Statistique à deux variables

L'étude d'une population peut porter sur deux caractères distincts, il est donc intéressant d'étudier si ces deux caractères sont réellement indépendants ou non...
On propose dans tout le cours l'exemple suivant :
En prévision du lancement d'un nouveau produit, une société a effectué une enquête auprès de clients éventuels pour fixer le prix de vente de ce produit. Les résultats sont donnés dans le tableau ci-dessous :

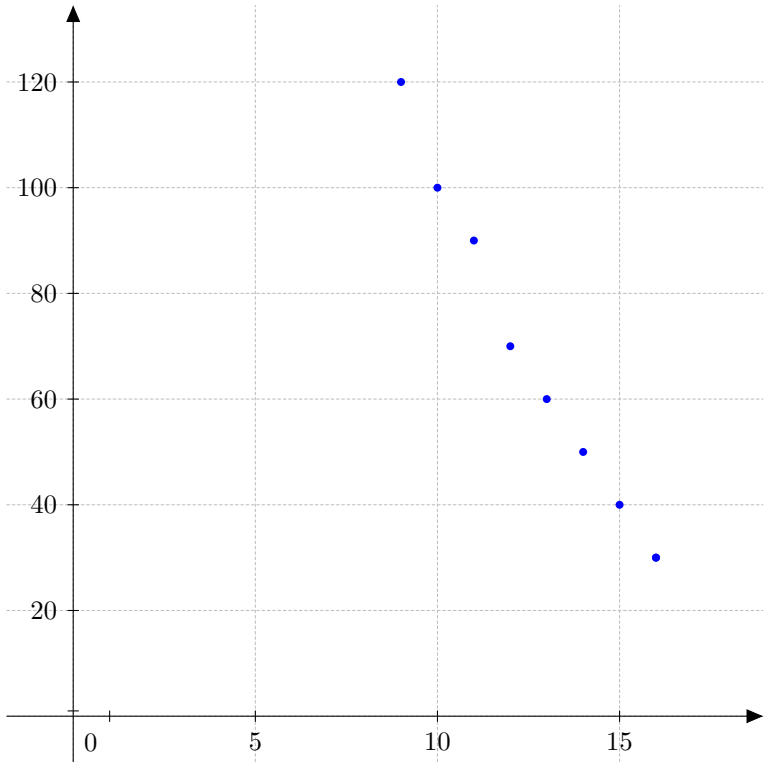
Prix de vente en euros x_i	9	10	11	12	13	14	15	16
Nombre d'acheteurs éventuels y_i	120	100	90	70	60	50	40	30

1 Nuage de points

Définition 1 :

Soit X et Y deux variables statistiques numériques observées sur n individus.
Dans un repère orthogonal, l'ensemble des n points de coordonnées (x_i, y_i) forme le nuage de points associé à cette série statistique.

Dans notre exemple, on représente dans un repère orthogonal le nuage des points M_i de coordonnées $(x_i\ ;\ y_i)$ correspondant au tableau précédent :



2 Point moyen

Définition 2 :

Soit une série statistique à deux variables, X et Y, dont les valeurs sont des couples $(x_i; y_i)$.
On appelle point moyen de la série le point G de coordonnées

- $x_G = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.
- $y_G = \dfrac{y_1 + y_2 + \cdots + y_n}{n}$.

Dans notre exemple :

- $x_G = \dfrac{9 + 10 + \cdots + 16}{8} = 12,5$.
- $y_G = \dfrac{120 + 100 + \cdots + 30}{8} = 70$.

On a donc $G(12,5; 70)$

3 Ajustement affine

3.1 Ajustement direct à la règle

L'idée est de tracer au juger une droite $\mathcal{D}$ passant le plus près possible des points du nuage et d'en trouver l'équation du type $y = ax + b$.

3.2 Méthode de Mayer

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage de point :
On partage les séries en deux, on obtient ainsi deux points moyen, la droite de Mayer est la droite passant par ces deux points.

Dans notre exemple, on cherche donc G_1,le point moyen de la sous série :

x_i	9	10	11	12
y_i	120	100	90	70

On obtient le point G_1 :

- $x_{G_1} = \dfrac{9 + 10 + 11 + 12}{4} = 10,5$.
- $y_{G_1} = \dfrac{120 + 100 + 90 + 70}{4} = 95$.

Soit $G_1 = (10,5; 95)$

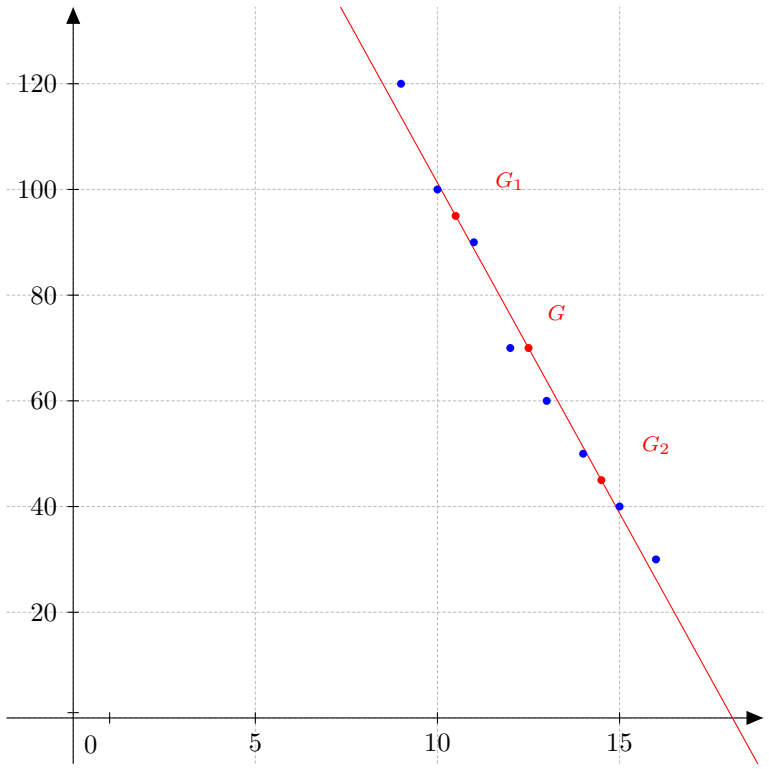
On fait de même pour le point G_2, avec la sous série :

x_i	13	14	15	16
y_i	60	50	40	30

On obtient le point G_2 :

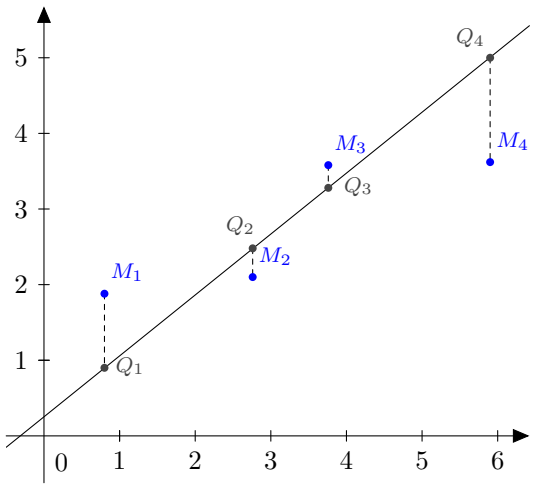
- $x_{G_2} = \dfrac{13 + 14 + 15 + 16}{4} = 14,5$.
- $y_{G_2} = \dfrac{60 + 50 + 40 + 30}{4} = 45$.

Soit $G_2 = (14,5; 45)$
La droite de Mayer est donc la droite qui passe par les points G_1 et G_2.
Ici, on trouve : $(G_1G_2) : y = -12,5x + 226,25$



3.3 Méthode des moindres carrés

On considère une série double représentée par un nuage de n points $M_i\,(x_i, y_i)$.
Pour une droite D quelconque, non parallèle aux axes, on note Q_i les projetés parallèlement à l'axe (yy') des points M_i sur la droite D.



Définition 3 :

On appelle droite de régression de y en x la droite D telle que :

$$S = \sum_{i=1}^{n} Q_i M_i^2 \text{ soit minimale.}$$

Propriété 1 :

Soit une série statistique à deux variables, X et Y, dont les valeurs sont des couples $(x_i; y_i)$. Une équation de la droite de régression D de y en x (appelée aussi droite d'ajustement de y en x par la méthode des moindres carrés) est donnée par :

$$D : y - \bar{y} = m\,(x - \bar{x}) \text{ avec } m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Définition 4 :

On appelle covariance des variables X et Y le réel noté σ_{xy} défini par :

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

En notant σ_x l'écart-type de la série X et m le coefficient directeur de la droite de régression D de y en x, on obtient :

$$m = \frac{\sigma_{xy}}{\sigma_x^2}$$

Dans notre exemple, on obtient :

- $\bar{x} = 12,5$ et $\bar{y} = 70$.
- $\sigma_{xy} = \frac{1}{8}(9 \times 120 + 100 \times 100 + ... + 16 \times 30) - 12,5 \times 70 = -66,25.$
- $\sigma_x^2 = \frac{1}{8}(9^2 + 10^2 + ... + 16^2) - 12,5^2 = 5,25$

On trouve donc : $m = \frac{\sigma_{xy}}{\sigma_x^2} = -\frac{66,25}{5,25} = -12,619.$
La droite de régression D de y en x est donc :
$D : y - 70 = -12,619\,(x - 12.5)$, ou encore $D : y = -12,619x + 227,738$

Définition 5 :

On appelle coefficient de corrélation linéaire de X et Y le réel défini par :

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

Dans notre exemple, on obtient :

- $\sigma_{xy} = \frac{1}{8}(9 \times 120 + 100 \times 100 + ... + 16 \times 30) - 12,5 \times 70 = -66,25.$
- $\sigma_x^2 = \frac{1}{8}(9^2 + 100^2 + ... + 16^2) - 12,5^2 = 5,25$ donc $\sigma_x = \sqrt{5,25} = 2,29$
- $\sigma_y^2 = \frac{1}{8}(120^2 + 100^2 + ... + 30^2) - 70^2 = 850$ donc $\sigma_y = \sqrt{850} = 29,15$

On a donc :

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} = -\frac{66,25}{2,29 \times 29,15} = -0,99$$

Si la valeur absolue du coefficient de corrélation linéaire est proche de 1, on parle de corrélation linéaire forte, on peut alors valablement effectuer un ajustement linéaire par D.