Terminales
Maths Complémentaires

Autrefois, les statistiques consistaient simplement à la collecte des informations. C'est à partir du 18ème siècle que le rôle prévisionnel des statistiques apparaît notamment par des prévisions sur la mortalité (utilisées par les compagnies d'assurances vie). Puis au 19ème siècle, Gauss introduit la méthode des moindres carrés pour prévoir des trajectoires d'astéroïdes.

Statistiques à deux variables

I – Ajustement affine d'un nuage de points

a – Définitions

Définitions 1

Sur une population, lorsqu'on étudie deux caractères quantitatifs x et y et que l'on note x_i et y_i les différentes valeurs prises par ces deux caractères, on présente ces données sous la forme d'un tableau comme ci-dessous que l'on nomme **série statistique à deux variables**.

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

Dans un repère, **le nuage de points** associé à cette série statistique est l'ensemble des points M_i de coordonnées $(x_i ; y_i)$.
Le **point moyen de ce nuage** est le point G de coordonnées $(\overline{x} ; \overline{y})$ où $\overline{x}$ est la moyenne des valeurs x_i et $\overline{y}$ est la moyenne des valeurs y_i.

Définitions 2

Dans certains cas, la forme du nuage de points peut laisser penser que les points M_i se répartissent autour d'une droite comme sur la figure ci-contre.

Dans ce cas, on pourra pratiquer un ajustement affine du nuage de points, c'est-à-dire tracer une droite "qui passe le plus près possible " des points du nuage.

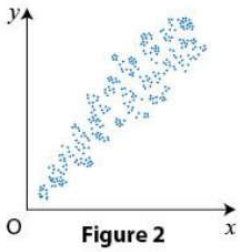


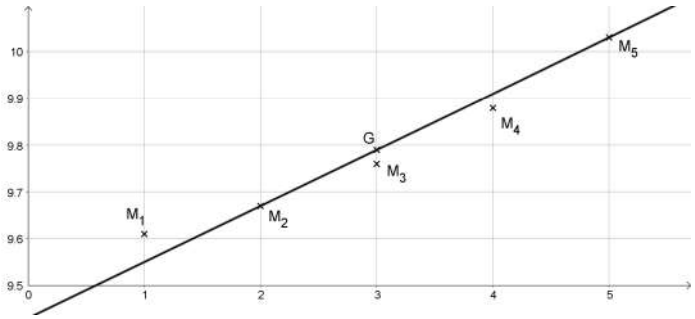
Figure 2

b – Exemple

Le tableau ci-dessous donne le montant du SMIC horaire, entre 2013 et 2019 :

Année	2015	2016	2017	2018	2019
Rang x_i	1	2	3	4	5
SMIC horaire y_i	9,61	9,67	9,76	9,88	10,03

$$\left.\begin{array}{l} \overline{x} = \dfrac{1+2+3+4+5}{5} = 3 \\ \overline{y} = \dfrac{9{,}61+9{,}67+9{,}76+9{,}88+10{,}03}{5} = 9{,}79 \end{array}\right\} \quad G\,(3 ; 9{,}79)$$



La forme allongée du nuage de points peut laisser penser qu'un ajustement affine est judicieux. Par exemple, la droite (GM_5) peut sembler un bon ajustement affine de ce nuage.

Remarque :

Dans le cas où l'on essaye d'effectuer un ajustement affine, on peut se demander : " quelle droite tracer ? " , " Y a-t-il une droite qui réalise un meilleur ajustement que les autres ? " …. . Les réponses sont abordées dans les paragraphes suivants.

II – Droite des moindres carrés

a – Variance, écart type et covariance

Définitions 3

Considérons une série statistique à deux variables x et y :

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

* **La variance des valeurs de x**, notée $V(x)$ par $V(x) = \frac{1}{n}\left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\right]$
* **L'écart-type des valeurs de x**, noté $\sigma(x)$, est la racine carrée de la variance : $\sigma(x) = \sqrt{V(x)}$
* **La variance des valeurs de y**, notée $V(y)$ par $V(y) = \frac{1}{n}\left[(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \ldots + (y_n - \overline{y})^2\right]$
* **L' écart-type des valeurs de y**, noté $\sigma(y)$, est la racine carrée de la variance : $\sigma(y) = \sqrt{V(y)}$
* **La covariance de la série $(x_i ; y_i)$**, notée $\text{cov}(x ; y)$ par

$$\text{cov}(x ; y) = \frac{1}{n}\left[(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})\right]$$

Exemple : avec les données de l'exemple du paragraphe 1 :

$V(x) = \frac{1}{5}\left[(1-3)^2 + \ldots + (5-3)^2\right] = 2$ donc $\sigma(x) = \sqrt{2}$

$V(y) = \frac{1}{5}\left[(9{,}61 - 9{,}79)^2 + \ldots + (10{,}03 - 9{,}79)^2\right] \approx 0{,}023$ donc $\sigma(y) = \sqrt{V(y)} \approx 0{,}15$

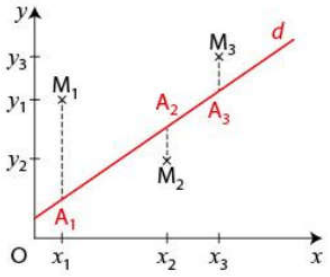
$\text{cov}(x ; y) = \frac{1}{5}\left[(1-3)(9{,}61 - 9{,}79) + \ldots + (5-3)(10{,}03 - 9{,}79)\right] = 0{,}21$

b – Droite des moindres carrés

Propriétés 4 : méthode des moindres carrés

Dans un repère, **la droite des moindres carrés** (ou **droite de régression**) de y en x associée au nuage de points $M_i(x_i ; y_i)$:

* passe par le point moyen $G(\overline{x} ; \overline{y})$ du nuage,
* a pour équation $y = a(x - \overline{x}) + \overline{y}$ avec $a = \frac{\text{cov}(x ; y)}{V(x)}$



Exemple : avec les données de l'exemple du paragraphe 1, la droite (d) des moindres carrés a pour équation :

$y = a(x - \overline{x}) + \overline{y}$ avec $a = \frac{\text{cov}(x ; y)}{V(x)} = \frac{0{,}21}{2} = 0{,}105$.

Donc $y = 0{,}105(x - 3) + 9{,}79 = 0{,}105x - 0{,}105 \times 3 + 9{,}79$ $\boxed{y = 0{,}105x + 9{,}475}$

La droite de régression permet, par exemple, d'estimer le SMIC horaire en 2022 ($x = 8$).

Dans ce cas $y = 0{,}105 \times 8 + 9{,}475 = 10{,}29$

Donc, si l'évolution se poursuit de la même manière, le taux horaire du SMIC devrait être de 10,29 € en 2022.

III – Coefficient de corrélation linéaire

a – Introduction

La décision d'ajuster un nuage de points par une droite ne peut pas simplement se prendre à partir de l'observation de la forme "allongée ou non " du nuage de points. D'un point de vue mathématique, il est nécessaire de quantifier cette décision. Pour cela, on utilise le coefficient de corrélation linéaire.

b – Définition et propriété

Définition 5

Considérons une série statistique à deux variables x et y :

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

On appelle **coefficient de corrélation linéaire entre x et y** le réel r défini par $\boldsymbol{r = \dfrac{\text{cov}\,(x\,;\,y)}{\sigma\,(x)\times\sigma\,(y)}}$

On rappelle que $\sigma\,(x)$ et $\sigma\,(y)$ désignent l'écart-type de x et de y, $\sigma\,(x) = \sqrt{V\,(x)}$ et $\sigma\,(y) = \sqrt{V\,(y)}$

Propriété 6

Pour toute série statistique à deux variables quantitatives : $\boldsymbol{-1 \leqslant r \leqslant 1}$

* **Si $r = 1$ ou si $r = -1$ la corrélation entre x et y est maximum** : les points du nuage sont alignés.
* si $r < 0$, la droite de régression a une **pente négative**.
* si $r > 0$, la droite de régression a une **pente positive**.

Remarque :

Plus r est proche de 1 ou de – 1, plus l'ajustement linéaire est judicieux.

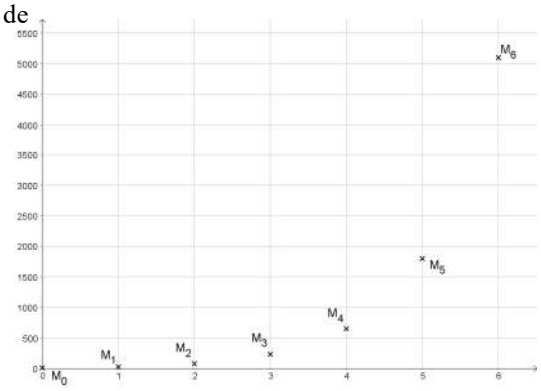
En pratique, on considère que l'ajustement linéaire est judicieux lorsque $r \geqslant 0{,}96$ (ou lorsque $r \leqslant -\ 0{,}96$).

Exemple : avec les données de l'exemple du paragraphe 1, le coefficient de corrélation entre x et y vaut $r = \dfrac{\text{cov}\,(x\,;\,y)}{\sigma\,(x)\times\sigma\,(y)} \approx \dfrac{0{,}21}{\sqrt{2}\times 0{,}15} \approx 0{,}99$. Ce coefficient est très proche de 1. L'ajustement affine était donc judicieux.

IV – Ajustement en utilisant un changement de variable (étude d'un exemple)

On s'intéresse à l'évolution de la croissance du nombre de bactéries à température ambiante sur un échantillon de glaces. On suppose que 10 bactéries sont déposées sur 1 gramme de glace. Voici les relevés du nombre de bactéries d'heure en heure ;

Heure (t_i)	0	1	2	3	4	5	6
Nombre de bactéries (N_i)	10	27	78	232	650	1800	5100



En utilisant la calculatrice (voir paragraphe V), on obtient : $\overline{t} = 3$ et $\overline{N} \approx 1128$.

Le coefficient de corrélation entre N et t vaut $r \approx 0{,}8$ ce qui montre qu'un ajustement affine n'est pas judicieux.

La forme du nuage de point (voir ci-contre) peut laisser penser que le nuage de points à la forme de la courbe d'une fonction exponentielle.

On pose alors $x_i = \text{e}^{t_i}$ et on obtient alors (valeurs arrondies au centième)

$x_i = \text{e}^{t_i}$	1	2,72	7,39	20,09	54,60	148,41	403,43
N_i	10	27	78	232	650	1800	5100

Le coefficient de corrélation entre N et x vaut $r' \approx 0{,}9999$ et la droite de régression entre N et x a pour équation : $N = 12{,}648\,x - 23{,}973$.

On peut donc écrire $N = 12{,}948\,\text{e}^{t} - 23{,}973$

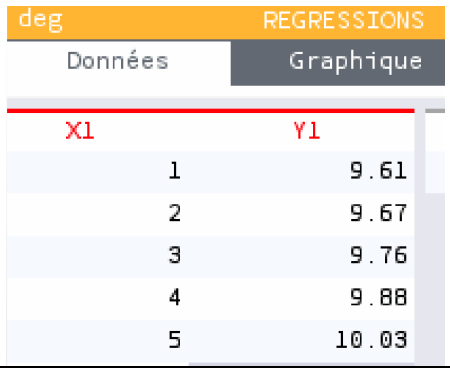
V – Utilisation de la calculatrice Numworks

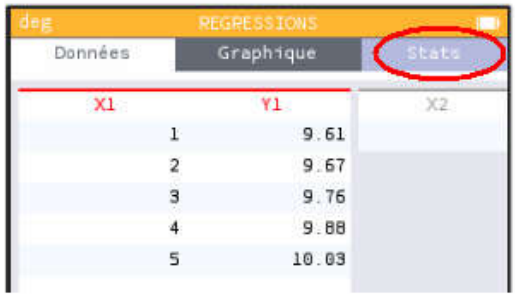
La calculatrice Numworks permet d'obtenir la plupart des résultats nécessaires pour la droites de régression (moyennes, variances, covariance, coefficient de corrélation, équation de la droite de régression ….).

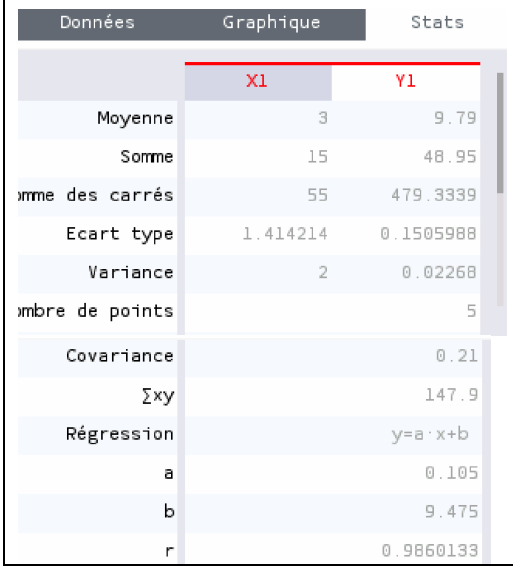
Voici la marche à suivre (à partir des données de l'exemple du paragraphe 1) :

Année	2015	2016	2017	2018	2019
Rang x_i	1	2	3	4	5
SMIC horaire y_i	9,61	9,67	9,76	9,88	10,03

Rentrer les données







Remarque **:** la touche "Graphique" permet d'obtenir le nuage de points ainsi que la droite de régression de y en x