Statistiques

1°. DISTRIBUTION MARGINALE

Définition

Soit (X , Y) une série statistique double sur un échantillon de taille n et $(x_i , y_i)_{1\le i\le n}$, les valeurs numériques prises respectivement par les variables X et Y .

- La distribution marginale de la variable X est la distribution des valeurs de $(x_i)_{1\le i\le n}$ prises par X
- La distribution marginale de la variable Y est la distribution des valeurs de $(y_i)_{1\le i\le n}$ prises par Y

Rappel Soit X une série statistique sur un échantillon de taille n ; de valeurs distinctes $x_1 , x_2 , \ldots\ldots , x_p$ d'effectifs respectifs $n_1 , n_2 , \ldots\ldots , n_p$. La moyenne , la variance et l'écart-type de la série sont définies respectivement par :

$\overline{X}=\frac{1}{n}\sum_{i=1}^{i=p} n_i x_i$	$v(X)=\frac{1}{n}\sum_{i=1}^{i=p} n_i (x_i-\overline{X})^2$ ou $v(X)=\frac{1}{n}\sum_{i=1}^{i=p} n_i x_i^2-(\overline{X})^2$	$\sigma_x=\sqrt{v(X)}$

Activité 1 page 98

2/ * Distribution marginale de X

x_i	9.6	12.8	18.4	31.2	36.8	47.2	49.6	56.8
n_i	9	13	30	21	13	6	4	4

* Calcul de la moyenne , la variance et l'écart-type

Tableau des calculs

x_i	n_i	$n_i.x_i$	$n_i.x_i^2$
9,6	9	86.4	829.44
12,8	13	166.4	2129.92
18,4	30	552	10157.8
31.2	21	655.2	20442.24
36.8	13	478.4	17605.12
47.2	6	283.2	13367.04
49.6	4	198.4	9840.64
56.8	4	227.2	12904.96
Total	100	2647.2	87277.16

Paramètres

Moyenne :

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{i=8} n_i x_i=\frac{1}{100}\times 2647.2=26.472$$

Variance :

$$v(X)=\frac{1}{n}\sum_{i=1}^{i=8} n_i x_i^2-(\overline{X})^2=\frac{1}{100}\times 87277.16-(26.472)^2\simeq 172$$

Ecart-type:

$$\sigma_x=\sqrt{v(X)}=\sqrt{172}\simeq 13.11$$

3/ * Distribution marginale de Y :

y_i	……	……	……	……
n_i				

* Calcul de la moyenne , la variance et l'écart-type :

y_i	n_i	$n_i.y_i$	$n_i.y_i^2$
70	6	420	29400
		…….	
104		4680	……
Total	……		898236

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{i=4} n_i y_i = ...$$

$$v(Y) = \frac{1}{n}\sum_{i=1}^{i=4} n_i y_i^2 - (\overline{Y})^2 = ...$$

$$\sigma_y = \sqrt{v(Y)} \simeq 9.87$$

Activité 2 page 99

1/ a/ Les valeurs prises par Y sont :
….. , ….. , ….. et ….. .

Distribution marginale de Y est :

y_i	…..	…..	6	…..
n_i			5	

b/ Calcul de la moyenne et l'écart-type

y_i	n_i	$n_i.y_i$	$n_i.y_i^2$
4	7	28	112
5	5	25	125
6	5	30	180
7	3	21	147
Total	20	104	564

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{i=4} n_i y_i = \frac{1}{20}\times 104 = 5.2$$

$$v(Y) = \frac{1}{n}\sum_{i=1}^{i=4} n_i y_i^2 - (\overline{Y})^2 = \frac{1}{20}\times 564 - (5.2)^2 = 1.16$$

$$\sigma_y = \sqrt{v(Y)} = \sqrt{1.16} \simeq 1.07$$

2/

x (distance parcourue en mille km)	Effectifs n_i
Moins de 50	
[50 , 60 [	
[60 , 70 [	
70 et plus	

3/

..... des voitures ont une puissance supérieur ou égal à 6 et ont parcouru une distance inférieure à 60000 km avant la première panne .

Activité 3 page 100

- Distribution marginale de X

x_i	42.5	47.5	52.5	57.5
n_i	22	33	24	21

$\overline{X} =$ $\sigma_x =$

- Distribution marginale de Y

y_i	137.5	157.5	162.5	167.5
n_i	30	25	23	22

$\overline{Y} =$ $\sigma_y =$

2°. COVARIANCE D'UNE SÉRIE STATISTIQUE DOUBLE

2°. 1- Cas d'un échantion simple

Activité 1 page 100

1/ La taille de l'échantillon est 12 .

2/ En enregistre les données du tableau dans une calculatrice en mode de fonctionnement statistique (voir page 104)

a/ $\overline{X} = 1276.308$; $\sigma_x = 109.6$ b/ $\overline{Y} = 1647.17$; $\sigma_y = 174.7$

3/ On a $\sum_{i=1}^{i=12} x_i . y_i = 25391882.43$ d'où $\frac{1}{12}\sum_{i=1}^{i=12} x_i . y_i - \overline{X}.\overline{Y} = 13605$

ce réel est appelé covariance de (X , Y) ; on note cov(X,Y) =13605

Définition Soit (X , Y) une série statistique double sur un échantillon de taille n .

On appelle covariance de (X , Y) le réel noté $\text{cov}(X , Y)$ défini par $\text{cov}(X , Y) = \frac{1}{n}\sum_{i=1}^{i=n} x_i . y_i - \overline{X}.\overline{Y}$

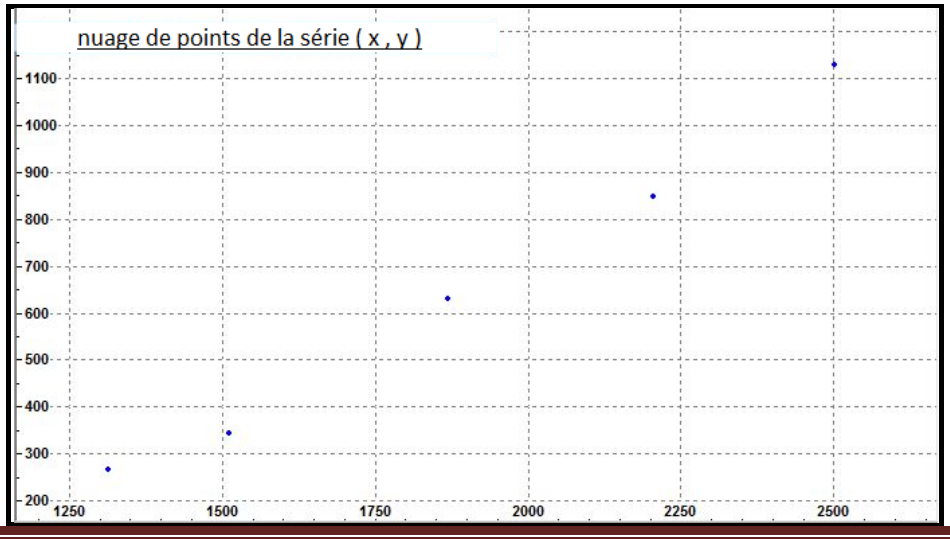
Interprétation

La covariance est positive si X et Y ont tendance à varier dans le même sens .

La covariance est négative si X et Y ont tendance à varier en sens contraire .

Activité 2 page 101

1/ Représentation du nuage de la série (X , Y)



2/ $\overline{X} = 1880.2$ et $\overline{Y} = 643.08$

$$\text{cov}(X , Y) = \frac{1}{n}\sum_{i=1}^{i=n} x_i . y_i - \overline{X}.\overline{Y} = \frac{1}{5} \times 6.7382 \times 10^6 - (1880.2 \times 643.08) = 1.385 \times 10^5$$

La covariance est positive donc X et Y varient dans le même sens .

Activité 3 page 102

Déterminer le signe de $\text{cov}(X , Y)$ puis calculer sa valeur .

2°. 2- Cas d'un échantion groupé

Définition Soit (X , Y) une série statistique double de taille n , et soit n_{ij} le nombre de fois qu'apparaît le couple (x_i , y_j) alors $\text{cov}(X , Y) = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{p} n_{ij} x_i . y_i - \overline{X}.\overline{Y}$

Exercice résolu page 102

3°. AJUSTEMENT D'UNE SÉRIE STATISTIQUE DOUBLE

Rappel Soit (X , Y) une série statistique double de valeur $(x_i , y_i)_{1 \le i \le n}$.

L'ensemble des points $M_i(x_i , y_i)$ dans un repère orthogonal est appelé nuage de points de la série .

Le point $G(\overline{x} , \overline{y})$ est appelé point moyen du nuage .

Nécessité de l'ajustement affine

Dans le cas d'un nuage de points de forme allongée, et afin de faciliter l'étude de la série, il est possible de remplacer ce nuage par une droite appelée droite d'ajustement affine .

On propose dans la suite deux méthodes de détermination d'un tel type d'ajustement .

3°. 1- Méthode de Mayer

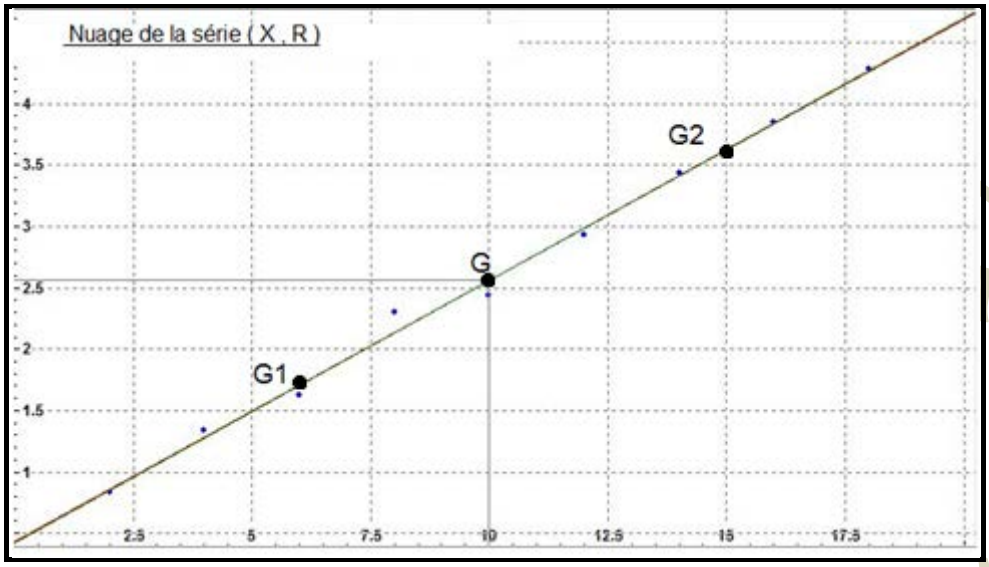
Principe de la méthode de mayer

On scinde le nuage de points de la série en deux parties contenant à peu prés le même nombre de points puis on détermine les points moyens G_1 et G_2 des deux nuages obtenus .

la droite (G_1G_2) appelé droite de Mayer est un ajustement affine du nuage des points de la série (X , Y).

Activité 2 page 107

1/ Nuage de la série (X , R).



2/ On divise la série en deux groupes comme-suit :

	1ier groupe					2ème groupe			
X	2	4	6	8	10	12	14	16	18
R	0.83	1.34	1.63	2.3	2.44	2.93	3.44	3.85	4.28

- On calcule les points moyens G_1 et G_2 des deux groupes :

  G_1(;) et G_2(;)

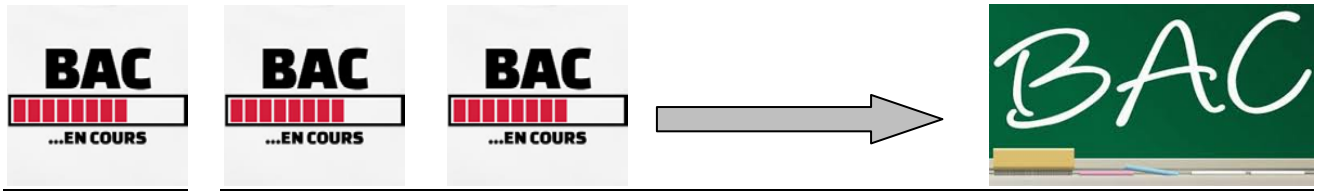
- Soit D : $R = a x + b$ la droite de Mayer alors $a = \dfrac{\quad}{\quad} = 0.21$

  Or G_1 appartient à la droite D donc ..... = 0.21 x ..... + b d'où b = ...

  est par suite D : $R = 0.21 x$ + ...

3/ D'après ce qui précède , on peut estimer la résistance thermique qu'on espère obtenir avec une épaisseur de polystyrène de 25 cm , soit $R = 0.21 \times 25$ + ... = ...

Exercice 3 page 114



3°. 2- Méthode d'ajustement par les moindres carrés

Définition

Soit (X , Y) une série statistique double sur un échantion de taille n .

- La droite des moindres carrés de Y en X , ou droite de régression de Y en X est :

$$D: y = a\,x + b \quad \text{où} \quad a = \frac{\text{cov}(X,Y)}{\sigma_x^2} \quad \text{et} \quad b = \overline{Y} - a\,\overline{X}$$

- La droite des moindres carrés de X en Y , ou droite de régression de X en Y est :

$$D': x = a'\,y + b' \quad \text{où} \quad a' = \frac{\text{cov}(X,Y)}{\sigma_y^2} \quad \text{et} \quad b' = \overline{X} - a'\,\overline{Y}$$

Conséquence Les droites des moindres carrés de Y en X et de X en Y passent par le point moyen G .

Exemple Le tableau ci-dessous (acivité 3 page 108) donne le chiffre d'affaire annuel en mille DT d'une société pendant huit années consécutives .

Rang de l'année (**x**)	1	2	3	4	5	6	7	8
Chiffre d'affaires en mille DT(y)	13.6	15	15.8	17	18	20	19	20

- Le nuage de points de la série (X,Y) est allongé (les points semblent alignés), donc un ajustement affine de cette série est justifié .
- Un ajustement affine de la série par les moindres carrés est alors la droite $D: y = a\,x + b$ telle que :

$a = \frac{\text{cov}(x,y)}{\sigma_x^2} = \text{———} \simeq 0.93$ et $b = \overline{y} - a\,\overline{x} =$ $\quad - 0.93 \times \quad = 13.115 \quad \Rightarrow \quad D: y = 0.93x + 13.115$

- On peut alors trouver une estimation du chiffre d'affaires de cette société à sa dixième année ; Il suffit de calculer la valeur de y pour $x = 10$, on trouve $y = 22.415$ mille DT

Activité 1 page 109

1/ $\overline{X} = \frac{1}{15}\sum_{i=1}^{i=15} n_i x_i \simeq 49.67$, $\sigma_x \simeq 21.69$; $\overline{Y} = \frac{1}{15}\sum_{i=1}^{i=15} n_i y_i \simeq 1.27$, $\sigma_y = 0.4$

2/ $\text{cov}(X\ ,\ Y) = \frac{1}{n}\sum_{i=1}^{i=n} x_i\,.\,y_i - \overline{X}.\overline{Y} = \frac{1}{15} \times 1081 \times -(49.667 \times 1.2747) \simeq 8.75$

3/ b/ Les points du nuage de la série (x , y) semblent alignés donc un ajustement affine est justifié .

c/ Un ajustement affine par les moindres carrés de la série (x , y) est alors :

$$D: y = a\,x + b \quad \text{avec} \quad a \simeq 0.019 \quad \text{et} \quad b \simeq 0.35 \quad \Rightarrow \quad D: y = 0.019\,x + 0.35$$

4/ Une estimation de la surface corporelle d'un sujet qui pèse 62kg est :$y = 0.019x62+0.35 = 1.528m^2$

3°. 3- Coefficient de corrélation linéaire

Définition soit (x ,y) une série statistique double. On appelle coefficient de corrélation linéaire le réel noté ρ_{xy} défini par : $\rho_{xy} = \dfrac{\text{cov}(x,y)}{\sigma_x \sigma_y}$

Propriétés * Le coefficient de corrélation linéaire est invariant par changement d'unité ou d'origine .

* $-1 \le \rho_{xy} \le 1$

Interprétation

* Si $|\rho_{xy}| \succ \dfrac{\sqrt{3}}{2}$, l'ajustement affine est justifié et les prédictions faites au moyen de cet ajustement sont raisonnables .
* Si $\rho_{xy} = -1$ ou $\rho_{xy} = 1$, alors il y'a dépendance totale entre x et y ,l'une est une fonction affine de l'autre (les points du nuage sont alignés)

Activité 4 page 111

Année	1997	1998	1999	2000	2001	2002
Rang de l'année (x)	1	2	3	4	5	6
Population scolaire en 3ième année (y)	67755	74581	79266	76138	80123	90087

1/ Le coefficient de corrélation linéaire est $\rho_{xy} = 0.9065$

2/ * On a $|\rho_{xy}| \succ \dfrac{\sqrt{3}}{2}$, donc un ajustement affine est justifié .

* Un ajustement affine par les moindres carrés de la série (X , Y) est alors :

$$D: y = a\,x + b \quad \text{avec} \quad a = 3575.9 \quad \text{et} \quad b = 65476 \quad \Rightarrow \quad D: y = 3575.9\,x + 65476$$

* Une estimation de la population scolaire en 3ième année secondaire au mois d'octobre 2010 correspond au rang x = 14 alors $y = 3575.9 \times 14 + 65476 = 115538.6 \approx 115539$

Activité 5 page 111

Couples (x_i , y_j)	(2,25)	(2,30)	(3,20)	(3,25)	(3,30)	(4,20)	(4,25)	(4,30)
Effectifs n_{ij}	5	25	8	20	3	30	7	2

$\rho_{xy} = -0.77 ; |\rho_{xy}| \prec \frac{\sqrt{3}}{2}$ donc un ajustement affine est non justifié (La corrélation entre x et y est faible).

3°. 4- Exemples d'ajustement non affine

Exemple 1 Exercice résolu page 111

Exemple 2 Activité 6 page 113

1/ Tableau des valeurs de la série (X , Y) .

x	100	400	900	1600	2500	3600	4900	6400	8100
y	0.26	0.29	0.33	0.385	0.472	0.575	0.7	0.84	0.99

2/ Le coefficient de corrélation linéaire est $\rho_{xy} = 0.9981$.

On a $|\rho_{xy}| \succ \frac{\sqrt{3}}{2}$,donc un ajustement affine est justifié .

Par la méthode des moindres carrés la droite de régression de y en x est :

$$D : y = 9.2 \times 10^{-5} x + 0.25$$

3/
$$y = \frac{R}{V} \Rightarrow R = V.y = V.(9.2 \times 10^{-5} x + 0.25)$$
$$= V.(9.2 \times 10^{-5} V^2 + 0.25)$$
$$= 9.2 \times 10^{-5} V^3 + 0.25 V$$

4/ Une évaluation de la valeur de R pour une vitesse de 100km/h est :

$$R = 9.2 \times 10^{-5} \times 100^3 + 0.25 \times 100 = 117 \text{ kw}$$

Exemple 3 Exercice n°3 Bac 2013 / Session principal