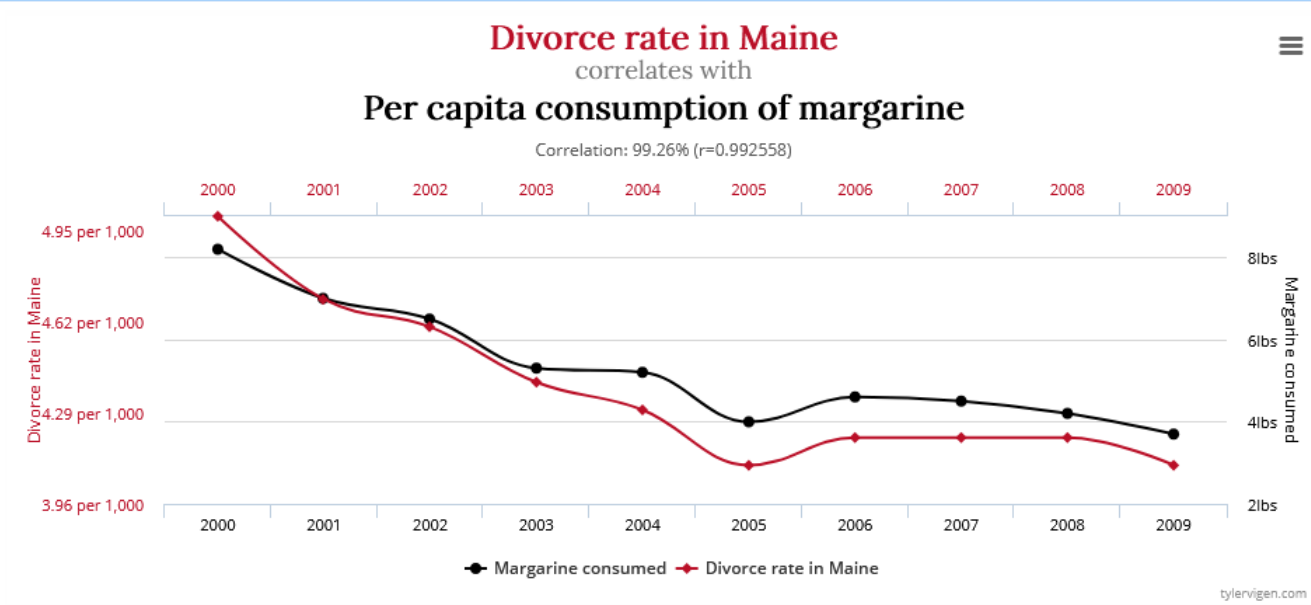
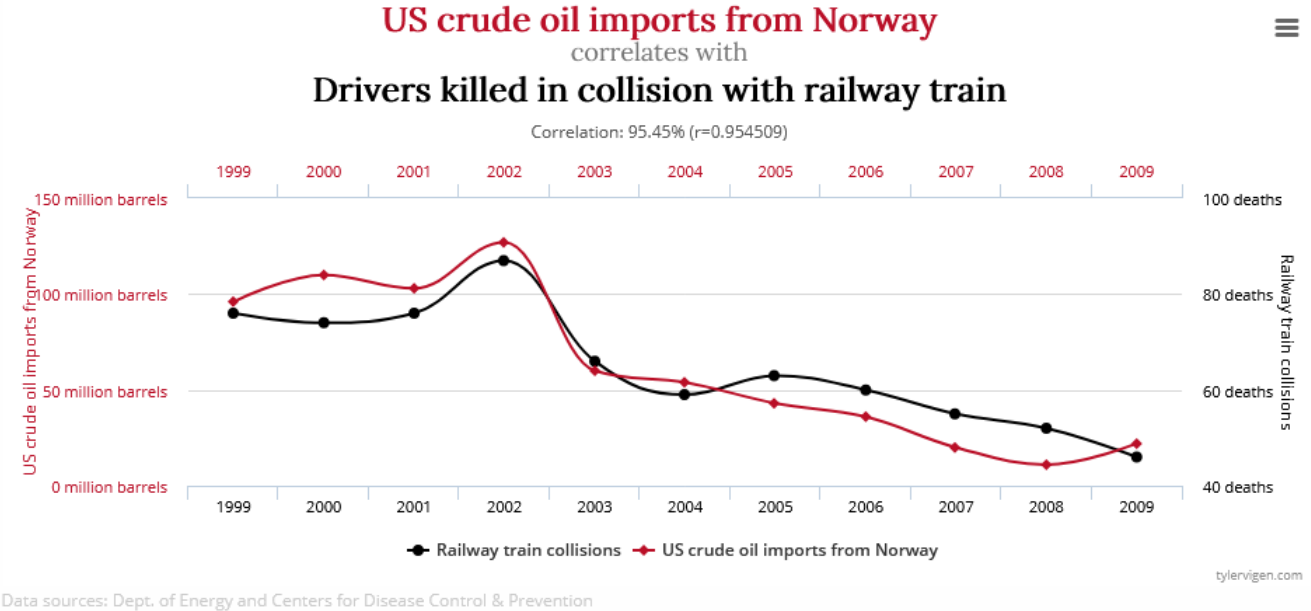
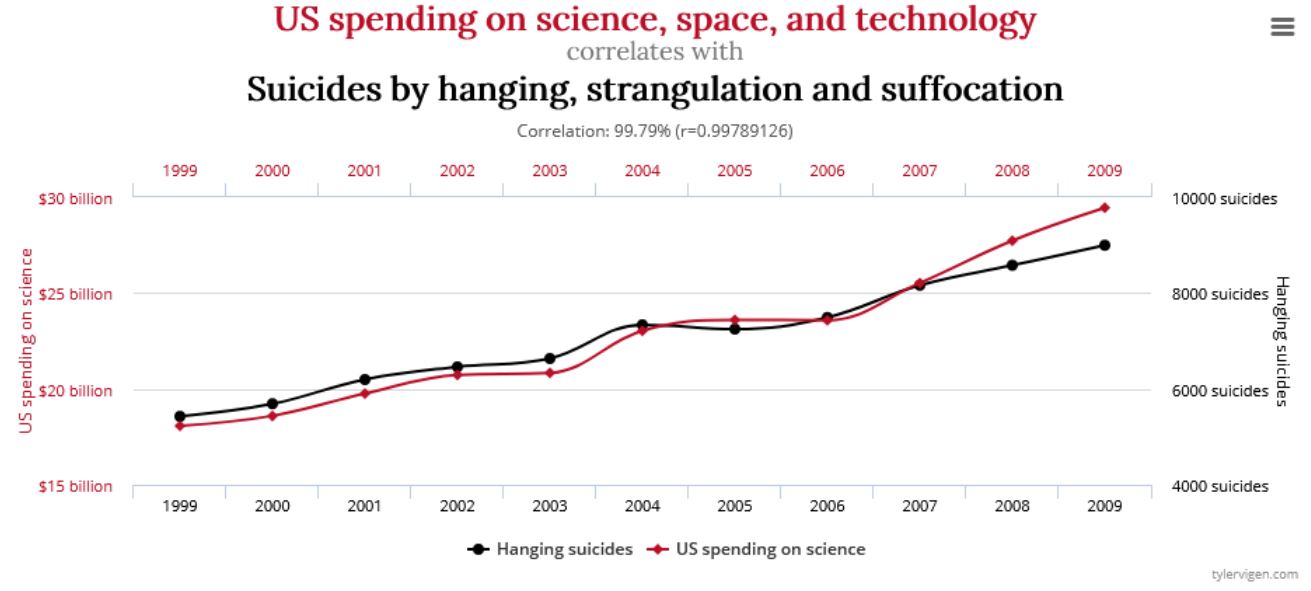


Data sources: National Vital Statistics Reports and U.S. Department of Agriculture



Data sources: Dept. of Energy and Centers for Disease Control & Prevention



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Sources : http://tylervigen.com/spurious-correlations

I. Ajustements d'un nuage de points

Dans certains cas, il semble exister un lien entre deux caractères d'une série statistique : la vente des crèmes solaires semble liée à celle des crèmes glacées sans qu'aucune des deux soit la cause ou la conséquence de l'autre (toutes deux sont certainement des conséquences d'un autre phénomène : l'ensoleillement).
Dans ces cas-là, il peut être intéressant d'étudier simultanément deux caractères d'une même population : les résultats peuvent alors être présentés sous différentes formes (tableaux, graphiques, etc).

DÉFINITIONS

Sur des individus d'une population, on réalise simultanément N observations de 2 caractères quantitatifs X et Y.

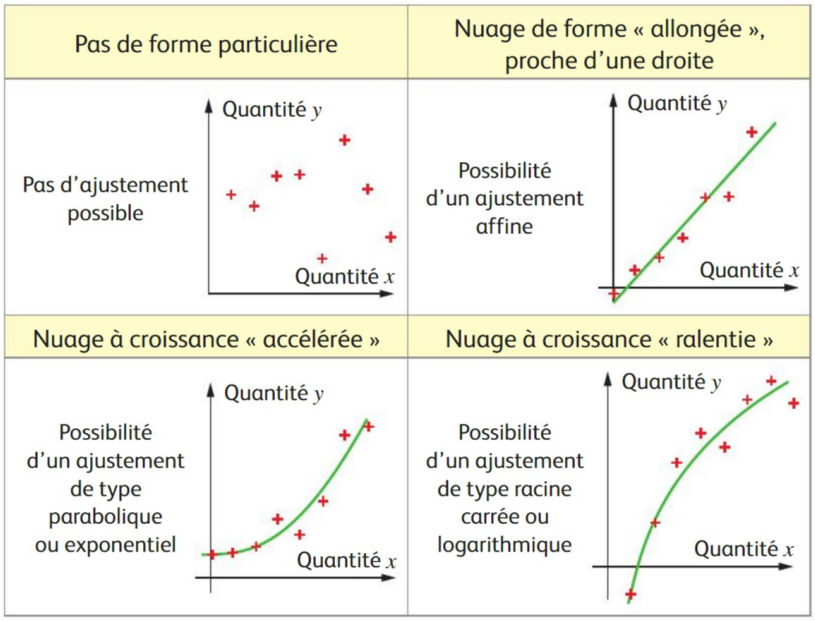
On obtient N couples $(x_1; y_1), \ldots, (x_N; y_N)$ où x_1 et y_1,…, x_N et y_N sont les valeurs observées de X et de Y : on parle de ***série statistique à deux variables***.

Le plan étant muni d'un repère, nous pouvons associer aux couples $(x_i; y_i)$ les points $M_i(x_i; y_i)$. L'ensemble de ces points constitue le ***nuage de points*** associé à cette série statistique.

Le nuage étant dessiné, on peut essayer de trouver une fonction f telle que la courbe d'équation $y = f(x)$ « passe le plus près possible » des points du nuage. C'est ce qu'on appelle un ***problème d'ajustement***.

Lorsqu'un ajustement est « possible », on dit que les variables X et Y sont ***corrélées*** : cependant, cela ne signifie pas nécessairement qu'il existe un lien de causalité entre ces deux variables.

Quelques exemples :



Source : Déclic Tle Maths complémentaires, éd. Hachette, 2020

Pour un ajustement affine, on parle souvent de ***régression linéaire***.

En latin, gradus signifie « pas » ou « marche ». *Régression* signifiait donc à l'origine « marcher en arrière ». Le statisticien anglais Francis Galton, cousin de Charles Darwin, introduisit ce terme en 1885. Travaillant sur l'hérédité, il cherchait à « expliquer » la taille des fils en fonction de celle de leur père : il constata que lorsque le père était plus grand que la moyenne, son fils avait tendance à être plus petit que lui et, a contrario, que lorsque le père était plus petit que la moyenne, son fils avait tendance à être plus grand que lui. Il y avait donc *régression* au sens courant du terme… Ce travail amena Galton à développer sa théorie *regression toward mediocrity*.

Lorsqu'on pense pouvoir réaliser un ajustement affine d'un nuage, il peut être intéressant d'orienter notre recherche en plaçant le point dont l'abscisse est la moyenne $\overline{x}$ des abscisses x_i, et l'ordonnée la moyenne $\overline{y}$ des ordonnées y_i :

DÉFINITION

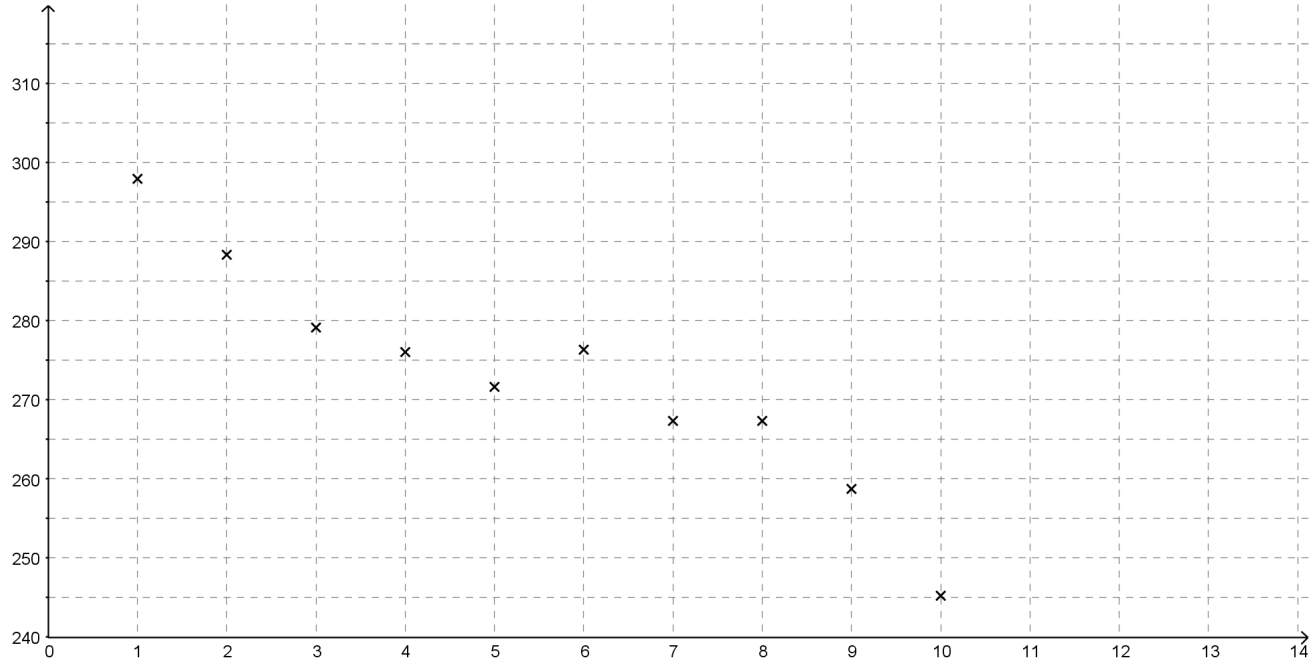
On appelle ***point moyen*** d'un nuage de N points $M_i(x_i ; y_i)$ le point $G(\overline{x} ; \overline{y})$.

EXEMPLE C1

On a copié ci-dessous le tableau d'une feuille de calcul donnant le nombre de mariages célébrés en France métropolitaine entre 2000 et 2009 (source : INSEE).

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Rang de l'année x_i	1	2	3	4	5	6	7	8	9	10
Nombre de mariages en milliers y_i	297,9	288,3	279,1	276	271,6	276,3	267,3	267,3	258,7	245,2

On souhaite estimer le nombre de mariages célébrés en France métropolitaine en 2014.



Partie A : première estimation

1. Calculer le taux d'évolution global du nombre de mariages célébrés en France entre 2005 et 2009 (on arrondira le résultat à 0,0001 %).

2. En déduire le taux d'évolution annuel moyen du nombre de mariages célébrés en France entre 2005 et 2009. On arrondira le résultat à 0,0001 %.

3. En considérant que le nombre de mariages célébrés continue à diminuer de ce taux moyen, soit d'environ 2,9 %, déterminer le nombre de mariages que l'on peut prévoir en France en 2014.

Partie B : deuxième estimation

On considère les points A(1 ; 297,9) et B(10 ; 245,2).

1. Déterminer l'équation réduite de la droite (AB), sous la forme $y = a\,x + b$ (arrondir a et b à 0,01 près).

2. Avec cet ajustement affine, déterminer le nombre de mariages que l'on peut prévoir en France métropolitaine pour l'année 2014.

3. Déterminer les coordonnées du point moyen, noté G. Vérifier si G appartient à la droite (AB).

Partie C : troisième estimation

On partage le nuage de points en deux groupes de même effectif (années de rangs 1 à 5 ; années de rangs 6 à 10).

1. Déterminer les coordonnées des points moyens G_1 et G_2 de chaque groupe.

2. On choisit la droite (G_1G_2) comme ajustement affine. C'est ce qu'on appelle ***la droite de Mayer***.

a. Vérifier que $G \in (G_1G_2)$.

b. Avec cet ajustement affine, déterminer le nombre de mariages que l'on peut prévoir en France métropolitaine pour l'année 2014.

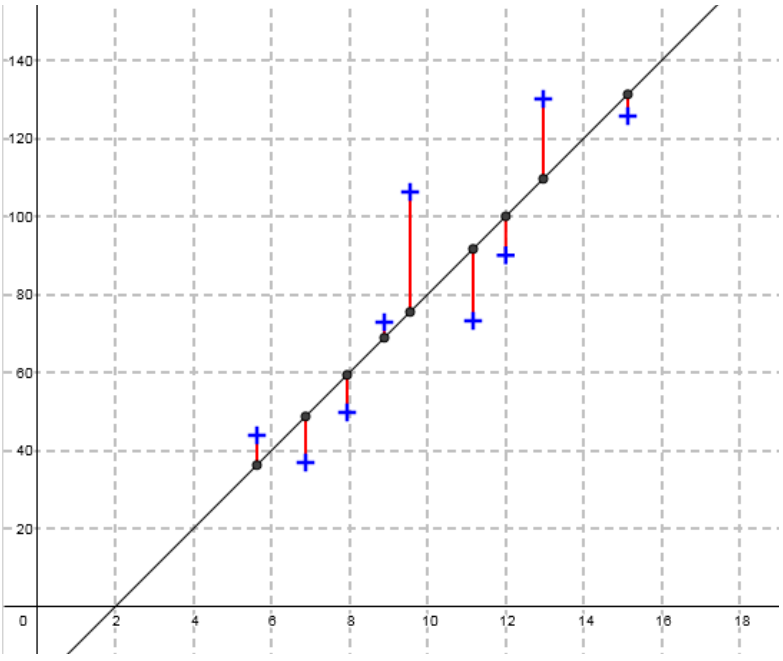
Johann Tobias Mayer (1723 - 1762) était un mathématicien, cartographe et astronome allemand.

Passionné d'architecture, son père lui enseigna les mathématiques et dès l'âge de 16 ans, le jeune Johann présentait des plans de fortifications militaires. Deux ans plus tard, il publiait des résolutions de problèmes géométriques ardus.

En 1746, Mayer travaille au service cartographique de Nuremberg et se spécialise en astronomie. Il sera professeur à Göttingen (1750) et nommé directeur de son célèbre observatoire (1754).

Mayer établit des tables des cycles lunaires et évalua les erreurs dues aux imperfections des réglages des instruments de mesure. En 1748, il utilisa pour la première fois, à la manière d'Euler à la même époque dans l'étude des orbites de Saturne et de Jupiter, mais indépendamment de lui, une méthode d'ajustement pour étudier la position d'un point sur la Lune et publia des tables de la Lune permettant aux navigateurs de faire le point à un demi-degré près.

Sources : Wikipédia ; http://serge.mehl.free.fr ; http://www.bibmath.net



Lorsqu'on cherche un ajustement affine, on essaie de minimiser l'écart entre cette droite d'ajustement et les points du nuage. Mais comment mesurer cet « écart » ? Une idée serait de calculer la somme des distances « verticales » entre les points du nuage et ceux de la droite : il faudrait donc sommer des valeurs absolues d'écarts entre points, ce qui n'est pas pratique. C'est pourquoi les scientifiques préfèrent sommer les carrés de ces écarts.

→

DÉFINITION-PROPRIÉTÉ (ADMISE)

La ***droite de régression par la méthode des moindres carrés*** (de y en x) est la droite d'équation $y=a\,x+b$ telle que la somme $\sum_i (y_i-(a\,x_i+b))^2$ est minimale.

Les coefficients a et b vérifient : $a=\dfrac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sum_i (x_i-\overline{x})^2}$ et $b=\overline{y}-a\,\overline{x}$.

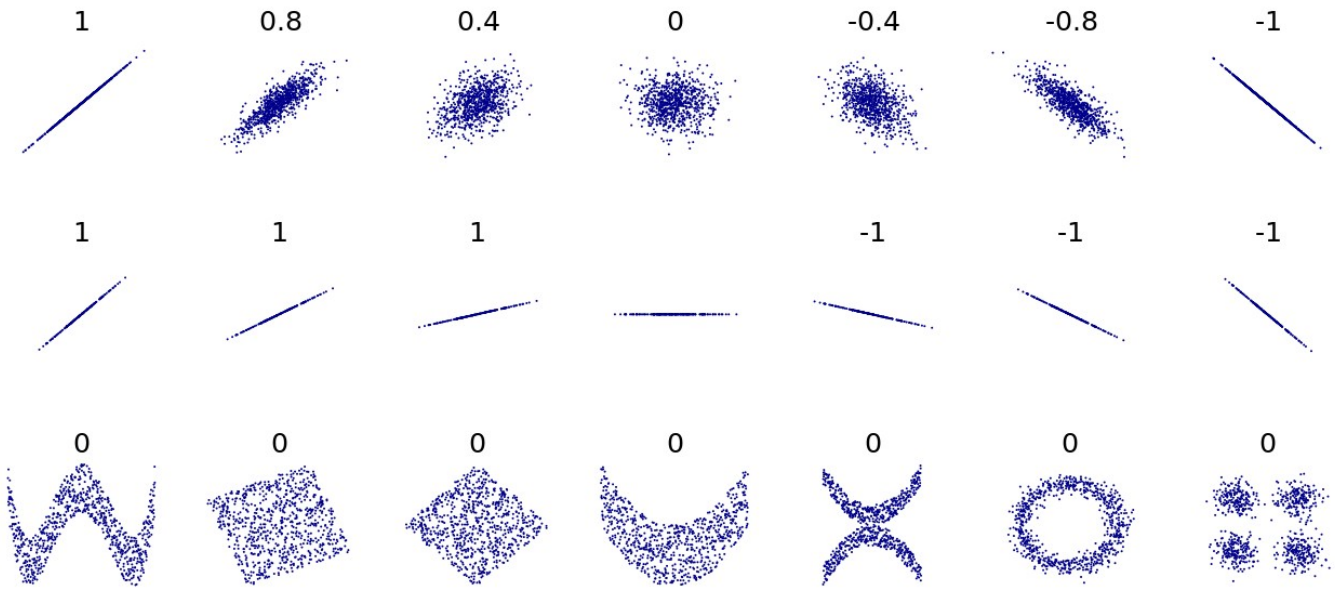
REMARQUES :

- $\overline{y}=a\,\overline{x}+b$ donc le point moyen appartient à cette « droite des moindres carrés ».
- On écrit souvent $a=\dfrac{\mathrm{cov}(X;Y)}{V(X)}$ où $V(X)$ est la variance de X et $\mathrm{cov}(X;Y)$ la covariance de X et Y, égale à : $\mathrm{cov}(X;Y)=\dfrac{1}{n}\sum_i (x_i-\overline{x})(y_i-\overline{y})$.

DÉFINITION-PROPRIÉTÉ (ADMISE)

On appelle ***coefficient de corrélation*** le nombre r égal à $\dfrac{\mathrm{cov}(X;Y)}{\sqrt{V(X)\,V(Y)}}$.

On a :
- $-1\leqslant r\leqslant 1$
- plus $|r|$ est proche de 1, plus X et Y sont corrélées linéairement
- si $r=-1$ ou $r=1$, tous les points du nuage sont alignés
- si $r=0$, les points sont « très dispersés » autour de la droite des moindres carrés.



↳ Exemples de coefficients de corrélation :
corrélation linéaire pour les deux premières lignes, non linéaire pour la troisième ligne
(source : Wikipédia)

REMARQUE : certains[1] estiment qu'un ajustement linéaire est pertinent lorsque $r>0{,}87$. Mais attention : l'interprétation d'un coefficient de corrélation dépend du contexte et des objectifs. Une corrélation de 0,9 peut être très faible si l'on vérifie une loi physique en utilisant des instruments de qualité, mais peut être considérée comme très élevée dans les sciences sociales où il peut y avoir une contribution plus importante de facteurs de complication. (Wikipédia)

EXEMPLE A 1


p. 263 capacité 3

Le tableau suivant donne la population française (métropole + DOM) à chaque recensement depuis 1999 (source : INED).

Année du recensement	1999	2006	2013	2019
Rang x_i depuis 1999	0	7	14	20
Population française y_i (en millier)	60 148	63 186	65 564	66 993

On note D : $y=a\,x+b$ la droite de régression par les moindres carrés associée à la série statistique $(x_i\,;y_i)$.

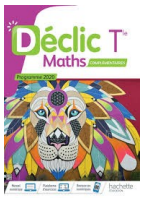
1. Déterminer par le calcul les coefficients *a* et *b*, en arrondissant *a* à 10^{-2} et *b* à l'entier.

Puis calculer le coefficient de corrélation *r*, à 0,001 près. L'ajustement affine est-il adapté ?

2. Vérifier les résultats précédents : **a.** à l'aide de la calculatrice ; **b.** à l'aide d'un tableur.

1 Voir http://serge.mehl.free.fr/chrono/Pearson.html#cor.

II. Ajustement et changement de variable



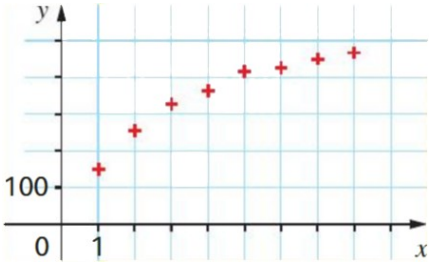
← voir page 264 du manuel Déclic

EXEMPLE A2

p. 265 capacité 4

Anna a créé un site web. Le tableau ci-dessous présente l'évolution du nombre hebdomadaire de visiteurs de ce site au cours des huit premières semaines suivant sa création.

Rang x_i de la semaine	1	2	3	4	5	6	7	8
Nombre y_i de visiteurs	152	253	327	361	412	426	451	465



Après avoir représenté le nuage de points associé à cette série, Anna constate que l'augmentation du nombre de visiteurs ralentit les dernières semaines : un ajustement affine n'est pas adapté. Elle envisage plutôt un ajustement de type « logarithmique » et pose $z=\ln(x)$.

1. a. Recopier et compléter le tableau de valeurs suivant, en arrondissant à 0,001 près.

x_i	1	2	3	4	5	6	7	8
$z_i=\ln(x_i)$	0	…	…	…	…	…	…	…
y_i	152	253	327	361	412	426	451	465

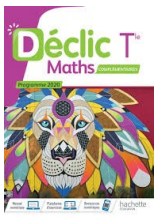
b. À l'aide de la calculatrice, déterminer l'équation $y=a\,z+b$ de la droite des moindres carrés ajustant le nuage de points $(z_i\,;y_i)$ et le coefficient de corrélation r associé. Arrondir les valeurs a et b à l'unité.

c. En déduire un ajustement du nuage de points. $(x_i\,;y_i)$.

2. En utilisant le modèle obtenu à la question 1.c., estimer le nombre de visiteurs lors de la 10e semaine.

3. Selon le modèle, à partir de quelle semaine peut-on penser que le nombre de visiteurs dépassera 600 ?

→ BILAN DU CHAPITRE & TRAVAIL EN AUTONOMIE ←



• Fiche bilan → p.266

• QCM 18 questions corrigées → p.267