Leçon N° 4 : Statistiques à deux variables

En premier lieu, il te faut relire les cours de première sur les statistiques à une variable, il y a tout un langage à se remémorer : **étude d'un échantillon d'une population, mode, moyenne et médiane puis réaliser une classification, ensuite sur la série étudiée, calculer la variance et l'écart type pour savoir si la série est dispersée ou peu dispersée, enfin trouver les quartiles et faire un diagramme en boîte avec positionnement de la médiane dans la boîte** etc…. En terminale, nous allons faire des **statistiques sur deux variables** en essayant de les relier entre elles par une relation simple. Soit donc deux séries statistiques (x_i) et (y_i) i variant de 1 à n (n entier quelconque, généralement, 5 ou 6 jusqu'à 10 quelquefois). Nous représenterons ces données dans un repère du plan (P) par des points $M_i(x_i ; y_i)$ afin de constituer ce que nous appelons **un nuage de points.**

Définition :

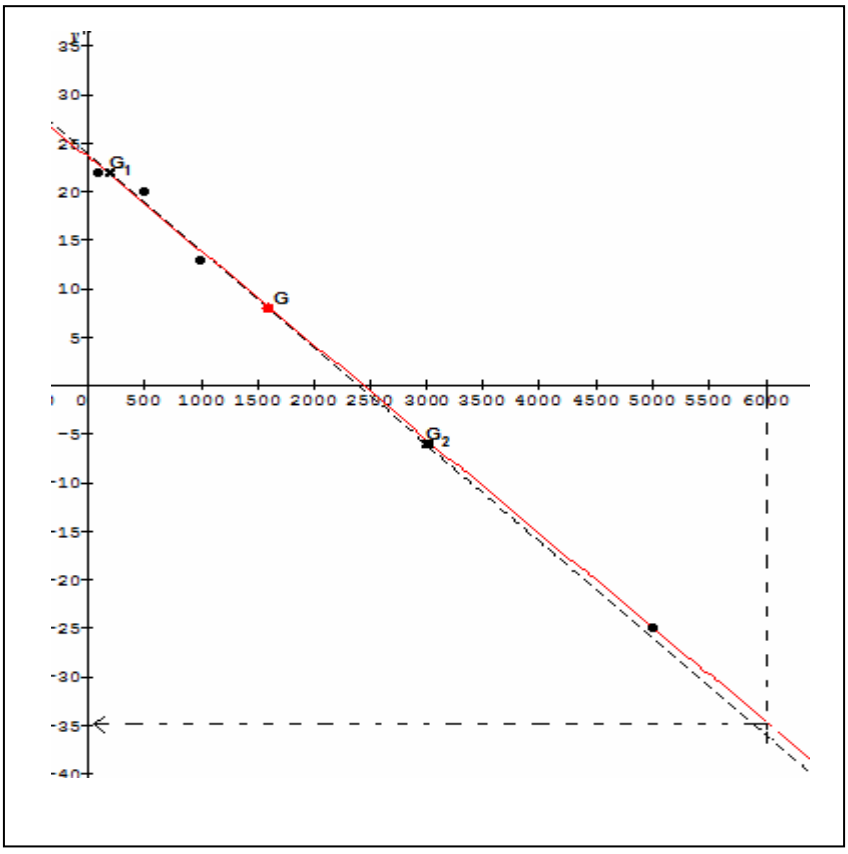
On appelle point moyen d'un nuage, le point G($\overline{x}$; $\overline{y}$ *)* $\overline{x}$ et $\overline{y}$ *moyennes calculées dans chaque série*.

Nous regardons ensuite si nous pouvons tracer une droite d'équation y = ax+b passant le plus prés possible des points. Si cela est possible, nous dirons que nous avons réalisé un ajustement affine du nuage de points et donc trouver une relation simple de la forme y = ax + b entre les deux variables.

Montrons un exemple ;

Dans un avion, en pleine ascension nous notons la température extérieure en degrés avec l'altitude correspondante en mètres, nous avons le tableau suivant :

Altitude (x_i)	0	100	500	1000	3000	5000
Températures (y_i)	24°	22°	20°	13°	– 6°	– 25°



Déterminons le point moyen G : $\overline{x}$ =1600 m ; $\overline{y}$ = 8°.**G(1600 ;8°).**

Pour réaliser un ajustement affine, nous avons une première méthode qui donne « **la droite de Mayer** ». Nous partageons le nuage de points en deux sous-nuages puis cherchons les points moyens de ces nuages G_1 et G_2, la droite cherchée est la droite (G_1G_2).

X	0	100	500
y	24°	22°	20°

X	1000	3000	5000
y	13°	-6°	-25°

G_1(200 ;22) G_2(3000 ;-6°)

(G_1G_2) a une équation de la forme y = ax + b.

a = $\frac{-6-22}{3000-200}$ = – 0,01. Pour trouver b, nous utilisons un des points : 22 = –0,01(200)+b et donc 22 = – 2 + b c'est-à-dire b = 24. **(G_1G_2) y = – 0,01x + 24**.

Au programme, il est demandé d'utiliser **la méthode dite « des moindres carrés »** qui s'est imposée à la place de la méthode de Mayer. Les coefficients sont donnés par la calculette après avoir rentré les données concernant les deux séries statistiques.
Cela donne ici : **a ≈– 9,8 10^{-3} soit –0,0098 et b≈23,65**.

Remarque ; la calculette parle d'un coefficient r, coefficient de corrélation qui indique si l'alignement est valable ou pas. ***Règle : si |r|≈1, alors l'alignement est de bonne qualité***. Ici, r≈–0,999.
La droite trouvée, tracée en rouge sur le graphique, a donc pour équation :
(D) y≈ – 0,0098x + 23,65 .
Les deux droites sont proches l'une de l'autre.
Elles passent par le point moyen G(1600 ; 8°). Nous pouvons le vérifier facilement pour (G_1G_2) :
8 = 1600(−0,01)+24
Si la calculette donne un coefficient de corrélation r dont la valeur absolue est éloigné de 1, cela veut dire qu'un ajustement affine ne se justifie pas car soit, les points ne sont pas assez alignés soit, il y a une grande dispersion des données et un autre type d'ajustement s'impose.
En résumé :
Lorsque nous avons deux séries statistiques, nous pouvons représenter ces données dans un repère du plan (P), cela donne un nuage de points et souvent les points sont alignés dans une certaines direction. Il est possible alors à la machine de trouver les coefficients a et b de la droite d'ajustement (« Méthode des moindres carrés ») . Cette droite (D) passe par le point moyen G(x ;y) du nuage.

Utilité : **Cette droite va permettre des prévisions à court terme par le calcul.**

Pour la température, nous pouvons la prédire pour 6000m par exemple :

Y ≈ (−0,01)6000 + 24 = −36°

Remarque : le problème étudié ci-dessus a fait l'objet de recherche en physique et effectivement, une loi a été trouvée disant que la température baisse de 1° tous les 100m soit si on appelle t la température et t_s la température au sol, x étant en mètre : **$t = -0,01x + t_s$**
(Exemple : t_s = 10°, pour x = 500, t_1 = 5° et pour x = 600, t_2 = 4°)

TERMINALE STG	FICHE LES STATISTIQUES

Exercice 1

Nous voulons étudier l'évolution de la population d'une commune.

Un relevé a été fait donnant le tableau suivant :

Années	1980	1990	2000	2002	2010
x	0	10	20	22	30
Population y	2030	2500	3000	3200	3400

Calculer les coordonnées du point moyen G. Représenter ce nuage de points. A la calculette, déterminer les coefficients a et b de la droite d'ajustement par la méthode des moindres carrés. Donner ensuite l'équation de la droite d'ajustement affine et tracer la sur le graphique. Vérifier que G appartient à cette droite.

Quelle prévision pour 2020 cette droite permet-elle de faire ?

Exercice 2

Le PDG d'une entreprise fait analyser la production d'un produit sur 10 ans. Nous avons le tableau suivant :

Années x	1	2	3	4	5	6	7	8	9	10
Production y	49	48	50	50	56	57	62	65	65	68

Représenter graphiquement ces données.

Pourquoi un ajustement affine est - il possible ?

Placer G le point moyen.

Tracer la droite (D1) passant par G et le dernier point (10 ; 68). ON considère qu'elle réalise un ajustement linéaire valable du nuage. Donner l'équation de (D1).

Utiliser votre calculette pour déterminer a et b les coefficients de la droite (D2) d'ajustement affine par la méthode des moindres carrés. Tracer (D2).

Faire une prévision pour 15 ans en utilisant (D1) et (D2). Quelle est l'erreur en % commise en prenant (D1) à la place de (D2).

Exercice3 (Avec Excel)

Nous avons le tableau suivant :

	A	B	C
1	x_i	y_i	ax_i+b
2	20	50	?
3	30	68	?
4	50	108	?
5	70	150	?
6	80	175	?
7	100	220	?
8	120	250	?

Entrer ces données dans une feuille de calcul Excel.

En utilisant les commandes :

=droitereg(B2 :B11 ;A2 :A11) et

=ordonnee.origine(B2 :B11 ;A2 :A11)

déterminer a te b les coefficients de la droite (D) d'ajustement par la méthode des moindres carrés.

Calculer alors $ax_i + b$

Faire un graphique dans la feuille pour illustrer ceci.

(En sélectionnant la colonne x_i et $ax_i + b$, nous pouvons tracer (D))

Exercice4(Type Bac)

Un couple de restaurateur étudie une formule Brunch-Culture. Ils ont recensé le nombre de personnes intéressées en fonction du prix fixé.

Soit x_i le prix en euros et y_i le nombre de personnes correspondant à ce prix.

x_i	y_i
18	47
20	45
23	42
25	40
28	36
30	30
33	25
35	22
38	18
40	15

1-a Représenter graphiquement ces données.

1-b Peut-on émettre l'hypothèse d'une relation simple entre x et y. Si oui, quelle genre de formule proposez-vous ?

2 Déterminer les coordonnées du point moyen G du nuage représentés précédemment.

3 On choisit de faire un ajustement affine par la droite (D) de coefficient directeur $-1,5$ passant par G. Donner l'équation réduite de cette droite (D) puis tracer la. Lire sur le graphique à partir de quel prix, personne ne viendra utiliser la formule proposée. Vérifier par le calcul.

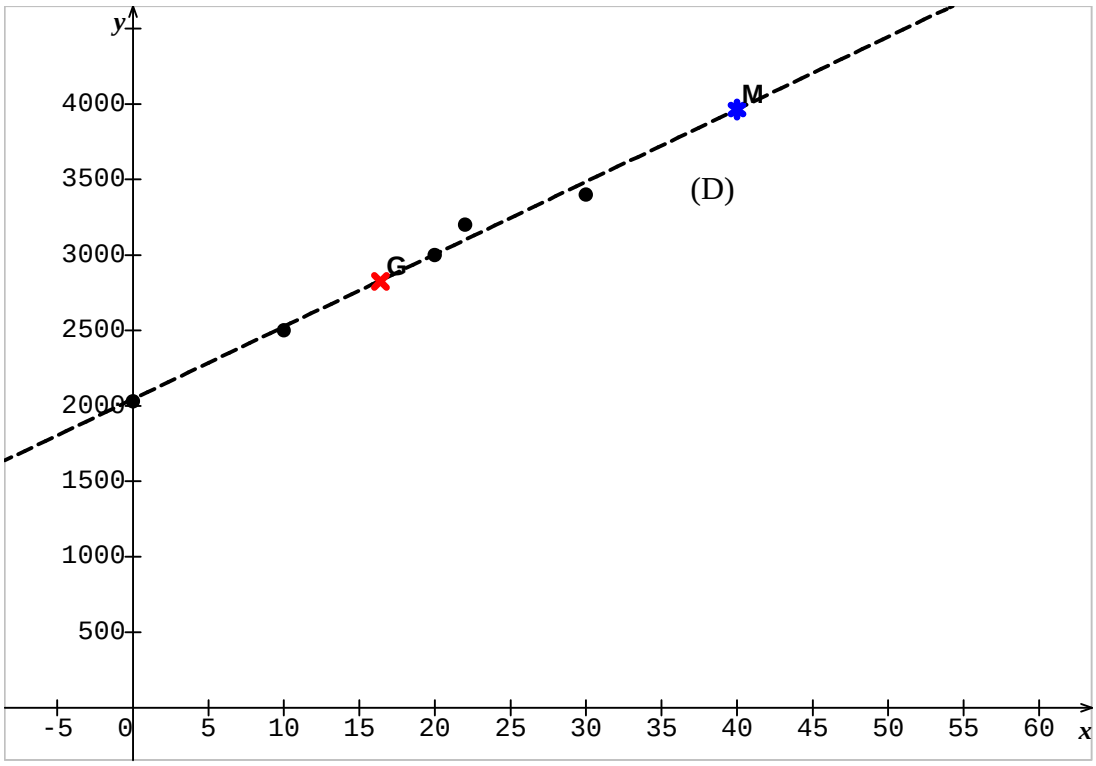
4 Quelle prévision donne (D) si on choisit x = 25€. Quel est en % l'erreur commise avec la réalité ?

Correction

Exercice 1

x_i représente le nombre d'années à partir de 1980 et y_i donne le nombre d'habitants de la commune. $\overline{x}$ = 16,4 et $\overline{y}$ = 2826. **Le point moyen G aura pour coordonnées (16,4 ;2826)**.

Représentons le nuage de points :



Le point G est bien au centre du nuage. Les points sont relativement alignés et la calculette donne : **a≈47,6 soit a≈48 et b≈2045,2 soit b≈2045**. Le coefficient de corrélation r entre x et y est de 0,992 donc l'ajustement affine est valable. La droite d'ajustement (D) aura pour équation ;

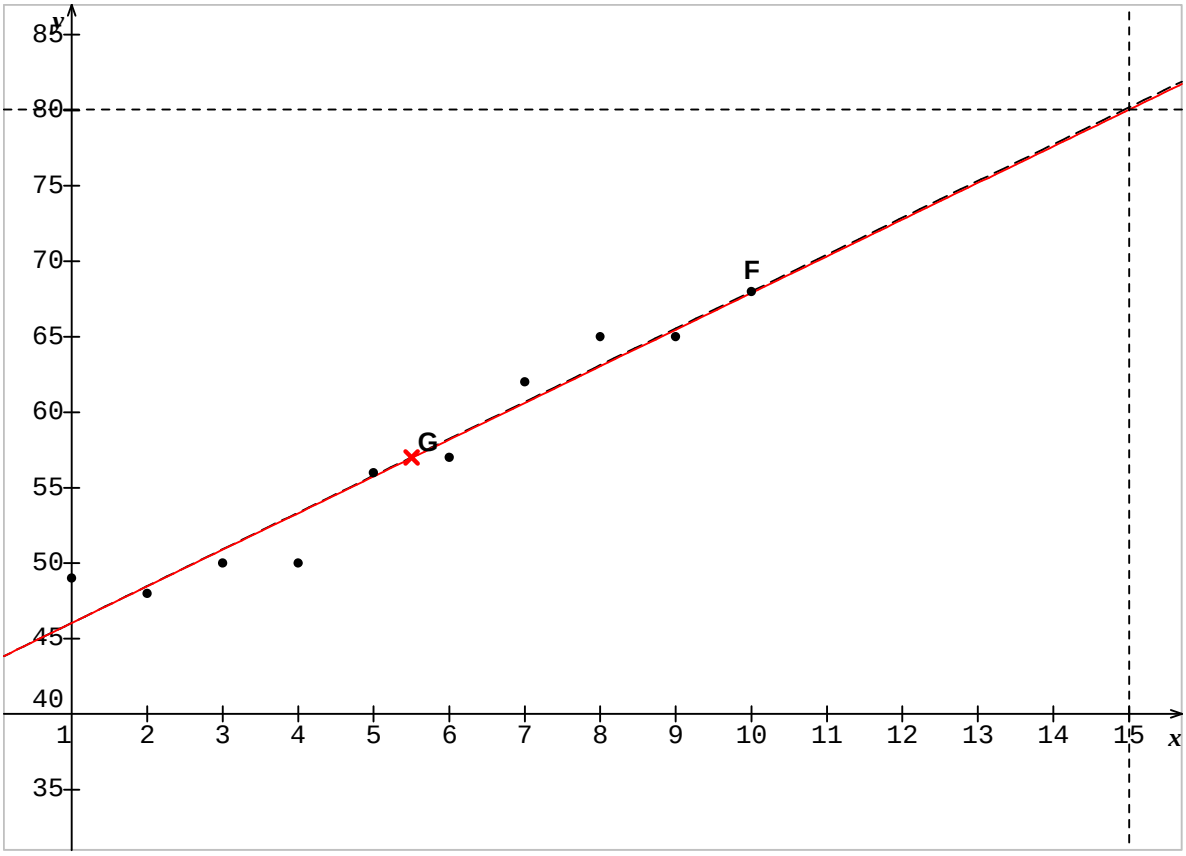
(D) $y \approx 48x + 2045$

(Pour la calculette CASIO, nous entrons les données dans le module STAT puis on choisit REG et enfin F1). Vérifions que G appartient à la droite (D) : 48(16,4) + 2045 = 2832, il y a une différence de 6 habitants car nous avons pris une valeur approchée pour a et b. en fait, si nous prenons 47,603 pour a et 2045,296 pour b alors 47,603(16,4) + 2045,296 = 2825,985 donc en fait 2826.

Nous pouvons alors effectuer une prévision pour 2020 c'est-à-dire x = 40 (2020 – 1980), cela donne une idée du nombre d'habitants pour l'avenir. Y ≈ 48(40) + 2045 ≈ **3965 personnes**.

Exercice 2

Nous représentons les données dans un repère du plan (P).



Pour les axes, nous pouvons prendre 1 comme origine des abscisses et 40 pour origine des ordonnées, Nous plaçons les données et nous remarquons que les points sont assez alignés et donc un ajustement affine se justifie parfaitement. Calculons les coordonnées du point G : La moyenne des x_i est 5,5 et celle des y_i est 57, donc **G(5,5 ; 57)**.

Appelons F, le dernier point F(10 ; 68) et traçons la droite (GF) qui sera la droite (D1). Cherchons l'équation cartésienne de (D1) : elle est de la forme $y=a_1x+b_1$.

$a_1 = \frac{68-57}{10-5,5} = \frac{11}{4,5} \approx 2,44$. Pour b_1, utilisons F, 68 =2,44(10) + b_1 et donc $b_1 \approx 43,6$.

La droite (D1) aura pour équation : **y ≈ 2,44x + 43,6.**

Si nous entrons les données dans la calculette (puis calc ; F2 ; REG F3 et enfin x F1), nous avons : LinearReg (ajustement affine) ; a ≈ 2,436 ; b ≈ 43,6 et r ≈ 0,97.

L'équation de (D2) est donc : **y ≈ 2,436x + 43,6.** Notons que les deux équations se ressemblent et (D1) et (D2) se confondent pratiquement sur le graphique (Tracé rouge et tracé noir en pointillés).
Faisons les prévisions pour x = 15 :
Avec (D1), y ≈2,44(15) + 43,6 ≈ **80,2** et avec (D2), y ≈ 2,436(15) + 43,6 ≈ **80,14.**
Si nous prenons (D1) à la place de (D2) alors l'erreur commise en % est :

$\frac{80,2-80,14}{80,14}$ ≈ **0,07%** (7,4 10^E-04 sur la calculette)

Remarque : conformément aux données du problème, si nous arrondissons à l'unité alors la réponse devient **80** et les **deux droites donnent la même prévision**.

Exercice 3

A B C

x	y	$ax_i + b$
20	50	48,6
30	68	69,2
50	108	110,5
70	150	151,8
80	175	172,4
100	220	213,6
120	250	254,9

(série 1) (série 2)

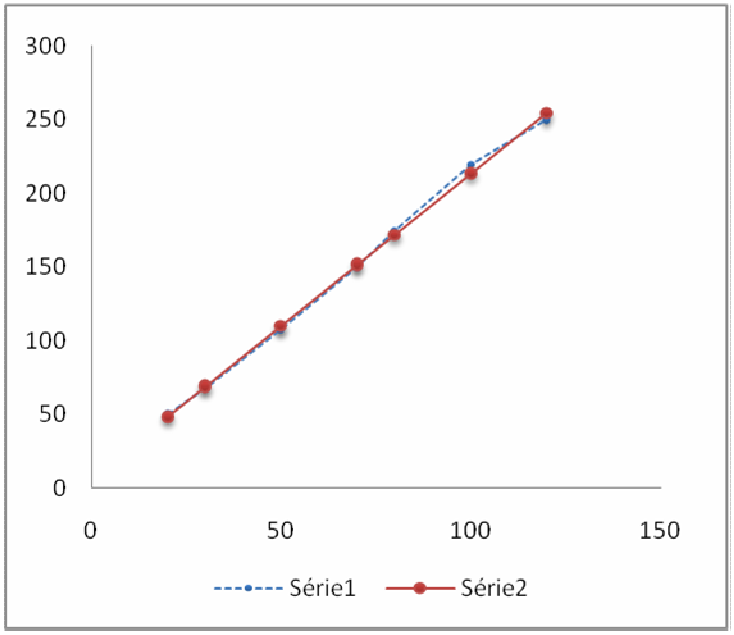
a= 2,0631295
b= 7,33273381

(Calcul par Excel des coefficients a et b de (D))

a est calculé avec :" =DROITEREG(B2:B8;A2:A8)"
entré dans la cellule suivant a=
b est calculé avec :"=ORDONNEE.ORIGINE(B2:B8;A2:A8)"
entré dans la cellule suivant b=

Dans la dernière colonne, nous avons calculé avec x, a et b,

La droite d'ajustement a donc pour équation ;
y ≈ 2,1 x + 7,3 (approximation au dixième) (tracé rouge sur le graphique)

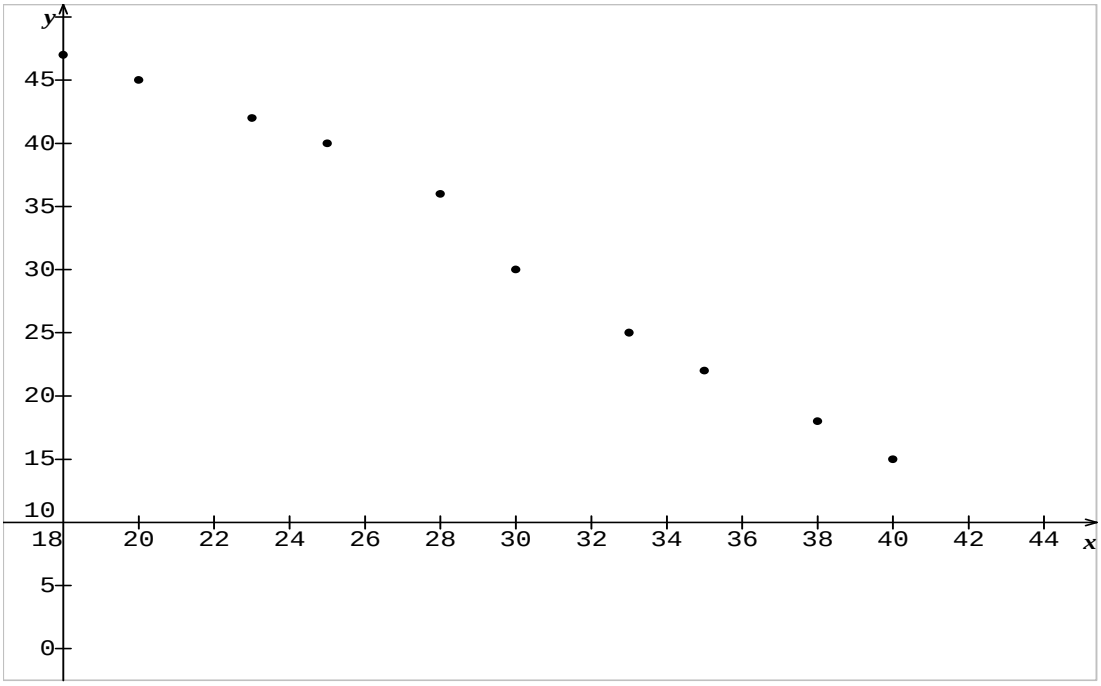


Nous avons ici un nuage ascendant et l'ajustement par une droite est valable

Exercice 4

1-a Pour faire le graphique, nous pouvons prendre comme origine (0 ; 0) mais aussi 18 pour l'axe des abscisses et 10 pour l'axe des ordonnées.

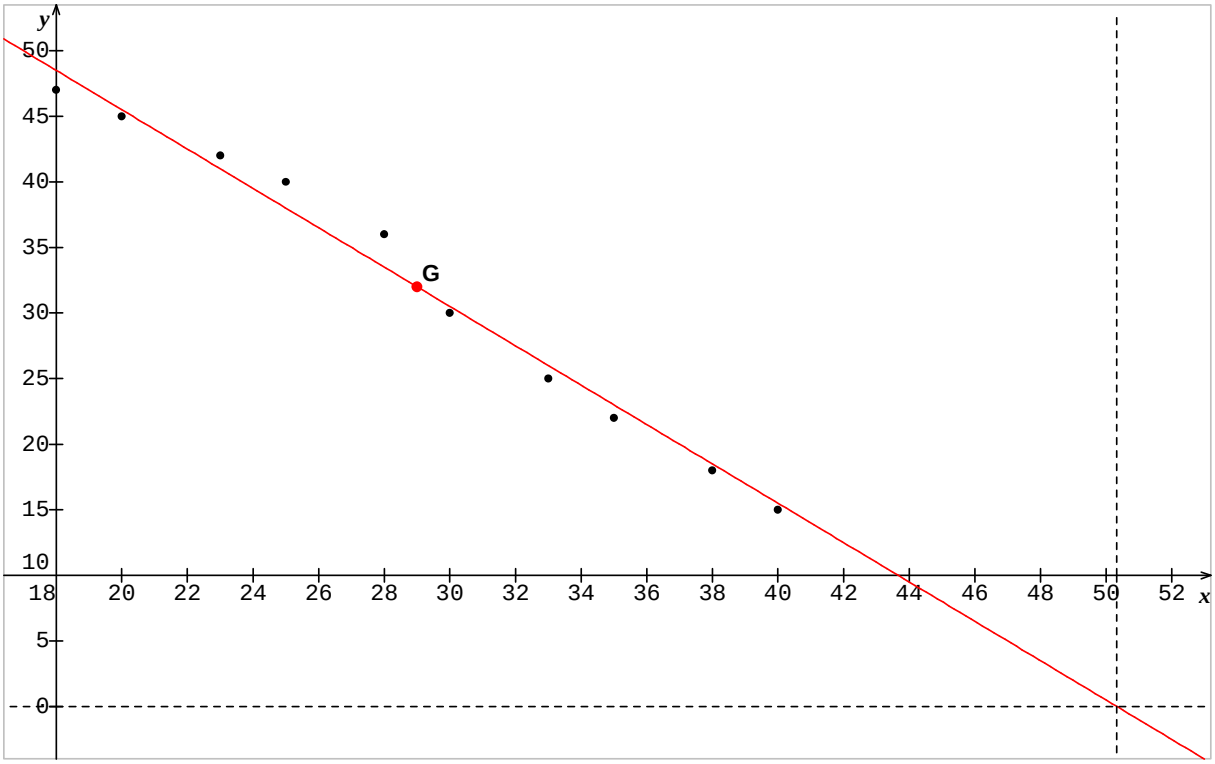
Nous allons obtenir un nuage de points descendant dans ce cas car quand le prix de la formule augmente, le nombre de personnes intéressées diminue.



1-b Oui, un ajustement par une droite se justifie car les points sont presque alignés. Nous utiliserons une fonction affine de la forme y = ax + b.

2- Calculons les coordonnées de G : **G(29 ; 32)**.

3- L'équation de la droite choisie sera de la forme y = – 1,5x + b . La droite passe par G, utilisons les coordonnées de G pour calculer b. 32 = – 1,5(29) + b donc b = **75,5**. L'équation de la droite sera donc : **y = – 1,5x + 75,5**.



Le graphique nous montre, que pour avoir y = 0 (0 personne intéressée), il faut prendre **x≈50**. Voyons par le calcul en utilisant l'équation de la droite, cherchons donc x tel que y = 0 : $-1{,}5x + 75{,}5 = 0$ soit $x = \frac{75}{1{,}5}$ ≈**50,33€**.

4-Si nous prenons x_i = 25 personnes alors y_i dans la série vaut 40€, le calcul avec (D) donne : $y = -1{,}5(25)+75{,}5 = 38$€. L'erreur en pourcentage sera $\frac{38-40}{40}$ =− 0,05 soit **− 5%.**