Statistiques à deux variables, cours, terminale, Mathématiques complémentaires

1 Vocabulaire des statistiques à deux variables

Définition :

- Soient x et y deux caractères quantitatifs d'une même population. Á chaque individu de la population on associe un couple $(x_i; y_i)$ où x_i et y_i pour $i \in \{1; \ldots; n\}$ avec n entier naturel sont les valeurs prises respectivement par x et y. L'ensemble de ces couples constitue une *série statistique à deux variables* x et y.
- Dans un repère $(O; \vec{i}; \vec{j})$, l'ensemble des points M_i de coordonnées $(x_i; y_i)$ est appelé .... associé à la série statistique.
- Soit une série statistique à deux variables x et y de moyennes $\bar{x}$ et $\bar{y}$.
  Le point G de coordonnées $(\bar{x}; \bar{y})$ avec
  $\bar{x} = ...$
  et $\bar{y} = ...$
  est appelé le .... du nuage de points associé à la série statistique.

Exemple :

Un magasin réalise une étude sur l'influence du prix de vente sur le nombre de machines à laver vendues au cours d'une année. Le tableau suivant donne les résultats de cette étude :

Prix x_i en euros	300	350	400	448	500	600
Nombre de machines vendues	208	190	160	152	124	102

Le nuage de points associé à cette série est constitué des points $M_1\,(300; 208)$, $M_2\,(350; 190)$,...,$M_6\,(600; 102)$.

Le point moyen associé à ce nuage de points est le point G de coordonnées $(\bar{x}; \bar{y})$ données par :

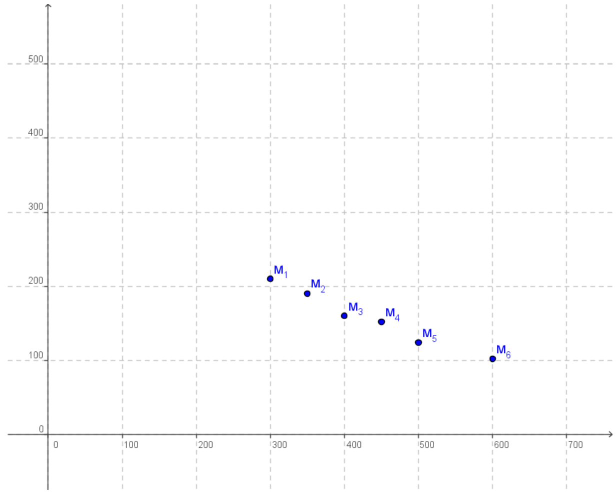
$\bar{x} = ...$

$\bar{y} = ...$

2 Ajustement d'un nuage de points

Définition :

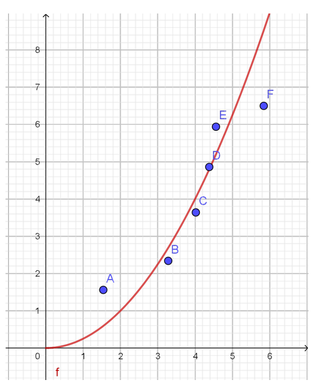
- on appelle .... du nuage de points toute courbe « résumant approximativement » le nuage.
- Toute droite « résumant approximativement » le nuage est appelée .... du nuage de points.



Remarque :

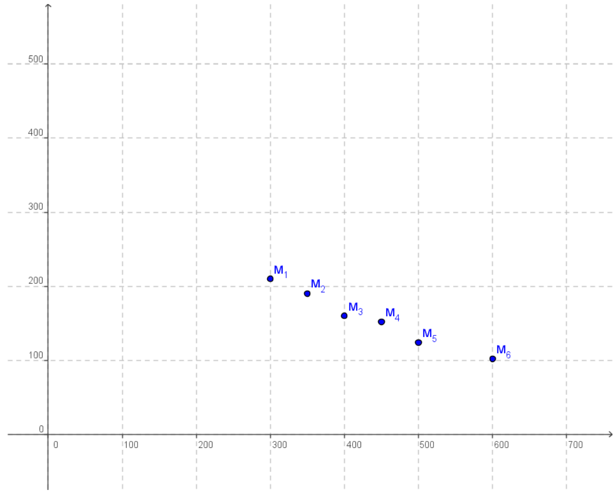
Certains nuages peuvent ne pas sembler être approchables par une quelconque courbe auquel cas les deux variables ne sont pas reliées entre elles.

L'exemple suivant suggère un ajustement par une fonction .... On parle alors d'ajustement quadratique :



Méthode graphique au jugé :

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



Propriété :
Soient A et B de coordonnées $(x_A; y_A)$ et $(x_B; y_B)$ deux points tels que ....
Alors la droite (AB) n'est pas parallèle à l'axe des ordonnées, elle a donc une équation de la forme .... et on a :

....

Exemple de savoir faire [Détermination de l'équation d'une droite dont on connaît les coordonnées de deux points] :
Soit $\mathcal{D}$ la droite passant par les points $M_1\,(300; 208)$, $M_2\,(350; 190)$ Son équation est de la forme ....

....

donc son équation est $y =$ ....

Or $M_1 \in \mathcal{D}$ donc ses coordonnées vérifient l'équation d'où

.... et ....

c'est à dire ....

L'équation est donc ....

Méthode des moindres carrés :

Avec les notations de la figure ci-dessous, étant donné un nuage de n points M_i, il existe une droite passant par le point moyen G et telle que la somme des carrés des écarts (ou *résidus*) $P_1M_1^2 + P_2M_2^2 + \ldots + P_nM_n^2$ soit minimale. Cette droite est appelée *droite de régression de y en x*. On peut montrer que son équation réduite est $y = mx + p$ avec :

$$m = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_p - \bar{x})(y_p - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_p - \bar{x})^2}$$

et

$$p = \bar{y} - m\bar{x}$$

En pratique, on utilisera la calculatrice pour l'obtenir.

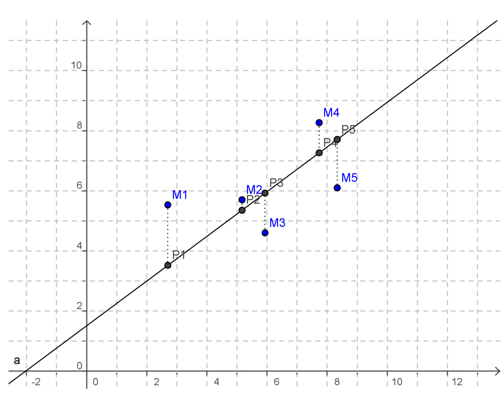
Exemple :

On reprend l'exemple précédent. Recherche de l'équation réduite à l'aide des formules :

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
	Total		

D'où :

$m = \ldots$ et $p = \ldots$



Utilisation de la calculatrice :

Recherche de l'équation réduite avec la calculatrice :

- TI 82 et plus :
  Aller dans le menu $\boxed{\text{STAT}}$ puis $\boxed{\text{EDIT}}$. Entrer les valeurs x_i dans la colonne L_1 et les valeurs y_i dans la colonne L_2. Quitter ($\boxed{\text{2nde}}$ $\boxed{\text{QUIT}}$) puis menu $\boxed{\text{STAT}}$ et $\boxed{\text{CALC}}$. Choisir $\boxed{\text{LinReg(ax+b)}}$ puis $\boxed{\text{2nd}}\boxed{\text{L1}}\boxed{,}\boxed{\text{2nd}}\boxed{\text{L2}}$ pour indiquer les deux colonnes à utiliser. Valider ensuite par $\boxed{\text{ENTER}}$.
- CASIO Graph 25 et plus :
  Aller dans le menu $\boxed{\text{STAT}}$ puis entrer les valeurs x_i dans la colonne 1 et les valeurs y_i dans la colonne 2. Sélectionner ensuite $\boxed{\text{CALC}}$. Choisir $\boxed{\text{SET}}$ et vérifier que la ligne « 2Var XList » est mise à « List1 » et que la ligne « 2Var YList » est mise à « List2 », sinon choisir le menu $\boxed{\text{LIST}}$ pour indiquer les numéros de liste adaptés. Taper ensuite sur $\boxed{\text{EXIT}}$ puis choisir $\boxed{\text{REG}}$ puis $\boxed{\text{X}}$.

Propriété et définition :

On a vu que la droite de régression linéaire de y en x rend minimal le nombre $P_1M_1^2 + \cdots + P_nM_n^2$. Le minimum est alors égal à $n(\sigma_y)^2(1 - (\frac{\sigma_{xy}}{\sigma_x\sigma_y})^2)$.
avec $\sigma_x = \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$,
$\sigma_y = \frac{1}{n}((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2)$,
et $\sigma_{xy} = \frac{1}{n}((x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}))$.
On appelle *coefficient de corrélation linéaire* le nombre défini par :

$$r = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

Plus la valeur absolue de ce nombre est proche de .... , plus la droite de régression linéaire est proche du nuage de points. On dit alors que les deux variables sont *corrélées*.

Remarques :

- En pratique, on lira le coefficient de corrélation linéaire sur la calculatrice en même temps que la droite des moindre carrés obtenue : il est noté r.
- Deux variables peuvent être corrélées sans que les variables sans qu'il n'y ait de relation de cause à effet entre elles.