Les statistiques

I. Point moyen – Covariance

Sur une population donnée, étudions deux caractères.
Pour chacun des n individus de cette population, notons x_i et y_i les valeurs prises par chacun de ces caractères, et présentons les données à l'aide de la série statistique à deux variable suivante :

Valeur x_i	x_1	x_2	…	x_n
Valeur y_i	y_1	y_2	…	y_n

a) Nuage de points - Point moyen

Définition : Dans un repère orthogonal, l'ensemble des points A_i de coordonnées $(x_i\,;\,y_i)$ (avec $1 \le i \le n$) est appelé le **nuage de points** associé à cette série statistique à deux variables.

Notons $\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$.

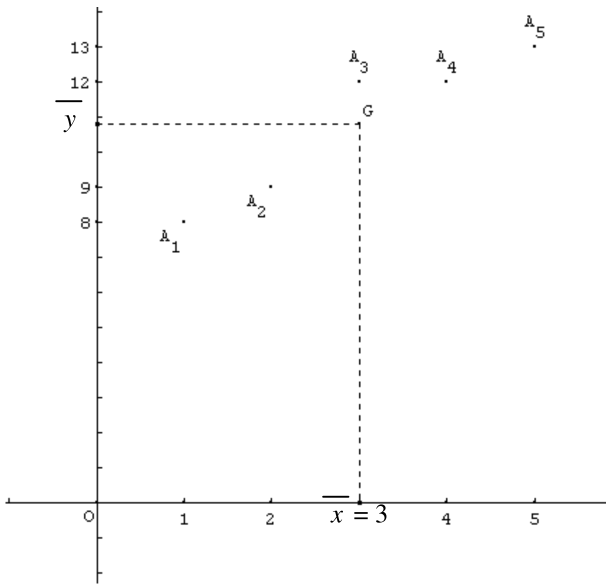
Définition : Le point G de coordonnées $(\overline{x}\,;\,\overline{y})$ est appelé le **point moyen** du nuage de points associé à cette série statistique à deux variables.

Obtention des coordonnées du point moyen grâce à la calculatrice :

Texas Instrument (TI – 80)	Casio Graph 25
$\boxed{\text{STAT}}$ 1 : Edit… permet d'entrer les valeurs de x dans L1, puis celles de y dans L2 $\boxed{\text{STAT}}$ CALC 2 : 2-VAR Stats puis (2nd $\boxed{\text{L1}}$ $\boxed{,}$ 2nd $\boxed{\text{L2}}$) $\boxed{\text{ENTER}}$ Ceci nous donne $\overline{x}$ et $\overline{y}$.	Dans le menu STAT, entrer les valeurs de x dans List 1, puis celles de y dans List 2. $\boxed{\text{CALC}}$ $\boxed{\text{SET}}$, entrer dans 2VarXList : List 1 2VarYList : List 2 $\boxed{\text{EXE}}$ puis Calc 2-Var On obtient alors $\overline{x}$ et $\overline{y}$.

Exemple :
La série statistique double suivante indique les notes mensuelles d'un élève au cours des cinq premiers mois de l'année scolaire numérotés de 1 à 5.

Mois x_i	1	2	3	4	5
note y_i	8	9	12	12	13



$\overline{x} = 3$ et $\overline{y} = 10{,}8$
donc le point moyen G du nuage représenté ci-dessus a pour coordonnées (3 ; 10,8).

b) variance - covariance

Pour étudier la dispersion de chaque variable x et y, on peut calculer leurs variances :

$$V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2 \quad \text{et} \quad V_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y})^2$$

Mais il est utile d'introduire une quantité qui fasse intervenir à la fois les valeurs de x et de y.

Définition : On appelle covariance de x et y le nombre :

$$C_{x\,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i\right) - \overline{x}\;\overline{y}\,.$$

La seconde expression est plus commode pour les calculs à la main.

→ Dans l'exemple précédent, $C_{x\,y} = \frac{1}{5}(1 \times 8 + 2 \times 9 + 3 \times 12 + 4 \times 12 + 5 \times 13) - 3 \times 10{,}8$

$= 35 - 32{,}4 = 2{,}6.$

II. Ajustement affine par la méthode des moindres carrés

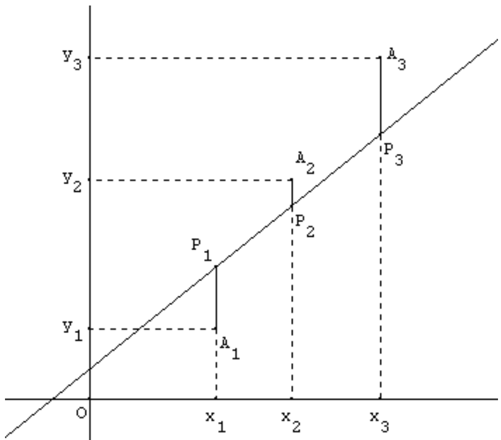
Lorsque les points du nuage paraissent presque alignés, on peut chercher une relation de la forme $y = ax + b$ qui exprime de façon approchée y en fonction de x, autrement dit, une fonction affine f telle que l'égalité $y = f(x)$ s'ajuste au mieux avec les données. Graphiquement, cela signifie qu'on cherche **une droite qui passe au plus près de tous les points du nuage**.

Une telle relation permettrait notamment de faire des **prévisions**.

Pour mesurer la qualité d'une telle formule, on considère, pour chaque valeur x_i, la différence entre la valeur observée, c'est à dire y_i, et la valeur calculée par la formule, c'est à dire $ax_i + b$. On souhaite que toutes les différences : $y_i - ax_i - b$ appelées **erreurs**, ou **résidus**, ou **perturbations**, soient les plus petites possible.

La méthode la plus couramment employée, dite **méthode des moindres carrés**, consiste à choisir a et b de façon que la **somme des carrés des résidus soit la plus petite possible**.

On considère par la suite un nuage de points $A_i(x_i\,;\,y_i)$ (avec $1 \leq i \leq n$).



Définition : Il existe une droite unique associée au nuage de points $A_i\,(x_i\,;\,y_i)$, avec $i = 1, 2, \ldots, n$, telle que la somme S des $A_iP_i^2$ soit minimale.

- Cette droite passe par le point moyen $G(\overline{x}\,,\,\overline{y})$ du nuage.
- Elle a une équation $y = ax + b$ avec $a = \dfrac{C_{x\,y}}{V_x}$ et $b = \overline{y} - a\,\overline{x}$.

Définition : Cette droite s'appelle la **droite de régression** de y en x.

Utilisation de la calculatrice pour la détermination de l'équation de la droite de régression :

Texas Instrument (TI – 80)	Casio Graph 25
$\boxed{\text{STAT}}$ 1 : Edit… permet d'entrer les valeurs de x dans L1, puis celles de y dans L2 $\boxed{\text{STAT}}$ CALC 3 : LINREG(aX+b) puis (2nd $\boxed{\text{L1}}$ $\boxed{,}$ 2nd $\boxed{\text{L2}}$) $\boxed{\text{ENTER}}$ Ceci nous donne les nombres a et b	Dans le menu STAT, entrer les valeurs de x dans List 1, puis celles de y dans List 2. $\boxed{\text{CALC}}$ $\boxed{\text{SET}}$, entrer dans 2VarXList : List 1 2VarYList : List 2 $\boxed{\text{EXE}}$ puis Calc Reg 1-Linear Ceci nous donne les nombres a et b

Exercice : Considérons la série statistique à deux variables (x_i ; y_i), pour $i = 1, 2, \ldots, 6$:

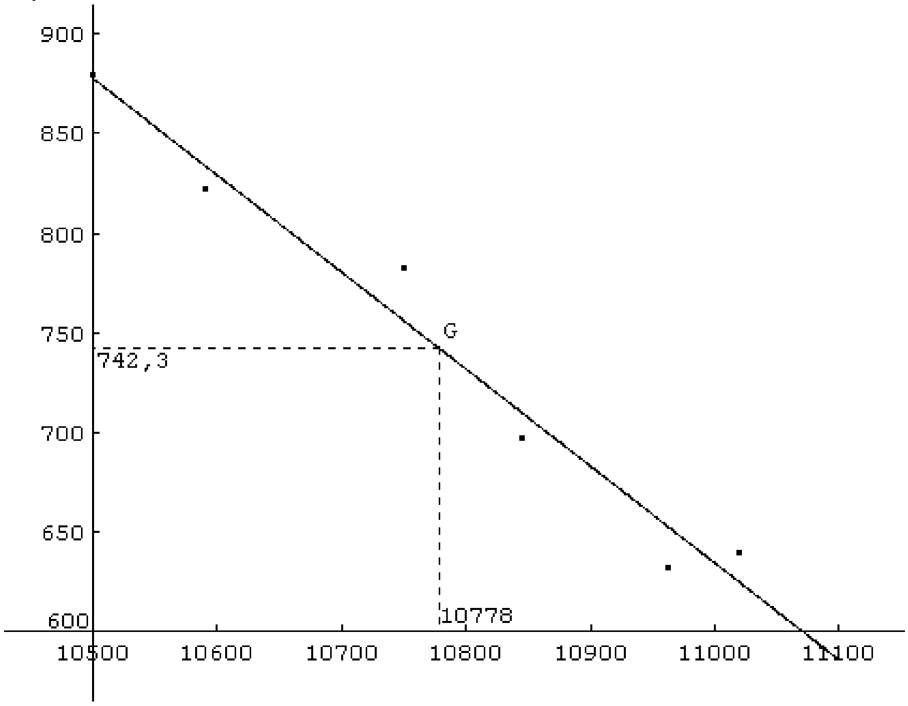
x_i	10 500	10 590	10 750	10 845	10 963	11 020
y_i	880	822	783	697	632	640

1. Placer, dans un repère orthogonal, le nuage de points $A_i(x_i ; y_i)$ et constater que la forme « allongée » du nuage justifie un ajustement affine.
2. Placer le point moyen G de ce nuage.
3. Tracer la droite d de régression de y en x.

Solution :

2. Le point moyen G a pour coordonnées ($\overline{x}$; $\overline{y}$).
On obtient grâce à la calculatrice :
$\overline{x} = 10\,778$ et $\overline{y} = 742{,}3$.

3. La droite d passe par G ;
pour la tracer, il suffit de connaître son coefficient directeur a.
D'après la calculatrice, $a \approx -0{,}487$.



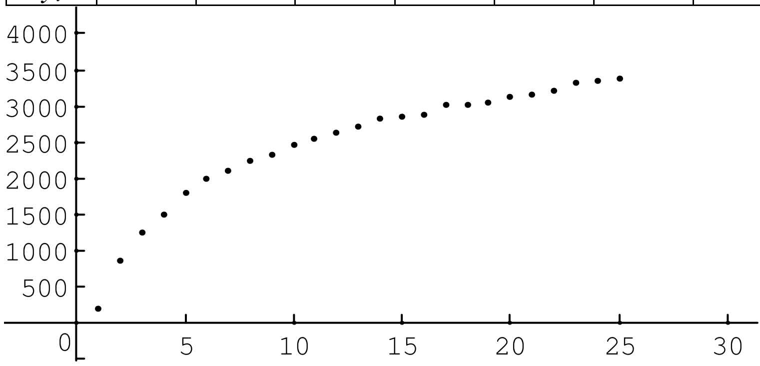
III. Exemple d'ajustement non affine

Ajustement logarithmique

On considère les données suivantes :

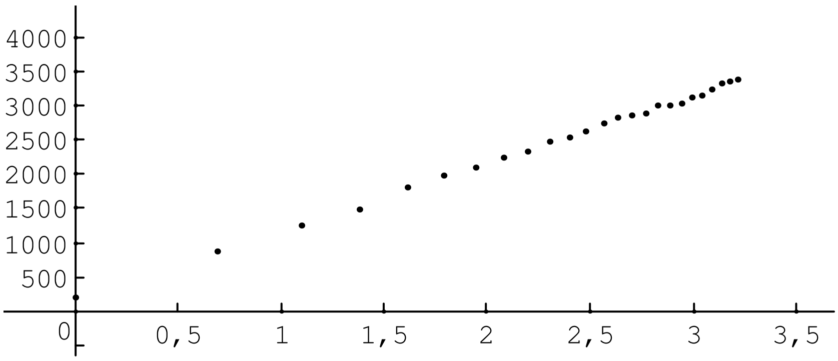
x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	198	881	1256	1489	1804	1983	2104	2247	2312	2468	2541	2639

x_i	13	14	15	16	17	18	19	20	21	22	23	24	25
y_i	2728	2811	2850	2890	3005	3010	3087	3125	3155	3221	3333	3365	3392



La forme du nuage suggère un ajustement de la forme $y = a \ln x + b$.

On peut le vérifier en posant $x' = \ln x$ et en plaçant les points de coordonnées (x' ; y) dans un nouveau repère :



Ces points sont presque alignés, ce qui permet d'envisager un ajustement affine du type $y = ax' + b$.
Alors la formule $y = a \ln x + b$ sera un bon ajustement de y en x.

Avec la calculatrice, on obtient : $y = 989 \ln x + 180$ (en arrondissant les coefficients à l'unité).

IV. Alignement de points et lien de causalité

On considère le tableau suivant :

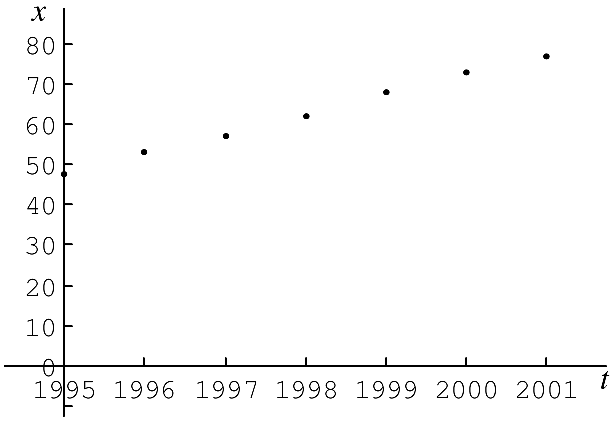
t_i : Année	1995	1996	1997	1998	1999	2000	2001
x_i : Nombre d'inscrits dans un club de belote	48	53	57	62	68	73	77
y_i : Nombre de hamburgers vendus dans un restaurant de Moscou	7000	7450	8000	8500	9050	9550	10000

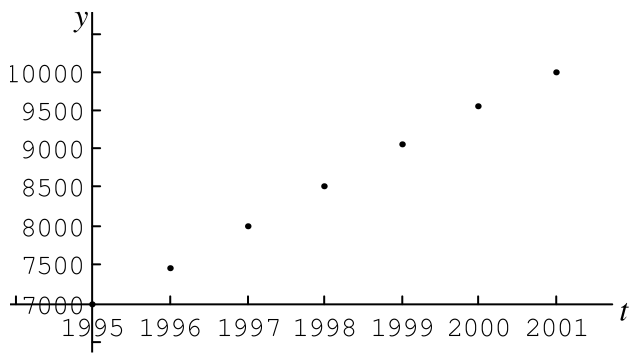
1. Placer dans un repère orthogonal le nuage de points $(t_i\,;\,x_i)$ et constater que sa forme allongée justifie un ajustement affine.

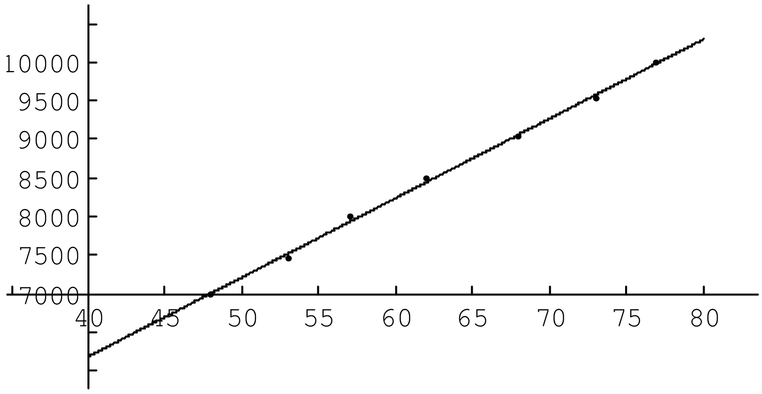
2. placer dans un autre repère orthogonal le nuage de points $(t_i\,;\,y_i)$ et constater que sa forme allongée justifie sa forme affine.

3. **a)** placer dans un troisième repère le nuage de points $(x_i\,;\,y_i)$ et constater que sa forme allongée justifie un ajustement affine.

b) Vérifier que la droite de régression de y en x admet comme équation $y = 103x + 2057$ et tracer cette droite.







Commentaire : Les x_i et les y_i sont liés approximativement par la relation $y_i = 103x_i + 2057$.
Mais il n'y a évidemment aucune relation de causalité entre les x_i et les y_i.