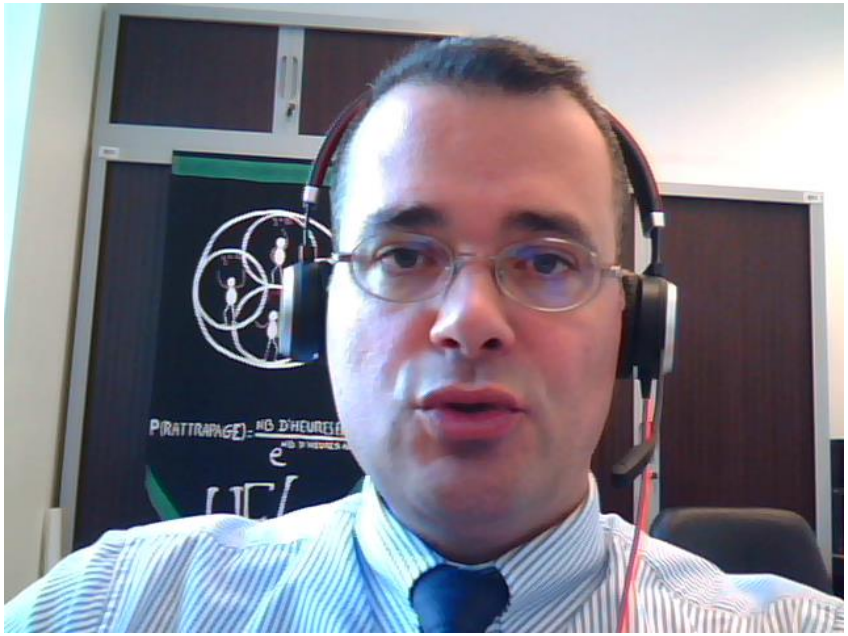
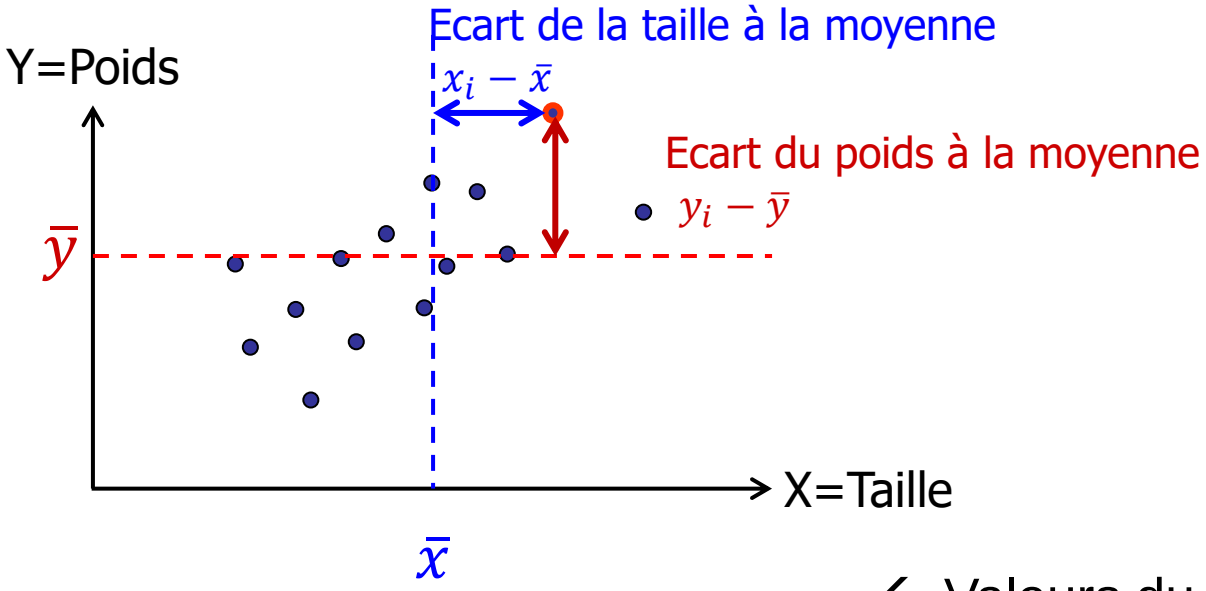
Tester l'association linéaire entre deux variables quantitatives : corrélation, régression linéaire simple



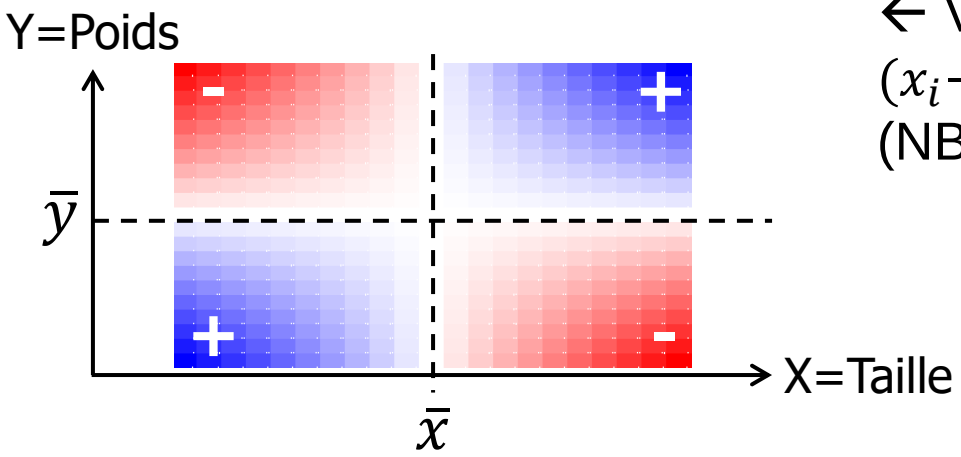
Introduction

- Etudier la relation bivariée, entre deux variables quantitatives X et Y, mesurées sur les mêmes individus
- Exemples :
  - Phénomènes différents mesurées chez les mêmes individus : poids et taille
  - Mesures successives d'un même phénomène : pression artérielle le matin et le soir
  - Etude de la variabilité inter-opérateur : mesure de pression artérielle par l'infirmière X et l'infirmière Y
- Questions :
  - Sont-elles indépendantes, en relation linéaire, ou +/- associées ?
  - Varient-elles dans le même sens ou non ?

Rappel sur la Covariance
Exemples : poids et taille



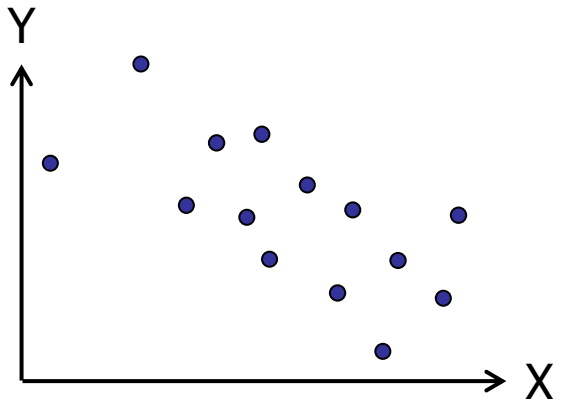
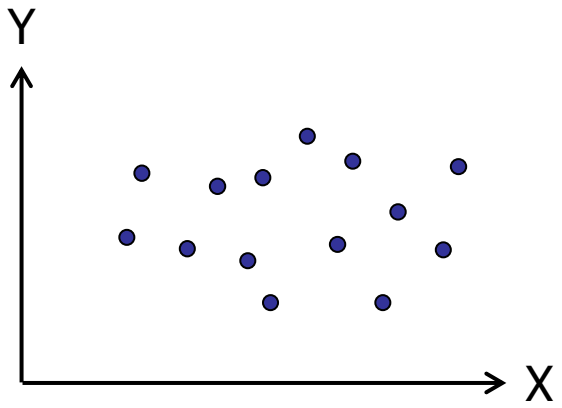
1 point =
1 individu *i* $\{x_i, y_i\}$



← Valeurs du produit
$(x_i - \bar{x}).(y_i - \bar{y})$ pour chaque individu i
(NB : nul sur les lignes des moyennes)

=> Dans ce cas, la moyenne de ce produit devrait être nettement positive.

Rappel sur la Covariance
Autres exemples

Variables :	Graphique :	Valeur de $moyenne[(x_i - \bar{x}).(y_i - \bar{y})]$
Y = montants des aides sociales perçues (€) X = revenus du travail (€)		Fortement négative
Y = distance travail (km) X = taille (cm)		Nulle

Valeur de
$moyenne[(x_i - \bar{x}).(y_i - \bar{y})]$

Rappel sur la Covariance
Calcul dans un échantillon

- Covariance sur l'échantillon (empirique), avec une ligne pour chacun des *n* individus *i* :

$$cov_{ech}(X,Y)$$

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

$$= (\sum_{i=1}^{n} x_i.y_i)/n - \bar{x}.\bar{y}$$

- Calcul analogue à l'estimateur <u>biaisé</u> de la variance :

$$cov_{ech}(X,X) = var_{ech}(X) = s^2{}_{ech}(X)$$

Rappel sur le Coefficient de corrélation linéaire de Pearson

Définition

- Coefficient insensible aux ordres de grandeur contrairement à la covariance
- Coefficient théorique de corrélation linéaire dans la population (rhô) : ρ
- Coefficient empirique de corrélation linéaire dans un échantillon :

$$r = \frac{\mathrm{COV}_{ech}(X,Y)}{S_{ech\,X} . S_{ech\,Y}}$$

  Attention : utiliser les estimateurs biaisés des écarts types
- Varie entre -1 et 1

Rappel sur le Coefficient de corrélation linéaire de Pearson

Interprétation

- En population :
  - Absence de relation linéaire entre X et Y ⇔ ρ=0
  - X et Y indépendantes ⇒ ρ=0
  - Y=aX+b (relation strictement linéaire a≠0) ⇔ ρ=1 si a>0 ou -1 si a<0
  - X et Y corrélées linéairement
    - dans le même sens ⇔ ρ ∈]0 ; 1]
    - en sens contraire ⇔ ρ ∈ [-1 ; 0[
    - fortement ⇔ |ρ| ∈ [0,5 ; 1]
    - faiblement ⇔ |ρ| ∈]0 ; 0,5[
- Sur l'échantillon :
  - idem, en tenant compte des fluctuations liées à l'échantillonnage
- ATTENTION : ρ=0 signifie absence de relation linéaire, mais ρ=0 ne signifie pas toujours indépendance !
  Exemples de nuages de points avec ρ=0 mais X et Y non indépendantes :

Tester l'absence de relation linéaire entre deux variables quantitatives

- Dans un échantillon
  - Taille n
  - Observation de X et Y
  - Calcul de r, par exemple non-nul
- Extrapolation en population :
  - X et Y sont-elles en relation linéaire ?
  - Le coefficient ρ est-il non-nul ?
    Test de nullité du coefficient de corrélation
  - Soit la relation $Y=aX+b+\varepsilon$, le coefficient *a* est-il non-nul ?
    Test de nullité de la pente de la droite de régression
  - Calcul différent, mais résultat identique
  - ATTENTION : ne teste pas l'indépendance stricto sensu… => importance de l'exploration graphique !

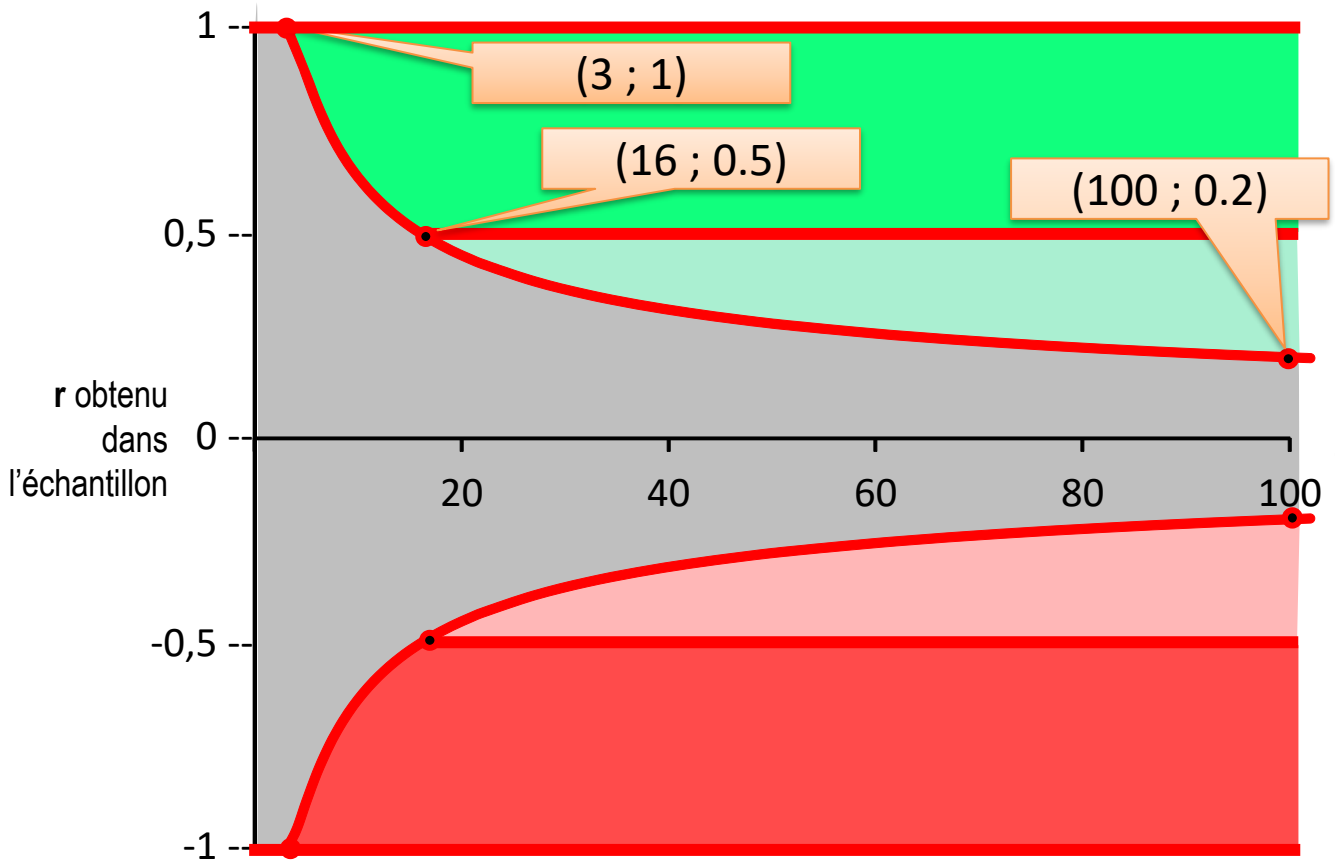
Test de nullité du coeff. de corrélation

- Test bilatéral
- Hypothèse H_0 :
  - $\rho=0$
  - Autrement dit, X et Y sont ne sont pas corrélées linéairement
- Condition de validité :
  - (…) acquises si X et Y suivent une distribution normale
  - (ce sont celles de la régression linéaire simple, cf. post)
- Réalisation (autre approche existe) :
  La quantité t suit une loi de Student à ν=n-2 ddl
  - Si $|t|>t\alpha\nu$ alors rejet de H_0 :
    X et Y sont linéairement corrélées
  - Sinon, non-rejet de H_0 : on n'observe pas de corrélation linéaire significative entre les variables en population

$$t = r.\sqrt{\frac{n-2}{1-r^2}}$$

Synthèse : interprétation de r

Interprétation en fonction de r (en Y) et de l'effectif n (en X)
(graphe obtenu d'après la table de Student, test bilatéral, α=5%):



Corrélation signif^{t} ≠0, forte et positive

Corrélation signif^{t} ≠0, faible et positive

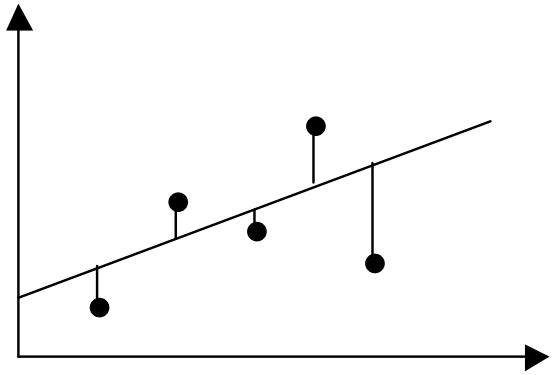
Corrélation non-signif^{t} ≠0

Corrélation signif^{t} ≠0, faible et négative

Corrélation signif^{t} ≠0, forte et négative

Régression linéaire simple de Y sur X

- On souhaite « modéliser » la relation entre X et Y par une équation Y=aX+b+ε
(ε : fluctuation aléatoire)



- Prédit Y et cherche à minimiser l'erreur « verticale » (c'est le plus fréquent)

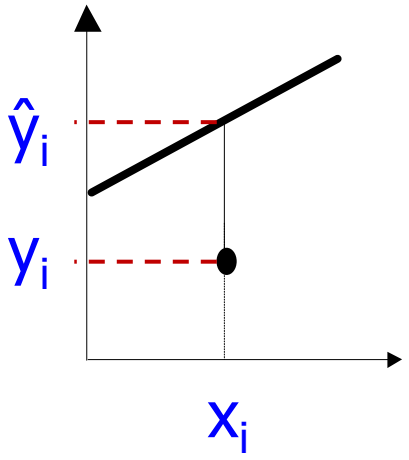
- On calcule a et b sur l'échantillon :

$$a = r.\frac{s_y}{s_x} = r.\frac{s_{ech\ Y}}{s_{ech\ X}} = \frac{cov_{ech\ X,Y}}{s^2{}_{ech\ X}}$$

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

Régression de Y sur X

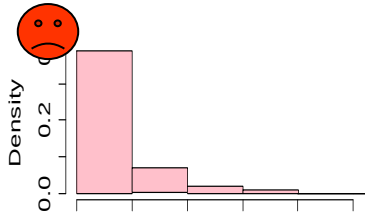
- On obtiendra ainsi sur l'échantillon :
  - a, b coefficients calculés
  - $y=ax+b$ équation de la droite de régression
  - $Y=aX+b+\varepsilon$ modèle de régression
- Et pour chaque individu i, avec xi :
  - y_i valeur observée
  - $\hat{y}_i=a.x_i+b$ valeur prédite par la régression (sur la droite)
  - $\hat{y}_i-y_i$ erreur (verticale) de prédiction ou « résidu »



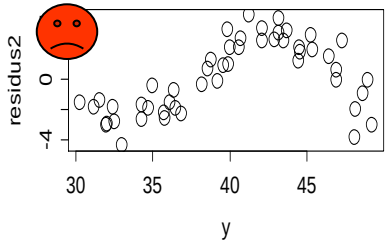
- Condition de validité : à valider a posteriori
  - Analyse des résidus (relation linéaire)
  - Recherche d'individus influents

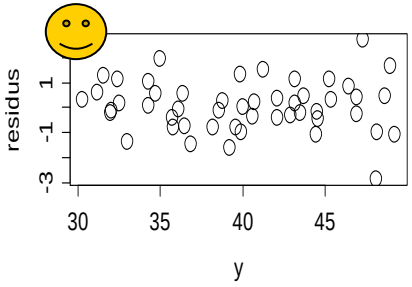
Analyse des résidus

- Rappel : une valeur par individu, $a.x_i + b - y_i$
- Distribution univariée des résidus
  - Toujours : moyenne=0 et écart type déterminé (cf. post)
  - Objectif : distribution normale (d'emblée acquis si Y ~ normale) => tracer un histogramme
- Résidus en fonction de $\hat{Y}$
  - Moyenne ne dépend pas de $\hat{Y}$
  - Variance ne dépend pas de $\hat{Y}$ (homoscédasticité)

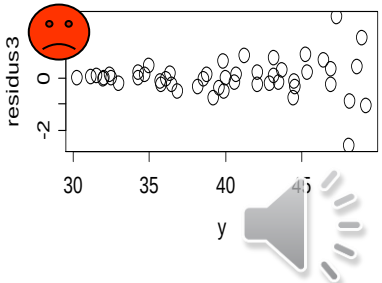


Moyenne dépend de $\hat{Y}$



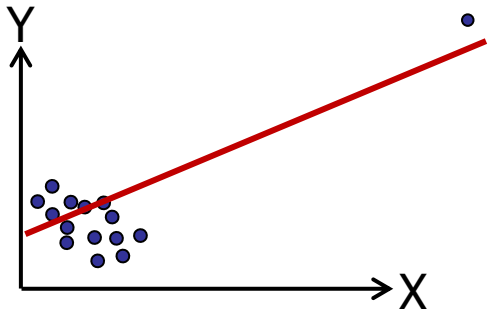


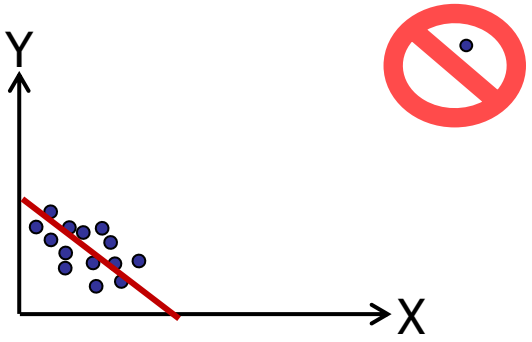
Dispersion dépend de $\hat{Y}$



Recherche d'individus influents

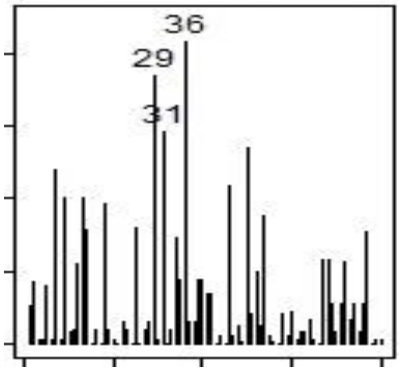
- Méthode graphique :



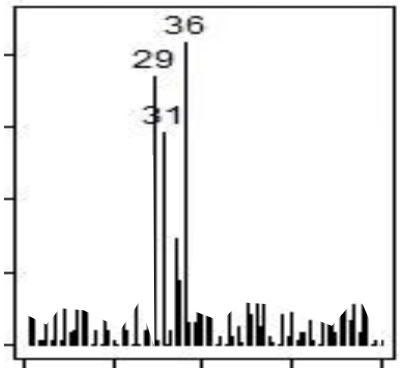


- Distance de Cook (conseillée) :

Aucun individu ne se détache, OK



Deux ou trois individus influents…



Prédiction

- Pour une nouvelle valeur x_0 (avec y_0 inconnu), on peut tenter de prédire y_0 par $\hat{y}_0$
- Prédiction probabiliste : un individu n'aura pas exactement la valeur prédite, mais :
- L'erreur de prédiction suit une loi normale centrée et d'écart type s_{YX} :

$$s_{yx} = s_y.\sqrt{1-r^2}$$

|r| élevée ⇔ prédiction précise

- IC de la prédiction à 95% :

$$IC_{y_0;1-\alpha} = \hat{y}_0 \pm u_\alpha.s_{yx}$$

Commentaire sur l'erreur de prédiction

$$s_{yx} = s_y . \sqrt{1 - r^2}$$

- Si r=0

$$s_{yx} = s_y$$

=> aucune réduction d'incertitude sur la valeur de y

- Si |r|=1

$$s_{yx} = 0$$

=> y est sur la droite, aucune incertitude

- Cas général : l'existence d'une relation linéaire réduit l'incertitude sur y, sachant x.

Epilogue

- Test de nullité de la pente : équivalent au test de nullité du coefficient de corrélation de Pearson
- Test de nullité de l'ordonnée à l'origine : peu utilisé en pratique
- Possibilité de forcer un modèle y=ax +0
- Régression de X sur Y :
  - Minimise l'erreur de prédiction de X : régression « horizontale »
  - Formules différentes
  - Rarement utilisée
- Régression orthogonale :
  - Minimise la distance (orthogonale) entre les points et la droite
  - Formules différentes
  - Rarement utilisée