Statistique à 2 variables

1. Exemples

Nous sommes souvent confrontés à des données entre lesquelles nous essayons d'établir des liens telles que :

- La taille et le poids d'un groupe d'individus.
- le budget vacances et les revenus des familles
- Le poids des récoltes et la durée d'ensoleillement ou la quantité de pluie reçue
- ……

Mais comment à partir de ces données, tirer des conclusions, exprimer le lien qui les unit ?
Prenons deux exemples qui nous serviront référence.

1.1 Exemple 1

Des élèves ont présenté un examen de mathématique et un examen de physique.
Les résultats sont les suivants :

Elèves N° :	1	2	3	4	5	6	7	8	9	10	11
Cote en math. (sur 20) : x_i	4	5	5	7	10	11	13	15	15	17	17
Cote en physique (sur 20): y_i	5	9	14	7	9	11	12	12	14	5	17

Au vu de ces résultats, on peut se poser différentes questions :
Les élèves forts en physique sont-ils forts en math ?
Y a-t-il un lien entre la cote de mathématique et celle de physique ? Et si oui, comment exprimer ce lien ?

1.2 Exemple 2

Un professeur de mathématiques a filmé avec une caméra numérique son fils en train de lancer un ballon. En regardant cet enregistrement avec arrêts sur images, il repère les données présentées ci-dessous.

durée (en s)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Hauteur (en cm)	84	121	149	167	175	174	163	143	114	75

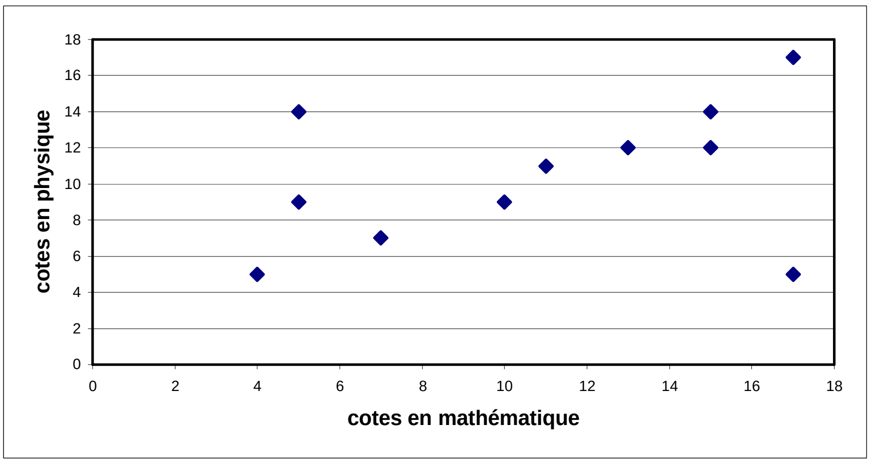
Comment, à partir de ces données déterminer une fonction h(t) qui exprime la hauteur du ballon en fonction du temps ?

2. Représentations graphiques

Pour se faire une meilleure idée des problèmes, représentons les données sur des graphiques.

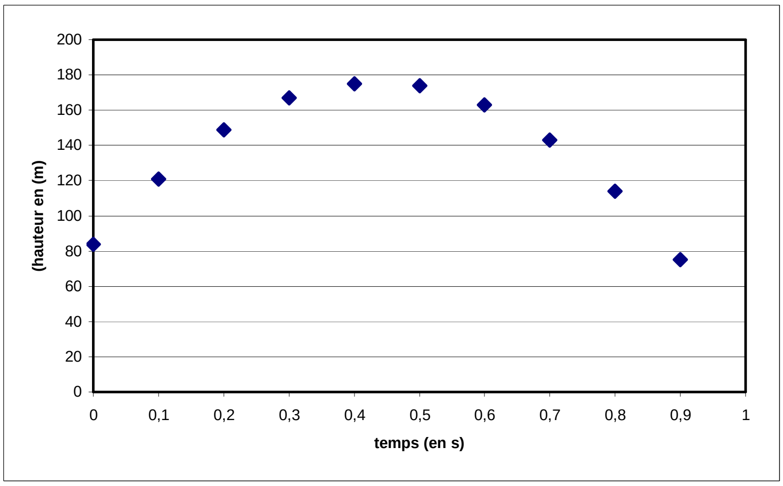
2.1 Exemple 1

Pour le premier exemple, nous obtenons le graphique suivant (nuage de points)



2.2 Exemple 2

En procédant comme pour le premier exemple, nous obtenons :



2.3 Observations :

En regardant ces graphiques, nous constatons que le type de fonction à ajuster sera différent selon les données : pour le premier exemple, si on peut ajuster une fonction, il s'agirait plutôt d'une fonction du premier degré tandis que dans le second, on choisirait une fonction du second degré.

2.4 Généralisation

Lorsque nous avons un nuage de points (x_i, y_i), différentes situations peuvent se présenter.

a) Les points sont disposés de façon quelconque : on dira que les caractères x et y sont indépendants.

b) Les points sont disposés autour d'une certaine courbe : on pourra faire un "ajustement graphique", c'est à dire tracer au mieux cette courbe.
La courbe la plus simple que l'on puisse obtenir est une droite. Parfois il s'agira d'une parabole, d'une fonction du troisième degré, ou d'autres fonctions que nous serons amenés à étudier ultérieurement.

c) Tracer cette courbe à main levée est très arbitraire, c'est pourquoi, nous allons développer des procédures de calcul.

3. Ajustement linéaire

Dans un premier temps, nous allons considérer des situations du genre du premier exemple, où le type de fonction à ajuster est une fonction du premier degré.

3.1.1 Ajustement graphique "à vue"

On trace une droite qui nous semble la plus près possible des points du nuage. En prenant deux points de la droite, nous pouvons obtenir rapidement son équation. Si nous utilisons cette technique en classe dans le premier exemple, nous constatons que plusieurs élèves peuvent obtenir des droites et donc des équations différentes. C'est donc une méthode rapide mais très approximative.

3.1.2 Droite de Mayer

On ordonne les points de la série de manière croissante selon les xi et on divise le nuage en deux parties contenant le même nombre d'éléments (à une unité près si les données sont en nombre impair). Dans chaque sous-nuage, on calcule le point moyen. La droite qui passe par ces deux points est appelée droite de Mayer

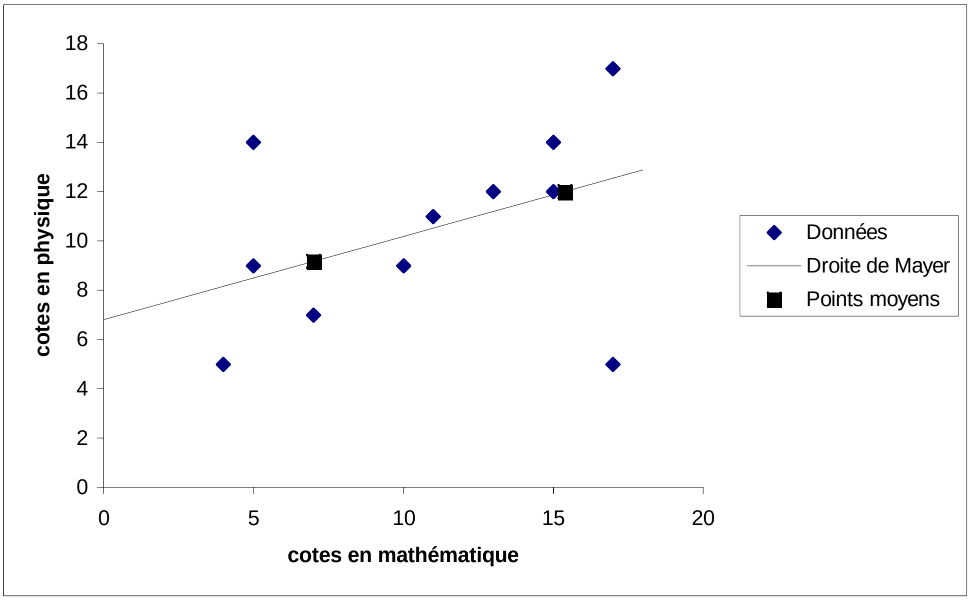
Dans notre exemple, nous obtenons

P_1 $(\overline{x}_1, \overline{y}_1) = \left(7, \frac{55}{6}\right)$: point moyen des 6 premiers points

P_2 $(\overline{x}_2, \overline{y}_2) = \left(\frac{77}{5}, 12\right)$: point moyen des 5 derniers points

La droite P_1P_2 (droite de Mayer) a pour équation y = 0.3373x + 6.8056

Le graphique ci-dessous reprend à la fois les données et la droite de Mayer

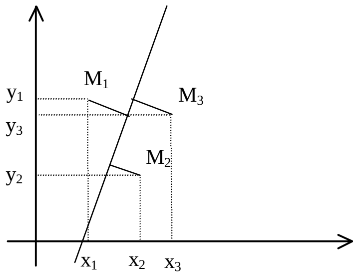


Ces 2 méthodes, si elles nous permettent de déterminer une droite qui s'approche des données ne nous permettent pas de déterminer "la meilleure droite d'approximation". Nous allons maintenant définir des critères qui vont permettre de décider quelle est cette "meilleure droite".

3.1.3 Ajustement par la méthode des moindres carrés.

Prenons une situation où nous avons n points : M_1 , M_2 , ...M_n de coordonnées respectives : (x_1 , y_1) (x_2 , y_2)...(x_n, y_n) et soit $d \equiv y = ax + b$: la droite cherchée.

(le graphique a été réalisé pour le cas de 3 points)

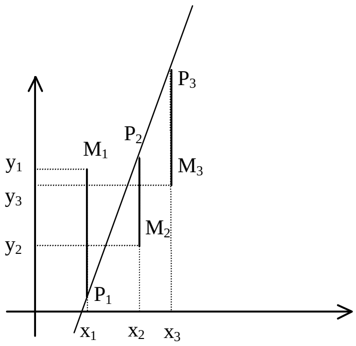


<u>1ère idée</u> : chercher a et b pour que la somme des distances des points M_i à la droite d soit la plus petite possible.

En réalisant ces calculs, on obtient une situation très compliquée. (Il faut alors minimiser $\sum_{i=1}^{n} \overline{M_i M'_i}$ où M'_i est la projection orthogonale de M_i sur la droite.

<u>2ème idée</u> : on peut aussi envisager de minimiser la somme des carrés des distances des M_i à la droite cherchée, mais là aussi cela reste très compliqué

<u>3ème idée</u> : On décompose le problème en 2 parties plus simples
Considérons d'abord P_i les points d'intersection des droites parallèles à OY menées par M_i
(le graphique ci-contre illustre une situation où on n'a que 3 points)
On veut que la somme des carrés des distances $\overline{M_i P_i}$ soit la plus petite possible: c'est pourquoi on appelle cette méthode la méthode des moindres carrés. La droite obtenue porte le nom de droite de régression de y en x.
Ensuite, on considérera les points d'intersection des droites parallèles à OX menées par M_i (voir graphique page suivante) et on minimisera la somme des carrés des distances $\overline{M_i Q_i}$ pour obtenir la droite de régression de x en y.
Si ces 2 droites sont proches, (confondues lorsque les points sont alignés), c'est que l'ajustement linéaire est "bon" pour la situation étudiée. Il faudra donc définir un coefficient qui mesure l'écart entre les deux droites : c'est le coefficient de corrélation linéaire.



Détermination de la droite de régression de y en x :

On veut minimiser la somme :S = $\overline{M_1P_1}^2 + \overline{M_2P_2}^2 + ... + \overline{M_nP_n}^2 = \sum_{i=1}^{n} \overline{M_iP_i}^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

Nous allons minimiser cette somme dans le cas où on a trois points. (On peut généraliser cette démonstration.)

S = $\sum_{i=1}^{3} (y_i - ax_i - b)^2 = (y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2$

$= 3b^2 - 2b(y_1 + y_2 + y_3 - ax_1 - ax_2 - ax_3) + y_1^2 + y_2^2 + y_3^2 + a^2(x_1^2 + x_2^2 + x_3^2) - 2a(x_1 y_1 + x_2 y_2 + x_3 y_3)$

En supposant a fixé, on cherche parmi toutes les valeurs de b celle qui minimise la somme. On voit qu'il s'agit d'un trinôme du second degré en b, le minimum est donc atteint pour

$$b = \frac{2(y_1 + y_2 + y_3 - ax_1 - ax_2 - ax_3)}{3 \cdot 2} = \frac{y_1 + y_2 + y_3}{3} - a\frac{x_1 + x_2 + x_3}{3} = \overline{y} - a\overline{x}$$

Nous avons ainsi exprimé b en fonction de a et de la moyenne des x_i et y_i

Et d ≡ $y = ax + \overline{y} - a\overline{x} \Leftrightarrow y - \overline{y} = a(x - \overline{x})$

Nous observons que le point moyen de coordonnée $(\overline{x}, \overline{y}) \in d$

Dans notre exemple : $\overline{x} = \frac{119}{11} = 10{,}81818$ et $\overline{y} = \frac{115}{11} = 10{,}4545$ et b = 10,4545 - 10,8181 a

Si nous déterminons la valeur de a, nous aurons immédiatement celle de b.
En reportant la valeur de b trouvée dans la somme à minimiser, nous obtenons :

$S_a = [y_1 - ax_1 - (\overline{y} - a\overline{x})]^2 + [y_2 - ax_2 - (\overline{y} - a\overline{x})]^2 + [y_3 - ax_3 - (\overline{y} - a\overline{x})]^2$

$= [(y_1 - \overline{y}) - a(x_1 - \overline{x})]^2 + [(y_2 - \overline{y}) - a(x_2 - \overline{x})]^2 + [(y_3 - \overline{y}) - a(x_3 - \overline{x})]^2$

En ordonnant cette expression selon les puissances de a, nous obtenons :

$S_a = a^2 [(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2] - 2a [(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y})] + (y_1 - \overline{y})^2 +$

$(y_2 - \overline{y})^2 + (y_3 - \overline{y})^2$: une fonction du second degré en a. Elle admet un minimum pour

$$a = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}$$

$$= \frac{\frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y})}{3}}{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{3}}$$

Le dénominateur de la fraction est la variance de la variable x et le numérateur est appelé covariance de x et y.

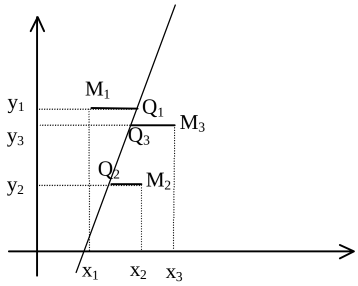
et donc $a = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Ce résultat peut se généraliser au cas de n observations.

Nous avons donc $a = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$: coefficient angulaire de la droite de régression de y en x.

Cette droite minimise la $\sum_{i=1}^{n} \overline{M_iP_i}^2$. Comme nous avons vu qu'elle passe par le point moyen $(\overline{x}, \overline{y})$ elle a donc pour équation : $y - \overline{y} = a(x - \overline{x})$

De même, la droite de régression d' de x en y minimise la $\sum_{i=1}^{n} \overline{M_i Q_i}^2$



On démontre de même que son coefficient angulaire a' = $\dfrac{\text{var}(y)}{\text{cov}(x, y)}$ et que d' passe par le point moyen ($\overline{x}$, $\overline{y}$); elle a donc pour équation :

y - $\overline{y}$ = a' (x - $\overline{x}$)

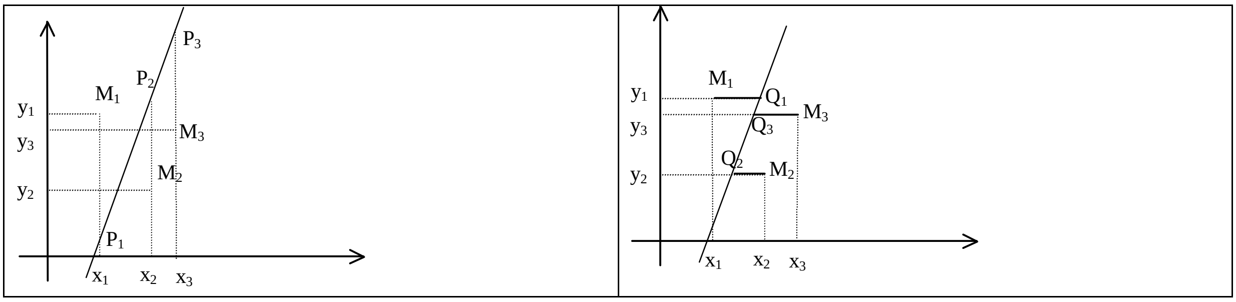
On définit $r^2 = \dfrac{a}{a'} = \dfrac{(\text{cov}(x,y))^2}{\text{var}(x)\ \text{var}(y)}$

D'où l'on déduit : le coefficient de corrélation linéaire $r = \dfrac{\text{cov}(x, y)}{\sigma(x)\ \sigma(y)}$

On peut montrer que 0 < |r| < 1
Ce coefficient permet de mesurer "l'écart" entre d et d'. Si les 2 droites sont confondues, a = a' et r = 1.
Si la dépendance linéaire est mauvaise, |r| est éloigné de 1.
On considère que si |r| < 0.87, il n'y a pas de dépendance linéaire.
Il y a une corrélation positive entre les 2 variables lorsque les variations des deux variables se produisent dans le même sens (les pentes des droites de régression sont positives)
Il y a une corrélation négative entre les 2 variables lorsque les variations des deux variables se produisent dans le sens contraire (les pentes des droites de régression sont négatives)

Dans la pratique, ces calculs sont rarement effectués. La plupart des calculatrices scientifiques actuelles permettent de déterminer les droites de régression de façon très aisée.

3.2 Synthèse sur les droites de régression



1. Covariance

   $$\text{cov}(x, y) = \frac{1}{n}\sum_{1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

2. La droite de régression de y en x : d ≡ y = ax + b minimise la $\sum_{i=1}^{n} \overline{M_i P_i}^2$

   d ∋ le point moyen : ($\overline{x}, \overline{y}$) et a = $\dfrac{\text{cov}(x, y)}{\text{var}(x)} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ ⇒ d ≡ y - $\overline{y}$ = a (x - $\overline{x}$)

3. La droite de régression de x en y : d' ≡ y = a'x + b' minimise la $\sum_{i=1}^{n} \overline{M_i Q_i}^2$

   d' ∋ le point moyen : ($\overline{x}, \overline{y}$) et a' = $\dfrac{\text{var}(y)}{\text{cov}(x, y)} \quad \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}$ ⇒ d' ≡ y - $\overline{y}$ = a' (x - $\overline{x}$)

4. Le coefficient de corrélation r

$r^2 = \frac{a}{a'} \Rightarrow r = \frac{\operatorname{cov}(x,y)}{s(x)\,s(y)}$ $(0 < |r| < 1)$

Si $|r| < 0.87$: la dépendance linéaire est mauvaise.
Si $0.87 < |r| < 1$ l'ajustement linéaire est bon et une des droites de régression peut être prise comme ajustement.

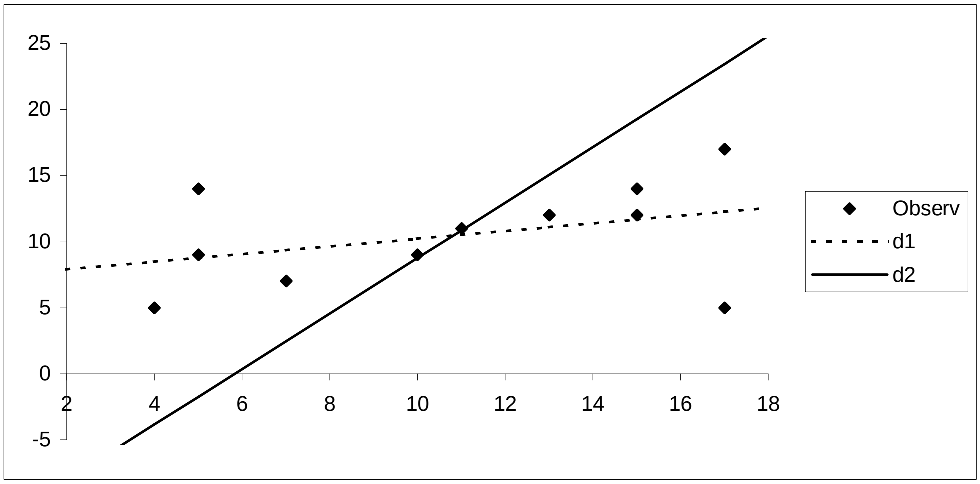
3.3 Application

Revenons maintenant à l'exemple 1 proposé plus haut. Nous allons calculer les droites de régression pour cet exemple et le coefficient de corrélation linéaire.

x_i (cotes en math.)	y_i (cote en phys.)
4	5
5	9
5	14
7	7
10	9
11	11
13	12
15	12
15	14
17	5
17	17

La calculatrice nous fournit directement les résultats :
$\overline{x} = 10.8181$ $\overline{y} = 10.4545$
ainsi que : a = 0,28867 b = 7,3316 et r = 0.37098

Nous avons ainsi la droite de régression de y en x :
$d_1 \equiv y = 0.28\,x + 7.33$
Pour déterminer la droite de régression de x en y, nous nous servirons de la relation :
$r^2 = \frac{a}{a'} \Leftrightarrow a' = \frac{a}{r^2}$ qui nous donne : $a' = \frac{0.28867}{0.37098^2} = 2.09749$
Or cette droite comprend le point moyen $(\overline{x}, \overline{y})$
Et on a l'équation de $d_2 \equiv y - 10.45 = 2.1\,(x - 10{,}8) \Leftrightarrow y = 2.1\,x - 12.23$
Le graphique ci-dessous reprend les points (x_i, y_i) ainsi que les droites d_1 et d_2.

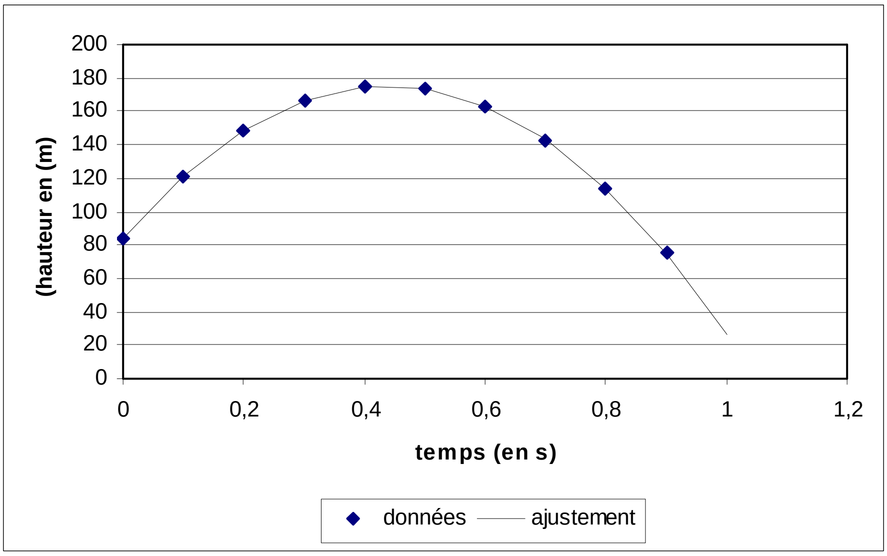


Le coefficient de corrélation r = 0.37 étant éloigné de 1, le décalage entre ces deux droites est important : on considère donc qu'il n'y a pas de dépendance linéaire.

4. Ajustement quadratique.

Dans le second exemple proposé, nous voulons ajuster une courbe du second degré.
Nous pouvons également procéder par tâtonnement : par exemple en résolvant un système de 3 équations à 3 inconnues (en prenant 3 des points des données) qui se ramène rapidement à un système de deux équations à 2 inconnues puisque l'on connaît l'ordonnée à l'origine. Comme dans le cas d'un ajustement linéaire, selon les points choisis, nous aurons plusieurs paraboles possibles.
Dans le cas qui nous occupe, on peut se référer aux formules de cinématique vues au cours de physique.
A nouveau, comme dans le point précédent, il va falloir choisir parmi toutes les possibilités la "meilleure courbe". Cependant, nous ne développerons pas ici les méthodes mathématiques qui permettent de choisir celle-ci, nous nous contenterons d'utiliser notre calculatrice.

Dans l'exemple proposé, nous obtenons : $h(t) = -475\ t^2 + 417.015t + 84.218$ qui est comme nous pouvons le vérifier également assez proche des valeurs trouvées par tâtonnements.
Graphiquement, nous obtenons :



5. Exercices.

5.1 Exercice 1

Une firme productrice de véhicules spéciaux entreprend une étude statistique de ses coûts de production. Une collecte de données est résumée dans le tableau suivant :

Représenter graphiquement ces données déterminer les droites de régression associées à celles-ci.

Nombre d'unités produites : x_i	Coût global de production. : y_i (en 10^3 €.)
1000	350
2000	500
3000	575
4000	75
5000	925
6000	1025
7000	1175
8000	1275
9000	1350
10000	1500
11000	1575

5.2 Exercice 2

Le tableau ci-contre donne quelques chiffres sur le tourisme en Europe :

Pays	Nombre total de touristes arrivant (en millions.)	Recette totale (en 10^6 €)
Allemagne	4,9	450
Espagne	4,1	70
France	5,5	400
Italie	8,6	500
Suisse	4,6	250

On demande
1° de construire le nuage de points et de dire si un ajustement linéaire paraît vraisemblable.
2° d'établir les équations des droites de régression de y en x et de x en y
3° de dessiner ces droites
4° de calculer le coefficient de corrélation et d'indiquer ce que signifie ici ce coefficient.

5.3 Exercice 3

La direction commerciale d'une entreprise industrielle a augmenté régulièrement ses dépenses publicitaires pendant plusieurs années.
a) A partir du tableau ci-contre, comparer la progression du chiffre d'affaires avec les dépenses en déterminant les droites de régression et le coefficient de corrélation.
b) Tracer le nuage de points correspondant et les droites de régression.
c) Estimez les dépenses publicitaires à consentir pour atteindre 1 000 000€ comme chiffre d'affaire. Cette estimation est-elle fiable ?
d) Peut-on estimer le chiffre d'affaires auquel on peut s'attendre si l'entreprise augmente son budget publicitaire jusque 3 000 € ? Justifier

Année	Dép. pub. (en €)	Chiffres d'affaires (en €)
1998	1830	881525
1999	1871	894275
2000	1998	919775
2001	1999	932 525
2002	2000	938 900
2003	2001	951 650
2004	2002	970 775

5.4 Exercice 4

<u>But</u> : doser les nitrites dans l'eau afin de déterminer l'indice de pollution organique de l'eau avant et après lagunage (station de Sart-Bernard) (Expériences réalisées dans les classes de biologie appliquée du collège)
On établit donc une gamme étalon en nitrates allant de 0 à 25 mg de nitrates par litre.
Les résultats des mesures étalons sont donnés dans le graphique ci-contre.
a) Donner une expression de l'absorbance en fonction de la concentration (droite de régression)
b) Exprimer la concentration en fonction de l'absorbance.
b) Si l'absorbance vaut 0.40, estimer la valeur de la concentration en nitrates.

Concentration (mg/l)	Absorbance
0	0
5	0.14
10	0.28
15	0.36
20	0.55
25	0.68

5.5 Exercice 5

Dans un ancien rapport sur la conjoncture économique, on relève le tableau ci-contre :

France	taux des salaires horaires	indice d'ensemble des prix de détail
Mars 1950	105	105.3
Juin 1950	107	103.9
Sept. 1950	114	109.2
Déc. 1950	120	113.6
Mars 1951	127	119.7
Juin 1951	138	127
Sept. 1951	156	131.6
Déc. 1951	160	141.3
Mars 1952	162	146.9
Juin 1952	163	142.7
Sept. 1952	164	147.9

a) Déterminer les droites de régression de cette série statistique.
b) Calculer son coefficient de corrélation ? Que pouvez-vous conclure ?
c) Si l'indice des prix vaut 155, peut-on prévoir le taux des salaires ? Cette prévision est-elle fiable ? Justifier.

5.6 Exercice 6

Un test permet de mesurer l'aptitude à la lecture d'enfants en fonction de leur âge.
x_i représente l'âge et y_i la durée de lecture (en secondes)

a) Déterminer les équations des droites de régression
b) L'hypothèse : plus l'âge augmente, plus la durée diminue est-elle vérifiée ? Justifier.

x_i	y_i
7	58
8	59
9	46
10	33
11	31
11	33
9	44
7	74
7	60
8	60
10	41
10	31

5.7 Exercice 7

Une société a mis au point un produit. Une enquête menée auprès de 500 personnes a montré une relation entre le prix x proposé (en €) pour ce produit et le nombre de clients disposés à l'acheter à ce prix. Les résultats de cette enquêtes sont donnés dans le tableau ci-contre.

a) Construire le nuage de points

b) Déterminer les droites de régression et les tracer sur le graphique.

c) Interpréter le coefficient de corrélation.

d) Si on proposait le produit à 30€, à combien peut-on estimer le nombre de personnes qui en achèteraient ? Cette estimation est-elle fiable ? Pourquoi ?

Prix (en €)	Nombre de clients
40	60
35	80
32	130
28	200
24	240
20	350
16	390
12	420
10	440
8	500

5.8 Exercice 8

Prenons la situation fictive où l'on mesure la productivité d'un groupe de travailleurs de huit heures à quinze heures :

Dans ce tableau, on voit la productivité croître tout au long de la journée pour diminuer sous l'influence de la fatigue en fin de journée.

On aimerait savoir si la productivité est fonction de l'heure de la journée.

Représenter graphiquement la situation et utiliser l'ajustement le plus approprié.

Heure du jour	Productivité
8	12,0
9	13,0
10	14,0
11	14,4
12	14,8
13	14,5
14	14,0
15	13,0