Chapitre 2 : Statistique descriptive bivariée

2.1 Définitions

Variable appariées : Si deux variables sont définies pour les mêmes individus, on dit que ce sont des variables appariées.

Exemple apparié : Revenu de l'homme et revenu de la femme au sein des couples hétérosexuels dijonnais

Exemple pas apparié : Revenu des hommes dijonnais et des femmes dijonnaises

variables dépendante et indépendante Si l'une des deux variables appariées est "manipulable" par l'expérimentateur, on l'appelle *variable indépendante*. Dans ce cas, l'autre variable est appelée *variable dépendante*.

Exemples

variable indépendante	variable dépendante
dosage d'un traitement	intensité de la douleur manifestée par les malade
sexe des personnes interrogées	taille

Notation : S'il y a une variable indépendante, on la note X (et on note Y la variable dépendante)

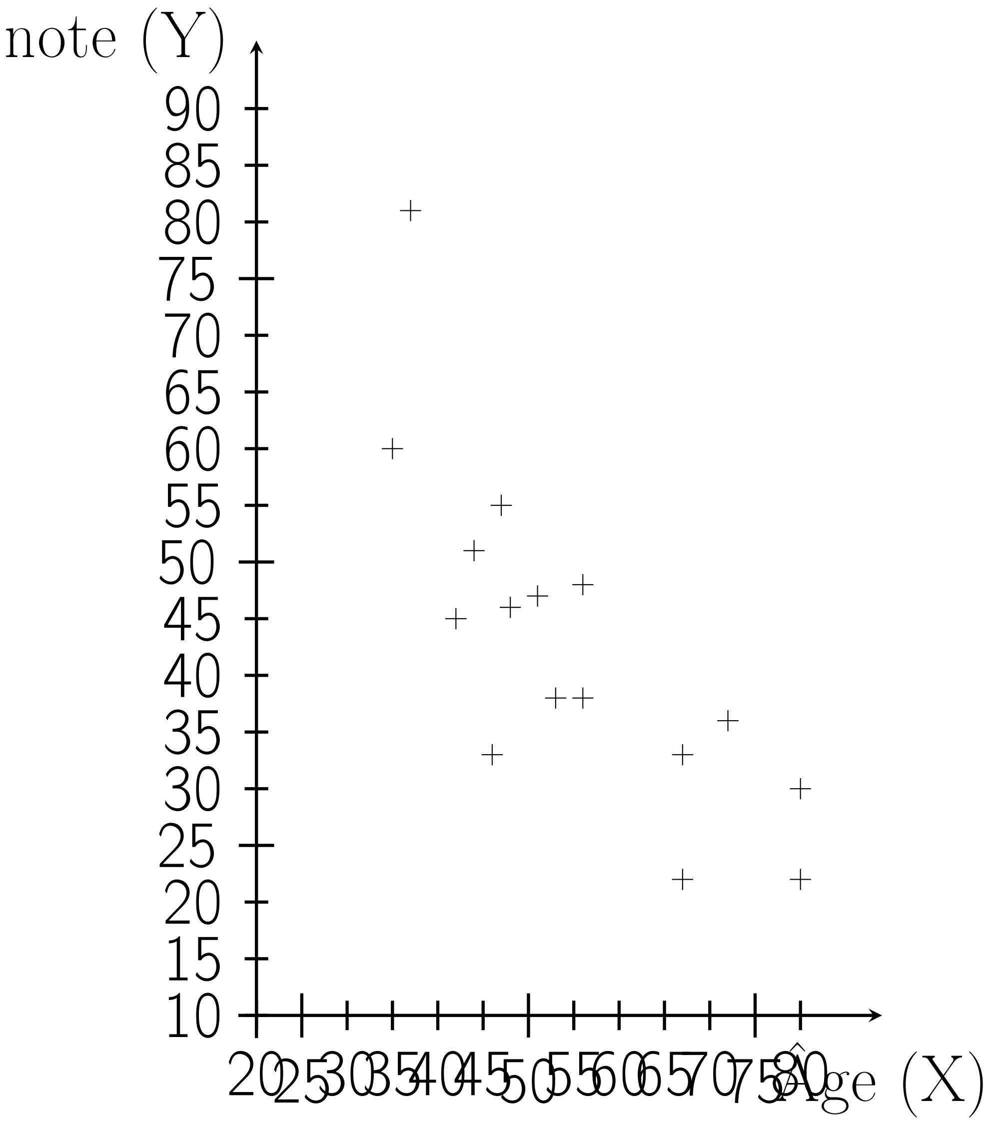
Exemple : Dans la suite, on considère les données de l'exercice 17

2.2 Représentation graphique : "Nuage de points"

Construction :

- On gradue les axes : X horizontalement et Y verticalement.
- On place un point par individu :
  - position horizontale : valeur de X
  - position verticale : valeur de Y

Exemple :



2.3 Coefficients de corrélation

Ils mesurent l'intensité du lien entre deux variables appariées

A) Coefficient de corrélation linéaire

Covariance $\mathrm{Cov}(X;Y) = m(X \times Y) - m(X) \times m(Y)$

Coefficient de corrélation linéaire $\mathrm{r}(X;Y) = \frac{\mathrm{Cov}(X;Y)}{s(X) \times s(Y)}$

Exemple (exercice 17) :

$m(X) = \frac{80+56+67+\cdots+35}{16} = \frac{881}{16}$

$m(Y) = \frac{22+38+33+\cdots+60}{16} = \frac{685}{16}$

$m(X \times Y) = \frac{80\times22+56\times38+67\times33+\cdots+35\times60}{16} = \frac{35206}{16}$

Donc $\mathrm{Cov}(X;Y) = \frac{35\,206}{16} - \frac{881}{16} \times \frac{685}{16} \simeq -157$

$m(X^2) = \frac{80^2+56^2+67^2+\cdots+35^2}{16} = \frac{51567}{16}$

$\mathrm{Var}(X) = \frac{51\,567}{16} - \left(\frac{881}{16}\right)^2 \simeq 191$

$m(Y^2) = \frac{22^2+38^2+33^2+\cdots+60^2}{16} = \frac{32671}{16}$

$\mathrm{Var}(Y) = \frac{32\,671}{16} - \left(\frac{685}{16}\right)^2 \simeq 209$

Donc $s(X) \simeq \sqrt{191} \simeq 13{,}8$, et $s(Y) \simeq \sqrt{209} \simeq 14{,}5$.

Et enfin $r(X;Y) \simeq \frac{-157}{13{,}8\times14{,}5} \simeq -0{,}78$.

Propriétés et interprétation :

- $r(X;Y)$ est toujours entre -1 et 1.
- Si $r(X;Y) \geqslant 0{,}75$ ou $r(X;Y) \leqslant -0{,}75$ alors
  - il y a un fort lien entre les variables
  - ce lien est « linéaire » (points presque le long d'une droite)
- Si $r(X;Y)$ n'est pas trop proche de zéro, alors
  - Si $r(X;Y) > 0$ alors Y tend à augmenter quand X augmente
  - Si $r(X;Y) < 0$ alors Y tend à diminuer quand X augmente.

B) Coefficient de corrélation des rangs

dit « coefficient de corrélation de Spearman »

i) Calcul des « rangs » On ajoute deux lignes au tableau de données :

- une ligne « x'_i » pour les rangs de X. On y met
  - 1 dans la colonne où X a la plus petite valeur
  - 2 dans la colonne où X a la $2^{\text{ème}}$ plus petite valeur
  - etc
- une ligne « y'_i » pour les rangs de Y. On y met
  - 1 dans la colonne où Y a la plus petite valeur
  - 2 dans la colonne où Y a la $2^{\text{ème}}$ plus petite valeur
  - etc

Exemple :

Âge (X)	80	56	67	72	67	51	80	42	56	47	53	37	48	46	44	35
Note (Y)	22	38	33	36	22	47	30	45	48	55	38	81	46	33	51	60
Rang x'_i	15,5	10,5	12,5	14	12,5	8	15,5	3	10,5	6	9	2	7	5	4	1,0
Rang y'_i	1,5	7,5	4,5	6	1,5	11	3	9	12	14	7,5	16	10	4,5	13	15

Ex aequo : Par exemple si deux colonnes ex aequo devraient avoir les rangs 10 et 11

- On leur attribue tous deux le rang 10,5
- La valeur suivante devrait avoir le rang 12

ii) Calcul du coefficient de corrélation de Spearman

Coefficent de Spearman Coefficient de corrélation linéaire des deux variables x' et y'

Méthode de calcul (p.3 du formulaire)

$$r_s(X;Y) \approx 1 - 6\,\frac{\sum_{i=1}^{n}(x'_i - y'_i)^2}{n(n^2-1)}$$

Exemple (exercice 17)

$$r_s(X;Y) = 1 - 6 \times \frac{(15{,}5-1{,}5)^2+(10{,}5-7{,}5)^2+(12{,}5-4{,}5)^2+\cdots+(1{,}0-15)^2}{16\left(16^2-1\right)} \simeq -0{,}77$$

Remarque

On peut simplifier le calcul en ajoutant au tableau une ligne « $\left(x'_i - y'_i\right)^2$ » (voir table 2.3 des notes de cours)

Propriétés et interprétation

- $r_s(X;Y)$ est toujours entre -1 et 1.
- Si $r_s(X;Y) \geqslant 0{,}75$ ou $r_s(X;Y) \leqslant -0{,}75$ alors
  - il y a un fort lien entre les variables
  - ce lien n'est pas forcément « linéaire » (points presque le long d'une courbe qui n'est pas forcément une ligne droite)
- Si $r_s(X;Y)$ n'est pas trop proche de zéro, alors
  - Si $r_s(X;Y) > 0$ alors Y tend à augmenter quand X augmente.
  - Si $r_s(X;Y) < 0$ alors Y tend à diminuer quand X augmente.

2.4 Regression linéaire

Si le coefficient de corrélation indique un lien linéaire entre les variables, on peut estimer une variable à partir de l'autre : (formules page 3 du formulaire)

- Détermination de Y à partir de X :
  - on utilise la droite $D_{Y|X}$ d'équation :
    $D_{Y|X}: \quad Y = aX + b$
    où $a = \dfrac{\mathrm{Cov}(X;Y)}{\mathrm{Var}(X)} = r(X;Y) \times \dfrac{s(Y)}{s(X)}$,
    et $b = m(Y) - a \cdot m(X)$
  - **Exemple** : Estimation de la note d'une personne de 80 ans :

on pose $a = \frac{Cov(X,Y)}{Var(X)} \simeq \frac{-156{,}988}{191{,}06} \simeq -0{,}822$ et
$b = m(Y) - a\, m(X) \simeq 42{,}812 - (-0{,}822) \times 55{,}062 \simeq 88{,}073$
D'où l'équation de la droite $D_{Y|X}$: $Y = -0{,}822\, X + 88{,}073$

Donc pour $x = 80$, on s'attend à

$y = -0{,}822 \times 80 + 88{,}073 = 22{,}313$.

- Détermination de X à partir de Y :
  - on utilise la droite $D_{X|Y}$ d'équation :
    $D_{X|Y}: \quad X = \ a'Y + b'$
    où $a' = \dfrac{\mathrm{Cov}(X;Y)}{\mathrm{Var}(Y)} = r(X;Y) \times \dfrac{s(X)}{s(Y)}$,
    et $b' = m(X) - a' \cdot m(Y)$
  - **Exemple** : Estimation de l'âge d'une personne qui a la note 25 :

on pose $a' = \frac{Cov(X,Y)}{Var(Y)} \simeq \frac{-156{,}988}{209{,}03} \simeq -0{,}751$ et
$b' = m(X) - a'\, m(Y) \simeq 55{,}062 - (-0{,}751) \times 42{,}812 \simeq 87{,}214$
D'où l'équation de la droite $D_{X|Y}$: $X = -0{,}751\, Y + 87{,}214$

Donc pour $y = 25$, on s'attend à

$x = -0{,}751 \times 25 + 87{,}214 = 68{,}439$.