STATISTIQUES à 2 variables

I. Série statistique à deux variables

1) Nuage de points

On considère deux variables statistiques x et y observées sur une même population de n individus.
On note $x_1, x_2, \dots, x_n$ les valeurs relevées pour la variable x et $y_1, y_2, \dots, y_n$ les valeurs relevées pour la variable y.
Les couples $(x_1\ ;\ y_1), (x_2\ ;\ y_2), \dots, (x_n\ ;\ y_n)$ forment une série statistique à deux variables.

Dans ce chapitre, on va s'intéresser au lien qui peut exister entre ces deux variables.

Définition : Dans un repère orthogonal, l'ensemble des points M_i de coordonnées $(x_i\ ;\ y_i)$, avec $1 \leq i \leq n$, est appelé le **nuage de points** associé à la série statistiques $(x_1\ ;\ y_1), (x_2\ ;\ y_2), \dots, (x_n\ ;\ y_n)$ à deux variables.

2) Point moyen

Définition : Le point G de coordonnées $(\bar{x}\ ;\ \bar{y})$, où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des x_i et des y_i, est appelé le **point moyen** du nuage de points associé à la série statistique $(x_1\ ;\ y_1), (x_2\ ;\ y_2), \dots, (x_n\ ;\ y_n)$ à deux variables.

Méthode : Représenter un nuage de points

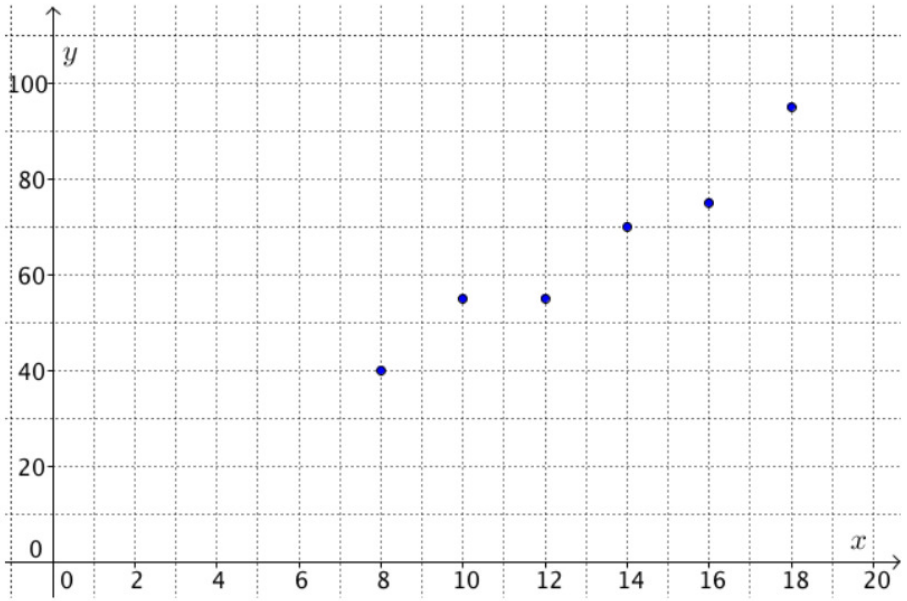
Vidéo https://youtu.be/Nn6uckb3RvE

Le tableau suivant présente l'évolution du budget publicitaire et du chiffre d'affaire d'une société au cours des 6 dernières années :

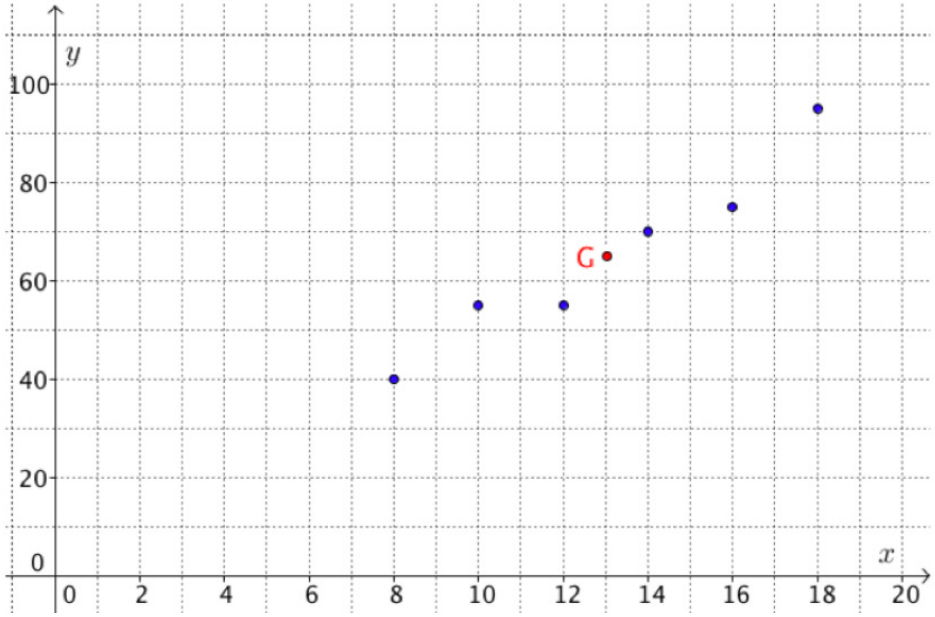
Budget publicitaire en milliers d'euros x_i	8	10	12	14	16	18
Chiffre d'affaire en milliers d'euros y_i	40	55	55	70	75	95

1) Dans un repère, représenter le nuage de points $(x_i\ ;\ y_i)$.
2) Déterminer les coordonnées du point moyen G du nuage de points.

1)



2) $\bar{x}$ = (8 + 10 + 12 + 14 + 16 + 18) : 6 = 13
$\bar{y}$ = (40 + 55 + 55 + 70 + 75 + 95) : 6 = 65.
Le point moyen G du nuage de points a pour coordonnées (13 ; 65). On peut placer ce point dans le repère.



II. Ajustement affine

1) Interpolation, extrapolation

L'objectif est, à partir des valeurs d'une série statistique à deux variables, d'obtenir des approximations pour des valeurs inconnues de cette série.

Exemples :
- On donne une série exprimant la population d'une ville en fonction des années et on souhaite faire des prévisions pour les années à venir.
Les prévisions sortent du domaine d'étude de la série, on parle dans ce cas d'**extrapolation**.
- On donne une série exprimant la température extérieure et la consommation électrique correspondante. Les températures étudiées s'échelonnent entre -10°C et 10°C avec un pas de 4°C.
Sans faire de nouveaux relevés, on souhaite estimer la consommation électrique pour toutes les températures entières comprises entre -10°C et 10°C.
Les calculs sont dans le domaine d'étude de la série, on parle dans ce cas d'**interpolation**.

Définitions : L'**interpolation** et l'**extrapolation** sont des méthodes qui consistent à estimer une valeur inconnue dans une série statistique.
- Pour une interpolation, le calcul est réalisé dans le domaine d'étude fourni par les valeurs de la série.
- Pour une extrapolation, le calcul est réalisé en dehors du domaine d'étude.

La méthode d'extrapolation est parfois contestable car en dehors du domaine d'étude fourni par les valeurs de la série. Rien ne nous assure en effet que le modèle mathématique mis en œuvre soit encore valable.

2) Droite d'ajustement

Pour obtenir de telles estimations, il faudra déterminer une droite passant « le plus près possible » des points du nuage.
L'interpolation ou l'extrapolation consistent à effectuer l'estimation par lecture graphique sur la droite ou par calcul à l'aide de l'équation de la droite.

Définition : Lorsque les points d'un nuage sont sensiblement alignés, on peut construire une droite, appelé **droite d'ajustement (ou droite de régression)**, passant « au plus près » de ces points.

Dans la suite, nous allons étudier différentes méthodes permettant d'obtenir une telle droite.

3) Méthode de Mayer

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage de point.

Méthode : Déterminer la droite d'ajustement par la méthode des points moyens

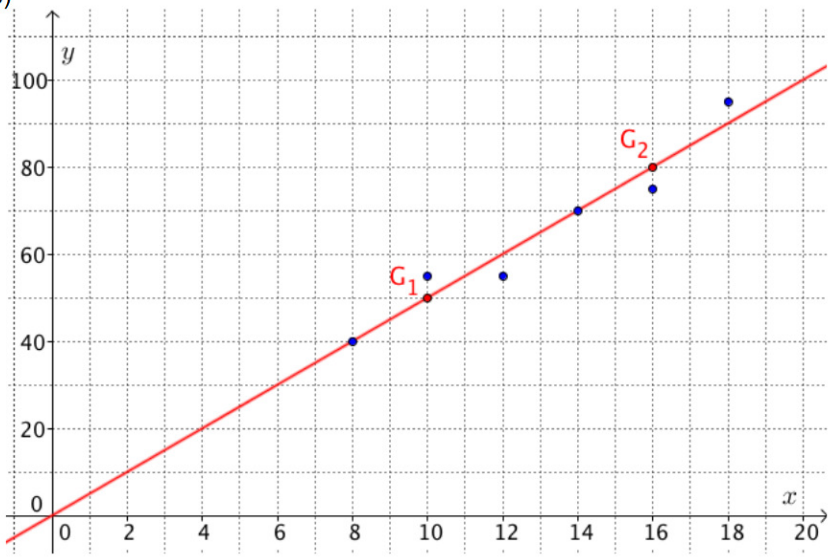
Vidéo https://youtu.be/ESHY4QPgriw

On reprend les données de la méthode du paragraphe I.
1) Soit G_1, le point moyen associé aux trois premiers points du nuage et G_2 le point moyen associé aux trois derniers points du nuage.
  a) Calculer les coordonnées de G_1 et G_2.
  b) On prend (G_1G_2) comme droite d'ajustement. Tracer cette droite.
2) À l'aide du graphique :
  a) Estimer le chiffre d'affaire à prévoir pour un budget publicitaire de 22 000 €.
  b) Estimer le budget publicitaire qu'il faudrait prévoir pour obtenir un chiffre d'affaire de 100 000 €.
  c) La méthode utilisée dans les questions 2a et 2b consiste-t-elle en une interpolation ou une extrapolation ?

1) a) $\overline{x_1} = (8 + 10 + 12) : 3 = 10$
$\overline{y_1} = (40 + 55 + 55) : 3 = 50.$
Le point moyen G_1 a pour coordonnées (10 ; 50).

$\overline{x_2} = (14 + 16 + 18) : 3 = 16$
$\overline{y_2} = (70 + 75 + 95) : 3 = 80.$
Le point moyen G_2 a pour coordonnées (16 ; 80).

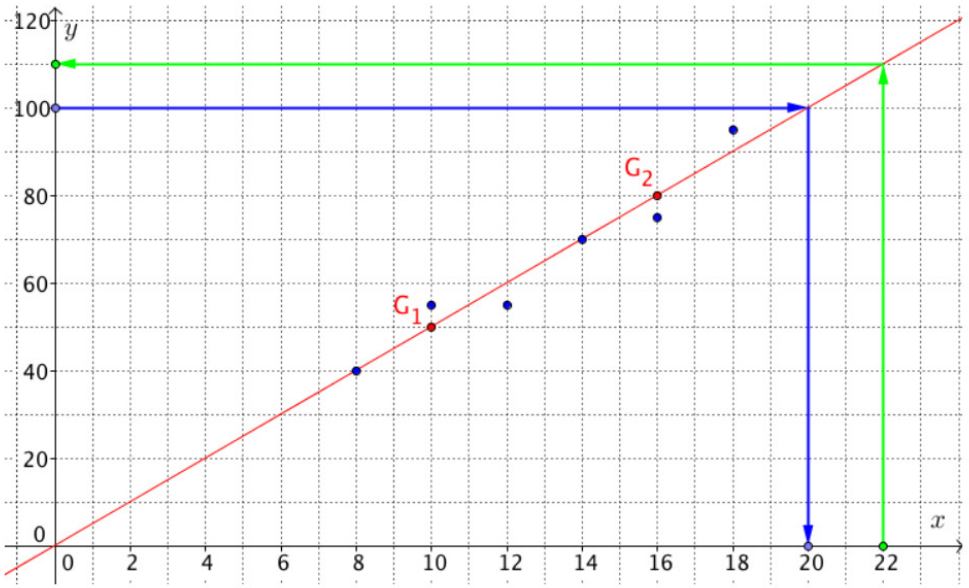
b)



2) On lit graphiquement :

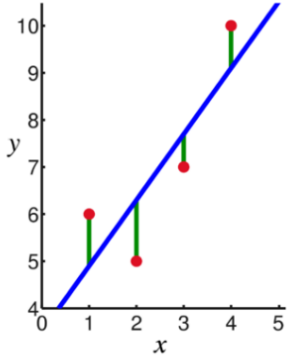
a) Le chiffre d'affaire à prévoir pour un budget publicitaire de 22 000 € est de 110 000 €.

b) Le budget publicitaire qu'il faudrait prévoir pour obtenir un chiffre d'affaire de 100 000 € est de 20 000€.



c) Les lectures graphiques sont réalisées ici en dehors du domaine d'étude, on parle donc d'extrapolation.

4) Méthode des moindres carrés



Cette méthode porte le nom de « moindre carrés » car elle consiste à rechercher la position de la droite d'ajustement tel que la somme des carrés des longueurs donnant les distances respectives (en vert) entre la droite et les points soit minimale.

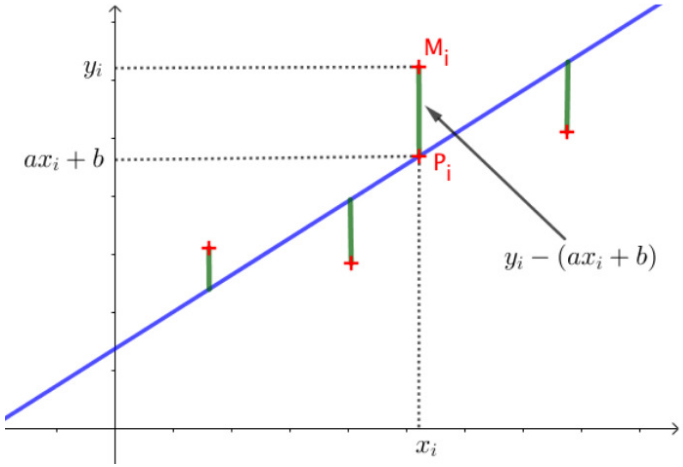
Le principe consiste donc à déterminer les coefficients a et b d'une droite d'équation $y = ax + b$ de sorte qu'elle passe le « plus près possible » des points du nuage.

Pour chaque abscisse x_i, on calcule la distance M_iP_i entre le point du nuage et le point de la droite, soit :

$$M_iP_i = |y_i - (ax_i + b)|$$

Il s'agit dans ce cas, de la droite d'ajustement de y en x.

A noter : Il existe également une droite d'ajustement de x en y en calculant les distances obtenues par projection horizontale.



Dans la méthode des moindres carrés, on recherche a et b pour lesquels la somme des carrés des distances est minimale, soit :

$$M_1P_1 + \cdots + M_nP_n = \left(y_1 - (ax_1 + b)\right)^2 + \cdots + \left(y_n - (ax_n + b)\right)^2 \text{ est minimale.}$$

Pour cela, on peut appliquer la propriété suivante :

Propriété : La droite d'ajustement de y en x a pour équation $y = ax + b$, avec :

- $a = \dfrac{cov(x,y)}{var(x)}$
- $b = \bar{y} - a\bar{x}$

où $cov(x;y) = \frac{1}{n}\left((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\right)$ est la covariance de (x,y)

et $var(x) = \frac{1}{n}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$ est la variance de x.

- Admis -

Méthode : Déterminer la droite d'ajustement par la méthode des moindres carrés

Vidéo https://youtu.be/vdEL0MOKAIg

On considère la série statistique à deux variables données dans le tableau suivant :

x_i	5	10	15	20	25	30	35	40
y_i	13	23	34	44	50	65	75	90

1) Dans un repère, représenter le nuage de points $(x_i\,;\,y_i)$.

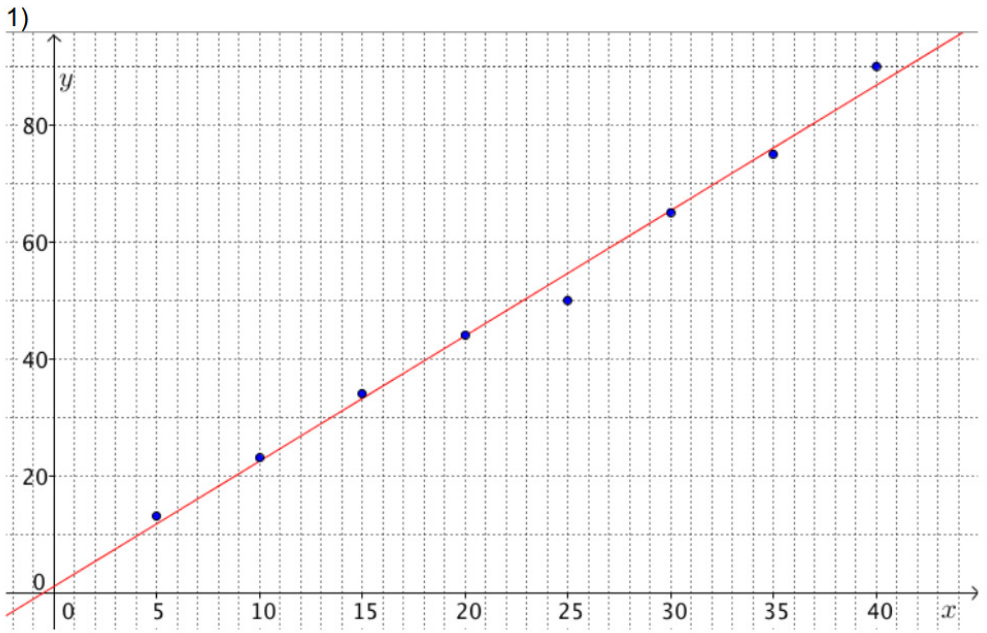
2) a) Déterminer une équation de la droite d'ajustement par la méthode des moindres carrés.

b) Vérifier à l'aide de la calculatrice.

c) Représenter la droite d'ajustement de y en x.

3) Estimer graphiquement la valeur de x pour y = 70. Retrouver ce résultat par calcul. S'agit-il d'une interpolation ou d'une extrapolation ?

1)



2) a) On commence par calculer, les moyennes $\bar{x}$ et $\bar{y}$:

$$\bar{x} = \frac{5 + 10 + \cdots + 40}{8} = 22{,}5$$

$$\bar{y} = \frac{13 + 23 + \cdots + 90}{8} = 49{,}25$$

Par la méthode des moindres carrés, la droite d'ajustement de y en x a pour équation $y = ax + b$ avec :

$$a = \frac{cov(x;y)}{var(x)}$$

$$= \frac{\frac{1}{8}\left((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_8 - \bar{x})(y_8 - \bar{y})\right)}{\frac{1}{8}\left((x_1 - \bar{x})^2 + \cdots + (x_8 - \bar{x})^2\right)}$$

$$= \frac{(5 - 22{,}5)(13 - 49{,}25) + (10 - 22{,}5)(23 - 49{,}25) + \cdots + (40 - 22{,}5)(90 - 49{,}25)}{(5 - 22{,}5)^2 + (10 - 22{,}5)^2 + \cdots + (40 - 22{,}5)^2}$$

$$\approx 2.138$$

Et $b = \bar{y} - a\bar{x} \approx 49{,}25 - 2{,}138 \times 22{,}5 = 1{,}145$

Une équation de la droite d'ajustement est : $y = 2{,}138x + 1{,}145$
Pour le tracé, on considère l'équation : $y = 2{,}1x + 1{,}1$

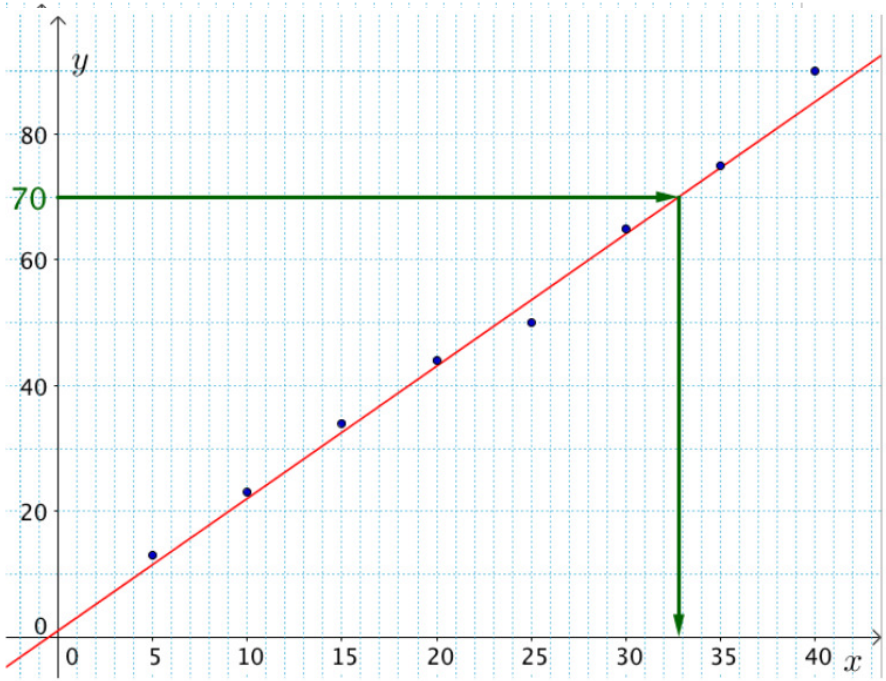
b) Avec TI :
- Appuyer sur « **STAT** » puis « **Edite** » et saisir les valeurs de x_i dans L1 et les valeurs de y_i dans L2.
- Appuyer à nouveau sur « **STAT** » puis « **CALC** » et « **RegLin(ax+b)** »
- Saisir **L1,L2**

Avec CASIO :
- Aller dans le menu « **STAT** ».
- Saisir les valeurs de x_i dans *List1* et les valeurs de y_i dans *List2*.
- Sélectionner « **CALC** » puis « **SET** ».
- Choisir *List1* pour *2Var XList* et *List2* pour *2Var YList* puis « **EXE** ».
- Sélectionner « **REG** » puis « **X** » et « **aX+b** ».

La calculatrice nous renvoie : $a = 2.138095238$ et $b = 1.142857143$
Une équation de la droite d'ajustement est : $y = 2{,}1x + 1{,}1$

Pour tracer la droite, il suffit de calculer les coordonnées de deux points de la droite d'ajustement :
- Si $x = 0$ alors $y = 2{,}1 \times 0 + 1{,}1 = 1{,}1$ donc le point de coordonnées $(0\,;1{,}1)$ appartient à la droite d'ajustement.
- Si $x = 10$ alors $y = 2{,}1 \times 10 + 1{,}1 = 22{,}1$ donc le point de coordonnées $(10\,;22{,}1)$ appartient à la droite d'ajustement.

c)



3) - Pour $y = 70$, on lit graphiquement $x \approx 33$.

- Par calcul, si $y = 70$, alors $70 = 2{,}1x + 1{,}1$

Soit $2{,}1x = 70 - 2{,}1$

$2{,}1x = 68{,}9$

$x = \frac{68{,}9}{2{,}1} \approx 32{,}8$

- Les calculs sont réalisés dans domaine d'étude, on parle donc d'interpolation.

5) Coefficient de corrélation

Définition : Le **coefficient de corrélation** de x et y est donné par :

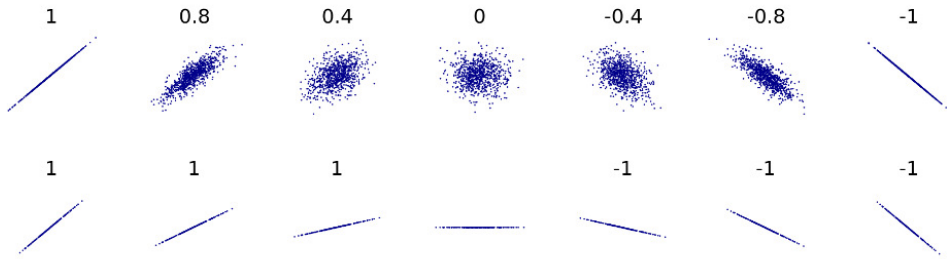
$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

Interprétation :

Le coefficient de corrélation ρ_{xy} est un nombre compris entre -1 et 1 qui mesure la relation entre les deux variables x et y. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte.

- Si $\rho_{xy} > 0$, les valeurs prises par y ont tendance à croître quand les valeurs de x augmentent.
- Si $\rho_{xy} < 0$, les valeurs prises par y ont tendance à décroître quand les valeurs de x augmentent.
- Si $\rho_{xy} = 0$, les variations des variables x et y sont indépendantes.

Exemples de coefficients de corrélation :



Méthode : Calculer un coefficient de corrélation

Vidéo https://youtu.be/FxREenh3fgE

En reprenant les données de la méthode précédente, calculer le coefficient de corrélation et interpréter le résultat.

$$cov(x,y) = \frac{1}{8}\left((5 - 22{,}5)(13 - 49{,}25) + \cdots + (40 - 22{,}5)(90 - 49{,}25)\right) \approx 280{,}625$$

$$var(x) = \frac{1}{8}\left((5 - 22{,}5)^2 + \cdots + (40 - 22{,}5)^2\right) \approx 131{,}25$$

$$var(y) = \frac{1}{8}\left((13 - 49{,}25)^2 + \cdots + (90 - 49{,}25)^2\right) \approx 604{,}4375$$

Soit :

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} \approx \frac{280{,}625}{\sqrt{131{,}25 \times 604{,}4375}} \approx 0{,}996$$

Le coefficient de corrélation est proche de 1 donc la corrélation entre les deux variables est fortes. Les points du nuage sont proches de la droite d'ajustement.

III. Ajustement par changement de variable

Lorsque le nuage de points n'est à priori pas modélisable par une droite, on peut réaliser un ajustement linéaire en effectuant un changement de variable.

Méthode : Effectuer un ajustement se ramenant par changement de variable à un ajustement affine

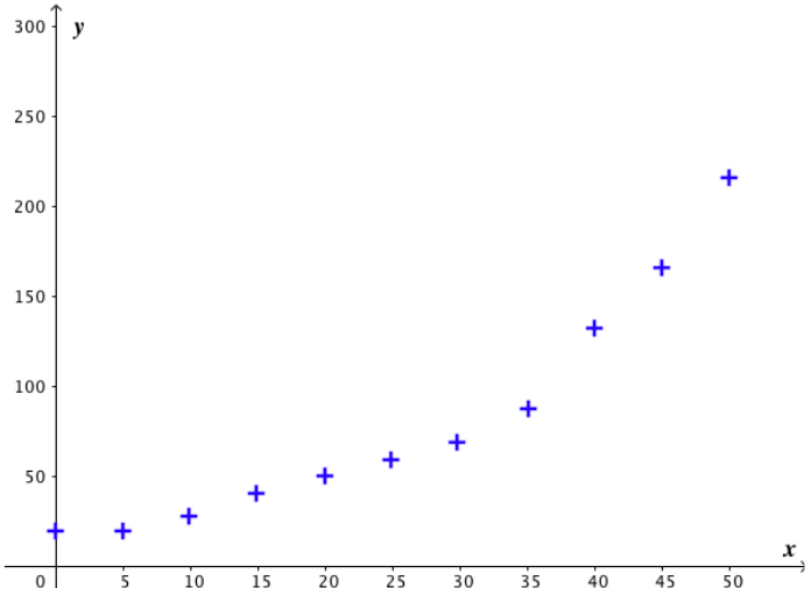
Vidéo https://youtu.be/nVDL0razCIY

On a relevé la population d'une grande métropole sur 50 ans tous les 5 ans. Les résultats sont présentés dans le tableau suivant :

Année x_i	0	5	10	15	20	25	30	35	40	45	50
Population en milliers y_i	19,4	19,4	27,6	40,3	50	59	69	87	132	166	216

1) Représenter le nuage de points dans un repère.
2) a) On effectue le changement de variable $z = \ln y$. Réaliser un nouveau tableau présentant les valeurs prises par les variables x et z.
b) Représenter un nouveau nuage de points à partir des données des variables x et z.
c) A l'aide la calculatrice, déterminer une équation de la droite d'ajustement de z en x par la méthode des moindres carrés. Représenter la droite d'ajustement.
3) a) En déduire la relation qui lie y et x puis tracer la courbe représentative de la fonction f définie par $y = f(x)$ dans le repère contenant le premier nuage de points.
b) En admettant que le modèle mathématique reste valable en dehors du domaine d'étude, extrapoler le nombre d'habitant 5 ans après l'étude.

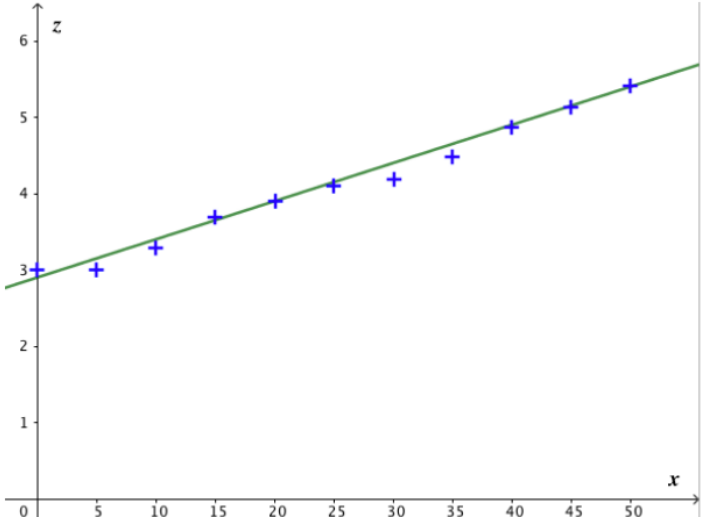
1)



2) a)

x_i	0	5	10	15	20	25	30	35	40	45	50
z_i	3	3	3,3	3,7	3,9	4,1	4,2	4,5	4,9	5,1	5,4

b)



c) Une équation de la droite d'ajustement est :
$z = 0{,}05x + 2{,}9$

On trace la droite : voir ci-dessus.

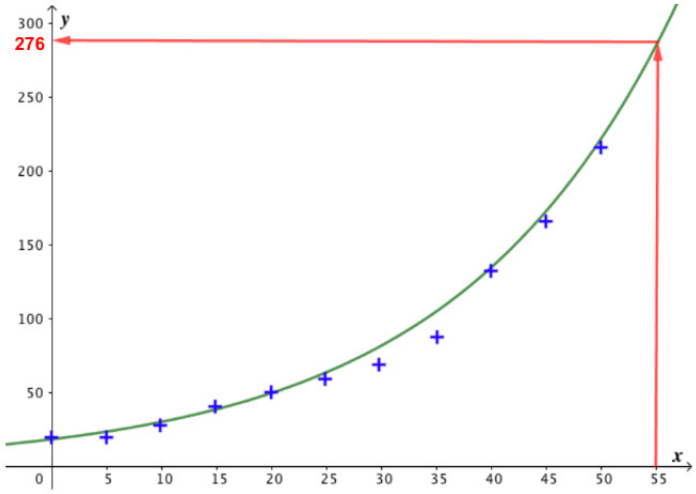
L1	L2
0	3
5	3
10	3.3
15	3.7
20	3.9
25	4.1
30	4.2
35	4.5
40	4.9
45	5.1
50	5.4

RégLin
y=ax+b
a=0.0492727273
b=2.868181818
r²=0.9876412049
r=0.9938013911

3) On a : $z = 0{,}05x + 2{,}87$
et : $z = \ln y$, soit :

$\ln y = 0{,}05x + 2{,}87$
$e^{\ln y} = e^{0{,}05x+2{,}87}$
$y = e^{0{,}05x+2{,}87}$
$y = e^{0{,}05x} e^{2{,}87}$
$y = 17{,}64\, e^{0{,}05x}$

Donc :
$f(x) = 17{,}64\, e^{0{,}05x}$ est l'expression de la fonction permettant d'ajuster le nuage de points $(x_i\ ;\ y_i)$.



b) $y = 17{,}64\ e^{0{,}05\times 55} \approx 276$
On peut supposer que 5 années après la fin de l'étude, la population de la ville sera proche de 276 000 habitants.