\frac{f_{ik}}{f_{i.}}$

Exemple :le tableau suivant donne la loi conditionnelle de Y sachant que $X \in [11,19[$

Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$
$f_{j\|X\in[11,19[}$	$2/61$	$8/61$	$35/61$	$16/61$

Interprétation ; par exemple $(8/61) * 100$ est le pourcentage des personnes agées entre 11 et 19 ans qui ont une durée de sommeil entre 7 et 9heures.

1.2.3 Fréquence marginale conditionnelle de x_i sachant que $Y = y_j$

La fréquence marginale conditionnelle de x_i sachant que $Y = y_j$ notée $f_{i|j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$.
La loi marginale de X sachant que $Y = y_j$ est représentée par le tableau suivant ;

X	x_1	...	x_l
$f_{i\|j}$	$\frac{f_{1j}}{f_{.j}}$	...	$\frac{f_{lj}}{f_{.j}}$

Exemple :le tableau suivant donne la loi conditionnelle de X sachant que $Y \in [7,9[$

X	$[1,3[$	$[3,11[$	$[11,19[$	$[19,31[$	$]31,59]$
$f_{i\|Y\in[7,9[}$	0	$3/52$	$8/52$	$26/52$	$15/52$

Interprétation ; Parmi ceux qui ont une durée de sommeil entre 7 et 9heures

$(26/52) * 100$ ont entre 19 et 31 ans.

1.3 Indépendance entre deux variables statistiques

Deux variables X et Y sont dites indépendantes si

$$f_{ij} = f_{i.} * f_{.j}, \quad i \in \{1, ..., l\} \quad \text{et} \quad j \in \{1, ..., k\}.$$

On peut vérifier l'indépendance entre X et Y à partir des distributions conditionnelles ;

$$\forall i \in \{1, ..., l\}, \ f_{i|j} = f_{i.}, \quad \text{pour tout} \quad j \in \{1, ..., k\}.$$

et

$$\forall j \in \{1, ..., k\}, \ f_{j|i} = f_{.j}, \quad \text{pour tout} \quad i \in \{1, ..., l\}$$

Exemple : $f_{11} = 0$, $f_{1.} = 38/234$ et $f_{.1} = 24/234$ donc on n'a pas l'indépendance entre l'age et la durée du sommeil.

1.4 Paramètres marginaux

1.4.1 Moyennes marginales

La moyenne marginale de X est $\bar{x}$; $\bar{x} = \frac{N}{\sum_i n_{i.} x_i} = \sum_i f_{i.} x_i$.
La moyenne marginale de Y est $\bar{y}$; $\bar{y} = \frac{N}{\sum_j n_{.j} y_j} = \sum_j f_{.j} y_j$.
Exemple : Pour calculer les moyennes marginales, on doit calculer les centres de classes pour les deux variables X et Y.
*La loi marginale de X

X	$[1, 3[$	$[3, 11[$	$[11, 19[$	$[19, 31[$	$]31, 59]$
centre de classes c_i	2	7	15	25	45
$f_{i.}$	$38/234$	$41/234$	$61/234$	$51/234$	$43/234$

Pour calculer la moyenne on doit calculer le produit $f_{i.} * c_i$ et ensuite faire la somme.
$\sum_i f_{i.} * c_i = 2 * 38/234 + 7 * 41/234 + 15 * 61/234 + 25 * 51/234 + 45 * 43/234$
$= 19.179$
*La loi marginale de Y

Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$
centre de classes c_j	6	8	10	13
$f_{.j}$	$24/234$	$52/234$	$77/234$	$81/234$

Calculons d'abord le produit $f_{.j} * c_j$ ensuite la somme ;
$\sum_j f_{.j} * c_j = 6 * 24/234 + 8 * 52/234 + 10 * 77/234 + 13 * 81/234 = 10.18$

1.4.2 Variances marginales

La variance marginale de X est V(X) ; $V(X) = \sum_i f_{i.} x_i^2 - \bar{x}^2$.
$V(X) = 2^2 * 38/234 + 7^2 * 41/234 + 15^2 * 61/234 + 25^2 * 51/234 + 45^2 * 43/234 - 19.179^2 = 576.222 - 367.834 = 208.388.$
La variance marginale de Y est V(Y) ;$V(Y) = \sum_j f_{.j} y_j^2 - \bar{y}^2$.
Exemple : $V(Y) = 6^2 * 24/234 + 8^2 * 52/234 + 10^2 * 77/234 + 13^2 * 81/234 - 10.18^2 = 109.3205 - 103.6324 = 5.6881$

1.4.3 Les moyennes conditionnelles

La moyenne conditionnelle de X sachant que $Y = y_j$ notée par $\bar{x}_j$

$$\bar{x}_j = \sum_{i=1}^{l} \frac{f_{ij}}{f_{.j}} x_i$$

. **Exemple :**la loi conditionnelle de X sachant que $Y \in [7,9[$ est donné parle tableau suivant :

X	$[1,3[$	$[3,11[$	$[11,19[$	$[19,31[$	$]31,59]$
$f_{i\|Y\in[7,9[}$	0	$3/52$	$8/52$	$26/52$	$15/52$

$\bar{x}_j = 2 * 0 + 7 * 3/52 + 15 * 8/52 + 25 * 26/52 + 45 * 15/52 = 28.192.$
La moyenne conditionnelle de Y sachant que $X = x_i$ notée par $\bar{y}_i$,

$$\bar{y}_i = \sum_{j=1}^{l} \frac{f_{ij}}{f_{i.}} y_j$$

. **Exemple :**La loi conditionnelle de Y sachant que $X \in [11,19[$

Y	$[5,7[$	$[7,9[$	$[9,11[$	$[11,15]$
$f_{j\|X\in[11,19[}$	$2/61$	$8/61$	$35/61$	$16/61$

$\bar{y}_i = 12 * 2/61 + 8 * 8/61 + 10 * 35/61 + 13 * 16/61 = 10.59.$

1.4.4 Les variances conditionnelles

La variance de X sachant $Y = y_j$ notée $V(X|Y = y_j)$,

$$V(X|Y = y_j) = \sum_{i=1}^{l} f_{i|j}(x_i - \bar{x}_j)^2 = \sum_{i=1}^{l} f_{i|j}x_j^2 - \bar{x}_j^2$$

.
Exemple : $V(X|Y = y_j) = 2^2 * 0 + 7^2 * 3/52 + 15^2 * 8/52 + 25^2 * 26/52 + 45^2 * 15/52 - 28.192^2 = 934.077 - 794.789 = 139.2881.$
La variance de Y sachant $X = x_i$ notée $V(Y|X = x_i)$,

$$V(Y|X = x_i) = \sum_{j=1}^{k} f_{j|i}(y_j - \bar{y}_i)^2 = \sum_{j=1}^{k} f_{j|i}y_j^2 - \bar{y}_i^2$$

.
Exemple : $V(Y|X = x_i) = 12^2 * 2/61 + 8^2 * 8/61 + 10^2 * 35/61 + 13^2 * 16/61 - 10.59^2 = 114.82 - 112.148 = 2.671.$
Remarque : Si X et Y sont indépendantes alors $\bar{x}_j = \bar{x}$ et $\bar{y}_i = \bar{y}$.

1.5 Covariance et coefficient de corrélation

1.5.1 Covariance

La covariance du couple statistique (X, Y) noté $Cov(X, Y)$ est donnée par la formule suivante ;

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{l} \sum_{j=1}^{k} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum_{i=1}^{l} \sum_{j=1}^{k} f_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

On montre que

$$Cov(X, Y) = \sum_{i=1}^{l} \sum_{j=1}^{k} f_{ij}x_i y_j - \bar{x}\bar{y}.$$

Propriétés :
1) $Cov(X, Y) \in \mathbb{R}$.
2) $Cov(X, Y) = Cov(Y, X)$.
3) $Cov(X, X) = V(X)$.

4) $V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$.
Exemple : $Cov(X,Y) = 2*(6*0+8*0+10*2/234+13*36/234)+7*(6*0+8*3/234+10*12/234+13*26/234)+15*(6*2/234+8*8/234+10*35/234+13*16/234)+25*(6*0+8*26/234+10*22/234+13*3/234)+45*(6*22/234+8*15/234+10*6/234+13*0)-19.179*10.18 = 169.1239-195.2422 == -26.118.$

1.5.2 Le coefficient de corrélation linéaire

C'est un indice qui mesure l'intensité de la liaison entre X et Y, noté par ρ_{XY} et défini par l'expression

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

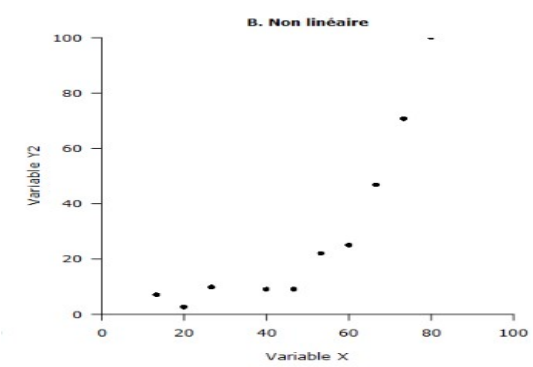
tels que $\sigma_X = \sqrt{V(X)}$ et $\sigma_Y = \sqrt{V(Y)}$.
Propriétés :
1) $\rho_{XY} \in [-1,1]$.
2) $\rho_{XY} = \rho_{YX}$.
3) $|\rho_{XY}| = 1$ signifie qu'il existe une liaison parfaitement linéaire entre X et Y.
4) Si X et Y sont indépendants alors $\rho_{XY} = 0$.
Remarque : $\rho_{XY} = 0$ n'implique pas forcement que X et Y sont indépendantes.
Exemple : $\rho_{XY} = \frac{-26.118}{\sqrt{(208.388)}\sqrt{(5.6881)}} = \frac{-26.118}{14.43*2.38} = \frac{-26.118}{34.3434} = -0.7604.$

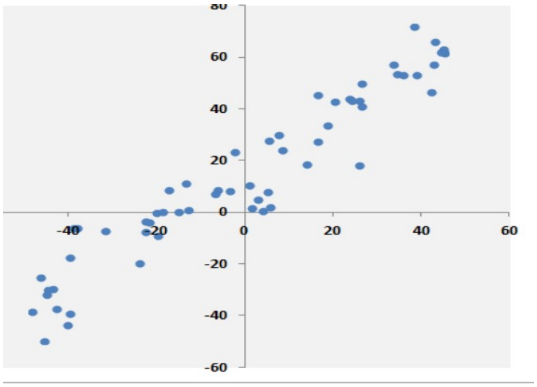
1.6 La régression linéaire simple

1.6.1 Droite de régression de Y en X

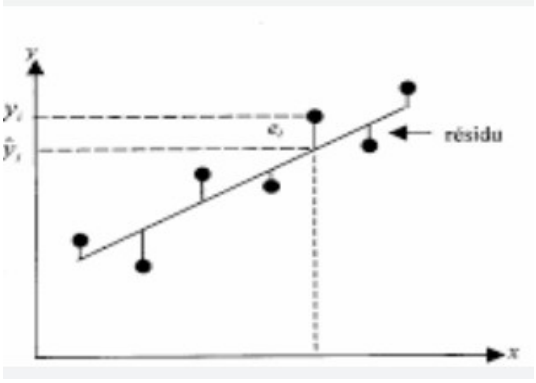
Si la forme du nuage des points du digramme de dispersion est allongée dans le sens de la première bissectrice (ou dans le sens inverse) et la valeur du coefficient de corrélation est assez importante, on peut exprimer la variable Y comme fonction affine de X, théoriquement on écrit

$$Y = aX + b.$$





Les paramètres a et b sont déterminés de telle sorte que la droite de régression passe au plus près possible de tous les points du diagramme de dispersion.
La méthode utilisée pour l'estimation de ces deux paramètres est appelée "Méthode des moindres carrées".
Elle consiste à trouver les paramètres a et b qui minimise le carré des distances $y_i - \hat{y}_i$ tel que $\hat{y}_i$ est la projection parallèle à (OY) d'un point (x_i, y_i) sur la droite de régression.



Il faut donc minimiser la somme S(a,b) ;

$$S(a,b) = \sum_{i=1}^{N} (y_i - a * x_i - b)^2$$

S(a,b) atteint son minimum lorsque ses dérivées partielles s'annullent ;

$$\begin{cases} \frac{\partial S(a,b)}{\partial a} = 0 \\ \frac{\partial S(a,b)}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \hat{a} = \frac{Cov(X,Y)}{V(X)} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

D'après la forme de $\hat{b}$, on constate que la droite d'ajustement passe le point moyen appelé aussi le centre de gravité du nuage des points.
l'équation sera alors

$$\hat{y} = \hat{a}x + \hat{b}$$

Remarques

1) Le carré du coefficient de corrélation linaire $R^2 = \rho_{XY}^2$ appelé le coefficient de détermination est le taux de variation de la variable Y expliqué par la droite de régression.

2) Une valeur de R^2 proche de 1 signifie un bon ajustement.

3) Si $R^2 = 1$ alors les points du diagramme de dispersion sont parfaitement alignés.

4) $\rho_{XY} > 0$ signifie que X et Y varient dans le même sens.

5) $\rho_{XY} < 0$ signifie que X et Y varient dans deux sens différents.

Exemple : Calculons $\hat{a}$ et $\hat{b}$;

$$\begin{cases} \hat{a} = \frac{-26.118}{5.69} = -4.59 \\ \hat{b} = 10.18 - 19.179 * (-4.59) = 98.21 \end{cases}$$

Alors l'équation de la droite de régression sera de la forme ;

$$\hat{y} = -4.59 * x + 98.21.$$

Exercices

Exercice 1 *Le tableau suivant représente la répartition de 25 étudiants suivant le nombre d'absences annuel X et la note finale Y.*

X/Y	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$
0	*1*	*0*	*3*	*2*
1	*0*	*2*	*3*	*4*
2	*3*	*1*	α	*0*
3	*2*	*1*	*0*	*0*
4	*1*	*0*	*1*	*0*

1) Interpréter α et calculer sa valeur.

2) Déterminer les distributions marginales des variables X et Y.

3) Déterminer la distribution conditionnelle de Y sachant que X=3 ;

4) les variables X et Y sont elles indépendantes ? justifier.

5) Calculer le coefficient linéaire ρ, conclure.

6) Donner l'équation de la droite de régression de Y e X.

Solution : 1) α est le nombre des étudiants qui ont deux absences annuelles et leurs notes est entre 10 et 15, $\alpha = 25-(1+3+2+2+3+4+3+1+2+1+1+1) = 1$.

2) La distribution marginale de X ;

X	0	1	2	3	4
$f_{i.}$	6/25	9/25	5/25	3/25	2/25

La distribution marginale de Y ;

Y	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$
$f_{.j}$	$7/25$	$4/25$	$8/25$	$6/25$

3) La distribution marginale de Y sachant que X=3 notée $f_{j|X=3}$;

Y	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$
$f_{j\|X=3}$	$\frac{2/25}{3/25}=2/3$	$\frac{1/25}{3/25}=1/3$	0	0

4) X et Y ne sont pas indépendantes car si on prend $Y \in [10,15[$ et $X=3$, $f_{ij}=0 \neq f_{i.}(=3/25)*f_{.j}(=8/25)$ (contre exemple).
5) Pour calculer ρ, on doit calculer $Cov(X,Y)$;
$Cov(X,Y) = (1/25)[(2.5*0+7.5*2+12.5*3+17.5*4)+2*(2.5*3+7.5*1+12.5*1)+3*(2.5*2+7.5*1)+4*(2.5*1+12.5*1)]-(9/25+2*5/25+3*3/25+4*2/25)*(2.5*7/25+7.5*4/25+12.5*8/25+17.5*6/25) = 11-1.44*10.1 = -3.544.$
$\rho = \frac{Cov(X,Y)}{\sigma_X*\sigma_Y}$.
Calculons d'abords $V(X) = (9/25+2^2*5/25+3^2*3/25+4^2*2/25)-1.44^2 = 1.4464$ $\sigma_X = \sqrt{1.4464} = 1.203$.
$V(Y) = (2.5^2*7/25+7.5^2*4/25+12.5^2*8/25+17.5^2*6/25-10.1^2) = 32.24$
$\sigma_Y = \sqrt{32.24} = 5.68$.
$\rho = \frac{-3.544}{1.203*5.68} = -0.52$.
On peut dire qu'il ya une corrélation linéaire négative entre X et Y, c'est à dire que X varie dans le sens contraire de Y.
6) l'équation de régression ;
$$\begin{cases} \hat{a} = \frac{-3.544}{5.69} = -0.62 \\ \hat{b} = 10.1 - 1.44*(-0.62) = 10.99 \end{cases}$$
Alors l'équation de la droite de régression sera de la forme ;

$$\hat{y} = -0.62*x + 10.99.$$

Exercice 2 *On a demandé à 10 étudiants de la filière informatique le nombre d'heures X passées à préparer l'examen d'informatique et on a relevé pour chacun d'eux la note Y (/20) obtenus à l'examen. On a eu le tableau suivant :*

ω_i	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
$X(\omega_i)$	*5*	*1*	*4*	*10*	*6*	*8*	*8*	*9*	*10*	*3*
$Y(\omega_i)$	*7*	*6*	*5*	*15*	*12*	*10*	*14*	*19*	*16*	*7*

1) Calculer les moyennes marginales de X et Y.
2) Calculer $Cov(X,Y)$.
3) Calculer le coefficient de corrélation ρ de du couple (X,Y).
4) Calculer l'équation de la droite de régression de Y en fonction de X.
5) Estimer la note d'un étudiant qui a passé 7 heures à préparer son examen.

Solution : 1) $\bar{x} = 64/10 = 6.4$.
$\bar{y} = 111/10 = 11.1$.

2) $Var(X) = 496/10 - 6.4^2 = 8.64$.
$Var(Y) = 1441/10 - 11.1^2 = 20.89$.
$Cov(X,Y) = 827/10 - 6.4 * 11.1 = 11.66$.
3) $\rho = \frac{11.66}{\sqrt{8.64} * \sqrt{20.89}} = 0.8679$.
4) $\begin{cases} \hat{a} = \frac{11.66}{8.64} = 1.35 \\ \hat{b} = 11.1 - \hat{a} * 6.4 = 2.46 \end{cases}$
L'équation de régression sera de la forme :$\hat{y} = 1.35 * x + 2.46$.
5)L'estimation de la note est alors $\hat{y} = 1.35 * 7 + 2.46 = 11.91$.