Série statistique double

Distributions marginales

Activité 1

Un relevé statistique des tailles X (en cm) et des poids Y (en kg) d'un échantillon de 100 élèves a permis de construire le tableau suivant :

X \ Y	[40, 45[	[45, 50[	[50, 55[	[55, 60[
[150, 155[	18	10	2	0
[155, 160[	3	16	5	1
[160, 165[	0	5	13	5
[165, 170[	0	2	6	14

Donner la distribution marginale de X et la distribution marginale de Y.

Calculer $\overline{X}$; $\overline{Y}$; $V(X)$; $V(Y)$; $\sigma(X)$ et $\sigma(Y)$.

Distribution marginale de X:

Classes	[150, 155[	[155,160[	[160,165[	[165,170[	Total
Effectifs					**100**

Distribution marginale de Y:

Classes	[40, 45[	[45,50[	[50,55[	[55,60[	Total
Effectifs					**100**

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{4} c_i n_i = \text{...}$$

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{4} c_i n_i = \text{...}$$

$$V(X) = \overline{X^2} - \overline{X}^2 = \frac{1}{N}\sum_{i=1}^{4} c_i^{\,2} n_i - \overline{X}^2 = \text{...}$$

$$V(Y) = \overline{Y^2} - \overline{Y}^2 = \frac{1}{N}\sum_{i=1}^{4} c_i^{\,2} n_i - \overline{Y}^2 = \text{..}$$

$$\sigma(X) = \sqrt{V(X)} = \text{......................} \qquad \sigma(Y) = \sqrt{V(Y)} = \text{............................}$$

Ajustement affine

I. Méthode des moindres carrées

----Covariance

Cas d'un échantillon simple

$$\boxed{Cov(X\ ,\ Y) = \overline{XY} - \overline{X}.\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}.\overline{Y}}$$ *où* $\overline{X}$ *et* $\overline{Y}$ *sont les moyennes arithmétiques respectives des distributions* $x_{i\ \ 1\le i\le n}$ *et* $y_{i\ \ 1\le i\le n}$ *de X et Y.*

Cas d'un échantillon groupé (voir exercice résolu page:213)

$$\boxed{Cov(X\ ,\ Y) = \overline{XY} - \overline{X}.\overline{Y} = \frac{1}{N}\sum_{j=1}^{q}\sum_{i=1}^{p} x_i y_j n_{ij} - \overline{X}.\overline{Y}}$$

Interprétation:

La covariance est positive si X et Y ont tendance de varier dans le même sens.
La covariance est négative si X et Y ont tendance de varier dans des sens contraires.

----Coefficient de corrélation linéaire

Soit une série statistique à deux caractères quantitatifs X et Y non constants observés dans une population donnée d'effectif total n.

On appelle coefficient de corrélation linéaire de la série double (X, Y), le réel r défini par : $\boxed{r = \frac{Cov(X,\ Y)}{\sigma(X).\sigma(Y)}}$ *où* $\sigma(X)$ *et* $\sigma(Y)$ *sont les écarts-types respectifs des variables statistiques X et Y.*

Soit une série statistique à deux caractères quantitatifs X et Y.

Lorsque le coefficient de corrélation linéaire r du couple (X, Y) est proche, en valeur absolue, de 1 ($\frac{\sqrt{3}}{2} \le |\,r\,| \le 1$)*, le nuage de points de la série considérée a une forme allongée et il est possible d'approcher la liaison entre X et Y par deux relations affines représentées graphiquement par deux droites* D_1 et D_2 *passant par le point moyen* $G(\overline{X},\overline{Y})$ *du nuage de points.*

La droite D_1 *: appelée droite de régression de Y en X et ayant pour équation:*

$$\boxed{y = ax + b \ \text{ avec }\ a = \frac{cov(X,\ Y)}{V(X)} \ \text{ et }\ b = \overline{Y} - a\overline{X}}$$

La droite D_2 *:appelée droite de régression de X en Y et ayant pour équation:*

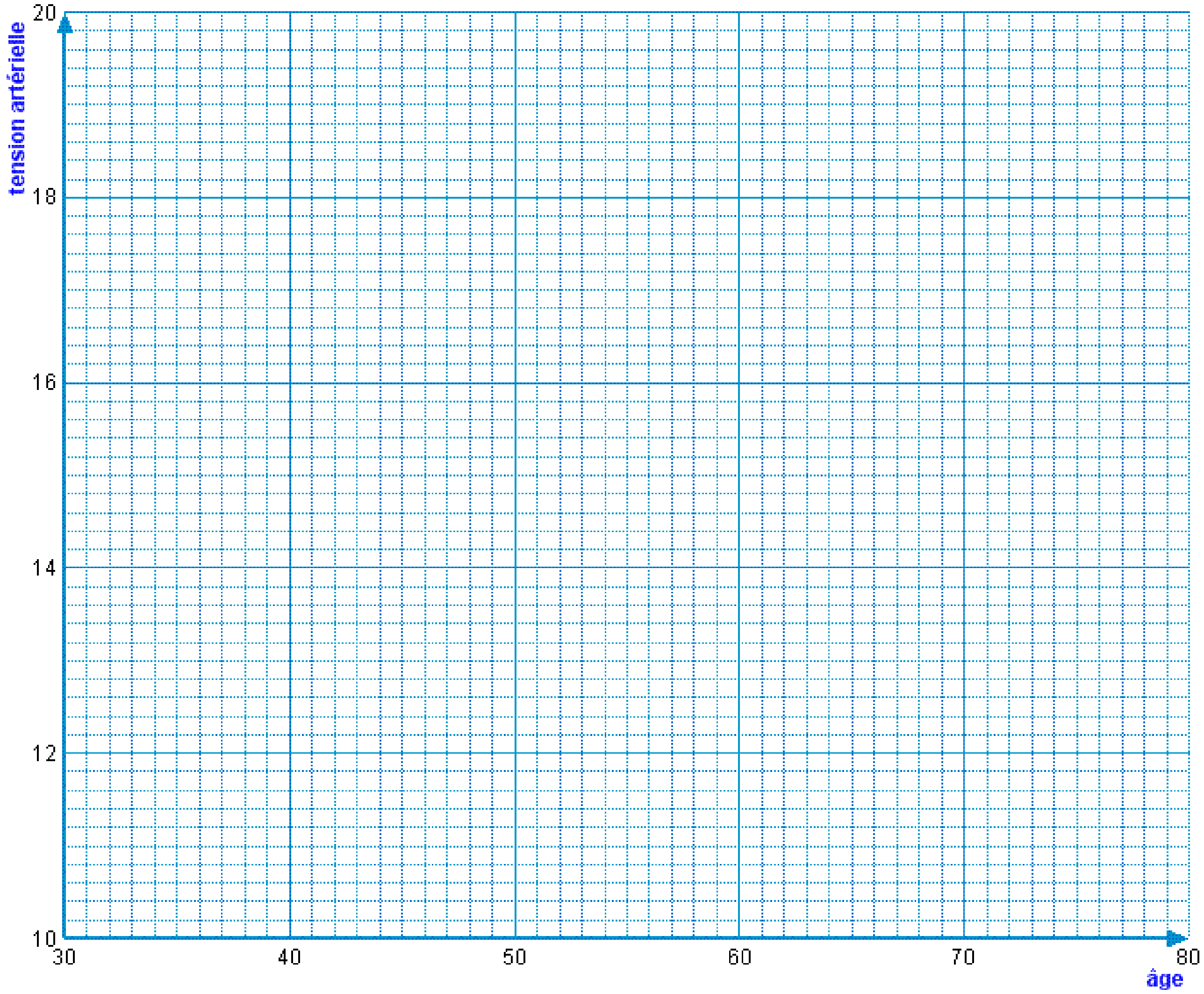
$$\boxed{x = a'y + b' \ \text{ avec }\ a' = \frac{cov(X,\ Y)}{V(Y)} \ \text{ et }\ b' = \overline{X} - a'\overline{Y}}$$

Activité2

Le tableau suivant donne l'âge X et la tension artérielle Y de 10 personnes.

X	58	40	74	34	65	49	53	51	36	40
Y	16,7	13,1	17,2	11,6	15,5	15,1	14,2	14,4	13,0	14,2

Construire le nuage de points de cette série statistique.



1) Déterminer la moyenne et la variance de chacune des variables X et Y.

2) Déterminer le coefficient de corrélation linéaire r des variables X et Y. Un ajustement linéaire entre X et Y est il justifié?

3) Déterminer une équation de la droite de régression de Y en X.

4) Estimer la tension artérielle d'une personne âgée de 45 ans.

Solutions

$\overline{X} = \dfrac{1}{10}\sum_{i=1}^{10} x_i = ...$ $\overline{Y} = \dfrac{1}{10}\sum_{i=1}^{10} y_i = ...$

$$V(X)=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{X}^2=\frac{1}{10}\sum_{i=1}^{10}x_i^2-\overline{X}^2=\ldots\ldots\ldots\ldots$$

$$V(Y)=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{Y}^2=\frac{1}{10}\sum_{i=1}^{10}y_i^2-\overline{Y}^2=\ldots\ldots\ldots\ldots$$

$$Cov(X,Y)=\frac{1}{10}\sum_{i=1}^{10}x_iy_i-\overline{X}.\overline{Y}=\ldots\ldots\ldots\ldots$$

$$r=\frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}=\frac{\ldots\ldots}{\sqrt{\ldots\ldots}\sqrt{\ldots\ldots}}=\ldots\ldots\ldots\ldots$$

……………………………………………………………………………………………

La droite de régression de Y en X admet pour équation: $y=ax+b$ *où*

$$a=\frac{Cov(X,Y)}{V(X)}=\frac{\ldots\ldots}{\ldots\ldots}=\ldots\ldots \quad et \quad b=\overline{Y}-a\overline{X}=\ldots\ldots$$

La tension artérielle d'une personne âgée de 45 ans est: ……………………………………

2. Méthode de Mayer

Activité3

Une banque a enregistré les nombres de retraits opérés dans un guichet automatique pendant une journée. Le tableau suivant donne les montants (en DT) des retraits et leurs effectifs.

Montant en DT : x_i	40	35	30	25	20	15	10	5
Effectifs de retraits: y_i	19	20	17	11	13	6	7	2

1) a) Construire, dans un repère orthogonal, le nuage des points représentant cette série statistique.

b) Quelle particularité peut –on remarquer au sujet de la forme du nuage ?

c) Déterminer, les cordonnées du point moyen G de ce nuage. Placer G.

2) On partage l'ensemble des points du nuage en deux parties. La première partie P_1 *correspond aux retraits inférieurs ou égaux à 25 DT et la deuxième partie* P_2 *correspond aux autres retraits.*

a) Déterminer les coordonnées des points moyens G_1 *et* G_2 *respectifs des parties* P_1 *et* P_2. *Placer* G_1 *et* G_2 *dans le même repère.*

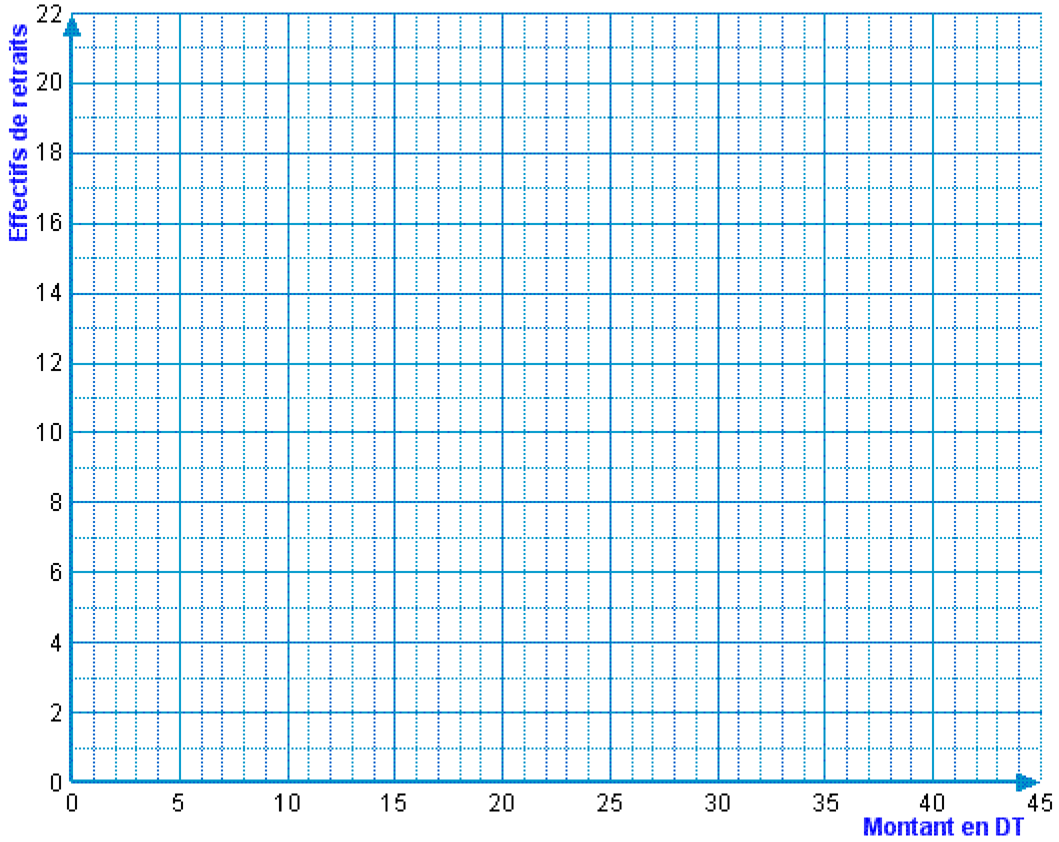
b) Donner une équation cartésienne de la droite (G_1G_2).

c) Vérifier que la droite (G_1G_2) *passe par le point G.*

3) Quel nombre de retraits de 50 DT peut –on prévoir en une journée ?

Solutions ------------------------

1) a)



b) ....

....

$$\overline{X} = \frac{1}{8}\sum_{i=1}^{8} x_i = ..., \quad \overline{Y} = \frac{1}{8}\sum_{i=1}^{8} y_i = ... \Rightarrow G(............, ...)$$

2) a) $\overline{X_1} = \frac{1}{4}\sum_{i=1}^{4} x_i = ..., \quad \overline{Y_1} = \frac{1}{4}\sum_{i=1}^{4} y_i = ... \Rightarrow G_1(.........., ...)$

$$\overline{X_2} = \frac{1}{4}\sum_{i=5}^{8} x_i = ..., \quad \overline{Y_2} = \frac{1}{4}\sum_{i=5}^{8} y_i = ... \Rightarrow G_2(..........., ...)$$

b)..

....

....

c) ....

3) ....

La droite (G_1G_2) *est appelée droite de* ***Mayer***

On dit qu'on a fait un ajustement linéaire à l'aide de la méthode de Mayer

Cas d'un échantillon groupé (Tableau à double entrée)

Activité4

On donne la série double suivante, relative aux voitures selon leur puissance Y et la durée des pneumatiques X (en milliers de Km).

1. Calculer le coefficient de corrélation linéaire.

2. Un ajustement par la méthode des moindres carrées est-il justifié?

Y \ X	2	3	4	
20	0	8	30	**38**
25	5	20	7	**32**
30	25	3	2	**30**
	30	**31**	**39**	**100**

----

1. **Distribution marginale de X:**

X	2	3	4
n_i	30	31	39

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{3} x_i n_i = \frac{60+93+156}{100} = 3,09$$

$$V(X) = \frac{1}{100}\sum_{i=1}^{3} x_i^{\,2} n_i - \overline{X}^2 = 0,6819\ ,\ \sigma(X) = 0,8258$$

Distribution marginale de Y:

Y	20	25	30
n_j	38	32	30

$$\overline{Y} = \frac{1}{100}\sum_{j=1}^{3} y_j n_j = 24,6$$

$$V(Y) = \frac{1}{100}\sum_{j=1}^{3} y_j^{\,2} n_j - \overline{Y}^2 = 622 - (24,6)^2 = 16,84$$

Covariance de (X,Y) et coefficient de corrélation:

Y \ X		2	3	4	
20	n_{ij}	0	8	30	2880
	$x_i y_j n_{ij}$	0	480	2400	
25	n_{ij}	5	20	7	2450
	$x_i y_j n_{ij}$	250	1500	700	
30	n_{ij}	25	3	2	2010
	$x_i y_j n_{ij}$	1500	270	240	
		1750	2250	3340	7340

n_{ij}
$x_i y_j n_{ij}$

$$\sum_{i=1}^{3}\sum_{j=1}^{3} x_i y_j n_{ij} = 7340$$

$$\overline{X}\cdot\overline{Y} = 3,09 \times 24,6 = 76,014$$

$$Cov(X,Y) = 73,4 - 76,014 = -2,614$$

Le coefficient de corrélation

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

$$= \frac{-2,614}{0,8258 \times 4,1036} \simeq -0,77$$

....

....