STATISTIQUES A DEUX VARIABLES

Objectifs :

- *Représenter graphiquement un nuage de points associé à une série statistique à deux variables.*
- *Déterminer une droite de régression, à l'aide de la calculatrice, d'un logiciel ou par calcul.*
- *Utiliser un ajustement affine pour interpoler ou extrapoler.*

1. Vocabulaire

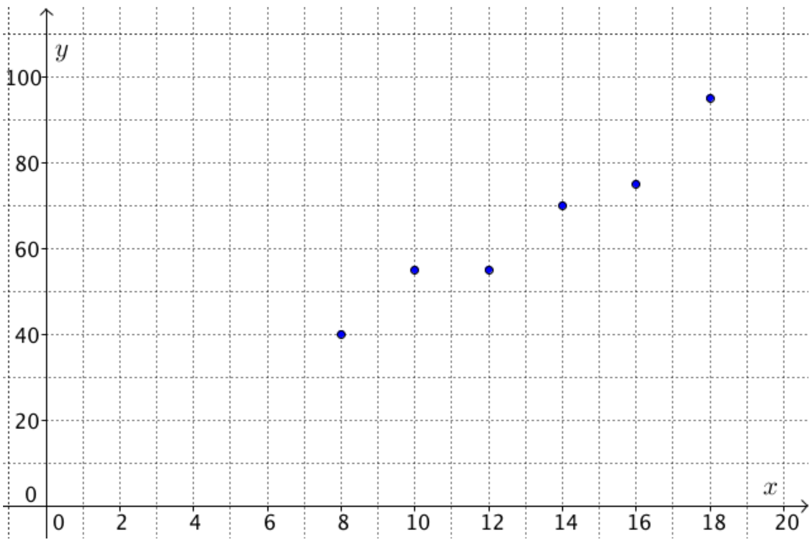
Définition 1. ***Nuage de points***

Soit une série statistique à deux variables x et y chacune de n valeurs $x_1, x_2, ..., x_n$ et $y_1, y_2, ..., y_n$.
Soit un repère orthogonal (*O* ; *I*, *J*), on appelle nuage de points l'ensemble des n points (non reliés) M_i de coordonnées $(x_i\ ;\ y_i)$ pour i entier variant de 1 à n.

Exemple : Le tableau suivant présente l'évolution du budget publicitaire et du chiffre d'affaire d'une société au cours des 6 dernières années :

Budget publicitaire en milliers d'euros x_i	8	10	12	14	16	18
Chiffre d'affaire en milliers d'euros y_i	40	55	55	70	75	95

Dans un repère, représenter le nuage de points $(x_i\ ;\ y_i)$.

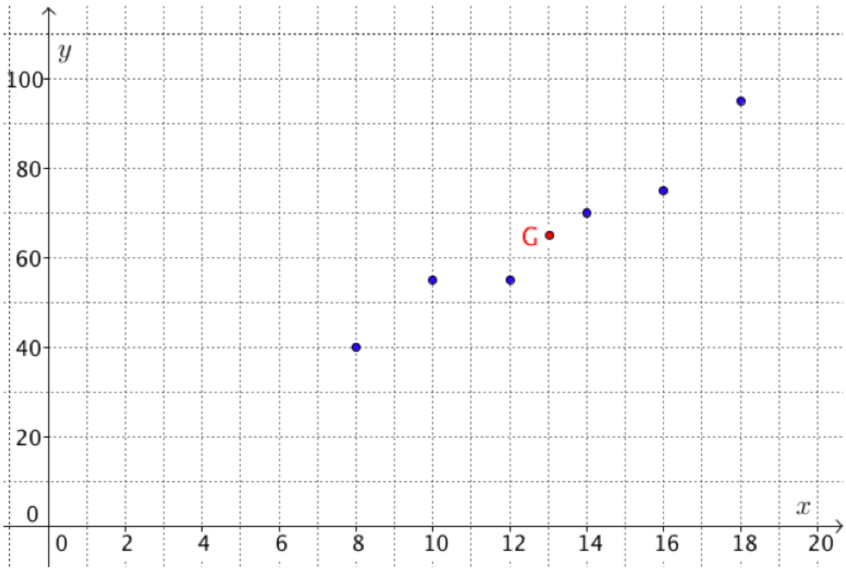


Définition 2. *Point moyen*
Soit une série statistique à deux variables x et y chacune de n valeurs x_1, x_2, ..., x_n et y_1, y_2, ..., y_n, et, $\overline{x}$ et $\overline{y}$ les moyennes arithmétiques respectives des n valeurs x_i et des n valeurs y_i.   **On appelle point moyen G du nuage, le point de coordonnées $(......\ ;\ ......)$.**

Exemple : Déterminer les coordonnées du point moyen G du nuage de points de l'exemple précédent.

$$\overline{x} = \frac{8+10+12+14+16+18}{6} = 11{,}2 \quad \text{et} \quad \overline{y} = \frac{40+55+55+70+75+95}{6} = 65$$

Le point moyen G du nuage de points a pour coordonnées (11,2 ; 65). On peut placer ce point dans le repère.



Méthodes pour représenter un nuage de points et déterminer le point moyen

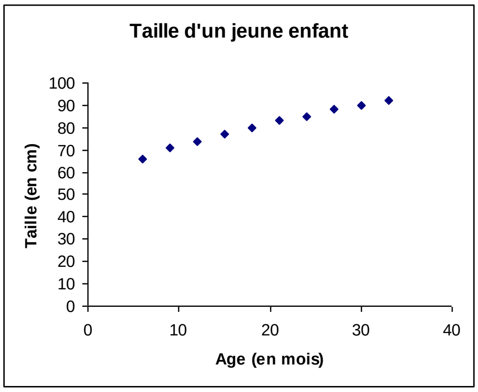
méthodes en vidéo

2. Ajustement d'un nuage de points

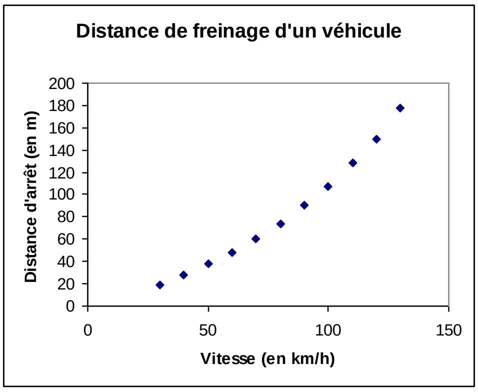
Définitions 3. Ajustement d'un *nuage de points*
● On appelle ajustement du nuage de points toute courbe « résumant approximativement » le nuage de points.   **● Toute droite « résumant approximativement » le nuage est appelée *droite d'ajustement affine* du nuage de points.**

Remarque : Il existe d'autres types d'ajustement : dans certains cas, on peut observer que visiblement une droite ne convient pas mais que le nuage de points semble être approché par un autre type de courbe, parabole par exemple.

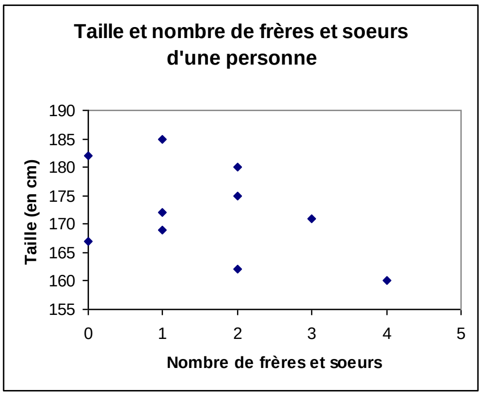
En outre, certains nuages peuvent ne pas sembler être approchables par une quelconque courbe auquel cas les deux variables ne sont pas reliées entre elles.



On peut tracer une droite $\mathscr{D}$ au voisinage des 10 points ; on dit alors que l'on a un ajustement affine.



Un ajustement affine ne convient pas ; on peut penser à « approcher » le nuage à l'aide d'une parabole.



Les points sont dispersés de façon quelconque ;
cela veut dire qu'il n'existe aucun lien entre x_i et y_i,
un ajustement n'est pas possible.

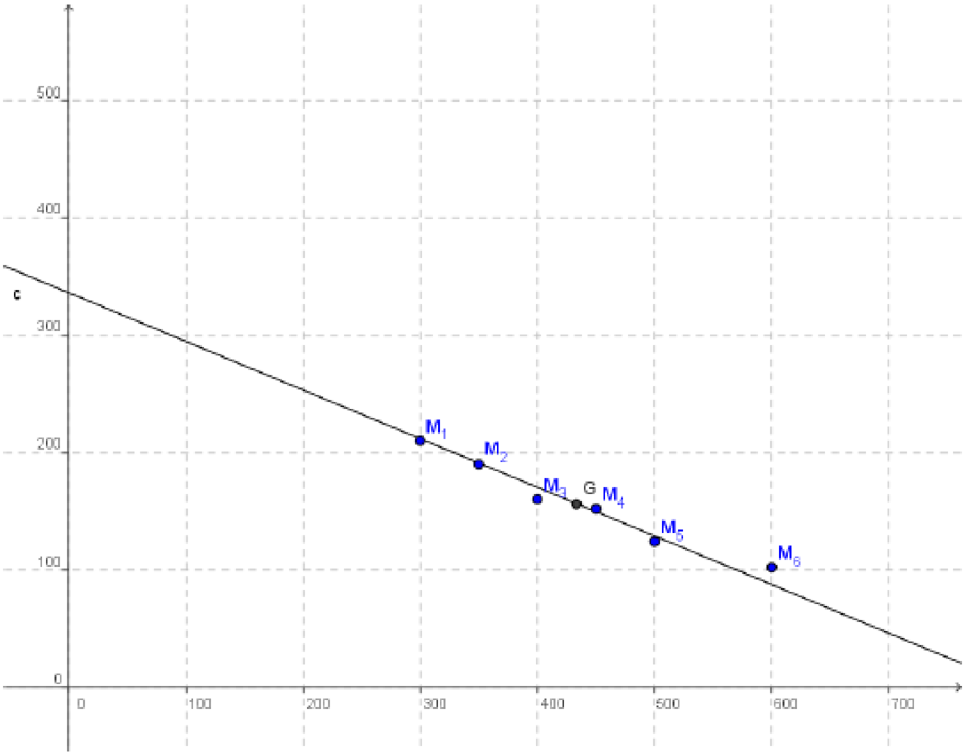
3. Détermination d'une équation d'une droite d'ajustement affine

Méthode graphique « au jugé »

On trace « au jugé » une droite qui « semble résumer » le nuage de points.
C'est une méthode simple mais qui dépend de la droite tracée.

Exemple : Le tableau suivant présente l'évolution du budget publicitaire et du chiffre d'affaire d'une société au cours des 6 dernières années :

Prix x_i en euros	300	350	400	450	500	600
Nombre de machines vendues y_i	210	190	160	152	124	102

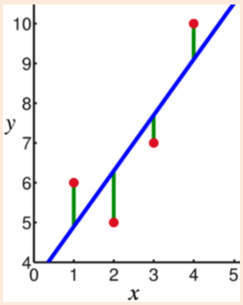


Méthode des moindres carrés

Cette méthode porte le nom de « moindre carrés » car elle consiste à rechercher la position de la droite d'ajustement tel que la somme des carrés des longueurs donnant les distances respectives (en vert) entre la droite et les points soit minimale.

Pour cela, on utilise la calculatrice qui va donner l'équation de la droite cherchée.

Cette droite est appelée droite de régression de y en x.



On peut montrer que son équation réduite est $y = mx + p$ avec

$$m = \frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\ldots+(x_n-\overline{x})(y_n-\overline{y})}{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2} \quad \text{et} \quad p = \overline{y} - m\,\overline{x}.$$

Exercice ❶ : On considère la série statistique à deux variables donnée dans le tableau suivant :

x_i	5	10	15	20	25	30	35	40
y_i	13	23	34	44	50	65	75	90

1) Dans un repère, représenter le nuage de points $(x_i\ ;\ y_i)$.

2) a) Déterminer une équation de la droite d'ajustement par la méthode des moindres carrés.
   b) Vérifier à l'aide de la calculatrice.
   c) Représenter la droite d'ajustement de y en x.

3) Estimer graphiquement la valeur de x pour $y = 70$. Retrouver ce résultat par calcul.
S'agit-il d'une interpolation ou d'une extrapolation ?

Définitions 4. ***Interpolation, extrapolation***

L'interpolation et l'extrapolation sont des méthodes qui consistent à estimer une valeur inconnue dans une série statistique.
- Pour une interpolation, le calcul est réalisé dans le domaine d'étude fourni par les valeurs de la série.
- Pour une extrapolation, le calcul est réalisé en dehors du domaine d'étude.

corrigé en vidéo

Remarques : ● La droite d'ajustement par la méthode des moindres carrés passe par le point moyen G.

● En utilisant la covariance $\text{cov}(X\ ;\ Y)$, égale à $\dfrac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})}{n}$, et la variance $V(X)$ égale à $\dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}$, on a : $m=\dfrac{\text{cov}(X\ ;\ Y)}{V(X)}$.

4. Coefficient de corrélation linéaire

Définition 5. ***Coefficient de corrélation linéaire***

On appelle coefficient de corrélation linéaire le nombre défini par :

$$r=\frac{\text{cov}(X\ ;\ Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

$$=\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}\times\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\ldots+(y_n-\bar{y})^2}}.$$

Interprétation : Le coefficient de corrélation r est un nombre compris entre -1 et 1 qui mesure la relation entre les deux variables x et y.

- Plus r est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte (c'est-à-dire que les points du nuage sont quasiment alignés. Dans ce cas, l'ajustement affine par la méthode des moindres carrés est adapté.
- Lorsque r est proche de 0, les points sont très dispersés autour de la droite. Dans ce cas, l'ajustement affine par la méthode des moindres carrés n'est pas adapté.

Méthode pour calculer un coefficient de corrélation linéaire

méthode en vidéo

Exercice ❷ : 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 y_i (en millier)	60 148	63 186	65 564	66 993

On note $\mathscr{D}$: $y = mx + p$ la droite de régression par les moindres carrés associée à la série statistique $(x_i\ ;\ y_i)$.

1) a) Déterminer par le calcul les coefficients m et p, en arrondissant m à 10^{-2} près, et p à l'entier près.
   b) Calculer le coefficient de corrélation r, à 0,0001 près.
   c) L'ajustement affine est-il adapté ?
2) Vérifier les calculs précédents :
   a) à l'aide de la calculatrice.
   b) à l'aide d'un tableur.

5. Ajustement se ramenant à un ajustement affine

Exercice ❸ : Une entreprise vend des perles pour la fabricationde bijoux de fantaisie. Le tableau ci-dessous indique la quantité vendue y_i (en tonne), pour un prix au kilogramme fixé x_i (en euro).

x_i	3	4	5,5	7	8,5	9,3	10	10,8
y_i	4,926	3,773	2,773	2,197	1,820	1,669	1,554	1,430

1) Tracer le nuage de points. Peut-on envisager un ajustement affine ?
2) On pose $z = \dfrac{100}{y}$. Déterminer les valeurs de la série statistique z. Arrondir à 10^{-3}.
3) a) Déterminer, à l'aide de la calculatrice, l'équation de la droite des moindres carrés des séries x et z.
   b) En déduire l'expression de y en fonction de x.
4) En extrapolant avec ce modèle, calculer la quantité de perles que vendrait l'entreprise si le prix montait à 24 € le kg.