Chapitre 2 : Statistique descriptive à deux variables

- Il arrive souvent que l'étude statistique porte simultanément sur deux caractères $X \ et \ Y$.
- Pour présenter les résultats, on peut utiliser deux modes de représentation :

➢ le tableau de contingence

➢ la représentation graphique.

1. Tableau de contingence (ou de corrélation) :

Soient :

➢ les valeurs possibles $x_1, x_2, \ \ldots, x_p$ pour le caractère statistique X (ou les modalités, ou les classes).

➢ les valeurs possibles $y_1, y_2, \ \ldots, y_q$ pour le caractère statistique Y (ou les modalités, ou les classes).

➢ l'effectif n_{ij} correspond à chaque observation $\left(X = x_i \ , \ Y = y_j\right)$ avec $\sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = n$.

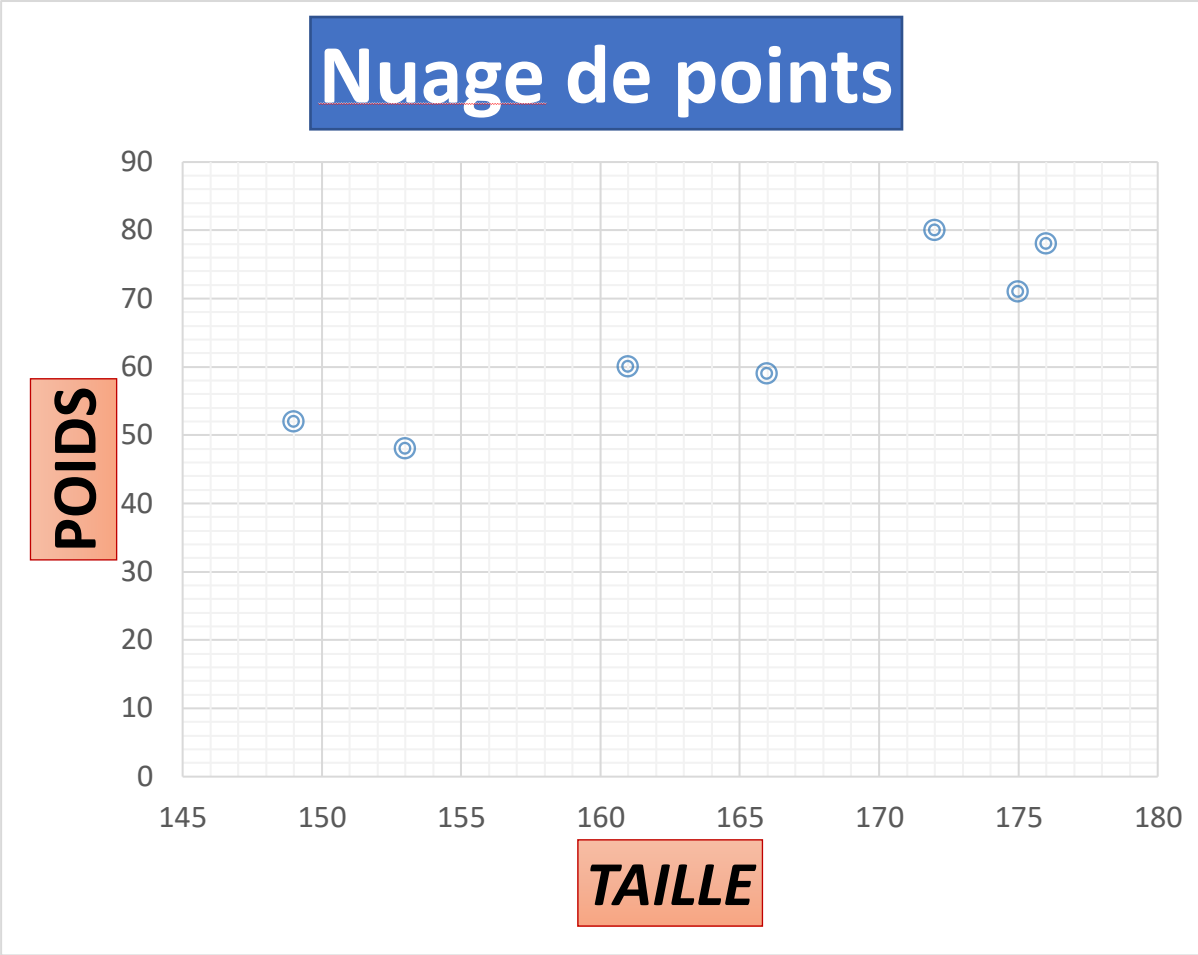
➢ si n désigne l'effectif total, la fréquence correspondante est $f_{ij} = \frac{n_{ij}}{n}$ avec $\sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij} = 1$.

➢ Ces renseignements se présentent souvent avec un tableau à double entrée s'appelant tableau de contingence.

Y / X	y_1	.....	y_j	.....	y_q
x_1	n_{11}	.....	n_{1j}	.....	n_{1q}
.	.		.		.
.	.		.		.
.	.		.		.
x_i	n_{i1}	.....	n_{ij}	.....	n_{iq}
.	.				.
.	.				.
.	.				.
.	.				.
x_p	n_{p1}	.....	n_{pj}	.....	n_{pq}

2. Nuage de points :

- La représentation graphique n'est valable que si les données sont quantitatives.
- Envisageons l'exemple où les caractères X et Y (quantitatifs) sont la taille (en cm) et le poids (en kg).
- Pour chaque sujet, on a un couple de valeurs (une taille, un poids).
- En portant en abscisse la taille et en ordonnée le poids, on définit chaque sujet par un point sur le diagramme.
- L'ensemble des points constitue le nuage de points.



3. Distributions marginales :

A partir de la distribution statistique du couple $(X\ ,\ Y)$, on peut déduire la distribution statistique concernant le caractère X seul, et celle qui est relative au caractère Y seul :

- $(X = x_i)$ a pour effectif : $\boldsymbol{n_{i.}} = \sum_{j=1}^{q} n_{ij}$ et pour fréquence: $\boldsymbol{f_{i.}} = \frac{\boldsymbol{n_{i.}}}{\boldsymbol{n}}$.
- $\left(Y = y_j\right)$ a pour effectif: $\boldsymbol{n_{.j}} = \sum_{i=1}^{p} n_{ij}$ et pour fréquence: $\boldsymbol{f_{.j}} = \frac{\boldsymbol{n_{.j}}}{\boldsymbol{n}}$.
- La détermination des effectifs $\boldsymbol{n_{i.}}$ et $\boldsymbol{n_{.j}}$ se fait à partir du tableau de contingence par addition suivant les lignes et les colonnes, et en reportant les résultats en marge du tableau.

Y / X	y_1	.....	y_j	.....	y_q	Total
x_1	n_{11}	.....	n_{1j}	.....	n_{1q}	$n_{1.}$
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
x_i	n_{i1}	.....	n_{ij}	.....	n_{iq}	$n_{i.}$
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
x_p	n_{p1}	.....	n_{pj}	.....	n_{pq}	$n_{p.}$
Total	$n_{.1}$	.....	$n_{.j}$	.....	$n_{.q}$	n

$$\sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = n$$

4. Distributions conditionnelles :

A- Distribution conditionnelle de Y pour $X = x_i$:

C'est la distribution des $n_{i.}$ observations vérifiant la condition $X = x_i$ et réparties selon les valeurs prises par Y.

Pour ceci, il suffit d'extraire du tableau de contingence la ligne correspondante à $X = x_i$.

$Y/X = x_i$	y_1	.....	y_j	.....	y_q	Total
Effectif	n_{i1}	.....	n_{ij}	.....	n_{iq}	$n_{i.}$

On obtient des fréquences conditionnelles en divisant les effectifs par $n_{i.}$:

$f_{j/i} = \frac{n_{ij}}{n_{i.}}$ avec $\sum_{j=1}^{q} f_{j/i} = 1$.

B- Distribution conditionnelle de X pour Y= y_j:

De la même manière, c'est la distribution des $n_{.j}$ observations vérifiant la condition $Y = y_j$.

$X/Y = y_j$	x_1	.....	x_i	.....	x_p	Total
Effectif	n_{1j}	.....	n_{ij}	.....	n_{pj}	$n_{.j}$

En divisant les effectifs par $n_{.j}$, on obtient les fréquences conditionnelles :

$f_{i/j} = \frac{n_{ij}}{n_{.j}}$ avec $\sum_{i=1}^{p} f_{i/j} = 1$.

5. Paramètres d'une série statistique double :

A- Moyennes et variances marginales :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i \quad et \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j$$

- Le point de coordonnées $(\bar{x}, \bar{y})$ s'appelle le point moyen.
- $S^2_{echX} = \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}\ x_i^2\right) - (\bar{x})^2$
- $S^2_{echY} = \left(\frac{1}{n}\sum_{j=1}^{q} n_{.j}\ y_j^2\right) - (\bar{y})^2$

B- Covariance :

- La notion de covariance généralise à deux variables la notion de variance.
- Sa formule est la suivante :

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\,(x_i - \bar{x})(y_j - \bar{y})$$
$$= \frac{1}{n}\left(\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j\right) - (\bar{x}\,\bar{y}).$$

- Propriétés :

1. $Cov(X,\,Y) = Cov(Y,\,X)$.
2. $Cov(X,\,X) = Var(X) = S^2_{echX}$.
3. $Cov(aX+b,\,cY+d) = ac\,Cov(X,\,Y)$ avec $a,\,b,\,c,\,d \in \mathbb{R}$.
4. $|Cov(X,\,Y)| \leq S_{echX}\,.\,S_{echY}$

C- Coefficient de corrélation :

La covariance n'a pas de signification concrète. Pour cela, on doit passer à un indicateur interprétable qui s'appelle le coefficient de corrélation linéaire :

$$r = \frac{Cov\,(X,\,Y)}{S_{echX}\,.\,S_{echY}}.$$

- Propriétés :

1. r est toujours entre -1 et 1: $-1 \leq r \leq 1$.
2. Le nuage de points $(x_i\,,\,y_j)$ est une droite si, et seulement si $r = 1$ (droite à pente positive) ou $r = -1$ (droite à pente négative).
3. Si $|r| \approx 1$, on dit qu'il existe une forte corrélation linéaire entre X et Y.

D- Coefficient de détermination :

- Le coefficient de détermination linéaire de Pearson, noté R, est une mesure de la qualité de la prédiction d'une régression linéaire.

$$R = r^2$$

- R est toujours entre 0 et 1.
- Plus le coefficient de détermination R se rapproche de 0, plus le nuage de points se disperse autour de la droite de régression. Au contraire, plus le R tend vers 1, plus le nuage de points se resserre autour de la droite de régression.

5. Ajustement : Droite de régression

- On considère un nuage de points $(x_i\ ,\ y_j)$.
- L'allure de ce nuage et des considérations sur le phénomène étudié peuvent suggérer une relation fonctionnelle entre X et Y, par exemple:

$$y = ax + b,\ \ y = ax^b\ ,\ \ y = a\ln(x) + b\ ,\ \ ....$$

- Quand les points du nuage sont à peu près alignés, on retient comme modèle $y = ax + b$ (ajustement affine), ou $y = ax$ (ajustement linéaire) quand la droite passe par l'origine.

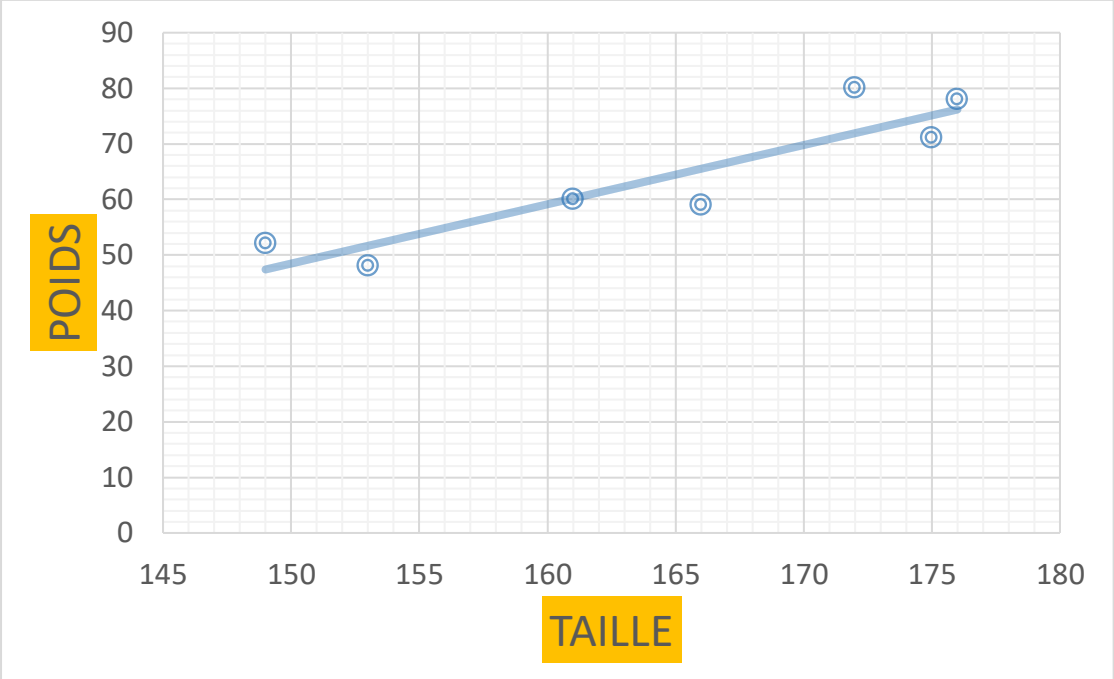
❖ Droite de régression de Y par rapport à X: au sens de moindre carrés

- La droite d'équation

$y = ax + b$ qui passe par le point moyen $M(\bar{x}\ ,\ \bar{y})$ et dont la pente

$a = \frac{Cov(X,Y)}{S^2_{echX}}$ s'appelle droite de régression de Y par rapport de X.

$$b = \bar{y} - a\bar{x}$$



Exercice : Pour une personne, on a fait varier l'intensité du travail fourni X (en kilojoules par minute) et on a relevé la fréquence cardiaque Y (nombre de battements par minute). On a obtenu les résultats suivants :

x_i	9,6	12,8	18,4	31,2	36,8	47,2	49,6	56,8
y_j	70	86	90	104	120	128	144	154

a) Représenter ces données par un nuage de points.

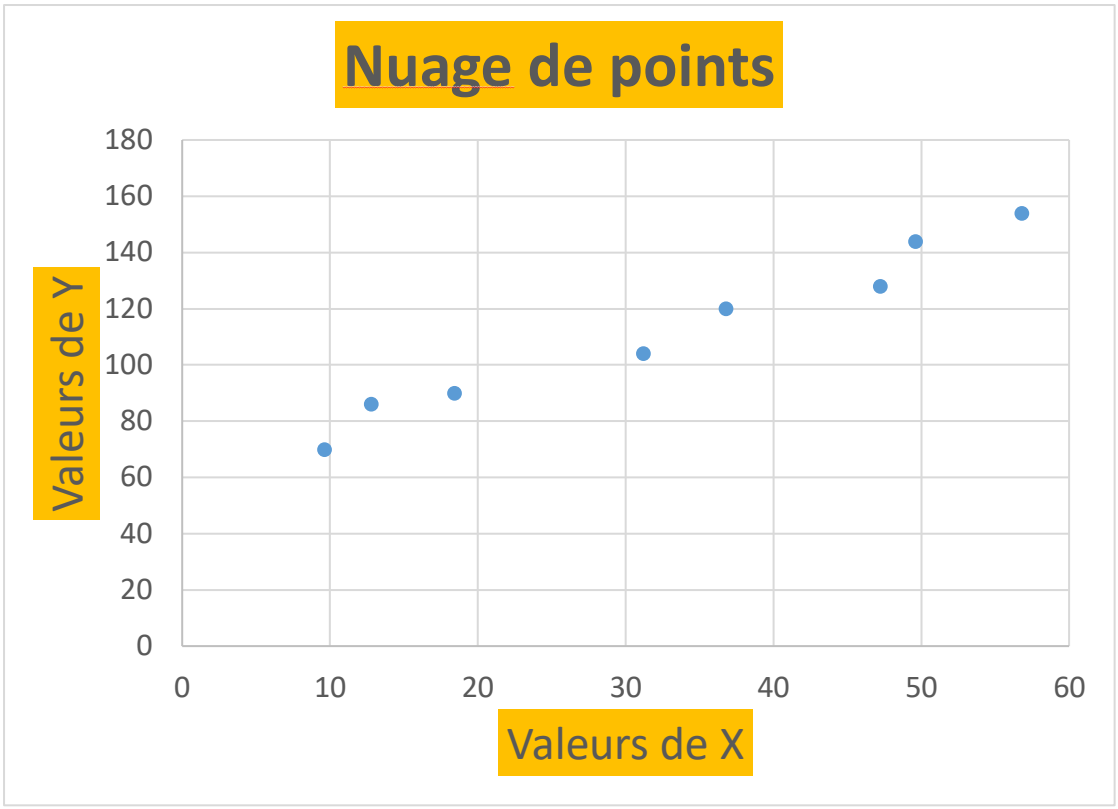
b) Calculer le coefficient de corrélation linéaire r.

c) Déterminer la droite de régression de Y par rapport à X.

d) Estimer la fréquence cardiaque lorsque l'intensité de travail fourni est 30 kilojoules par minute ; puis lorsqu'elle est de 75.

Solution :

a) **Nuage de points :** On observe que les points expérimentaux sont à peu près alignés, ce qui justifie l'hypothèse d'un modèle d'ajustement affine.



b) Calculs :

Puisque tous les effectifs sont égaux à 1 :

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 32{,}8 \;\; et \; \bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = 112.$$

x_i^2	**92,16**	**163,84**	**338,56**	**973,44**	**1354,24**	**2227,84**	**2460,16**	**3226,24**	**Somme= 10836,48**
y_i^2	4900	7396	8100	10816	14400	16348	20736	23716	Somme= 106412

$$S_{echX}^2 = \frac{1}{n}\sum_{i=1}^{8} x_i^2 - (\bar{x}^2) = \frac{10836{,}48}{8} - (32{,}8)^2 = 278{,}72$$

$$S_{echY}^2 = \frac{1}{n}\sum_{i=1}^{8} y_i^2 - (\bar{y}^2) = \frac{106412}{8} - (112)^2 = 757{,}5$$

$S_{echX} = 16{,}695$ et $S_{echY} = 27{,}523$

$x_i \times y_i$	672	1100,8	1656	3244,8	4416	6041,6	7142,4	8747,2	Somme=33020,8

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{8}\sum_{i=1}^{8} x_i y_i - (\bar{x}\bar{y}) = \frac{33020{,}8}{8} - (32{,}8 \times 112) = 454$$

Donc le coefficient de corrélation

$$r = \frac{Cov\,(X,\,Y)}{S_{echX}\,.\,S_{echY}} = \frac{454}{16{,}695 \times 27{,}523} = 0{,}988.$$

c) Droite de régression :

$y = ax + b$ avec $a = \frac{Cov\,(X,Y)}{S^2_{echX}} = \frac{454}{278{,}72} = 1{,}629$ et $b = \bar{y} - a\bar{x} = 112 - (1{,}629 \times 32{,}8) = 58{,}569$

Donc $y = 1{,}629x + 58{,}569$.

d) Estimation affine :

Pour estimer la fréquence cardiaque pour une intensité du travail fourni égale à 30, il suffit de remplacer x par 30 dans la droite de régression :

$$y = (1{,}629 \times 30) + 58{,}569 = 107{,}44.$$

La même chose pour $x = 75$:

$$y = (1{,}629 \times 75) + 58{,}569 = 180{,}74.$$