Chapitre 2
Statistique à deux variables

I. Nuage de points

1) *Point moyen*

Définition :
Sur une population donnée, on s'intéresse à deux caractères.
Pour chacun des *n* individus de cette population, notons x_i et y_i les valeurs prises par chacun de ces caractères $(1 \leqslant i \leqslant n)$.
Les *n* couples $(x_i ; y_i)$ forment une **série statistique à deux variables**.

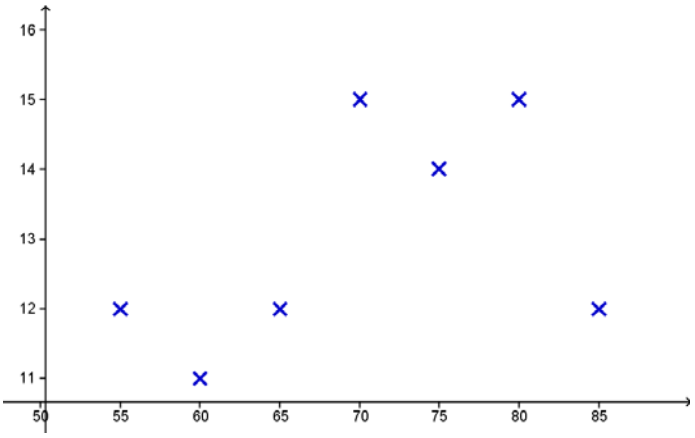
Exemple :
On mesure le poids x_i (en kg) et la tension artérielle systolique y_i (en mmHg) de 7 individus du même âge.
On a donc la série statistique à deux variables :

x_i	55	75	60	85	70	80	65
y_i	12	14	11	12	15	15	12

Définition :
Dans un repère orthogonal, l'ensemble des points $M_i(x_i ; y_i)$ est appelé le **nuage de points** associé à cette série statistique à deux variables.

Exemple :

Le graphique ci-contre représente le nuage de points associé à la série statistique de l'exemple précédent.



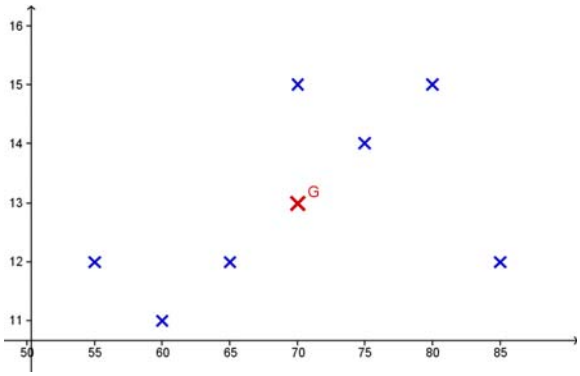
Définition :
Le point G de coordonnées $(\overline{x} ; \overline{y})$ avec :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

est appelé le **point moyen** du nuage de points associé à cette série statistique à deux variables.

Exemple :

Le point moyen du nuage ci-dessus est :
$G(70;13)$



2) Covariance

Définition :
On appelle **covariance** de x et y le nombre C_{xy} ou $cov(x,y)$, et défini par :

$$C_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

Remarques :

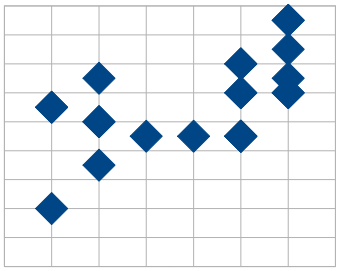
- La **variance** $V(X)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ qui sert à caractériser la dispersion d'un échantillon (et à calculer l'**écart type** $\sigma=\sqrt{V}$) vérifie donc $V(X)=C_{xx}$.
- La covariance est un nombre permettant d'évaluer le sens de variation (et ainsi la dépendance) de deux variables statistiques.

Exemple :
Dans le cas de la série statistique de l'exemple :

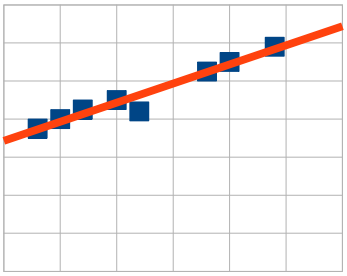
- Pour la série « poids » $X=\{55;75;60;85;70;80;65\}$ on a :
  $\overline{x}=70$; $V(X)=100$; $\sigma_x=10$
- Pour la série « tension artérielle » $Y=\{12;14;11;12;15;15;12\}$ on a :
  $\overline{y}=13$; $V(Y)\approx 2{,}29$; $\sigma_y\approx 1{,}51$
- Pour la série à deux variables X et Y : $C_{xy}=\frac{50}{7}\approx 7{,}14$

3) Ajustement

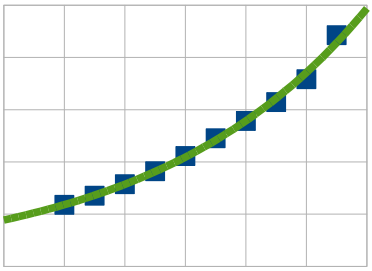
Lorsque deux variables X et Y sont liées l'une à l'autre, l'étude de la forme du nuage de points permet de modéliser cette relation.



Pas de relation



Relation affine



Relation exponentielle

Définition :
Effectuer un **ajustement de *y* en *x*** d'un nuage de points consiste à trouver une fonction *f* telle que la courbe d'équation $y=f(x)$ passe « le plus près possible » des points du nuage.

Remarque :
Pour les statisticiens, une formule telle que $y=f(x)$ est appelé un **modèle** (permettant d'étudier les relations entre les variables) : x est la variable explicative et y la variable expliquée.

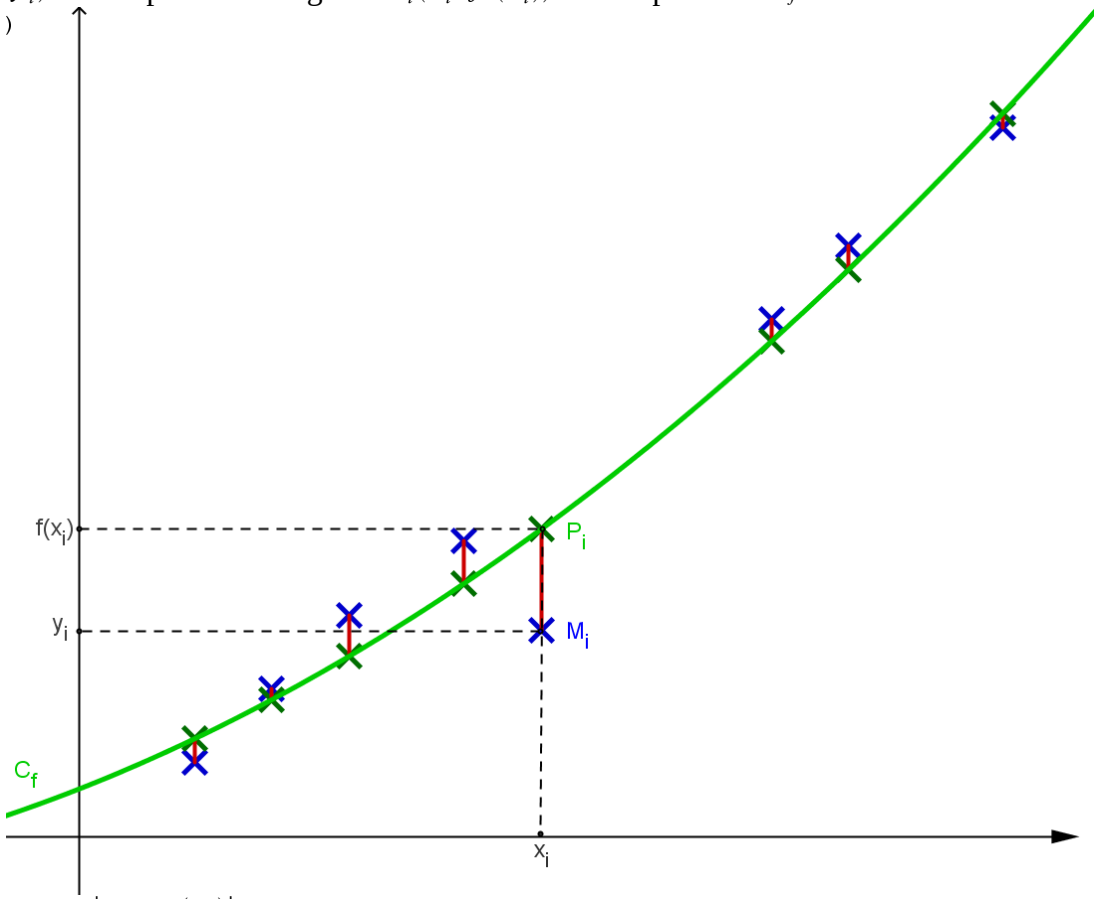
II. Méthode des moindres carrés

1) *Introduction*

Définition :
Effectuer un ajustement d'un nuage de points $(x_i;y_i)$ par la **méthode des moindres carrés** consiste à trouver une fonction *f* qui **minimise** la somme des carrés des écarts entre les valeurs y_i observées et les valeurs $f(x_i)$ données par le modèle.

La fonction *f* doit donc minimiser l'expression $\sum_{i=1}^{n}[y_i-f(x_i)]^2$.

Interprétation graphique :
$\mathscr{C}_f$ est la courbe représentative de la fonction *f*.
$M_i(x_i;y_i)$ est un point du nuage et $P_i(x_i;f(x_i))$ est un point de $\mathscr{C}_f$.



Alors $M_iP_i=|y_i-f(x_i)|$ et $(M_iP_i)^2=(y_i-f(x_i))^2$.

Donc $\mathscr{C}_f$ est la courbe qui minimise $\sum_{i=1}^{n}(M_iP_i)^2$, la somme des carrés des distances « verticales ».

Intérêt :
Pour une valeur donnée x_0 de la variable *X*, la fonction *f* permet de prévoir approximativement la valeur correspondante de *Y* (en calculant $f(x_0)$).

- Si x_0 appartient à l'intervalle d'observation des valeurs de *X*, on parle d'**interpolation**.
- Si x_0 n'appartient pas à l'intervalle d'observation des valeurs de *X*, on parle d'**extrapolation** (on suppose alors que le modèle est encore valable à l'extérieur de l'intervalle).

2) *Ajustement affine par moindres carrés*

Propriété (admise) :
Lors d'un ajustement affine de *y* en *x* par la méthode des moindres carrés, la droite *d* qui représente, dans un repère, la fonction *f* obtenue a pour coefficient directeur :

$$a=\frac{C_{xy}}{V(X)}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$ et passe par le point moyen $G=(\overline{x};\overline{y})$

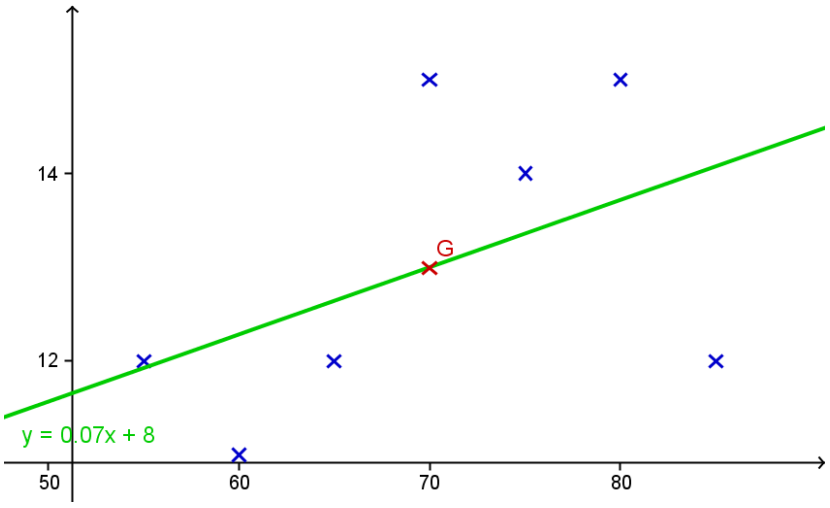
Remarque :

- La droite *d* est appelée la droite d'ajustement de *y* en *x* par la méthode des moindres carrés. Une équation de *d* est :
$$y=a(x-\overline{x})+\overline{y}$$
- On démontre que *d* est la seule droite pour laquelle la somme $\sum_{i=1}^{n}[y_i-(a\,x_i+b)]^2$ est minimale.

Exemple :
Pour la série :

x_i	55	75	60	85	70	80	65
y_i	12	14	11	12	15	15	12

On a donc :
$$a=\frac{C_{xy}}{V(X)}=\frac{\frac{50}{7}}{100}=\frac{50}{700}\approx 0{,}0714$$
et nous avions $G(70;13)$.



La droite d'ajustement par la méthode des moindres carrés a pour équation :
$$y=\frac{50}{700}(x-70)+13=\frac{5}{70}x+8$$

Utilisation de la calculatrice :