Etude de séries statistiques à deux variables

I. Généralités

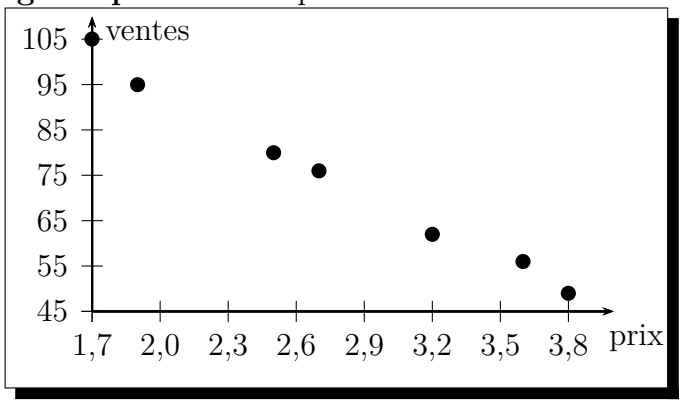
1) Nuage de points

En statistique à deux variables, on étudie deux caractères x et y sur une même population (à l'aide d'un échantillon) afin de savoir s'il existe une relation entre ceux-ci.

Exemple 1 :
Relevé du prix à l'unité de certains appareils et du nombre de ventes correspondantes, pendant une certaine période :

Prix x_i en milliers d'euros	1,7	1,9	2,5	2,7	3,2	3,6	3,8
Nombre y_i d'appareils vendus	105	95	80	76	62	56	49

On peut représenter le **nuage de points** correspondant :



Construction à la calculatrice :

Casio
Menu STAT
Entrer les valeurs xi dans List1
Entrer les valeurs yi dans List2
Choisir GRPH
Règler les paramètres avec SET
Choisir GPH1

T.I.
Touche STAT
Menu EDIT
Entrer les valeurs xi dans L1
Entrer les valeurs yi dans L2
2nd Y= (Stat Plot) pour vérifier les réglages
Règler les valeurs du repère avec la touche WINDOWS
Appuyer sur la touche GRAPH

2) Point moyen

Le **point moyen** d'un nuage est le point de coordonnées

$$x_G = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \text{ et } y_G = \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Exemple 1 :
Les coordonnées de G sont ici : $x_G = \dfrac{1,7 + 1,9 + \cdots + 3,8}{7} = \dfrac{19,4}{7} \simeq 2,77$
et $y_G = \dfrac{105 + 95 + \cdots + 49}{7} = \dfrac{523}{7} \simeq 74,71$.
G a donc pour coordonnées approximatives $(2,77\,;\,74,71)$.

II. Ajustement affine

1) Principe

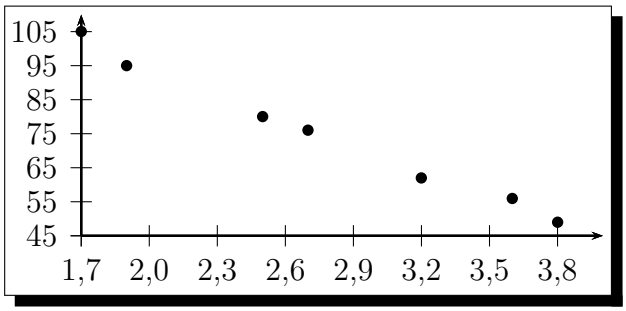
Lorsque le nuage semble suivre une certaine courbe, on cherche une fonction correspondant à cette courbe.
On dit alors qu'il existe une corrélation entre x et y (la valeur de x influe sur celle de y).
On peut alors prévoir la valeur de y correspondant à une valeur quelconque de x (ou inversement).
Si le nuage a une forme relativement rectiligne, on peut procéder à un **ajustement affine**, c'est-à-dire chercher une relation de la forme $y = ax + b$.

Remarques :
– il n'y a pas forcément de relation de cause à effet entre les variables, elles peuvent être des effets d'une autre cause (exemple : nombre de coups de soleil et de lunettes de soleil vendues) ;
– quelques exemples de corrélations ici et là.

2) Tracé à la règle

On peut simplement tracer une droite approximative, passant par le point moyen et ajustant le nuage et lire éventuellement son équation.

Exemple 1 :
Tracez une droite ajustant ce nuage.
Son coefficient directeur est environ :
$a \simeq -\dfrac{56}{2,1} \simeq -26,7.$
La droite passe par G donc $y_G = ax_G + b$ d'où :
$b = y_G - ax_G \simeq 74,71 - (-26,7) \times 2,77 \simeq 148,7$
L'équation approximative de cette droite est donc :
$y \simeq -26,7\,x + 148,7.$



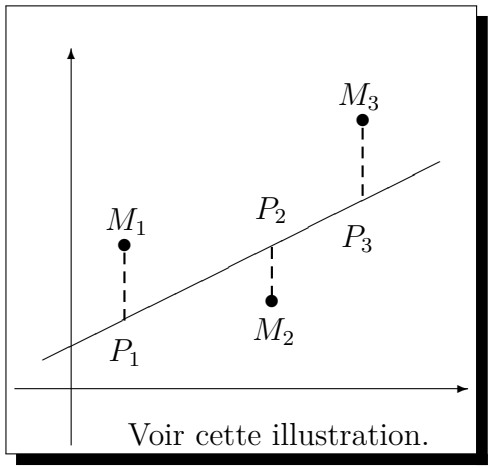
3) Méthode des moindres carrés, droites de régression

a) Droites de régression

Droite de régression de y en x :
Principe : soient $M_i(x_i ; y_i)$ les points du nuage et (d) une droite « approximant » ce nuage. Soient P_i les points de la droite (d) ayant la même abscisse que les M_i. Une façon d'avoir une bonne approximation consiste à s'assurer que la somme des carrés des distances M_iP_i soit minimale.

Théorème 1
Il existe une unique droite (d) pour laquelle $\sum_{i=1}^{N}(M_iP_i)^2$ est minimale.

Cette droite est appelée droite « des moindres carrés » ou **droite de régression** de y en x.
On écrit son équation sous la forme $y = ax + b$.
On utilise cette droite quand on considère que y dépend de x.



Droite de régression de x en y :
Cette fois-ci, on considère les points Q_i de la droite d de même ordonnée que les M_i et on cherche à minimiser $\sum_{i=1}^{N} M_iQ_i^2$. La solution est la droite de régression de x en y dont on écrira l'équation sous la forme $x = a'y + b'$
On utilise cette droite quand on considère que x dépend de y.

Détermination à la calculatrice :

Casio (menu STAT)	**T.I. (touche STAT)**
Configuration : CALC, SET, 2Var Freq=1.	Calc, Setup(*). Dans 2-Var Stats, Freq=1
Calculs ($\bar{x}$, σ, etc.) : CALC, 2VAR	Calc, 2-Var Stats L1,L2(**)
Droite de régression : CALC, REG, x	Calc, Linreg L1,L2

(*) Sur certaines TI, il n'y a pas de Setup. Voir alors (**).
(**) « L1,L2 » est optionnel sur les TI ayant un Setup. On l'obtient avec la touche OPTN ou 2nd Stats (Listes).

EXEMPLE 1 : dans l'exemple présent (vente d'appareils), la machine donne :
- la droite de régression de y en x a pour équation $y = -25,31x + 144,83$.
- la droite de régression de x en y a pour équation $x = -0,039y + 5,68$
  ce qui peut aussi s'écrire $y = -25,64x + 145,64$.

Remarque :
En général, les deux droites de régression sont distinctes, dans l'exemple présent, les équations sont similaires car le nuage a une forme de droite. Elles passent toutes deux par le point moyen G.

b) Coefficient de corrélation linéaire

Définition
Le **coefficient de corrélation linéaire** de la série double $(x\,;\,y)$ est tel que $\boxed{r^2 = aa'}$.

Propriété 1
Pour toute série statistique double, on a $\boxed{-1 \leqslant r \leqslant 1}$
(en fait, r est le cosinus d'un angle)

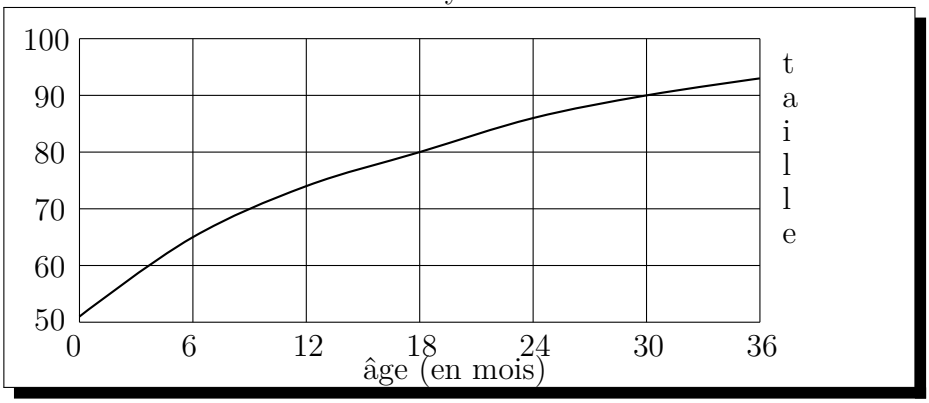
Propriété 2
Plus r est proche de ± 1, plus l'utilisation d'un ajustement affine est pertinente.
Si r est proche de 0, il y a peu de corrélation linéaire entre x et y et l'ajustement affine n'a pas de sens.

Détermination à la calculatrice :
r est souvent donné avec la droite de régression, voir les manipulations du b).
Sur T.I., il faut parfois activer l'affichage de r dans 2nd Catalog, mettre CorrelAff ou Diagnostic sur On.
Pour les machines qui ne le donnent pas, on se servira de $r = \pm\sqrt{aa'}$.

EXEMPLE 1 : $r \approx -0,99512 \simeq -1$ donc on a une forte corrélation linéaire (ce qui est clair au vu de la forme du nuage).

III. Exemples de changement de variable

Exemple 2 :
Considérons la série suivante donnant la taille moyenne d'un enfant entre 0 et 36 mois :



Relevons quelques valeurs et calculons r :

x	0	6	12	18	24	30	36
y	51	65	74	80	86	90	93

donne $r \simeq 0,969$ ce qui indique une bonne corrélation linéaire.

Néanmoins, on constate que la courbe n'est pas vraiment rectiligne, elle a la même allure que la fonction racine carrée (donc y se comporte à peu près comme $\sqrt{x}$). Posons donc $X = \sqrt{x}$ et regardons si y est une fonction (à peu près) linéaire de X : le coefficient de corrélation linéaire de la nouvelle série

$X = \sqrt{x}$	0	$\sqrt{6}$	$\sqrt{12}$	$\sqrt{18}$	$\sqrt{24}$	$\sqrt{30}$	$\sqrt{36}$
y	51	65	74	80	86	90	93

est $r \simeq 0,997$: la corrélation linéaire est bien meilleure. La droite de régression de y en X a pour équation $y = 7,24X + 49,55$ donc la courbe a à peu près pour équation $y = 7,24\sqrt{x} + 49,55$.

Par exemple, à $x = 15$ mois, la taille moyenne est d'environ $y = 77,6$ cm.

Changement de variable avec la calculatrice : Supposons que la liste 1 contienne les valeurs de x, la liste 2 celles de y. On veut que la machine calcule les valeurs après le changement de variable et les place dans la liste 3. Pour cela, éditez les listes statistiques, allez sur l'entête de la liste 3, et entrez (pour l'exemple 2) $\sqrt{}$ List1 (sur Casio) ou $\sqrt{}\,L_1$ (sur T.I.)
Pensez bien sûr ensuite à modifier le setup ou à utiliser LinReg L3,L2 (sur TI).

En cas de besoin, d'autres notices de calculatrices sur le site xmaths ; le manuel de la Casio Fx-CG20 à partir de la page 49 (stats) et 65 (stats à deux variables) et un exercice guidé sur la TI89

Exemple 3 :
Sur un parcours donné, la consommation y d'une voiture est donnée en fonction de sa vitesse moyenne x par le tableau suivant :

x (en km/heure)	80	90	100	110	120
y (en litres/100 km)	4	4,8	6,3	8	10

A l'aide de la méthode des moindres carrés, estimer la consommation aux 100 km (arrondie au dixième) de la voiture pour une vitesse de 130 km/h. Faîtes de même en remplaçant y par $z = \ln y$ (on écrira y sous la forme $y = A\,\mathrm{e}^{Bx}$). Comparez les coefficients de corrélation linéaire obtenus.
Réponses : La machine donne $y = 0,152x - 8,58$ donc si x = 130 alors $y = 0,152 \times 130 - 8,58 = 11,18 \simeq 11,2$ litres.

La machine donne $z = 0,0234x - 0,5080$ d'où : $y = \mathrm{e}^{0,0234x-0,508} = \mathrm{e}^{-0,508} \times \mathrm{e}^{0,0234x} \simeq 0,6017\,\mathrm{e}^{0,0234x}$.
Si $x = 130$ alors $y = 0,6017\,\mathrm{e}^{0,0234\times 130} \simeq 12,6$ litres.