

# Statistiques descriptives bivariées

## *Objectifs*

- Observer simultanément des individus d'une population sur deux caractères
- Mesurer un lien éventuel entre deux caractères en utilisant un résumé chiffré qui traduit l'importance de ce lien
- Qualifier ce lien :
  - en cherchant une relation numérique approchée entre deux caractères quantitatifs
  - en cherchant des correspondances entre les modalités de deux caractères qualitatifs

2 types de variables  $\Rightarrow$  3 types de croisements :

- qualitatif  $\times$  qualitatif
- qualitatif  $\times$  quantitatif
- **quantitatif  $\times$  quantitatif**

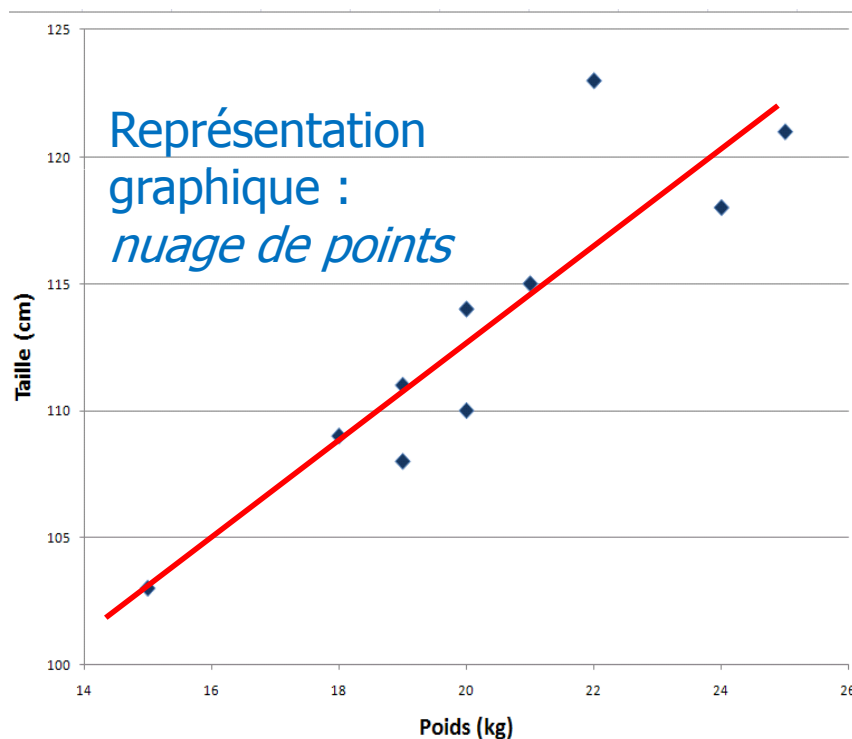
# Croisement quantitatif $\times$ quantitatif

# Croisement Quantitatif × Quantitatif

## Droite de régression

*Exemple : Etude du lien entre l'âge et le poids chez les enfants de 6 ans*

Enfant	1	2	3	4	5	6	7	8	9	10
Taille	121	123	108	118	111	109	114	103	110	115
Poids	25	22	19	24	19	18	20	15	20	21



On considère X et Y deux variables qualitatives sur une population de taille n.

On note  $\{x_i\}_{i=1,\dots,n}$  la série observée pour X et  $\{y_i\}_{i=1,\dots,n}$  la série observée pour Y.

L'objectif est de trouver une fonction f telle que

$$y_i \approx f(x_i).$$

On se restreint aux fonctions affines  $f(x) = ax + b$

Et on cherche les coefficients a et b qui minimisent *l'erreur quadratique moyenne*

$$EQ(a,b) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (ax_i + b))^2$$

# Croisement Quantitatif × Quantitatif

## Droite de régression

On obtient les coefficients :

$$\hat{a} = \frac{C_{xy}}{S_x^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

où  $c_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  est la **covariance** entre X et Y.

Garde les mêmes propriétés que la covariance théorique

$y = \hat{a}x + \hat{b}$  est appelée **droite de régression** de Y en X. Elle traduit les variations de Y qui peuvent être expliquées par X.

Attention la droite de régression de X en Y n'est nécessairement la même que celle de Y en X

Exemple : Etude du lien entre l'âge et le poids chez les enfants de 6 ans

$\bar{x}$	$\bar{y}$	$S_x^2$	$S_y^2$	$C_{xy}$
113,20	20,30	38,62	8,46	16,27

L'équation de la droite de Y en X :  $y = 0,42x - 27,38$

L'équation de la droite de X en Y :  $y = 1,92x - 74,15$

# Croisement Quantitatif × Quantitatif

## Covariance et coefficient de corrélation

A partir de la covariance, on définit le *coefficient de corrélation linéaire* (coefficient de Pearson) à valeurs dans  $[-1,1]$

$$r_{xy} = \frac{C_{xy}}{S_x S_y}$$

1.  $|r|$  est proche de 1 alors X et Y sont très liés entre eux par une droite affine.
2.  $r < 0$  : globalement X et Y varient en sens inverse .
3.  $r > 0$  : globalement X et Y varient dans le même sens .
4.  $|r| \cong 0$  : on ne peut rien dire sur un lien éventuel entre X et Y.

### Remarques :

- Le coefficient de corrélation correspond à la covariance des séries centrées et réduites

$$\hat{a} = \frac{C_{xy}}{S_x^2} = r_{xy} \frac{S_y}{S_x}$$

Exemple : Etude du lien entre l'âge et le poids chez les enfants de 6 ans

On trouve

$$r_{xy} = 0,90$$

- $r_{xy} \cong 1 \Rightarrow$  L'équation de droite est donc pleinement justifiée
- $r_{xy} > 0 \Rightarrow$  plus la taille est grande et plus le poids est important (et vice-versa)

On appelle *prévisions* les valeurs données par la droite de régression. Pour chaque point  $x_i$  de la série observée, on peut calculer la prévision (i.e. une valeur approchée de  $y_i$  par la droite de régression)

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

### Propriétés :

La variable Y et la partie de cette variable expliquée par la droite de régression ont la même moyenne :

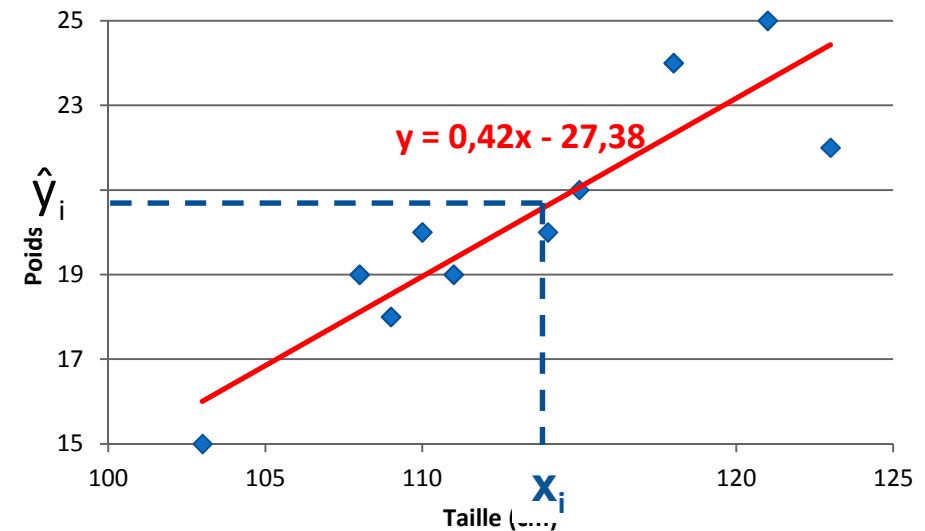
(\*)  $\bar{\hat{y}} = \bar{y}$

mais pas la même variance :

(\*\*)  $s_{\hat{y}}^2 = s_y^2 \times r_{xy}^2$

⇒ La variance de Y expliquée la droite de régression est plus petite que la variance de Y

⇒ La variance de Y expliquée la droite de régression est d'autant meilleure que le coefficient de Pearson est proche de 1 en valeur absolue.



(\*) démonstration en TD

(\*\*)  $\hat{y} = \hat{a}x + \hat{b} \Rightarrow s_{\hat{y}}^2 = (\hat{a})^2 s_x^2 = \left( r_{xy} \frac{s_y}{s_x} \right)^2 s_x^2 = r_{xy}^2 s_y^2$

## Les résidus

On appelle *résidu* l'écart entre la valeur observée  $y_i$  et la valeur prédite

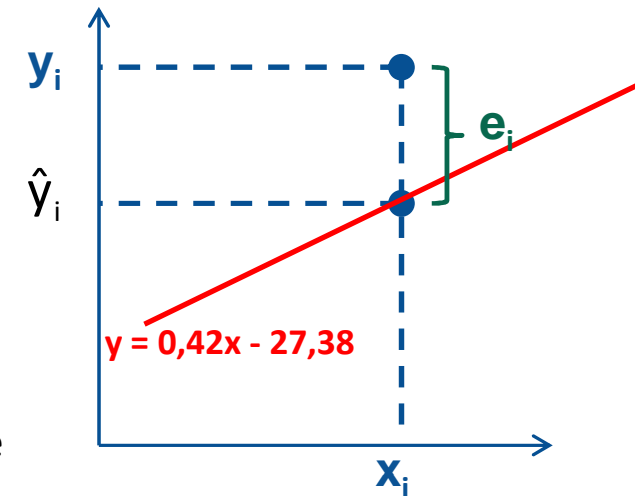
$$e_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + \hat{b})$$

On calcule alors l'erreur globale

$$EQ(\hat{a}, \hat{b}) = s_e^2 = s_y^2 (1 - r_{xy}^2)$$

⇒ L'erreur globale est proportionnelle à la variance de la variable Y

⇒ L'erreur est d'autant plus petite que le coefficient est proche de 1 en valeur absolue



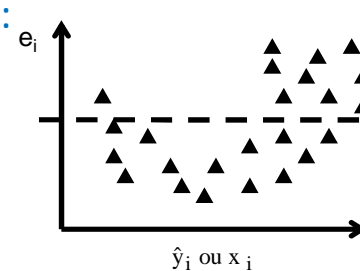
$$s_e^2 = s_{\{y - (\hat{a}x + \hat{b})\}}^2 = s_{y - \hat{a}x}^2 = s_y^2 + (\hat{a})^2 s_x^2 - 2\hat{a}C_{xy} = s_y^2 + \left(r_{xy} \frac{s_y}{s_x}\right)^2 s_x^2 - 2\left(r_{xy} \frac{s_y}{s_x}\right)(s_x s_y r_{xy}) = s_y^2 (1 - r_{xy}^2)$$

### Validité du modèle :

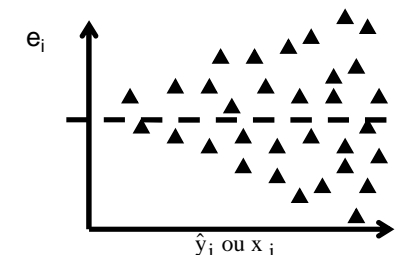
Un modèle est explicatif s'il ne reste plus « d'information » dans les résidus pouvant expliquer  $y$ .

On vérifie (graphiquement) les trois points suivants :

- La moyenne des résidus est nulle
- Les résidus ne sont pas corrélés
- La variance des résidus est constante



Résidus corrélés



Variance non constante

# Croisement Quantitatif × Quantitatif

## Décomposition de la variance

Nous avons vu que la variance de la variable Y n'est pas égale à la variance des valeurs prédites. Cependant elle peut se décomposer comme suit :

$$\underbrace{S_y^2}_{\text{variance totale}} = \underbrace{S_{\hat{y}}^2}_{\text{variance expliquée}} + \underbrace{S_e^2}_{\text{variance résiduelle}}$$

Démonstration  
en TD

En divisant cette égalité par la variance totale, on obtient le pourcentage de variance de y expliquée par le modèle, ce qu'on appelle encore le coefficient de détermination,

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = r_{xy}^2 \in [0,1]$$

Dans l'exemple précédent, on a la décomposition de la variance suivante :

<i>Variances</i>	
Régression	6,17
Résidus	1,44
Total	7,61

D'où  $R^2 = 6,17/7,61 = 0,81$ . Cela signifie que 81% de la variation des poids observés est expliquée par la droite de régression : poids = 0,42 × taile - 27,38



Un modèle peut s'avérer très précis pour ajuster les valeurs observées mais très mauvais en ce qui concerne la prévision de nouvelles valeurs.

Observation est *influente* si une faible variation entraîne une modification importante des caractéristiques du modèle.

### Détection des observations influentes (*atypiques/outliers*)

- On retire la  $i^{\text{ème}}$  observation de l'ensemble des données
- On ajuste un nouveau modèle sans la  $i^{\text{ème}}$  donnée
- On calcule  $y_{(-i)}$  la prévision de  $y_i$  avec le nouveau modèle
- On calcule le résidu,  $e_{(-i)} = y_i - y_{(-i)}$

✓ Un résidu important signale une observation influente

On a  $e_{(-i)} = \frac{e_i}{1 - h_{ii}}$  où  $h_{ii} = \frac{1}{n} + \frac{1}{(n-1)} \frac{(x_i - \bar{x})^2}{s_x^2}$

Le PRESS (*predicted residual sum of squares*) donne une indication sur les qualités prédictives du modèle

$$\text{PRESS} = \frac{1}{n} \sum_{i=1}^n e_{(-i)}^2 = \frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{(1 - h_{ii})^2}$$

Un *levier*

$$1/n \leq h_{ii} \leq 1$$

proche de 1 indique une observation influente

Sous l'hypothèse de normalité des résidus, les *résidus studentisés*,

$$\delta_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

doivent être compris (IDC) **entre  $\pm 2$**

- Les droites de régression n'explique que les liaisons linéaires.
- Si X et Y sont liées par une relation de la forme  $Y=aX^2$  alors  $r_{XY}=0$   
Le coefficient de corrélation linéaire de Pearson ne peut pas détecter cette liaison.
- Il n'existe pas de mesure universelle pour détecter des relations quelconques
- On essaie par des transformations de se ramener à une droite affine

Famille	Fonctions	Transformation	Forme affine
exponentielle	$y = a.e^{bx}$	$y' = \log(y)$	$y' = \log(a) + b.x$
puissance	$y = ax^b$	$y' = \log(y) \quad x' = \log(x)$	$y' = \log(a) + b.x'$
inverse	$y = a + \frac{b}{x}$	$x' = \frac{1}{x}$	$y' = a + b.x'$
logistique	$y = \frac{1}{1 + e^{-(a.x+b)}}$	$y' = \log\left(\frac{y}{1-y}\right)$	$y' = a.x + b$