Objectifs de la statistique bivariée

- Observer simultanément des individus d'une population sur deux caractères

$$\begin{array}{ccc} P & \to & M = M_1 \times M_2 \\ \omega & & c(\omega) = (c_1(\omega), c_2(\omega)) \end{array}$$ où M_1 *et* M_2 sont égaux à R (ensemble de valeurs numériques) ou N (ensemble de codes)

- Mesurer un lien éventuel entre deux caractères en utilisant un résumé chiffré qui traduit l'importance de ce lien.

$$\begin{array}{ccc} M^{card} & \to & R \\ v & & l_{card}(v) \end{array}$$ où $card$ est la taille de l'échantillon ou de la population et v le vecteur de tous les couples de réponses

- Qualifier ce lien :
  - en cherchant une relation numérique approchée entre deux caractères quantitatifs

$$\begin{array}{ccc} R & \xrightarrow{r} & R \\ x & & y = r(x) \end{array}$$ où r permet d'approximer c_2 en fonction de c_1 $$\begin{array}{ccc} P & \to & R \\ \omega & & c_2(\omega) = r(c_1(\omega)) + \varepsilon(c_1(\omega)) \end{array}$$

  - en cherchant des correspondances entre les modalités de deux caractères qualitatifs

Croisement qualitatif × qualitatif

Croisement qualitatif × qualitatif

Tableau de contingence

- Les seuls calculs possibles sur des caractères qualitatifs sont des

  effectifs et/ou des fréquences

- Chercher un lien entre deux caractères qualitatifs reviendra à étudier l'ensembles des effectifs des sous populations définies par les couples de modalités (x_i,y_j) prises respectivement par C_1 et C_2.

Tableau de contingence

$C_1 \backslash C_2$	y_1		y_j		y_l
x_1	$n_{1,1}$				$n_{1,1}$
	$n_{i,j}$ est le nombre d'individus ω tels que $C_1(\omega) = x_i$ et $C_2(\omega)= y_j$				
x_i	$n_{i,1}$		$n_{i,j}$		$n_{i,l}$
x_k	$n_{k,1}$		$n_{k,j}$		$n_{k,l}$

Exemple : Etude du lien entre la couleur des yeux et la couleur des cheveux

Yeux \ Cheveux	bruns	chatains	roux	blonds
bleus	11	10	0	8
verts	5	8	1	4
marrons	16	22	3	12

Croisement qualitatif × qualitatif

Effectifs marginaux

Pour faire des interprétations sur des correspondances entre des modalités de C_1 et des modalités de C_2, il faut compléter le tableau avec les effectifs de C_1 sans C_2 et des effectifs de C_2 sans C_1. Ces effectifs sont appelés effectifs marginaux (en marge de)

	y_1		y_j		y_l	
x_1	$n_{1,1}$				$n_{1,l}$	$\mathbf{n_{1,.}}$
x_i	$n_{i,1}$		$n_{i,j}$		$n_{i,l}$	$\mathbf{n_{i,.}}$
x_k	$n_{k,1}$		$n_{k,j}$		$n_{k,l}$	$\mathbf{n_{k,.}}$
	$\mathbf{n_{.,1}}$		$\mathbf{n_{.,j}}$		$\mathbf{n_{.,l}}$	$\mathbf{n}$

Effectifs marginaux

pour C_1: $n_{i,.} = \sum_{j=1}^{l} n_{i,j}$ pour C_2: $n_{.,j} = \sum_{i=1}^{k} n_{i,j}$

Effectif total

$$n = \sum_{j=1}^{l} n_{.,j} = \sum_{i=1}^{k} n_{.i,.} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{i,j}$$

<u>Exemple</u> : Etude du lien entre la couleur des yeux et la couleur des cheveux

Yeux \ Cheveux	bruns	chatains	roux	blonds	
bleus	11	10	0	8	**29**
verts	5	8	1	4	**18**
marrons	16	22	3	12	**53**
	32	**40**	**4**	**24**	**100**

Comparaison des effectifs non pertinente

Croisement qualitatif × qualitatif
Tableau de contingence des fréquences

Des effectifs ne sont pas directement comparables tandis que des fréquences sont toujours comparables

	y_1		y_j		y_l	
x_1	$f_{1,1}$				$f_{1,l}$	$\mathbf{f_{1,.}}$
x_i	$f_{i,1}$		$f_{i,j}$		$f_{i,l}$	$\mathbf{f_{i,.}}$
x_k	$f_{k,1}$		$f_{k,j}$		$f_{k,l}$	$\mathbf{f_{k,.}}$
	$\mathbf{f_{.,1}}$		$\mathbf{f_{.,j}}$		$\mathbf{f_{.,l}}$	**1**

$f_{i,j}$ est la proportion d'individus ω dans P tels que $C_1(\omega) = x_i$ et $C_2(\omega) = y_j$

Fréquences marginales

pour C_1: $f_{i,.} = \sum_{j=1}^{l} f_{i,j}$ pour C_2: $f_{.,j} = \sum_{i=1}^{k} f_{i,j}$

$$1 = \sum_{j=1}^{l} f_{.,j} = \sum_{i=1}^{k} f_{i,.} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{i,j}$$

<u>Exemple</u> : Etude du lien entre la couleur des yeux et la couleur des cheveux

Yeux \ Cheveux	bruns	chatains	roux	blonds	
bleus	0,11	0,1	0	0,08	0,29
verts	0,05	0,08	0,01	0,04	0,18
marrons	0,16	0,22	0,03	0,12	0,53
	0,32	0,4	0,04	0,24	1

Croisement qualitatif × qualitatif

Profils lignes et profils colonnes

L'analyse croisée consiste à chercher des correspondances entre des modalités de C_1 et des modalités de C_2.

Profils lignes	y_1		y_j		y_l	
x_1	$f_{1/1}$				$f_{l/1}$	$f_{1,.}$
x_i	$f_{1/i}$		$f_{j/i}$		$f_{l/i}$	$f_{i,.}$
x_k	$f_{1/k}$		$f_{j/k}$		$f_{l/k}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	

La ligne des fréquences marginales de C_2 est appelée *profil moyen.*

Profil ligne : répartition en fréquences du caractère C_2 dans une sous population définie par $P_{i,.} = \{\omega / C_1(\omega) = x_i\}$

$$f_{j/i} = \frac{n_{i,j}}{n_{i,.}}$$

comparable avec $f_{.j}$

Profil colonne : répartition en fréquences de C_1 dans une sous population définie par $P_{.,j} = \{\omega / C_2(\omega) = y_j\}$

$$f_{i/j} = \frac{n_{i,j}}{n_{.,j}}$$

comparable avec $f_{i,}$

Exemple

	bruns	châtains	roux	blonds
bleus	0,38	0,34	0,00	0,28
verts	0,28	0,44	0,06	0,22
marrons	0,30	0,42	0,06	0,23
	0,32	0,4	0,04	0,24

Profils ligne

	bruns	châtains	roux	blonds	
bleus	0,34	0,25	0,00	0,33	0,29
verts	0,16	0,20	0,25	0,17	0,18
marrons	0,50	0,55	0,75	0,50	0,53

Profils colonne

Croisement qualitatif × qualitatif

Un premier exemple caricatural

Exemple 1	y_1	y_2	y_3
x_1	10	20	30
x_2	100	200	300
x_3	1000	2000	3000

Ex 1 : Profils lignes	y_1	y_2	y_3
x_1	10/60=1/6	20/60=2/6	30/60=3/6
x_2	100/600=1/6	2/6	3/6
x_3	1000/6000=1/6	2/6	3/6
Fréq. marginales	(10+100+1000)/6660)=1/6	2/6	3/6

Ex 1 : Profils colonnes	Y_1	y_2	y_3	**Fréq marginales**
x_1	10/1110=1/111	1/111	1/111	(10+20+30)/6660=1/111
x_2	10/111	10/111	10/111	10/111
x_3	100/111	100/111	100/111	100/111

D'une modalité de C_1 à l'autre les répartitions des effectifs de C_2 sont proportionnelles.
Le caractère C_1 ne donne aucune information sur la répartition du caractère C_2.
Le caractère C_2 ne donne aucune information sur la répartition du caractère C_1.

Croisement qualitatif × qualitatif

Un deuxième exemple caricatural

Exemple 2	y_1	y_2	y_3
x_1	10	0	0
x_2	0	100	0
x_3	0	0	1000

Ex 2 : Profils lignes	y_1	y_2	y_3
x_1	1	0	0
x_2	0	1	0
x_3	0	0	1
Fréq. marginales	1/111	10/111	100/111

Ex 2 : Profils colonnes	Y_1	y_2	y_3	Fréq marginales
x_1	1	0	0	1/111
x_2	0	1	0	10/111
x_3	0	0	1	100/111

D'une modalité de C_1 à l'autre les répartitions des effectifs de C_2 sont totalement différentes.
Le caractère C_1 donne une information parfaite sur la répartition du caractère C_2.
Le caractère C_2 donne une information parfaite sur la répartition du caractère C_1.

Croisement qualitatif × qualitatif
Indépendance

C_1 et C_2 ne sont pas liés

$\Leftrightarrow$ les profils lignes sont égaux $\Leftrightarrow$ les profils colonnes sont égaux

$$\Leftrightarrow \quad \boxed{f_{i,j} = f_{i,.} \times f_{.,j}} \quad \forall i \in \{1,\dots,k\}, \forall j \in \{1,\dots,l\}$$

Tableau de contingence théorique si C_1 et C_2 sont indépendants

	y_1		y_j		y_l	
x_1	$f_{1,.} \times f_{.,1}$		$f_{1,.} \times f_{.,j}$		$f_{1,.} \times f_{.,l}$	$f_{1,.}$
x_i	$f_{i,.} \times f_{.,1}$		$f_{i,.} \times f_{.,j}$		$f_{i,.} \times f_{.,l}$	$f_{i,.}$
x_k	$f_{k,.} \times f_{.,1}$		$f_{k,.} \times f_{.,j}$		$f_{k,.} \times f_{.,l}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	1

Tableau de contingence observé

	y_1		y_j		y_l	
x_1	$f_{1,1}$		$f_{1,j}$		$f_{1,l}$	$f_{1,.}$
x_i	$f_{i,1}$		$f_{i,j}$		$f_{i,l}$	$f_{i,.}$
x_k	$f_{k,1}$		$f_{k,j}$		$f_{k,l}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	1

Croisement qualitatif × qualitatif

Test du chi-deux

Comment mesurer le lien de dépendance entre les C_1 et C_2 ? Comment mesurer la « distance » entre les deux tableaux? Mr Pearson a créée la *distance du χ^2* :

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{i,j} - t_{i,j}\right)^2}{t_{i,j}}$$

où $t_{i,j} = n \times f_{i,.} \times f_{.,j}$ est l'effectif théorique de la case (i,j).

1. La distance du χ^2 est d'autant plus grande que C_1 et C_2 sont liées entre eux.
2. La distance du χ^2 accorde plus d'importance aux différences entre les effectifs observés et effectifs théoriques sur les petits effectifs théoriques. S'écarter de 2% par rapport à 75% est moins significatif que de s'écarter de 2% par rapport à 5% .
3. La distance du χ^2 respecte le principe d'équivalence distributionnelle.
   - Si deux colonnes ont des effectifs proportionnels alors la fusion des modalités correspondante s du caractère C_2 ne change pas la distance du χ^2 entre C_1 et C_2 .
   - Si deux lignes ont des effectifs proportionnels alors la fusion des modalités correspondantes du caractère C_1 ne change pas la distance du χ^2 entre C_1 et C_2 .
4. Malheureusement la distance du χ^2 dépend aussi :
   - du nombre de modalités de C_1 et C_2.
   - du nombre d'individus.
5. On ne peut donc comparer deux distance du χ^2 que sur deux tableaux strictement équivalents en modalités et en nombre d'individus.

Croisement qualitatif × qualitatif

Coefficients normalisés

- *Coefficient de contingence* : $$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

  CC varie entre 0 et presque 1. Plus il est proche de 0 plus C_1 et C_2 sont indépendants et plus il est proche de 1 plus C_1 et C_2 sont liés. Par contre il dépend de k et l. On ne peut donc comparer que des tableaux de mêmes dimensions.

- *V de Cramer :* $$V = \sqrt{\frac{\chi^2}{n \times [\min(k,l) - 1]}}$$

  Même interprétation que le coefficient précédent avec l'avantage de ne plus dépendre de k et l. C'est le coefficient normalisé le plus utilisé.

- Il existe d'autres coefficients comme le coefficient phi de Pearson ou le PEM (Pourcentage de l'Écart Maximum).
- Mais il faut retenir :
  1. que ces coefficients ne varie proportionnellement avec l'importance du lien
  2. que plus ils sont proches de 0 plus C_1 et C_2 sont indépendants et plus ils sont proches de 1 plus C_1 et C_2 sont liés.
  3. qu'il faut comparer l'évolution dans le temps de ces coefficients sur des tableaux équivalents

	bruns	chatains	roux	blonds	
bleus	11	10	0	8	29
verts	5	8	1	4	18
marrons	16	22	3	12	53
	32	40	4	24	100

Tableau de contingence observé

	bruns	chatains	roux	blonds	
bleus	9,28	11,6	1,16	6,96	29
verts	5,76	7,2	0,72	4,32	18
marrons	16,96	21,2	2,12	12,72	53
	32	40	4	24	100

Tableau de contingence théorique

Croisement quantitatif × qualitatif

Croisement qualitatif × quantitatif (1)

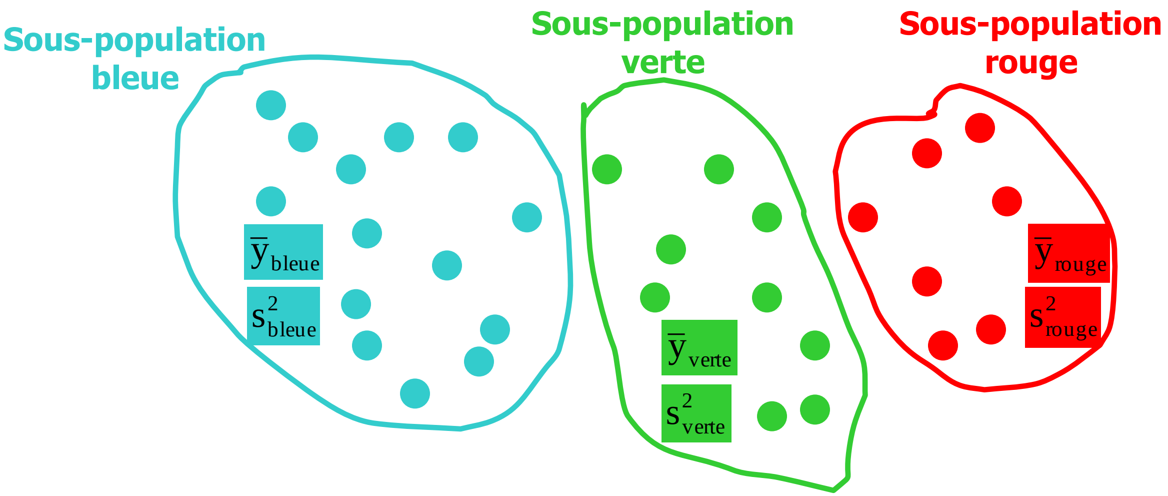
- Pour étudier le lien entre un caractère qualitatif à p modalités et un caractère quantitatif, on partitionne la population P en sous populations : une sous-population pour chaque modalité du caractère qualitatif
- On étudie le caractère quantitatif C_2 sur chaque sous-population en calculant la moyenne et la variance de C_2. On parle de *variation intra*,

Effectif de la sous-population ℓ

Moyenne de la sous-population ℓ

$$\mathrm{var}^{intra}(C_2) = \frac{1}{n}\sum_{l=1}^{p} n_l\, s_l^2 = \frac{1}{n}\sum_{l=1}^{p} n_l \times \frac{1}{n_l}\sum_{i=1}^{n_l}\left(y_i - \overline{y_l}\right)^2$$

Variance de la sous-population ℓ



<u>Décomposition de la moyenne</u>

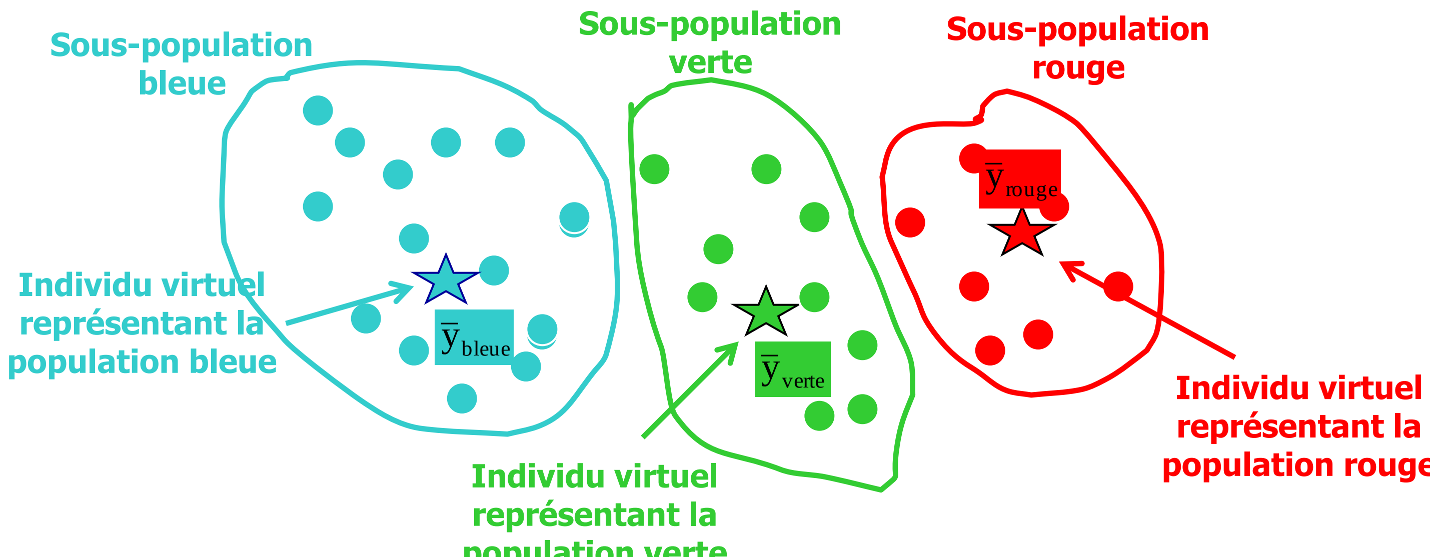
$$\frac{1}{n}\sum_{\ell=1}^{p} n_\ell\, \overline{y_\ell} = \bar{y}$$

	Effectifs	Moyenne de la taille (cm)	Variance de la taille
Hommes	23	162,3	202,0
Femmes	35	149,3	110,7
Pop. totale	58	154,4	185,3

Croisement qualitatif × quantitatif (2)

- Pour chaque sous-population, on crée un individu virtuel dont la valeur sur C_2 est égale à la moyenne des valeurs de C_2 des individus de la sous-population.
- On crée donc une nouvelle population formée de ces individus virtuels. Chaque individu aura un poids de n_i , l'effectif de chaque sous-population. On parle de *variation inter*,

$$\mathrm{var}^{inter}(C_2) = \frac{1}{n}\sum_{\ell=1}^{p} n_\ell (\overline{y_\ell} - \overline{y})^2$$



Croisement qualitatif × quantitatif (3)

On peut donc définir trois variances sur la caractère C_2.

1. une première qui explique les variations de C_2 dans toute la population : totale
2. une deuxième qui explique les variations de C_2 dans les sous-populations : intra
3. une troisième qui explique les variations de C_2 entre les sous-populations : inter

Nous avons la <u>décomposition de la variance</u> suivante :

$$\mathrm{var}^{totale}(C_2) = \underbrace{\mathrm{var}^{inter}(C_2)}_{\text{Variance expliquée}} + \underbrace{\mathrm{var}^{intra}(C_2)}_{\text{Variance résiduelle}}$$

On en déduit une mesure du lien entre C_1 et C_2 avec le *rapport de corrélation*

$$\boxed{\frac{\mathrm{var}^{inter}(C_2)}{\mathrm{var}^{totale}(C_2)}}$$

Cette expression varie entre 0 et 1. Plus sa valeur est proche de 1 plus les deux caractères sont liés

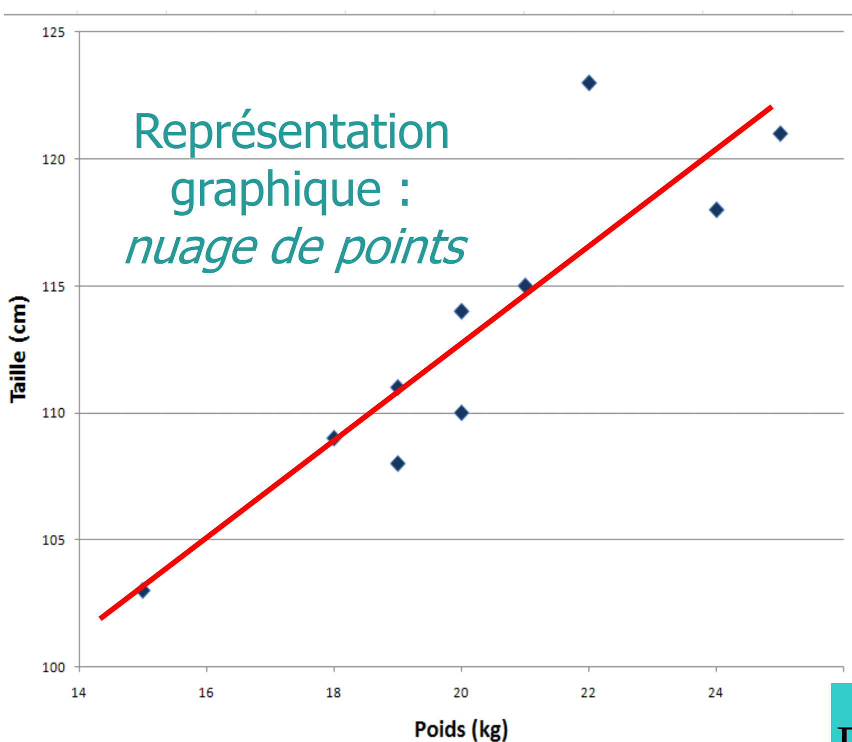
Croisement quantitatif $\times$ quantitatif

Croisement quantitatif × quantitatif (1)

Droite de régression

<u>Exemple</u> : 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 C_1 et C_2 deux caractères quantitatifs sur une population de taille n.

On note $\{x_i\}_{i=1,\dots,n}$ la série observée pour C_1 et $\{y_i\}_{i=1,\dots,n}$ la série observée pour C_2.

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 (2)

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 C_1 et C_2.

1. $y = \hat{a}x + \hat{b}$ est appelée *droite de régression* de C_2 en C_1. Elle traduit les variations de C_2 qui peuvent être expliquées par C_1.

2. Attention la droite de régression de C_1 en C_2 n'est nécessairement la même que celle de C_2 en C_1

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	r_{xy}
113,20	**20,30**	**38,62**	**8,46**	**0,90**

L'équation de la droite de C2 en C1 : y=0,42 x – 27,38

L'équation de la droite de C1 en C2 : y=1,92 x – 74,15

Croisement quantitatif × quantitatif (3)

Covariance et coefficient de corrélation

L'inégalité de Cauchy-Schwartz permet de montrer que $|c_{xy}| \leq s_x s_y$.

On peut alors définir 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 C_1 et C_2 sont très liés entre eux par une droite affine.
2. $r < 0$: globalement C_1 et C_2 varient en sens inverse .
3. $r > 0$: globalement C_1 et C_2 varient dans le même sens .
4. $|r| \cong 0$: on ne peut rien dire sur un lien éventuel entre C_1 et C_2.

Remarque : $\hat{a} = \frac{c_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$

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

On trouve

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

$r_{xy} \cong 1 \Rightarrow$ 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)

Croisement quantitatif × quantitatif (4)

Prévisions

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)

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

<u>Propriétés</u> :

1)
$$\overline{\hat{y}} = \overline{y}$$

Le caractère C_2 et la partie de ce caractère expliquée par la droite de régression ont la même moyenne.

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

1. La variance de C_2 expliquée la droite de régression est plus petite que la variance de C_2.
2. La variance de C_2 expliquée la droite de régression est d'autant meilleure que le coefficient de Pearson est proche de 1 en valeur absolue.

Croisement quantitatif × quantitatif (5)

Résidus

On appelle *résidus* l'écart entre la valeur observée y_i et la valeur prédite $\hat{y}_i$

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

On calcule alors l'erreur globale

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

1. L'erreur globale est proportionnelle à la variance du caractère C_2.
2. L'erreur est d'autant plus petite que le coefficient est proche de 1 en valeur absolue.

Propriétés :

1) La moyenne des résidus est nulle, $\bar{e} = 0$

2) Les résidus et la série explicative x sont non corrélés, $c_{ex} = \frac{1}{n}\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0$

Les résidus ne contiennent plus « d'information » pouvant expliquer y.

3) Formule de décomposition de la variance

$$\underbrace{s_y^2}_{\text{variance totale}} = \underbrace{s_{\hat{y}}^2}_{\text{variance expliquée}} + \underbrace{s_e^2}_{\text{variance résiduelle}} \longrightarrow \boxed{R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = r_{xy}^2 \in [0,1]}$$

coefficient de détermination

Croisement quantitatif × quantitatif (6)

1. Les droites de régression n'explique que les liaisons linéaires.
2. Si C_1 et C_2 sont liées par une relation de la forme $C_2 = a.(C_1)^2$ alors $r(C_1, C_2) = 0$
   Le coefficient de corrélation linéaire de Pearson ne peut pas détecter cette liaison.
3. Il n'existe pas de mesure universelle pour détecter des relations quelconques
4. 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	$y = ax^b$	$y' = \log(y)$ $x' = \log(x)$	$y' = \log(a) + b.x'$
inverse	$y = a + \dfrac{b}{x}$	$x' = \dfrac{1}{x}$	$y' = a + b.x'$
logistique	$y = \dfrac{1}{1 + e^{-(a.x+b)}}$	$y' = \log\left(\dfrac{y}{1-y}\right)$	$y' = a.x + b$