Série statistique à deux variables

La statistique descriptive à deux variables a pour but de mettre en évidence une relation éventuelle qui peut exister entre **deux variables** d'une population, considérées **simultanément**.

I) Définition

On appelle *série statistique à deux variables* X et Y, le relevé simultané de la valeur de deux caractères statistiques X et Y. Elle est donc constituée d'une liste de couples de nombres $(x_i\,;\,y_i)$.

Variable X	x_1	x_2	$\ldots$	x_n
Variable Y	y_1	y_2	$\ldots$	y_n

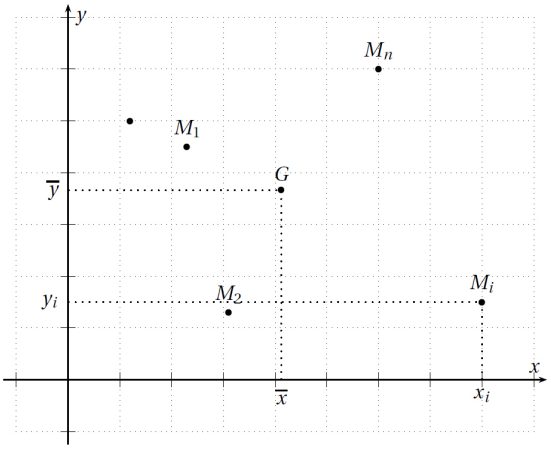
II)Nuage de points et point moyen

1) Le plan étant muni d'un repère orthogonal, on peut associer à chaque couple $(x_i\,;\,y_i)$ de la série statistique le point M de coordonnées $(x_i\,;\,y_i)$.
Le graphique ainsi obtenu constitue un *nuage de points*.

2) Le point moyen G du nuage de points est le point de coordonnées : $(\overline{x};\overline{y})$
où :

- l'abscisse est la moyenne de la série (x_i) : $\overline{x} = \frac{x_1+x_2+\ldots+x_n}{n}$
- l'ordonnée est la moyenne de la série (y_i) : $\overline{y} = \frac{y_1+y_2+\ldots+y_n}{n}$

On dit aussi que c'est le centre de gravité du nuage.



III) Ajustement

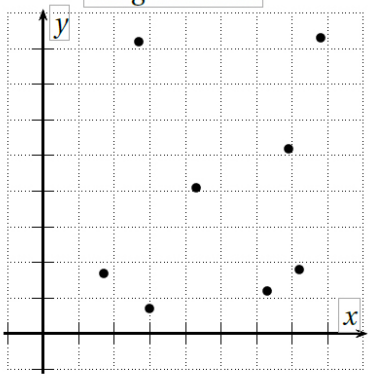
1) Corrélation

Il y a ***corrélation*** entre deux variables X et Y observées sur les individus d'une même population lorsqu'il y a une **relation** entre X et Y.

Remarque. L'existence d'une corrélation entre deux variables peut être décelée **dans un premier temps** à l'aide d'un nuage de points.

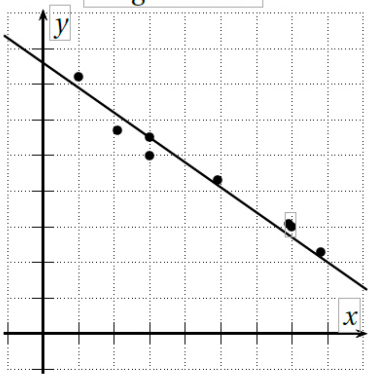
Exemples. Considérons les diagrammes suivants :

Diagramme A



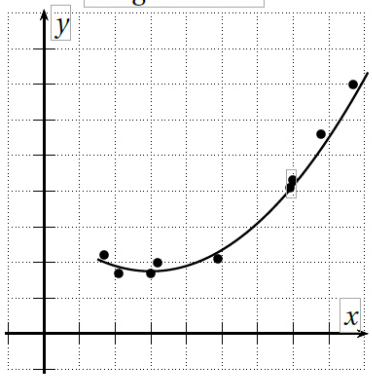
Absence de corrélation entre X et Y

Diagramme B



Corrélation entre X et Y

Diagramme C



Corrélation entre X et Y

2) Ajustement de y en x

Lorsque les valeurs de x sont connues, effectuer un ajustement de y en x d'un nuage de points consiste à trouver une fonction dont la courbe représentative d'équation $y = f(x)$ est la plus « proche » du nuage.

Remarques.

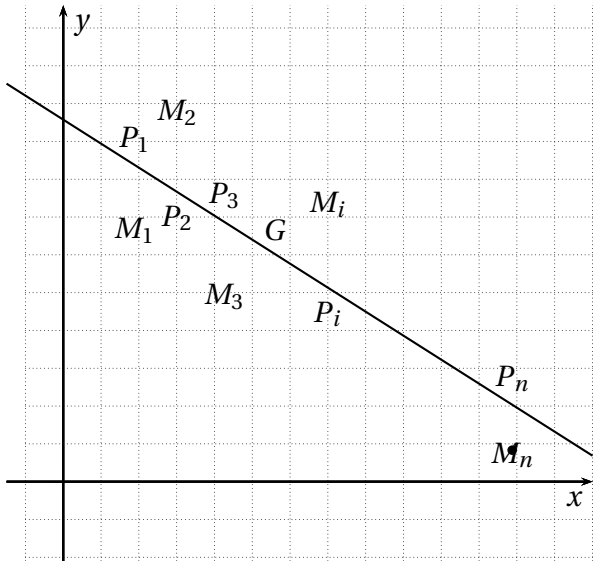
- Un ajustement permet de faire des estimations : *interpolation* (dans l'intervalle d'étude) et *extrapolation* (en dehors). Extroplation = prévision.
- Lorsque les points du nuage sont *presque alignés*, comme pour le diagramme B, on recherche une droite qui passe le plus près possibles des points. On effectue alors un ***ajustement affine***.
- On verra qu'il existe des ajustements qui ne sont pas affines, comme sur le diagramme C.

3) Ajustement affine

Une droite d'ajustement affine est une droite qui passe **au plus près** du nuage de points.

IV) Ajustement par la méthode des moindres carrés

On connaît les valeurs x_i, on cherche à obtenir une droite d'ajustement dont les valeurs y sont les plus proches possibles des y_i « verticalement ».
Les points M_1, M_2, …, M_n sont les points du nuage.
Les points P_1, P_2, …, P_n sont les points d'une droite $\mathscr{D}$ de mêmes abscisses que, respectivement, M_1, M_2, …, M_n, d'équation $y = ax + b$ qui est telle que la somme $S = M_1P_1^2 + M_2P_2^2 + \ldots + M_nP_n^2$ soit minimale.
On admet qu'une telle droite existe toujours et on dit que cette droite réalise un ajustement affine du nuage de y en x par la méthode dite *des moindres carrés*.



Elle passe toujours par le point moyen *G* du nuage.

1) Définition :

La droite $\mathscr{D}$ d'équation $y = ax + b$ obtenue par la méthode des moindres carrés est appelée *droite de régression de y en x.*

Remarque

- La droite de régression de y en x minimise la somme des carrés des distances en ordonnée

2) Corrélation :

Le nombre qui décrit **la validité** de la droite d'ajustement et qui mesure **le degré de dépendance** linéaire entre les variables x et y est le coefficient de corrélation (de Pearson), noté r.

3) Propriétés :

$-1 \leq r \leq 1$ r est toujours compris entre -1 et +1.
Si $r > 0$: entre x et y il y a une ***corrélation positive*** (dépendance linéaire positive).
Si $r < 0$: entre x et y il y a une ***corrélation négative***.
Si $r = 0$ ou voisin de 0 : il n' y a pas de dépendance linéaire entre x et y.
Si $r = 1$ ou $r = -1$: les points du nuage sont rigoureusement alignés ; la dépendance linéaire est parfaite.

Si $r \geq 0{,}95$ alors la corrélation linéaire entre X et Y est forte. Dans ce cas un ajustement affine est justifié (Les points du nuage sont dans une situation proche de l'alignement).

Une forte corrélation entre deux grandeurs x et y ne signifie pas nécessairement qu'il y a un **lien de causalité** entre ces grandeurs. Par exemple, il est possible que les deux grandeurs soient des **effets d'une même cause**. Par exemple, on peut constater une forte corrélation entre les notes en latin et en mathématiques dans un groupe d'étudiants, ce qui ne veut pas dire pour autant que la bonne note obtenue en latin favorise une bonne note en mathématiques ou vice-versa. Cf fiche Corrélation et causalité.pdf.

Remarque. On détermine a, b et r avec la calculatrice ou le tableur (voir tuto en vidéo).