Terminales – Maths complémentaires *Ch 7 : Statistiques à deux variables*

Objectif n° 1 : Nuage de points – Ajustement affine

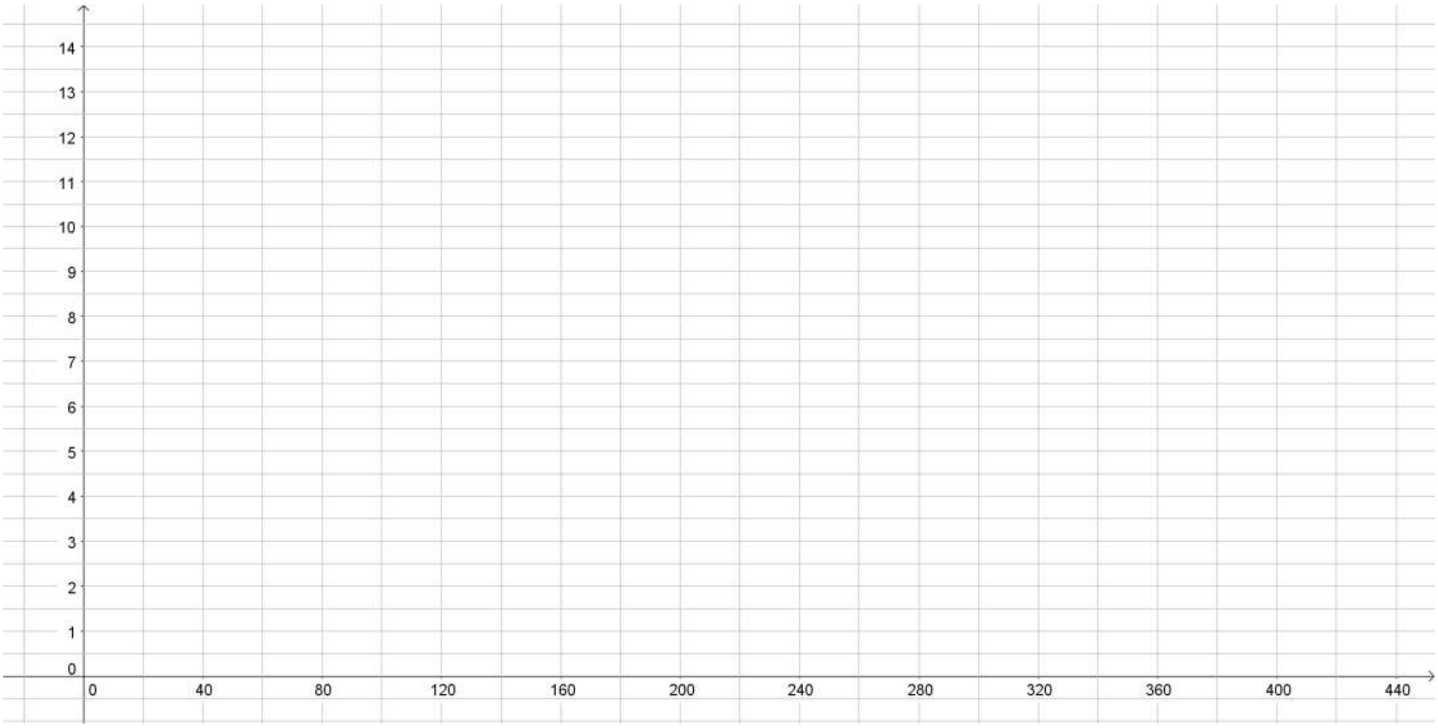
Exercice 1 :

Un laboratoire de recherche a relevé le nombre initial de rats dans une colonie et le taux de mortalité en pourcentage correspondant.

Voici ces relevés :

Nombre de rats (n)	40	80	120	160	200	240	280	320
Taux de mortalité (t)	0,9	2	2,5	3,5	4,9	6,2	9,3	10,9

1. Placer les points M_1 (40 ; 0,9) ; M_2 (80 ; 2) ; ….. ; M_8 (320 ; 10,9) dans le repère ci-dessous.



2. Le nuage de points obtenus fait penser à une droite. Tracer en rouge la droite (M_1M_8). Cette droite est **un ajustement affine** du nuage de points.

3. Considérons le point G dont l'abscisse est la moyenne des valeurs de n (cette moyenne sera notée $\overline{n}$) et dont l'ordonnée est la moyenne des valeurs de t (cette moyenne sera notée $\overline{t}$). Ce point G s'appelle **le point moyen** du nuage.

a. Déterminer les coordonnées $\overline{n}$ et $\overline{t}$ du point G.

b. Placer le point G sur le graphique précédent.

4. Tracer en bleu la droite (GM_8). Cette droite constitue un autre ajustement du nuage qui semble peut être meilleur que le précédent.

5. En utilisant la droite (GM_8), déterminer graphiquement :

a. le taux de mortalité pour un nombre initial de 220 rats (comme 220 est situé entre 40 et 320, on dit que l'on a effectué **une interpolation**).

b. le taux de mortalité pour un nombre initial de 400 rats (comme 400 est à l'extérieur de l'intervalle [40 ; 320], on dit que l'on a effectué **une extrapolation**).

Dans cet exercice, les deux caractères étudiés (le nombre de rats et le taux de mortalité) sont des nombres. Il s'agir de **caractères quantitatifs**. On parle alors **d'une série statistique à deux variables**.

On déduit de l'exercice 1 les définitions suivantes :

Définitions 1

Sur une population, lorsqu'on étudie deux caractères quantitatifs x et y et que l'on note x_i et y_i les différentes valeurs prises par ces deux caractères, on présente ces données sous la forme d'un tableau comme ci-dessous que l'on nomme **série statistique à deux variables**.

Valeurs x_i	x_1	x_2	…	x_n
Valeurs y_i	y_1	y_2	…	y_n

Dans un repère, **le nuage de points** associé à cette série statistique est l'ensemble des points M_i de coordonnées (x_i ; y_i).

Le **point moyen de ce nuage** est le point G de coordonnées ($\overline{x}$; $\overline{y}$) où $\overline{x}$ est la moyenne des valeurs x_i et $\overline{y}$ est la moyenne des valeurs y_i.

Définitions 2

Dans certains cas, la forme du nuage de points peut laisser penser que les points M_i se répartissent autour d'une droite comme sur la figure ci-contre.

Dans ce cas, on pourra pratiquer un ajustement affine du nuage de points, c'est-à-dire tracer une droite "qui passe le plus près possible " des points du nuage.

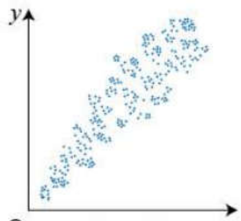
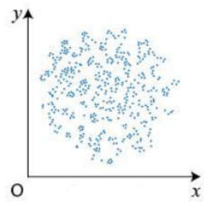


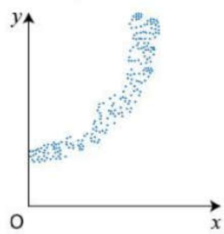
Figure 2

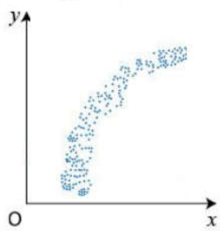
Remarques :

1. Dans certains cas, la forme du nuage de points peut laisser penser qu'il n'existe pas de dépendance entre les valeurs de x et celles de y.



2. Dans d'autres cas, les points du nuage de points peuvent sembler se répartir autour d'une courbe autre qu'une droite. Il sera alors parfois possible à ce moment là de se ramener à un ajustement affine à l'aide d'un changement de variable (on verra cela dans l'objectif 5).





3. Dans le cas où l'on essaye d'effectuer un ajustement affine, on peut se demander : " quelle droite tracer ? " , " Y a-t-il une droite qui réalise un meilleur ajustement que les autres ? " …. . On répondra à ces questions dans l'objectif 2.

Exercice 2 :

Le tableau suivant donne l'évolution du nombre d'adhérents à un club de rugby de 2020 à 2025.

Année	2020	2021	2022	2023	2024	2025
Rang de l'année x_i	1	2	3	4	5	6
Nombre d'adhérents y_i	70	90	115	145	170	220

1. Représenter, dans le repère ci-dessous, le nuage de points M_i (x_i ; y_i).

2. Déterminer les coordonnées $\overline{x}$ et $\overline{y}$ du point moyen G et le placer sur le repère.

3. Déterminer l'équation réduite de la droite (GM_6).

4. Si l'évolution se poursuit ainsi, estimer le nombre d'adhérents à ce club en 2028.

Exercice 3 :

Le tableau suivant donne l'évolution du montant net du SMIC mensuel, en euros, entre 2000 et 2005.

Année	2000	2001	2002	2003	2004	2005
Rang de l'année x_i	1	2	3	4	5	6
Montant net y_i	842	874	903	934	985	1039

1. Représenter, dans le repère ci-dessous, le nuage de points M_i (x_i ; y_i).

2. Déterminer les coordonnées $\overline{x}$ et $\overline{y}$ du point moyen G et le placer sur le repère.

3. Déterminer l'équation réduite de la droite (GM_5).

4. En 2020, le SMIC net mensuel était de 1219 €. Comparer ce montant avec celui extrapolé à l'aide de la droite précédente.

Exercice 4 :

Pour huit personnes, des médecins ont relevé:

- l'intensité du travail, x_i , en kilojoules par minute, fourni par le cœur,
- la fréquence cardiaque, y_i , en nombre de battements par minute.

Voici les résultats obtenus:

Valeurs x_i	9,6	12,8	18,4	31,2	36,8	47,2	49,6	56,8
Valeurs y_i	70	86	90	104	120	128	144	154

1. Représenter, dans un repère le nuage de points M_i (x_i ; y_i).

2. Déterminer et placer le point moyen G ($\overline{x}$; $\overline{y}$)

3. Pour ajuster ce nuage, on va utiliser la droite (GM_2). Tracer cette droite et déterminer son équation réduite.

4. Utiliser cette équation pour estimer :

a. le nombre de battements par minute correspondant à un travail de 40 kilojoules par minute.

b. l'intensité du travail correspondant à 100 battements par minute.

Objectif n° 2 : Droite des moindres carrés

Exercice 5 :

Une entreprise de locations de véhicules demande à un cabinet d'expertise d'étudier l'évolution de son chiffre d'affaire y_i (en milliers d'euros) durant les quatre dernières années.

Année	2022	2023	2024	2025
Rang de l'année x_i	1	2	3	4
Chiffre d'affaire y_i	580	510	500	490

1. Calculer $\overline{x}$ et $\overline{y}$, moyennes respectives des valeurs x_i et y_i.

2. Compléter le tableau suivant :

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	580					
2	510					
			Total			

3. On définit les notions suivantes :

* **La variance des valeurs de x**, notée $V(x)$ par $V(x) = \frac{1}{4}\left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + (x_4 - \overline{x})^2\right]$

* **L'écart-type des valeurs de x**, noté $\sigma(x)$, est la racine carrée de la variance : $\sigma(x) = \sqrt{V(x)}$

* **La variance des valeurs de y**, notée $V(y)$ par $V(y) = \frac{1}{4}\left[(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + (y_3 - \overline{y})^2 + (y_4 - \overline{y})^2\right]$

* **L' écart-type des valeurs de y**, noté $\sigma(y)$, est la racine carrée de la variance : $\sigma(y) = \sqrt{V(y)}$

* **La covariance de la série (x_i ; y_i)**, notée $\text{cov}(x\,;y)$ par

$$\text{cov}(x\,;y) = \frac{1}{4}\left[(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y}) + (x_4 - \overline{x})(y_4 - \overline{y})\right]$$

A l'aide des résultats obtenus dans le tableau de la question 2, déterminer :

a. $V(x)$ puis $\sigma(x)$	***b.*** $V(y)$ puis $\sigma(y)$	***c.*** $\text{cov}(x\,;y)$

De façon plus générale, on a les définitions suivantes :

Définitions 3

Considérons une série statistique à deux variables x et y :

Valeur x_i	x_1	x_2	…	x_n
Valeur y_i	y_1	y_2	…	y_n

* **La variance des valeurs de x**, notée $V(x)$ par $V(x) = \frac{1}{n}\left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\right]$

* **L'écart-type des valeurs de x**, noté $\sigma(x)$, est la racine carrée de la variance : $\sigma(x) = \sqrt{V(x)}$

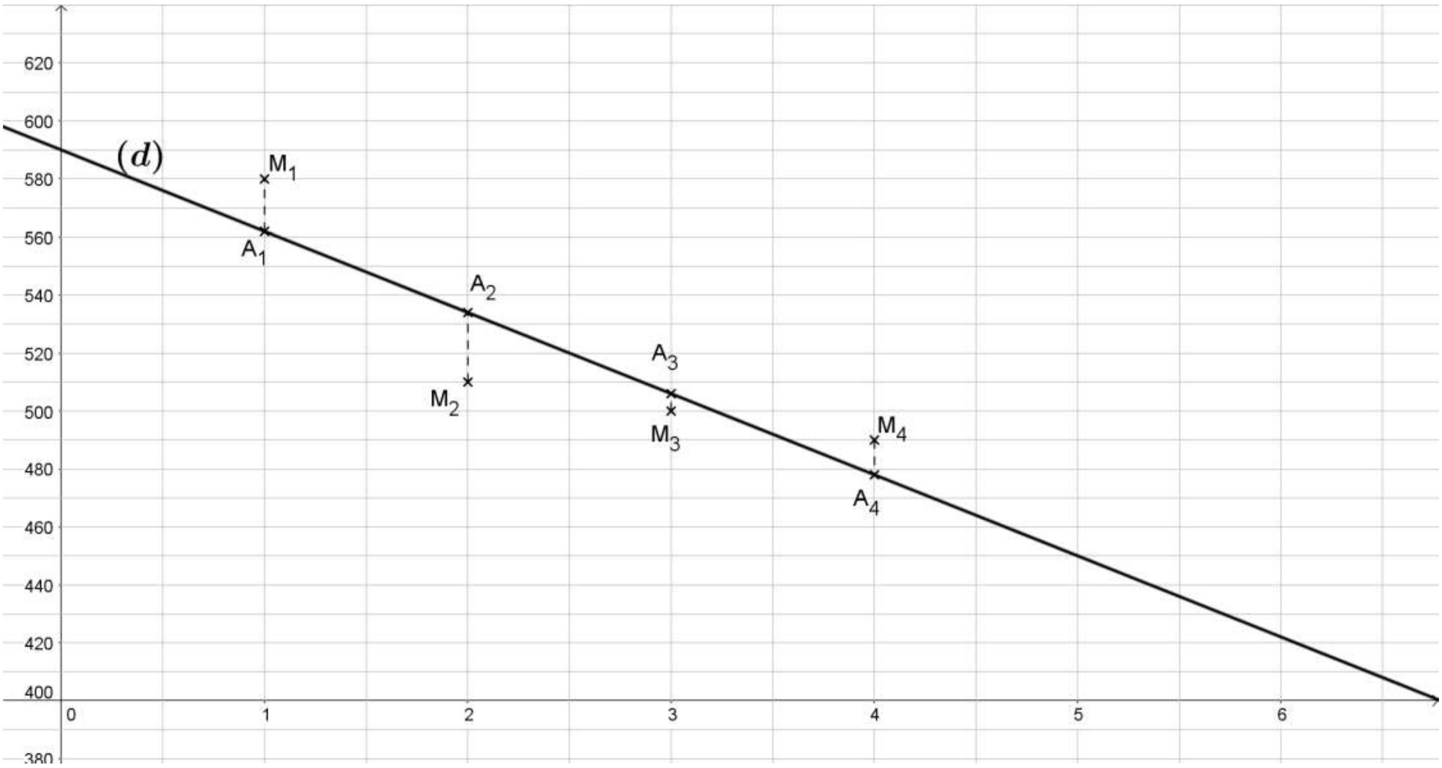
* **La variance des valeurs de y**, notée $V(y)$ par $V(y) = \frac{1}{n}\left[(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \ldots + (y_n - \overline{y})^2\right]$

* **L' écart-type des valeurs de y**, noté $\sigma(y)$, est la racine carrée de la variance : $\sigma(y) = \sqrt{V(y)}$

* **La covariance de la série (x_i ; y_i)**, notée $\text{cov}(x\,;y)$ par

$$\text{cov}(x\,;y) = \frac{1}{n}\left[(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})\right]$$

4. Sur le graphique ci-dessous, on a placé les points $M_i\ (x_i\ ;\ y_i)$. La droite (d) tracée sur ce graphique réalise un ajustement affine du nuage de points.



Pour que cette droite réalise " le meilleur ajustement possible du nuage de points, on a choisi le critère suivant : " *cette droite doit être telle que la somme* $M_1A_1^2 + M_2A_2^2 + M_3A_3^2 + M_4A_4^2$ *soit la plus petite possible* ".

On dit alors qu'on a appliqué la **méthode des moindres carrés**.

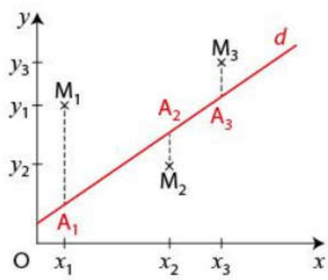
On a les propriété suivantes (admises) :

Propriétés 4 : méthode des moindres carrés

Dans un repère, **la droite des moindres carrés** (ou **droite de régression**) de y en x associée au nuage de points $M_i\ (x_i\ ;\ y_i)$:

* passe par le point moyen $G\ (\overline{x}\ ;\ \overline{y})$ du nuage,

* a pour équation $y = a\,(x - \overline{x}) + \overline{y}$ avec $a = \dfrac{\text{cov}\,(x\ ;\ y)}{V\,(x)}$



a. Sur le graphique précédent, placer le point moyen $G\ (\overline{x}\ ;\ \overline{y})$ du nuage et vérifier que G est sur la droite (d)

b. Déterminer l'équation réduite de la droite (d).

c. En supposant que l'évolution se poursuive ainsi, donner l'estimation du chiffre d'affaire de cette entreprise en 2027.

Exercice 6 :

Afin d'orienter ses investissements, une chaîne d'hôtels réalise des analyses sur le montant des frais de publicité (en milliers d'euros) par rapport au taux d'occupation des chambres (exprimé en %).

Frais x_i	22	24	25	27	30
Taux y_i	48	50	55	54	58

Les résultats seront arrondis au centième.

1. Représenter, dans un repère, le nuage de points associé à cette série et placer le point moyen G de ce nuage.

2. Présenter dans un tableau (voir modèle dans l'exercice 5) les calculs permettant d'obtenir l'équation de la droite de régression de y en x associé à ce nuage (et donner cette équation).

3. Tracer cette droite dans le repère précédent.

4. Estimer le taux d'occupation des chambres pour 28 000 € de frais de publicité.

Exercice 7 :

Le tableau ci-contre donne, pour les années 2010 à 2023 (les statistiques de 2013 et 2020 ne sont pas données), l'âge moyen y_i des femmes ayant accouché en France Métropolitaine.

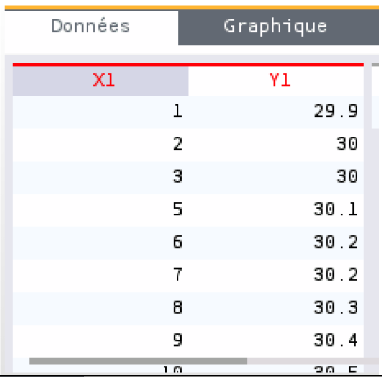
Année	Rang x_i	Age y_i
2010	1	29,9
2011	2	30
2012	3	30
2014	5	30,1
2015	6	30,2
2016	7	30,2

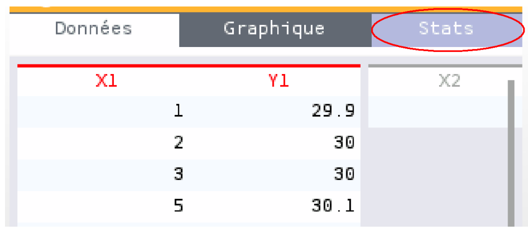
Année	Rang x_i	Age y_i
2017	8	30,3
2018	9	30,4
2019	10	30,5
2021	12	30,8
2022	13	30,9
2023	14	30,9

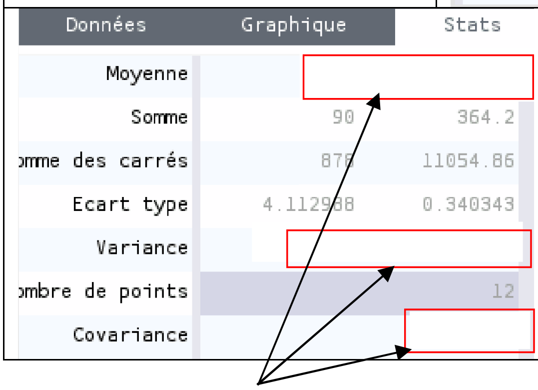
Source : Etat civil

Le nombre de données étant élevé, on va utiliser les fonctionnalités de la Numworks pour déterminer les résultats utiles pour l'obtention de la droite de régression de y en x.

Rentrer les données







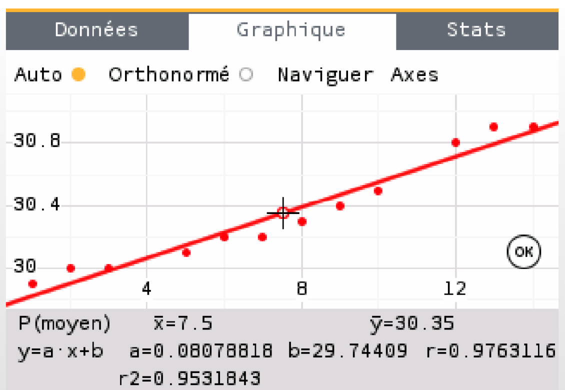
1. Compléter les données manquantes et déterminer l'équation de la droite de régression de y en x. (*arrondir au centième*)

2. Utiliser cette droite pour estimer :

a. en quelle année l'âge moyen des femmes des femmes ayant accouché en France pourrait dépasser 35 ans.

b. l'âge moyen des femmes ayant accouché en France en 2020.

Remarque : la touche "Graphique" permet d'obtenir le nuage de points ainsi que la droite de régression de y en x



Objectif n° 3 : Coefficient de corrélation linéaire

Introduction :

La décision d'ajuster un nuage de points par une droite s'est prise jusqu'à présent à la seule vue du nuage de points, selon que sa forme est "allongée ou pas ".

D'un point de vue mathématique, il est nécessaire de quantifier cette décision. Voici comment :

Définition 5

Considérons une série statistique à deux variables x et y :

Valeur x_i	x_1	x_2	...	x_n
Valeur y_i	y_1	y_2	...	y_n

On appelle **coefficient de corrélation linéaire entre x et y** le réel r défini par

$$r = \frac{\operatorname{cov}(x\,;y)}{\sigma(x)\times\sigma(y)}$$

On rappelle que $\sigma(x)$ et $\sigma(y)$ désignent l'écart-type de x et de y, $\sigma(x) = \sqrt{V(x)}$ et $\sigma(y) = \sqrt{V(y)}$

Propriété 6

Pour toute série statistique à deux variables quantitatives : $-1 \leqslant r \leqslant 1$

* **Si $r = 1$ ou si $r = -1$ la corrélation entre x et y est maximum** : les points du nuage sont alignés.

* si $r < 0$, la droite de régression a une **pente négative**.

* si $r > 0$, la droite de régression a une **pente positive**.

Remarque :

Plus r est proche de 1 ou de – 1, plus l'ajustement affine est judicieux.

En pratique, on considère que l'ajustement linéaire est affine lorsque $r \geqslant 0{,}96$ (ou lorsque $r \leqslant -0{,}96$).

Exercice 8 :

1. Reprenez les données de l'exercice 5 et calculer le coefficient de corrélation r entre x et y.

2. L'ajustement linéaire était-il judicieux ?

Exercice 9 :

Lors d'une randonnée dans une région montagneuse, Jenny a mesuré à différents reprises l'altitude h_i, en km, à laquelle elle se trouvait et la température t_i en °C.

Altitude h_i	Température t_i
400	8,2
800	6,2
1200	3
1600	0,2
2000	– 2

1. A l'aide de la calculatrice, déterminer la covariance de h et t ainsi que les écart-types de h et de t.

2. En déduire le coefficient de corrélation entre h et t.

3. Peut-on envisager un ajustement affine de ce nuage de points ? Si oui, donner l'équation de la droite de régression de t en h (*arrondir au millième*).

4. Estimer alors la température que devrait avoir Jenny à 2500 m d'altitude.

Objectif n° 4 : Ajustement à l'aide d'un changement de variable

Exercice 10 :

Au cours d'une séance d'essai, un pilote automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule.

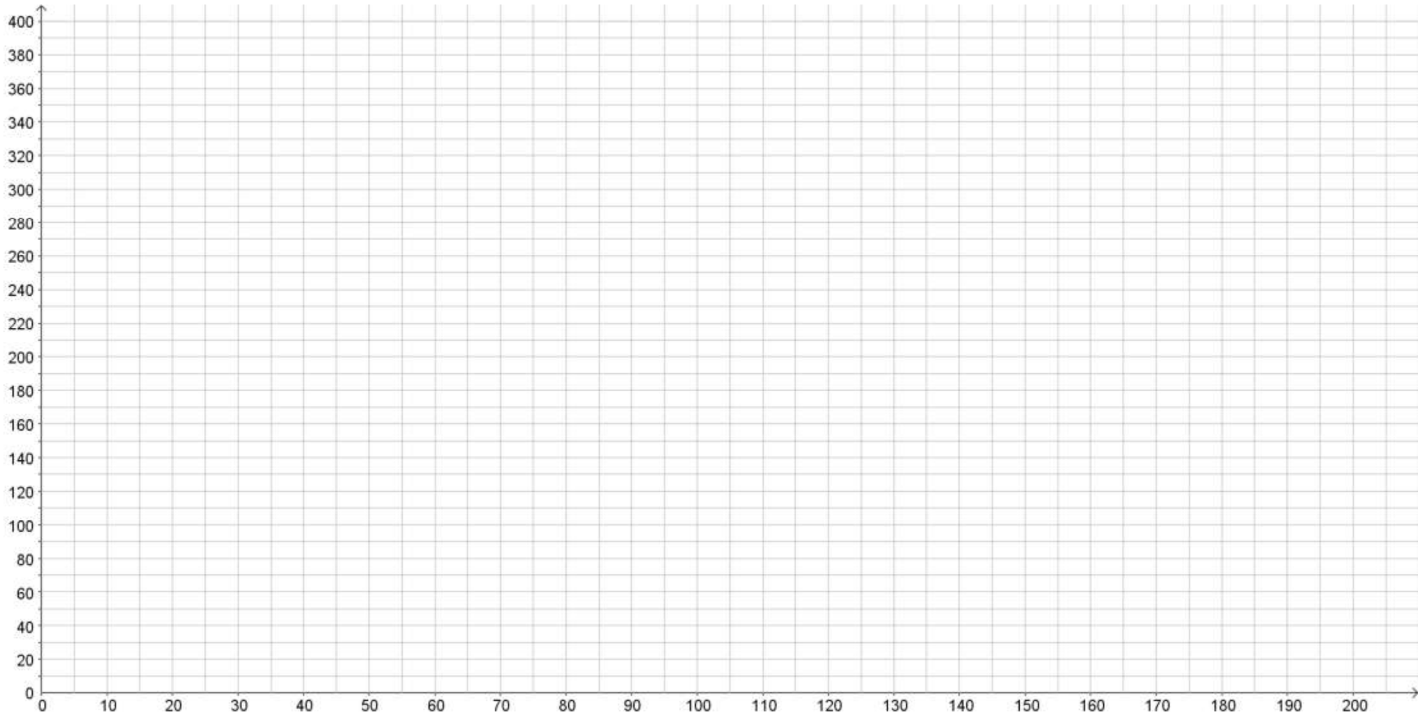
Au moment du top sonore, on mesure la vitesse v_i (en km.h^{-1}) de l'automobile, puis la distance d_i, en m, nécessaire pour arrêter le véhicule.

On a obtenu les résultats suivants :

Vitesse v_i	0	20	43	62	80	115	130	150
Distance d'arrêt d_i	0	3,5	20,5	35,9	67,8	135,8	168,5	250

1. A l'aide de la calculatrice, calculer le coefficient de corrélation r_1 entre v et d. Un ajustement affine est-il judicieux ?

2. Représenter ci-dessous le nuage de points associé à cette série.



3. L'allure du nuage de points peut faire penser à la courbe d'une fonction du 2nd degré. Posons alors $x = v^2$.

a. Compléter le tableau ci-dessous :

$x_i = (v_i)^2$								
d_i	0	3,5	20,5	35,9	67,8	135,8	168,5	250

b. Calculer alors le coefficient de corrélation r_2 entre x et d.

c. Déterminer l'équation de la droite de régression de d en x (*arrondir au centième*)

4. a. En déduire une relation entre la distance d'arrêt d et la vitesse v .

b. Estimer la distance d'arrêt pour une vitesse de 180 km.h^{-1}

c. Lors d'une expérience, la distance d'arrêt d'un pilote a été de 80 m. Donner une estimation (arrondie au dixième) de sa vitesse.

Exercice 11 :

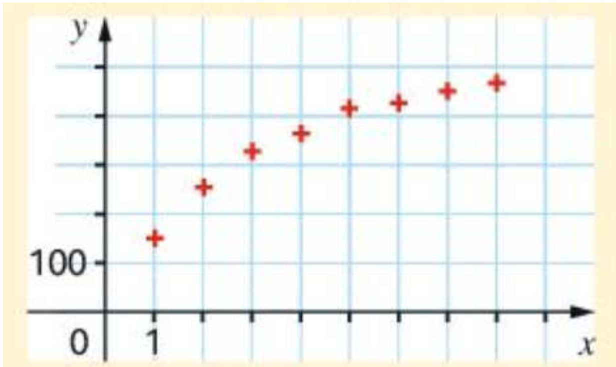
Le responsable d'un site internet s'intéresse au nombre de pages visitées au cours des huit semaines suivant l'ouverture de son site :

Semaine x_i	1	2	3	4	5	6	7	8
Nombre de pages visitées y_i (en milliers)	160	240	315	361	412	435	451	470

Tous les résultats seront arrondis au centième.

1. Un ajustement affine entre x et y est-il judicieux ?

2. En observant plus attentivement le nuage de points, le responsable du site remarque que la disposition des points semble suivre une courbe correspondant à une fonction logarithme.



a. On pose $z_i = \ln(x_i)$. Compléter le tableau ci-dessous (*arrondir au millième*):

$z_i = \ln(x_i)$								
y_i	160	240	315	361	412	435	451	470

b. A l'aide de la calculatrice, déterminer le coefficient de corrélation r' entre z et y. Le responsable du site a-t-il eu raison de ce changement de variable ?

c. Donner l'équation de la droite de régression de y en z (*arrondir au centième*)

d. Estimer alors une relation entre y et x.

e. Si l'évolution se poursuit ainsi, quel sera le nombre de pages visitées durant la $10^{\text{ème}}$ semaine ?

f. Si l'évolution se poursuit ainsi, au bout de combien de temps y aura-t-il plus de 600 000 pages visitées ?

Remarque : dans cet exercice, le nuage de points peut aussi faire penser à une fonction "racine carrée".
On aurait alors pu poser $z_i = \sqrt{x_i}$ (au lieu de $z_i = \ln(x_i)$). Dans ce cas, le coefficient de corrélation aurait été égal à environ 0,990, ce qui aurait permis un bon ajustement (à peine un peu moins bon que celui effectué en utilisant le logarithme).

Exercice 12 :

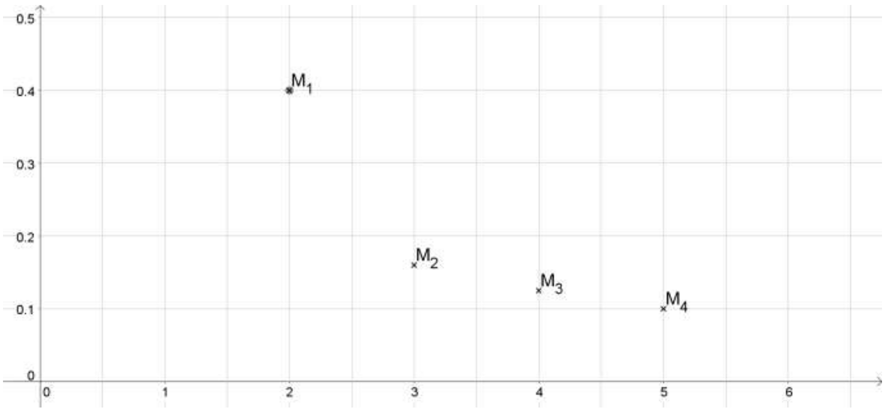
On considère la série statistique à deux variables donnée par le tableau ci-dessous :

Valeurs x_i	2	3	4	5
Valeurs y_i	0,4	0,16	0,125	0,1

1. A l'aide de la calculatrice, déterminer le coefficient de corrélation linéaire r entre x et y. Un ajustement affine est-il judicieux. ?

2. Voici le nuage de points associé à cette série.

Il peut faire penser à une fonction "inverse".



a. On pose $z_i = \frac{1}{y_i}$. Compléter le tableau ci-dessous :

Valeurs x_i	2	3	4	5
Valeurs $z_i = \frac{1}{y_i}$				

b. A l'aide de la calculatrice, déterminer le coefficient de corrélation linéaire $r\,'$ entre z et y.

c. Donner l'équation de la droite de régression de z en x (*arrondir au centième*)

d. Estimer alors une relation entre y et x.

e. Estimer la valeur de y si $x = 1$.