Chapitre 2

Statistiques à deux variables

Sommaire

2.1 Activité

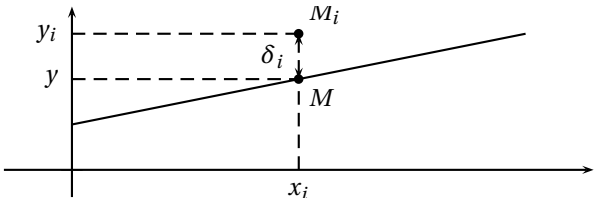
ACTIVITÉ 2.1 (Délit d'initié (d'après Yallouz Arie))**.**
La petite histoire racontée ci-dessous n'a bien sûr rien à voir avec des faits réels.
Un financier assure lors d'un procès intenté contre lui sous le motif de « délit d'initié », avoir basé ses achats d'actions d'une grande entreprise sur les données publiques résumées dans le tableau ci-dessous, donnant pour onze trimestres consécutifs un indicateur des ventes et un indicateur de la valeur de ses actions en bourse.

Trimestre i	1	2	3	4	5	6	7	8	9	10	11
Vente	100	95	67	80	87	55	70	90	82	50	90
Cours	85	77	74	62	67	69	56	63	71	68	54

1. Représenter graphiquement l'évolution du cours des actions en fonction de l'indicateur des ventes.
   Cette représentation graphique permet-elle d'anticiper le cours de l'action ?
2. Au cours du procès l'avocat du financier présente un graphique où l'on regarde comment évolue l'indicateur du cours boursier en fonction de l'indicateur de vente du trimestre *précédent*.
   (a) Représenter graphiquement l'évolution du cours des actions en fonction de la vente au trimestre précédent.
   (b) Quelle est la forme du nuage de points obtenu ?
   (c) On appelle A le point ayant la plus petite abscisse parmi les points du nuage et B celui ayant la plus grande abscisse.
   - Tracer la droite (AB).
   - Passe-t-elle suffisamment proche du nuage pour pouvoir prédire les évolutions du cours des actions ?
   - Donner une équation de la droite (AB) de la forme $y = mx + p$ (*on arrondira les coefficients au centième*).

   (d) N_1 désigne le nuage correspondant aux cinq premiers points et N_2 celui correspondant aux cinq derniers points.
   - Calculer les coordonnées du point moyen G_1 du nuage N_1 et celles du point moyen G_2 du nuage N_2.
   - Tracer la droite (G_1G_2).
   - Passe-t-elle suffisamment proche du nuage pour pouvoir prédire les évolutions du cours des actions ?
   - Donner une équation de la droite (G_1G_2) de la forme $y = ax + b$(*on arrondira les coefficients au centième*).
   - Vérifier que cette droite passe par le point moyen G du nuage de points.
3. Le financier assure avoir basé sa prévision du cours de bourse à l'aide d'une droite Δ d'équation : $y = 0,44x+31,86$.
   L'objet de cette question est de comparer les trois types d'ajustement.

(a) On ajuste les points (M_i) d'un nuage par une droite $\mathscr{D}$ d'équation $y = ax + b$, on note δ_i l'écart entre le point $M_i\left(x_i\,;\,y_i\right)$ du nuage et le point M de même abscisse x_i appartenant à la droite Δ.
Ainsi $\delta_i = y_i - y = y_i - (ax_i + b)$ (voir figure ci-contre).
Compléter le tableau 2.1 de la présente page.



(b) La somme $\sum_{i=1}^{N}\left[y_i - (ax_i + b)\right]^2$ est appelée somme des résidus quadratiques en y. Comparer les deux sommes.
Que remarquez-vous ?

TABLE 2.1 – Comparaison des ajustements

x_i (vente au trimestre i)	y_i (cours au trimestre $i+1$)	Avec la droite (AB) $\left[y_i - (mx_i + p)\right]^2$	Avec la droite (G_1G_2) $\left[y_i - (ax_i + b)\right]^2$	Avec la droite Δ $\left[y_i - (0{,}44x_i + 31{,}86)\right]^2$
100	77			
95	74			
67	62			
80	67			
87	69			
55	56			
70	63			
90	71			
82	68			
50	54			
Somme				

2.2 Bilan et compléments

2.2.1 Nuage de points, point moyen

On suppose que, suite à une étude, on s'intéresse à deux variables numériques discrètes sur une population. À chaque individu de cette population on associe ainsi un couple $\left(x_i\,;\,y_i\right)$, où x_i est la valeur de la première variable et y_i la valeur de la seconde.
L'ensemble des couples forme une série satistique double à deux variables, notée simplement $\left(x_i\,;\,y_i\right)$.
Si la première variable est le temps, on parle de série chronologique.

Définition 2.1. Soit $\left(x_i\,;\,y_i\right)$ une série satistique à deux variables.
- L'ensemble des points $M_i\left(x_i\,;\,y_i\right)$ est appelé le *nuage de points* de la série.
- Le *point moyen* de ce nuage est le point $G\left(\overline{x}\,;\,\overline{y}\right)$ où $\overline{x}$ est la moyenne des x_i et $\overline{y}$ la moyenne des y_i.
- On appelle *droite de régression*, ou ajustement affine, toute droite conçue pour passer *au plus près* des points du nuage.

Remarque. Un ajustement affine n'a de sens que si les points sont presque alignés (si le nuage a une forme allongée régulière). Il existe d'autres types d'ajustements quand le nuage de point n'a pas cette allure et nous en verrons quelques uns en exercice.

2.2.2 Droite de régression par la méthode des moindres carrés

La distance d'un point $M_i\left(x_i\,;\,y_i\right)$ à une droite d'équation $y = ax + b$ étant celle entre le point M_i et le point de la droite d'abscisse x_i et S la somme des carrés de ces distances, on admet qu'il existe une droite pour laquelle S est minimale :

Propriété 2.1. *La* droite de régression *de y en x par la méthode des moindres carrés est la droite :*
- *passant par le point moyen $G\left(\overline{x}\,;\,\overline{y}\right)$;*
- *de coefficient directeur : $a = \dfrac{cov(x\,;\,y)}{V(x)}$ où :*
  - $V(x) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$ *est la* variance *des x_i*
  - $cov(x\,;\,y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$ *est la* covariance *de x et y.*

L'équation réduite de cette droite de régression est alors $y = a(x - \overline{x}) + \overline{y}$.

On l'admettra.

Remarques. • La calculatrice permet d'obtenir directement les coefficients de la droite de régression de y en x par la méthode des moindres carrés, ce qui évite les longs calculs (voir le tableau 2.2 de la présente page).
• La droite de régression de y en x n'est pas la même que la droite de régression de x en y. La première minimise la somme des carrés des distances « verticales », la seconde la somme des carrés des distances « horizontales ».

TABLE 2.2 – Utilisation de la calculatrice

On peut retrouver tous ces paramètres statistiques en utilisant les listes d'une calculatrice.

<table>
<tr><th></th><th>TI-82</th><th>Casio Graph 25</th></tr>
<tr><td>Effacer les anciennes données</td><td>STAT
4 : ClrList 4
2nd L1
,
2nd L2
ENTER</td><td>Sélectionner le menu STAT
F6
DEL-A F4
YES F1
▷
DEL-A F4
YES F1</td></tr>
<tr><td>Entrer les nouvelles données.
On entre les valeurs des x_i dans la première colonne (L1 ou list 1) et les valeurs des y_i dans la deuxième colonne (L2 ou list 2) ;</td><td>STAT
1 : Edit ENTER
A l'écran :<table><tr><th>L1</th><th>L2</th><th>L3</th></tr><tr><td>30</td><td>12</td><td></td></tr><tr><td>40</td><td>19</td><td></td></tr><tr><td>50</td><td>24</td><td></td></tr><tr><td>60</td><td>30</td><td></td></tr><tr><td>…</td><td>…</td><td></td></tr></table></td><td>A l'écran<table><tr><th></th><th>List 1</th><th>List 2</th><th>List 3</th></tr><tr><td>1</td><td>30</td><td>12</td><td></td></tr><tr><td>2</td><td>40</td><td>19</td><td></td></tr><tr><td>3</td><td>50</td><td>24</td><td></td></tr><tr><td>4</td><td>60</td><td>30</td><td></td></tr><tr><td>…</td><td>…</td><td>…</td><td></td></tr></table></td></tr>
<tr><td>Calculer les paramètres statistiques</td><td>CALC ▷
4 : Linreg $(ax+b)$ ENTER
A l'écran :
Linreg $(ax+b)$
$y = ax+b$
$a =$
$b =$
$r =$</td><td>CALC F2
REG F3
X F1
A l'écran :
Linreg
$a =$
$b =$
$r =$
$r^2 =$
$y = ax+b$</td></tr>
</table>

2.3 Exercices

Tous les résultats seront arrondis à 10^{-3}.

EXERCICE 2.1.
Le tableau suivant recense, par clinique, le nombre de postes de personnel non médical en fonction du nombre de lits de la clinique :

Clinique	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
Nombre de lits X	122	177	77	135	109	88	185	128	120	146	100
Nombre de postes Y	205	249	114	178	127	122	242	170	164	188	172

1. Construire le nuage de points $M_i\left(x_i\,;\,y_i\right)$ correspondant à cette série (*unités graphiques : en abscisse 1 cm pour 10 lits ; en ordonnée 1 cm pour 20 postes*). D'après l'allure du nuage, un ajustement affine est-il justifié ?
2. Calculer les coordonnées du point moyen G du nuage et le placer sur le graphique.
3. (a) Sans utiliser les fonctions statistiques de votre calculatrice, mais en vous aidant éventuellement du tableau 2.3 page suivante ou d'un tableur :

TABLE 2.3 – Tableau pour la question3a

	x_i	y_i	$X_i = x_i - \overline{x}$	$Y_i = y_i - \overline{y}$	$Y_i X_i$	X_i^2
	122	205				
	177	249				
	77	114				
	135	178				
	109	127				
	88	122				
	185	242				
	128	170				
	120	164				
	146	188				
	100	172				
Somme						
Moyenne						

i. Déterminer $V(x) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$

ii. Déterminer $\text{cov}(x\,;y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$

iii. En déduire le coefficient directeur de la droite de régression $\mathscr{D}$ de y en x par la méthode des moindres carrés puis son équation. Tracer $\mathscr{D}$ sur le graphique.

(b) On dit que la droite de régression obtenue par la méthode des moindres carrés est très sensible aux valeurs extrêmes (ou aberrantes). Faites une expérience pour illustrer cette affirmation en changeant l'une des données (*on pourra pour cette question utiliser les fonctions statistiques de la calculatrice*).

4. Une clinique possède 25 lits. En utilisant les résultats de la question 3a, à combien peut-on estimer, par le calcul, le nombre de postes de personnel non médical ? Illustrer votre réponse sur le graphique.

EXERCICE 2.2.

La tableau suivant donne le PNB (en € par habitant) ainsi que le nombre d'hôpitaux (pour 1 million d'habitants) dans quelques pays européens :

Pays	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
X = PNB (en € par habitant)	5 100	7 800	11 200	15 800	20 100	22 500	26 200	28 900
Y = nombre d'hôpitaux par million d'habitants	620	1 080	1 550	2 100	3 000	3 250	3 800	4 200

1. Représenter le nuage de points associé à la série statistique $(X\,;Y)$ (*unités graphiques : en abscisse 1 cm pour 1 000 € ; en ordonnée 1 cm pour 200 hôpitaux ; on prendra pour origine le point (5 000 ; 600)*).
D'après l'allure du nuage un ajustement affine est-il justifié ?
2. Déterminer les coordonnées du point moyen G de ce nuage de points. Placer G sur le graphique.
3. *Droite de* MAYER
Dans cette question, on considère deux sous-nuages : celui constitué des points correspondants aux pays P_1, P_2, P_3 et P_4 et celui constitué des points P_5, P_6, P_7 et P_8.
   (a) Calculer les coordonnées des points moyens G_1 et G_2 des deux sous-nuages. Placer G_1 et G_2 sur le graphique.
   (b) Démontrer qu'une équation de la droite (G_1G_2), où les coefficients sont arrondis à 10^{-2}, est : $y = 0,15x - 199$. La représenter sur le graphique.
   (c) Compléter le tableau suivant :

X	5 100	7 800	11 200	15 800	20 100	22 500	26 200	28 900
Y	620	1 080	1 550	2 100	3 000	3 250	3 800	4 200
$0,15X - 199$								
$Y - (0,15X - 199)$								
$[Y - (0,15X - 199)]^2$								

   En déduire la somme des résidus quadratiques S associée à la droite de MAYER.
4. *Par les moindres carrés*
Déterminer une équation de la droite de régression $\mathscr{D}$ de y en x par la méthode des moindres carrés. La représenter sur le graphique.

5. La somme des résidus quadratiques S' associée à $\mathscr{D}$ est $S' \approx 35\,482,50$. Laquelle des deux droites réalise-t-elle le meilleur ajustement affine ?
6. *Estimations*
   À l'aide de l'équation de $\mathscr{D}$ et en détaillant les calculs répondre aux questions suivantes :
   (a) Un pays a un PNB de 23 400 € par habitant. Quelle estimation peut-on faire du nombre d'hôpitaux par million d'habitants dans ce pays ? (*Arrondir à l'unité*)
   (b) Un pays a 3 500 hôpitaux par million d'habitants. À combien peut-on estimer son PNB en € par habitant ? (*Arrondir à l'unité*)

EXERCICE 2.3.
Un hypermarché dispose de 20 caisses. Le tableau ci-dessous donne le temps moyen d'attente à une caisse en fonction du nombre de caisses ouvertes :

Nombre de caisses ouvertes X	3	4	5	6	8	10	12
Temps moyen d'attente (en minutes) y	16	12	9,6	7,9	6	4,7	4

1. Construire le nuage de points $M_i\left(x_i\,;\,y_i\right)$ correspondant à cette série statistique. (*Unités graphiques : en abscisse 1 cm pour une caisse ouverte ; en ordonnée 1 cm pour une minute d'attente*).
2. Calculer les coordonnées du point moyen G du nuage et le placer sur le graphique.
3. *Un ajustement affine*
   (a) Déterminer l'équation de la droite de régression $\mathscr{D}$ de y en x par la méthode des moindres carrés. La représenter sur le graphique.
   (b) Estimer, à l'aide d'un calcul utilisant l'équation de $\mathscr{D}$:
      i. le nombre de caisses à ouvrir pour que le temps moyen d'attente à une caisse soit de 5 minutes ;
      ii. le temps moyen d'attente à la caisse lorsque 15 caisses sont ouvertes.
      Pensez-vous dans ce cas que l'ajustement affine soit fiable ?
4. *Un ajustement non affine*
   On considère la fonction f définie sur $]0\,;+\infty[$ par : $f(X) = \dfrac{\lambda}{X}$ et $\mathscr{C}$ sa représentation graphique.
   (a) Déterminer λ de façon à avoir : $f(3) = 16$.
   (b) Tracer alors $\mathscr{C}$ dans le repère utilisé pour le nuage.
   (c) Estimer à l'aide d'un calcul utilisant la fonction f :
      i. le nombre de caisses à ouvrir pour que le temps moyen d'attente à une caisse soit de 5 min ;
      ii. le temps moyen d'attente à la caisse lorsque 15 caisses sont ouvertes.

EXERCICE 2.4.
Lors d'une période de sécheresse, un agriculteur relève la quantité totale d'eau en m^3 utilisée par son exploitation depuis le premier jour et donne les résultats suivants :

Nombre de jours écoulés : x_i	1	3	5	8	10
Volume utilisé en m^3 : y_i	2,25	4,3	8	17,5	27

1. Représenter la série statistique $\left(x_i\,;\,y_i\right)$.
   (*Unités graphiques : abscisse 1 cm pour un jour ; ordonnée 0,5 cm pour un m^3*).
2. Donnez l'équation de la droite Δ des moindres carrés sous la forme $y = mx + p$ où m et p sont les coefficients arrondis à 10^{-2}. La représenter sur le graphique.
3. Le nuage de points permet d'envisager un ajustement par une parabole $\mathscr{P}$ qui passe par les points $A(1\,;2,25)$ et $B(10\,;27)$, et qui a pour équation $y = ax^2 + b$ où a et b sont des réels.
   Déterminer a et b et donnez l'équation de la parabole $\mathscr{P}$. La représenter sur le graphique.
4. Dans cette question, on compare les deux ajustements à l'aide du tableau suivant :

x_i	1	3	5	8	10
y_i	2,25	4,3	8	17,5	27
$\lvert y_i - (mx_i + p)\rvert$	2,54	0,91	2,71		
$\lvert y_i - (ax_i^2 + b)\rvert$	0	0,05	0,25		

   Les sommes des deux dernières lignes évaluent, pour chaque ajustement, la somme des écarts entre les ordonnées des points du nuage et les ordonnées des points de même abscisse de l'ajustement.
   Donnez les arrondis à 10^{-1} des deux totaux.
   Déduisez l'ajustement qui paraît le mieux adapté.

EXERCICE 2.5.
Une étude fictive faite en France sur le taux d'équipement des ménages en automobile et l'âge des femmes lors de leur premier mariage donne les résultats suivants :

années	1 979	1 981	1 984	1 986	1 990	1 991	1 992	1 979	1 993	1 994	1 995
taux x_i	68,6	70	72,9	73,4	74,6	76,5	76,8	77	78	79,5	79
âge y_i	22,9	23,1	23,9	24,5	25	25,6	25,8	26,1	26,4	26,7	26,9

1. Représenter la série statistique.
   (Unités graphiques : abscisse 1 cm pour 1 % origine à 66 ; ordonnée 1 cm pour 1 an origine à 22)
   L'allure du nuage semble-t-il justifier un ajustement affine ?
2. Calculer les coordonnées du point moyen G du nuage puis déterminer l'équation réduite de la droite de regression de y en x par la méthode des moindres carrés.
3. Suivant cet ajustement affine, quel serait l'âge au premier mariage pour un taux de 90 % ? Ce calcul a-t-il un sens ?
4. Peut-on en déduire qu'il y a un lien entre le taux d'équipement des ménages en automobile et l'âge du premier mariage ? Comment expliquer la corrélation entre ces deux grandeurs ?

EXERCICE 2.6.
Au cours d'une séance d'essais, 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 v_i (en km/h) de l'automobile, puis la distance d_i (en m) nécessaire pour arrêter le véhicule.
Pour sept expériences, on a obtenu les résultats suivants :

vitesse v_i	20	43	62	80	98	115	130
distance d'arrêt d_i	3,5	20,5	35,9	67,8	101,2	135,8	168,5

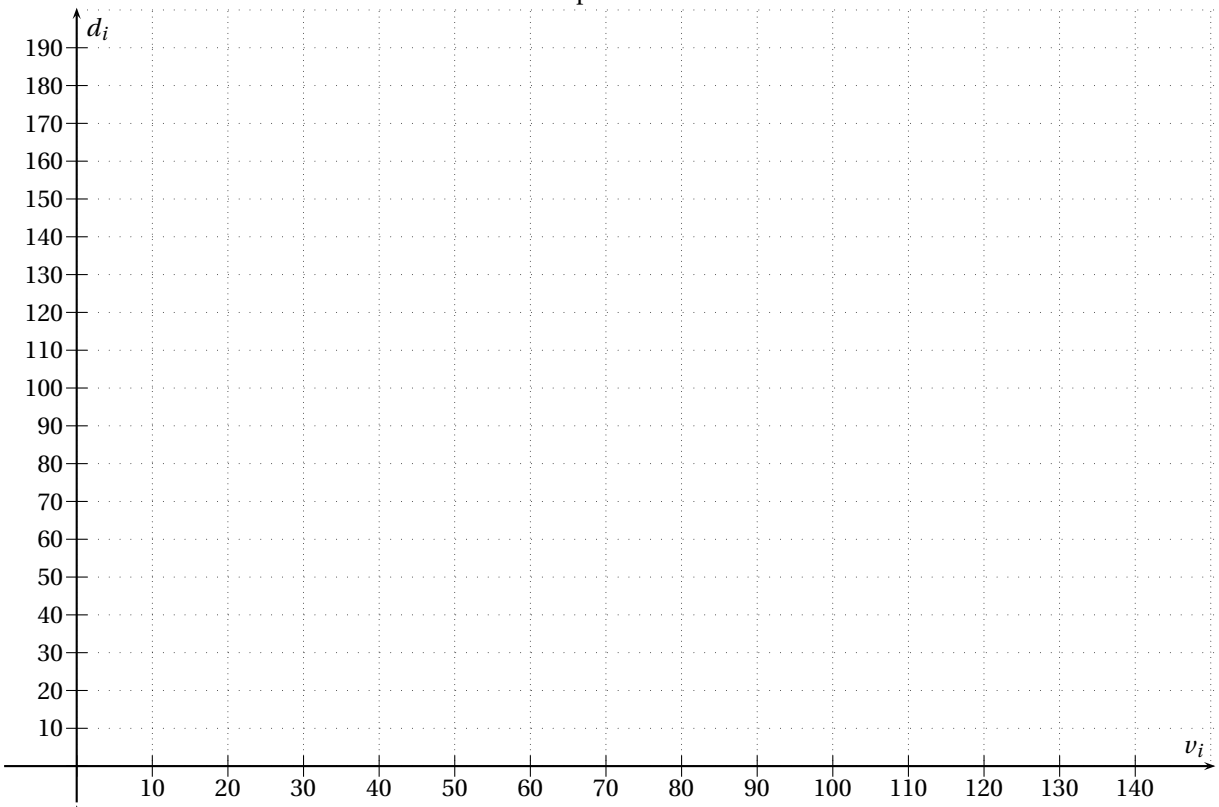
1. Dans le premier repère fourni sur la figure 2.1 page suivante, représenter le nuage de points de coordonnées $(v_i\,;d_i)$.
   Un ajustement affine semble-t-il pertinent ? Argumenter.
2. On pose $y_i = \sqrt{d_i}$.
   (a) Reproduire et compléter le tableau suivant (*on arrondira les résultats au centième*) :

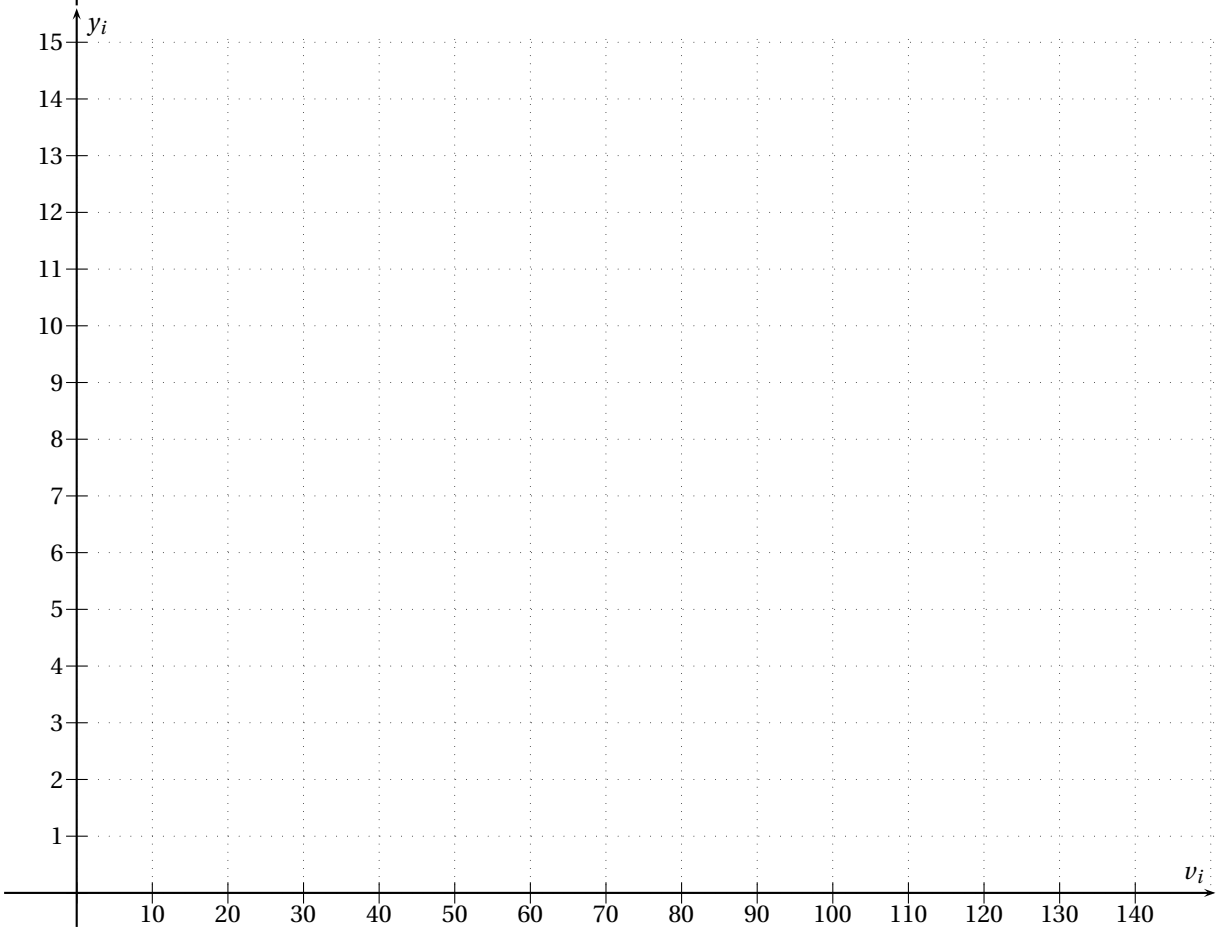
v_i	20	43	62	80	98	115	130
y_i							

   (b) Dans le second repère fourni sur la figure 2.1 page ci-contre, construire le nuage des points de coordonnées $(v_i\,;y_i)$ associé à cette nouvelle série double avec pour unités.
   La forme du nuage permet-elle d'envisager un ajustement affine ? Argumenter.
   (c) Donner l'équation de la droite de régression de y en v obtenue par la méthode des moindres carrés (*on arrondira les coefficients au millième*) et la tracer.
   (d) À l'aide de cette équation estimer (*arrondis au centième*) :
   - la vitesse v d'un véhicule lorsque sa distance d'arrêt est de 180 m ;
   - la distance d'arrêt d de ce véhicule s'il roule à 150 km/h.

   (e) Le manuel du code de la route donne, pour calculer la distance d'arrêt (en m), la méthode suivante : « Prendre le carré de la vitesse exprimée en dizaines de km/h ».
   Comparer les résultat obtenus au 2d à ceux que l'on obtiendrait avec cette méthode.

FIGURE 2.1 – Repères de l'exercice 2.6





EXERCICE 2.7.
La société MERCURE vend des machines agricoles. Suite à une restructuration en 1 998, elle a pu relancer sa production et ses bénéfices annuels ont évolué comme indiqué dans le tableau suivant :

Année	1 999	2 000	2 001	2 002	2 003	2 004
Rang de l'année x_i	0	1	2	3	4	5
Bénéfice en milliers d'euros y_i	64	75	100	113	125	127

1. (a) Représenter le nuage de points associé à la série statistique $(x_i\,;\,y_i)$ dans un repère orthogonal.
Les unités graphiques seront 2 cm pour une unité sur l'axe des abscisses et 1 cm pour 10 unités sur l'axe des ordonnées.
   (b) Calculer les coordonnées du point moyen G du nuage (*arrondir au dixième*). Placer le point G dans le repère.
2. En première approximation, on envisage de représenter le bénéfice y comme une fonction affine du rang x de l'année.
   (a) Donner une équation de la droite d'ajustement $(\mathscr{D})$ obtenue par la méthode des moindres carrés (*arrondir les coefficients au centième*).
   (b) Tracer cette droite $(\mathscr{D})$ dans le repère.
   (c) Quelle prévision ferait-on pour le bénéfice en 2 005 avec cette approximation ?
3. En observant le nuage de points, on envisage un deuxième modèle d'ajustement donné par $y = f(x)$ avec
$$f(x) = -2x^2 + 23x + 63.$$
   (a) Étudier les variations de la fonction f sur l'intervalle $[0\,;6]$.
   (b) Tracer la courbe représentative $(\mathscr{C}_f)$ de la fonction f dans le repère de la question 1).
   (c) Quelle prévision ferait-on pour le bénéfice en 2 005 avec ce deuxième modèle d'ajustement ?
4. En réalité, le bénéfice en 2 005 est en hausse de 0,9 % par rapport à celui de 2 004.
Des deux ajustements envisagés dans les questions précédentes, quel est celui qui donnait la meilleure prévision pour le bénéfice en 2005 ? Justifier la réponse.