université de BORDEAUX

geii Génie électrique et informatique industrielle / iut de BORDEAUX

Année universitaire 2024-2025

Cours de statistique descriptive

BUT GEII

Semestre 4

Auteur : Florent ARNAL
Adresse électronique : florent.arnal@u-bordeaux.fr
Site : http://flarnal.e-monsite.com

Table des matières

I Statistique descriptive à une variable

I.1 Notion de variable

Les données statistiques se présentent sous la forme d'individus auxquels on associe des caractères (appelés également "variables statistiques"). L'ensemble des individus constitue un échantillon (ou encore une série statistique), formant ainsi un sous-ensemble d'un groupe appelé "population".
Les caractères statistiques peuvent être de plusieurs natures :

• Les variables qualitatives :
Les valeurs possibles d'une variable qualitative sont appelées les modalités.
Exemples : couleur de cheveux, sexe.
Remarque : On qualifie d'ordinale une variable qualitative pour laquelle la valeur mesurée sur chaque individu est numérique (par exemple, une appréciation où l'on qualifie le produit de Passable (1) à très bon (5)).
On les représente en utilisant :
▷ un diagramme circulaire, un diagramme en anneau (donut) ;
▷ un diagramme en barres.

• Les variables quantitatives (que l'on peut mesurer). Parmi elles, on distingue :
▷ Les variables quantitatives discrètes prenant des valeurs (isolées) dans un ensemble fini ou dénombrable ($\mathbb{N}$, $\mathbb{N}^*$, ...).
Exemple : Notes.
▷ Les variables quantitatives continues prenant leurs valeurs dans un intervalle de $\mathbb{R}$ (voire $\mathbb{R}$).
Exemples : taille d'un individu, pH, masse.
On les représente en utilisant, notamment, un nuage de points, un histogramme, ...

I.2 Représentation paramétrique et graphique

I.2.1 Paramètres de position

1. Médiane d'une série

Définition 1 : La médiane notée M_{e} est un nombre réel tel que la moitié des observations lui sont inférieures (ou égales) et la moitié supérieures (ou égales).

Remarque 1 : La médiane partage la série statistique $(x_i)_{1\leq i\leq N}$ en deux groupes de "même effectif" (les valeurs du caractère étant rangées par ordre croissant).
Cas d'une série discrète d'effectif total N(les observations étant rangées par ordre croissant).
On convient que :

- Si N impair, la médiane est telle que $M_e = x_{\frac{N+1}{2}}$.
- Si N est pair, on convient que la médiane est $M_e = \frac{x_{\frac{N}{2}}+x_{\frac{N}{2}+1}}{2}$.

Exemple 1 (Exemples et influence de valeurs extrêmes)
On considère les notes liées à une analyse sensorielle sur le goût de deux produits "concurrents" soumis à un jury (panels de consommateurs).

$$(1;3;4;5;5;6;6;7) \text{ et } (1;3;4;5;8;8;10)$$

Déterminons la médiane de ces deux séries.
✍

2. Moyenne d'une série

Définition 2 : La moyenne (pondérée) d'une série statistique $(x_i;n_i)_{1\leq i\leq p}$ telle que $\sum_{i=1}^{p} n_i = n$ est le réel noté $\overline{x}$ défini par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$$

Remarque 2 : La moyenne et la médiane d'une distribution statistique s'expriment dans la même unité que les valeurs prises par le caractère étudié.
Il est à noter que la médiane présente l'avantage par rapport à la moyenne de ne pas dépendre des valeurs extrêmes qui peuvent être suspectes voire aberrantes.

Propriété 1 : Soit $(x_i)_{1\leq i\leq n}$ une série statistique. On a :

$$\sum_{1\leq i\leq n}(x_i-\overline{x})=0$$

✍

I.2.2 Paramètres de dispersion

1. L'étendue

Définition 3 : L'étendue d'une série statistique est la différence entre la plus grande et la plus petite valeur observées. On la note souvent R voire e.

Exemple 2 *Les séries* $(1;3;4;6;8;8)$ *et* $(1;7;7;8;8;8)$ *ont la même étendue* $R=7$.

Remarque 3 : Comme l'étendue ne permet pas de différencier les séries, elle n'est pas ici une très bonne mesure de la dispersion. D'une façon générale, quand il existe des valeurs extrêmes, l'étendue est une mesure médiocre de la dispersion. Il est donc utile de faire appel à de nouveaux paramètres comme la variance et l'écart-type.

2. La variance et l'écart-type

Définition 4 : La variance de la série $(x_i;n_i)_{1\leq i\leq p}$ telle que $\sum_{i=1}^{p}n_i=n$ est définie par :

$$\sigma_X^2=\frac{1}{n}\sum_{i=1}^{p}n_i\left(x_i-\overline{x}\right)^2$$

Le réel σ_X est appelé écart-type de la série (x_i).
Le réel $SCE_X\sum_{i=1}^{p}n_i\left(x_i-\overline{x}\right)^2$ est appelé Somme des Carrés des écarts (à la moyenne). On note :

$$\sigma_X^2=\frac{SCE_X}{n}$$

Remarque 4 : La variance d'une série statistique correspond à la moyenne des carrés des écarts à la moyenne.
L'écart-type permet ainsi de mesurer la dispersion autour de la moyenne d'une série statistique. Plus l'écart-type est faible, plus la série est homogène.
Il s'exprime dans la même unité que les valeurs observées.

Remarque 5 : [Formule de Koenig-Huygens]
En développant la somme des carrés des écarts, on a :
$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\sum_{i=1}^{n}\left({x_i}^2+\overline{x}^2-2\overline{x}x_i\right)=\sum_{i=1}^{n}{x_i}^2+n\overline{x}^2-2\overline{x}\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}{x_i}^2+n\overline{x}^2-2\overline{x}n\overline{x}$

On a donc : $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}{x_i}^2-n\overline{x}^2$ ce qui implique

$$\sigma_X^2=\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\overline{x}^2 \tag{1}$$

Exemple 3 *Considérons la série* $(5;7;9;11;13)$.
On a : $\overline{x}=9$ *et* $SCE=(5-\overline{x})^2+\cdots+(13-\overline{x})^2=40$.
Ainsi : $\sigma_X^2=\dfrac{SCE}{n}=8$ *et* $\sigma_X=2\sqrt{2}$.

I.2.3 Quartiles d'une série statistique

Les valeurs étant rangées par ordre croissant, on rappelle que la médiane partage la population en deux populations (basse et haute).

Définition 5 : On appelle premier quartile Q_1 la médiane de la partie basse de la population, troisième quartile Q_3 la médiane de la partie haute de la population.

Remarque 6 : Les quartiles correspondent aux valeurs prises par le caractère qui partagent la série en quatre groupes de même effectif.
Le premier quartile peut être considéré comme la plus petite valeur du caractère telle qu' au moins 25 % des valeurs lui sont inférieures ou égales.

Définition 6 : L'intervalle $[Q_1;Q_3]$ est appelé intervalle interquartile.
La longueur de cet intervalle Q_3-Q_1, parfois noté EIQ, est appelée écart interquartile.

Remarque 7 : L'intervalle interquartile est un paramètre de dispersion qui élimine les valeurs extrêmes qui peuvent être douteuses, ce qui est un avantage par rapport à l'étendue.
Cependant, il ne tient compte que de 50 % de la population ce qui engendre une perte parfois conséquente d'information.

Les quartiles ne sont qu'un cas particulier de la notion de quantiles (https://fr.wikipedia.org/wiki/Quantile).

Bilan : Résumés d'une série statistique par ses paramètres

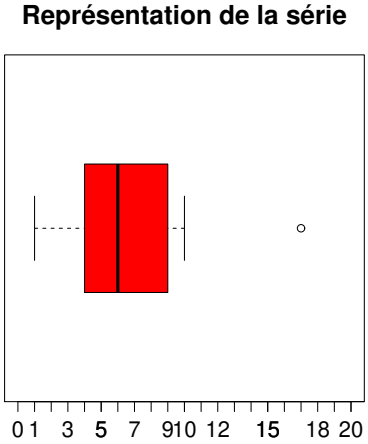
- Le couple (médiane ; étendue) est le plus "simple" à déterminer mais ne permet pas de situer les valeurs extrêmes par rapport à la moyenne.
- Le couple (médiane ; intervalle ou écart interquartile) est plus précis que le précédent car il est insensible aux valeurs extrêmes.

I.3 Représentation graphique d'une série à l'aide d'un boxplot

Pour représenter une série (variable) quantitative, il est souvent utile d'utiliser un boxplot (boîte à moustaches). On repère sur la boîte à moustaches d'une variable :

- l'échelle des valeurs de la variable, située sur l'axe vertical ;
- la valeur du premier quartile Q_1, correspondant au trait inférieur de la boîte ;
- la valeur de la médiane M_e, représentée par un trait horizontal à l'intérieur de la boîte ;
- la valeur du troisième quartile Q_3, correspondant au trait supérieur de la boîte ;
- les 2 «moustaches» inférieure et supérieure délimitent les valeurs dites adjacentes qui sont déterminées à partir de l'écart interquartile $(Q_3 - Q_1)$. La valeur extrême de la moustache inférieure correspond à la plus petite des valeurs supérieures ou égales à $Q_1 - 1,5 * (Q_3 - Q_1)$. La valeur extrême de la moustache supérieure correspond à la plus grande des valeurs inférieures ou égales à $Q_3 + 1,5 * (Q_3 - Q_1)$;
- d'éventuelles valeurs suspectes situées au-delà des valeurs extrêmes définies ci-dessus peuvent apparaître et sont représentées par des marqueurs (rond, étoile, etc.). Dans le cas où il n'existe pas de valeur suspecte, les valeurs extrêmes des moustaches correspondent aux valeurs minimales et maximales.

Exemple 4 *Considérons la série suivante :* $(1; 3; 4; 5; 6; 7; 9; 10; 17)$.
On a : $M_e = 6$; $Q_1 = 4$; $Q_3 = 9$. *Ainsi,* $Q_3 - Q_1 = 5$ *ce qui induit que :*
$Q_1 - 1,5 * (Q_3 - Q_1) = -3,5$ *et* $Q_3 + 1,5 * (Q_3 - Q_1) = 16,5$. *Il y a donc une valeur suspecte* (17). *Les valeurs extrêmes des «moustaches» étant égales à* 1 *et* 10.
On obtient la représentation graphique suivante :

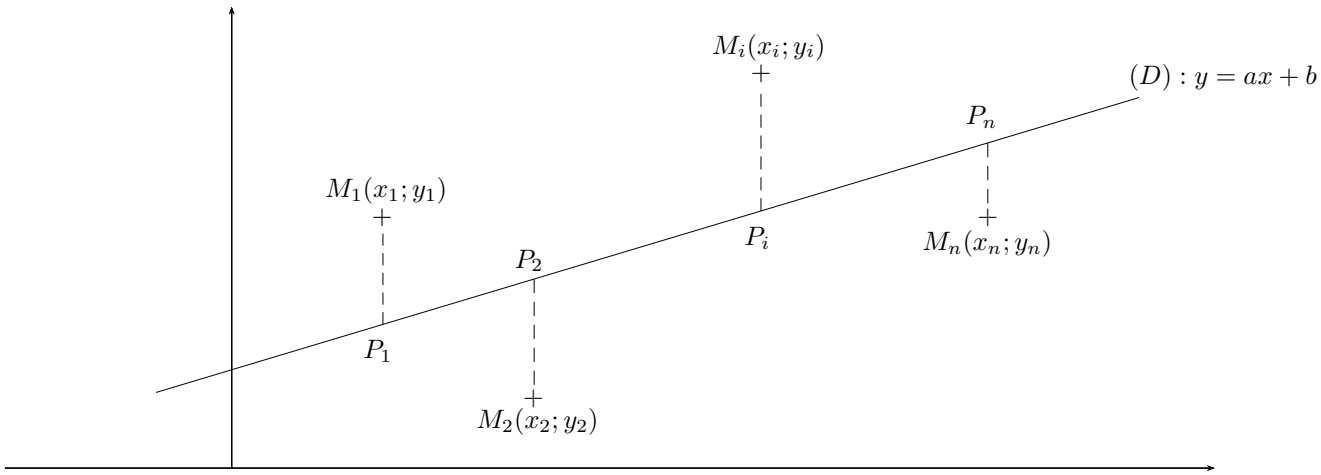


Remarque 8 : Dans la boîte à moustaches définie par TUKEY, la boîte a pour longueur la distance interquartile $(Q_3 - Q_1)$, et les moustaches sont basées usuellement sur 1,5 fois la longueur de la boîte. Dans ce cas, une valeur est suspecte (atypique) si elle dépasse de $1,5$ fois l'écart interquartile au dessous du premier quartile ou au dessus du troisième quartile.
Le choix de la valeur $1,5$ par TUKEY a une justification probabiliste.
En effet, si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir 99,3 % des observations. On ne devrait donc trouver que 0,7% d'observations suspectes (outliers).

II Statistique descriptive à 2 variables

II.1 Méthode des moindres carrés

On considère une série statistique double représentée par un nuage de points $M_i\,(x_i;y_i)_{1\leq i\leq n}$.
Soit (D) une droite d'ajustement. Pour tout entier naturel i tel que $1\leq i\leq n$, on note P_i le projeté de M_i sur la droite (D) parallèlement à l'axe des ordonnées.
Ajuster ce nuage de point par la méthode des moindres carrés, c'est déterminer la droite (D) pour que la somme $\sum_{i=1}^{n} {M_iP_i}^2$ soit minimale.
Illustration graphique :



II.2 Equations des droites de régression

Déterminons la droite d'équation $y = ax + b$ rendant minimale la somme $\sum_{i=1}^{n} {M_iP_i}^2$.

Ainsi, minimiser $\sum_{i=1}^{n} {M_iP_i}^2$ revient à déterminer le minimum de la fonction φ définie sur $\mathbb{R}^2$ par

$$\varphi(a,b) = \sum_{i=1}^{n} [(ax_i + b) - y_i]^2$$

On admet que si la fonction φ admet un extremum en $(x_0;y_0)$ alors ses dérivées partielles s'y annulent.
$\frac{\partial\varphi}{\partial a}(a,b) = 2\sum_{i=1}^{n} x_i[(ax_i + b) - y_i]$
$\frac{\partial\varphi}{\partial b}(a,b) = 2\sum_{i=1}^{n} [(ax_i + b) - y_i]$
$\frac{\partial\varphi}{\partial a}(a,b) = 0$ implique $\sum_{i=1}^{n} \left(a{x_i}^2 + bx_i - x_iy_i\right) = 0$ soit $a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_iy_i = 0$.
$\frac{\partial\varphi}{\partial b}(a,b) = 0$ implique $\sum_{i=1}^{n} [(ax_i + b) - y_i] = 0$ soit $a\sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i = 0$.
On a donc :

$$\sum_{i=1}^{n} x_iy_i = a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} x_i \quad \text{et} \quad \sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb$$

En posant $\alpha = \sum_{i=1}^{n} {x_i}^2$ et $\beta = \sum_{i=1}^{n} x_i$, on a :

$$\begin{pmatrix} \alpha & \beta \\ \beta & n \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_iy_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

En notant M la matrice définie par $M = \begin{pmatrix} \alpha & \beta \\ \beta & n \end{pmatrix}$, on a : $M\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_iy_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$.

On rappelle que :

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

$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} \left({x_i}^2 + \overline{x}^2 - 2\overline{x}x_i\right) = \sum_{i=1}^{n} {x_i}^2 + n\overline{x}^2 - 2\overline{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} {x_i}^2 + n\overline{x}^2 - 2\overline{x}n\overline{x}$

On a donc : $\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2$ ce qui implique

$$\mathbb{V}(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \overline{x}^2 \tag{2}$$

La matrice M est inversible si et seulement si $n\alpha - \beta^2 \neq 0$ soit $n\alpha - (n\overline{x})^2 \neq 0$.

M inversible équivaut donc à $\frac{1}{n}\alpha - \overline{x}^2 \neq 0$. D'après ce qui précède, on en déduit que :

M inversible ssi $\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \neq 0$.

Or : $\sum_{i=1}^{n} (x_i - \overline{x})^2 = 0 \Longleftrightarrow \forall i \in \{1; 2; ...; n\},\ x_i - \overline{x} = 0 \Longleftrightarrow \forall i \in \{1; 2; ...; n\},\ x_i = \overline{x}$.

Ainsi : M inversible si et seulement si les x_i ne sont pas tous égaux.

Dans ce cas, on a : $M^{-1} = \frac{1}{n\alpha - \beta^2}\begin{pmatrix} n & -\beta \\ -\beta & \alpha \end{pmatrix}$ avec $\begin{pmatrix} a \\ b \end{pmatrix} = M^{-1}\begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$. Ainsi :

$$a = \frac{1}{n\alpha - \beta^2}\left(n\sum_{i=1}^{n} x_i y_i - \beta\sum_{i=1}^{n} y_i\right) = \frac{n\sum_{i=1}^{n} x_i y_i - n^2\overline{x}\overline{y}}{n\sum_{i=1}^{n} {x_i}^2 - (n\overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2}$$

Quant à b, on a :

$$b = \frac{1}{n\alpha - \beta^2}\left(-\beta\sum_{i=1}^{n} x_i y_i + \alpha\sum_{i=1}^{n} y_i\right) = \frac{-n\overline{x}\sum_{i=1}^{n} x_i y_i + n\overline{y}\sum_{i=1}^{n} {x_i}^2}{n\sum_{i=1}^{n} {x_i}^2 - (n\overline{x})^2} = \frac{-\overline{x}\sum_{i=1}^{n} x_i y_i + \overline{y}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2}.$$

Cette relation peut s'écrire : $b = \dfrac{\overline{y}\left(\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2\right) - \overline{x}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}\right)}{\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2} = \overline{y} - a\overline{x}.$

En conclusion, φ peut admettre un extremum uniquement en $(a; b)$ tel que

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

En considérant $b = \overline{y} - a\overline{x}$, on obtient :

$\varphi(a, b) = \sum_{i=1}^{n} (a(x_i - \overline{x}) + (\overline{y} - y_i))^2 = \sum_{i=1}^{n} \left(a^2(x_i - \overline{x})^2 + (\overline{y} - y_i)^2 + 2a(x_i - \overline{x})(\overline{y} - y_i)\right)^2$.

$\varphi(a, b)$ peut donc s'écrire sous la forme :

$$\varphi(a, b) = a^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + 2a\sum_{i=1}^{n}(x_i - \overline{x})(\overline{y} - y_i) + \sum_{i=1}^{n}(\overline{y} - y_i)^2$$

On constate que, si $b = \overline{y} - a\overline{x}$ alors $\varphi(a, b)$ est un polynôme du second degré avec un coefficient du terme dominant positif. Cette fonction admet donc un minimum ce qui permet de conclure que les réels a et b obtenus précédemment permettent de minimiser $\sum_{i=1}^{n} M_i P_i{}^2$.

A noter que, tout comme la formule de la variance (2) , $\sum_{i=1}^{n}(x_i-\overline{x})(\overline{y}-y_i)$ peut s'écrire autrement en développant :
$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}(x_iy_i - x_i\overline{y} - \overline{x}y_i + \overline{xy}) = \sum_{i=1}^{n} x_iy_i - n\overline{xy} - n\overline{xy} + n\overline{xy} = \sum_{i=1}^{n} x_iy_i - n\overline{xy}$ On a donc :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - n\overline{xy} \tag{3}$$

Une équation de la droite d'ajustement de Y en X est donc $Y = aX + b$ avec

$$a = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

Théorème-Définition 1 :
On considère deux variables X et Y prenant respectivement les valeurs $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$.
La covariance de X et Y est le réel, noté $\mathbb{C}ov\,(X,Y)$ voire σ_{XY}, défini par :

$$\mathbb{C}ov\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y})$$

ou encore

$$\mathbb{C}ov\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}$$

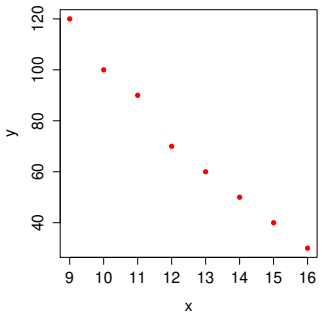
Propriété 2 : La droite de régression de Y en X a pour équation $Y = aX + b$ avec

$$a = \frac{\mathbb{C}ov\,(X,Y)}{\mathbb{V}(X)} \text{ et } b = \overline{y} - a\overline{x}$$

Exemple 5 *Considérons le nombre d'acheteurs potentiels d'un produit en fonction de son prix de vente.*

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

Le nuage de points associé à la série $(x_i;y_i)$ laisse apparaître un relation linéaire entre le nombre Y d'acheteurs potentiels et le prix X.



Calculons les coefficients de la droite de régression de Y en X :
$\overline{x} = 12,5$ *;* $\overline{y} = 70$ *;* $Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \overline{xy} = -66,25$ *et* $\sigma_X^2 = 5,25$.
Ainsi : $a = \dfrac{Cov(X,Y)}{\sigma_X^2} = \dfrac{-66,25}{5,25} \simeq -12,62$ *et* $b = \overline{y} - a\overline{x} \simeq 227,74$.
La droite de régression de Y en X a pour équation

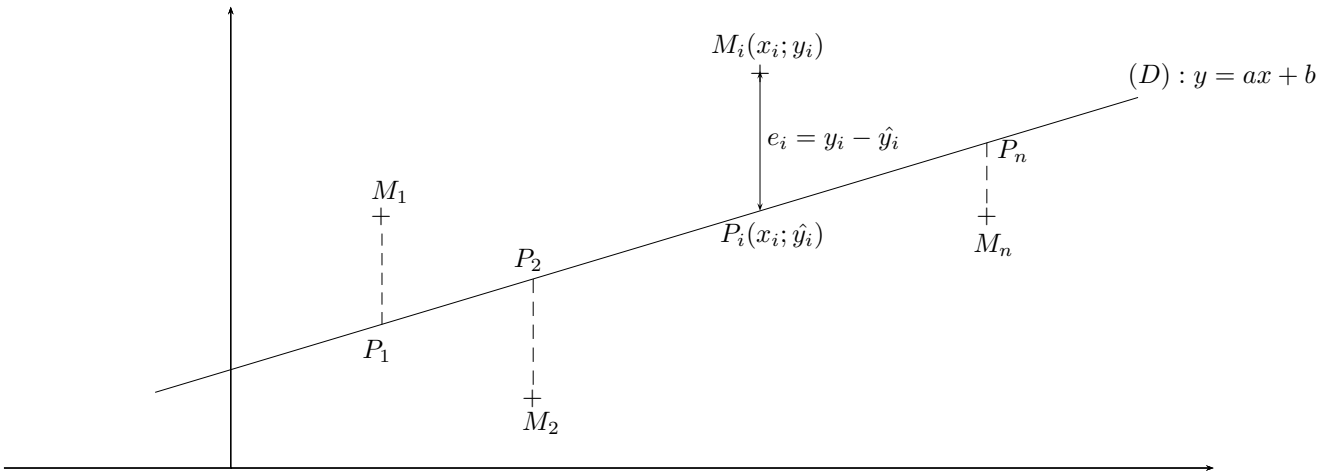
$$Y = -12,62X + 227,74$$

II.3 Résidus de la régression

Définition 7 : [Résidus]
On considère deux variables X et Y prenant respectivement les valeurs $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$.
On appelle résidus de Y par rapport à X les écarts , notés e_i, entre les valeurs observées de la variable dépendante y_i et les valeurs correspondantes $\hat{y}_i = ax_i + b$ calculées à l'aide de l'équation de la droite de régression de Y en X . On a donc, pour tout $i \in [\![1; n]\!]$:

$$e_i = y_i - \hat{y}_i$$

Illustration graphique :



Remarque 9 : Les valeurs $\hat{y}_i$ sont aussi appelées valeurs estimées de la variable dépendante Y.

Propriété 3 : Les résidus sont de somme et de moyenne nulle.

En effet :
✍

Remarque 10 : Les séries $(y_i)_{1\leq i\leq n}$ et les estimations $(\hat{y}_i)_{1\leq i\leq n}$ ont donc la même moyenne. On a donc :

$$\overline{y} = \overline{\hat{y}}$$

Exemple 6 *Calculons les résidus associés à l'exemple précédent.*
Dans un premier temps, nous déterminons les estimations en utilisant l'équation $\hat{y}_i = -12,62x_i + 227,74$.
On obtient : $114,16$; $101,54$; $88,92$; $76,30$; $63,68$; $51,06$; $38,44$; $25,82$.
Ainsi, les résidus étant définis par $e_i = y_i - \hat{y}_i$*, on obtient :*

$$5,84 \quad ; \quad -1,54 \quad ; \quad 1,08 \quad ; \quad -6,30 \quad ; \quad -3,68 \quad ; \quad -1,06 \quad ; \quad 1,56 \quad ; \quad 4,18$$

II.4 Coefficients de corrélation et de détermination

Ces coefficients permettent d'apprécier (partiellement) la pertinence et la qualité d'un ajustement.

Définition 8 : [Coefficient de corrélation]
On appelle coefficient de corrélation linéaire de la série $(x_i, y_i)_{1\leq i\leq n}$ le nombre réel, noté R ou $\rho(X,Y)$, défini par :

$$R = \frac{\mathbb{C}ov(X,Y)}{\sigma_X \sigma_Y}$$

Propriété 4 : Comme σ_X et σ_Y sont positifs alors R, a et $\mathbb{C}ov(X,Y)$ sont de même signe.

Définition 9 : (Les différentes SCE)

On considère $(x_i; y_i)_{1\leq i\leq n}$ une série statistique à deux variables constituée de n couples.

- La variabilité totale (des y_i) est définie par :

$$SCE_{\text{totale}} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- La variabilité expliquée SCE_{exp} (par l'ajustement affine) est définie par :

$$SCE_{\text{exp}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

  Elle correspond à la SCE des estimations $\hat{y}_i$.
- La variabilité résiduelle, associée aux résidus et notée SCE_{res}, est définie par :

$$SCE_{\text{res}} = \sum_{i=1}^{n} {e_i}^2$$

$SCE_{\text{totale}} = \sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2$.
En développant, on a :
✍

Or, $\sum_i (y_i - \hat{y}_i)\ (\hat{y}_i - \overline{y}) = \sum_i [(y_i - \overline{y}) - a(x_i - \overline{x})]\ [a(x_i - \overline{x})]$
✍

On en déduit la propriété suivante :

Propriété 5 :

$$SCE_{\text{totale}} = SCE_{\text{exp}} \ + \ SCE_{\text{res}}$$

On peut dès lors expliciter les différentes SCE ainsi :
$SCE_{\text{totale}} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = n\mathbb{V}(Y)$
$SCE_{\text{exp}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 = na^2\mathbb{V}(X) = n\left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}(X)}\right)^2 \mathbb{V}(X) = n\frac{\mathbb{C}ov^2(X,Y)}{\mathbb{V}(X)} = nR^2\mathbb{V}(Y)$
$SCE_{\text{res}} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = n\mathbb{V}(Y)(1 - R^2)$.
✍

Définition 10 : [Coefficient de détermination]
Le coefficient de détermination d'une série statistique à deux variables X et Y correspond à R^2 ou $\rho^2(X,Y)$.

Il s'avère que :
✍
$\frac{SCE_{\text{exp}}}{SCE_{\text{totale}}} =$

La définition du coefficient de détermination conduit à la propriété suivante :

Propriété 6 : [Coefficient de détermination et SCE]
Le coefficient de détermination R^2 d'une série statistique à deux variables X et Y vérifie :

$$R^2 = \frac{SCE_{\text{exp}}}{SCE_{\text{totale}}}$$

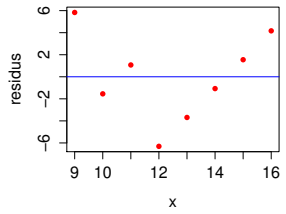
Il correspond ainsi à la part de variabilité de Y expliquée par la régression.

La relation $SCE_{\text{totale}} = SCE_{\text{exp}} + SCE_{\text{res}}$ induit la propriété suivante :

Propriété 7 : Le coefficient de détermination R^2 d'une série statistique à deux variables X et Y vérifie :
- $R^2 \leq 1$
- $-1 \leq R \leq 1$

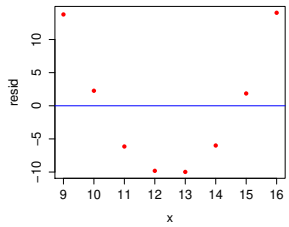
À retenir :
L'ajustement sera d'autant meilleur que R^2 est proche de 1.
En pratique : on considère que si $0,8 \leq |R| \leq 1$, il y a une forte corrélation.
Si R^2 est proche de 1, on peut dire qu'il existe une corrélation importante entre les deux variables.
Mais ceci n'implique pas nécessairement l'existence d'une relation directe de cause à effet entre les deux variables.
La pertinence d'un ajustement affine peut être vérifiée par l'étude des résidus dont la représentation graphique ne doit faire apparaître aucune tendance.
Sinon, malgré un coefficient de corrélation linéaire (ou de détermination) élevé, un autre ajustement sera souvent plus cohérent ...

Exemple 7 *Voici ci-dessous une représentation graphique des résidus calculés précédemment. Il ne laisse apparaître aucune tendance.*



Le coefficient de détermination $\rho^2(X,Y)$ est environ égal à $0,98$. L'ajustement affine est donc pertinent.

Exemple 8 *Le graphique ci-dessous représente les résidus associés à un ajustement affine avec un coefficient de détermination $\rho^2(X,Y)$ est environ égal à $0,99$.*



On voit clairement que les résidus sont positifs aux extrêmités et négatifs au centre.
Le modèle linéaire n'est pas pertinent (un ajustement quadratique le serait).

À retenir (Covariance et indépendance) :

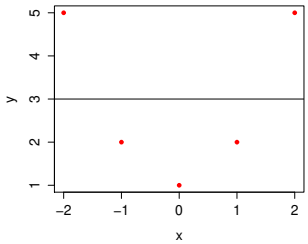
1. Si X et Y sont indépendantes alors $\mathbb{C}ov(X,Y)=0$ (on a également $\rho(X,Y)=0$).
2. Par contre, la réciproque n'est pas vraie comme le montre l'exemple ci-dessous.
   Si $\mathbb{C}ov(X,Y)=0$ alors cela n'implique pas que X et Y sont indépendantes.

Exemple 9

Considérons deux variables X et Y définies par le tableau ci-dessous :

x_i	-2	-1	0	1	2
y_i	5	2	1	2	5

Le nuage de points associé à la série $(x_i;y_i)$, représenté ci-dessous, ne laisse apparaître aucune liaison linéaire entre X et Y.



On a : $\mathbb{C}ov(X,Y)=0$ mais X et Y ne sont pas indépendantes car $Y=X^2+1$.