Statistiques descriptives bivariées

- Objectifs des statistiques descriptives bivariées
- Qualitatif x Qualitatif
- Quantitatif x Quantitatif
- Qualitatif x Quantitatif

Objectifs des statistiques bivariées

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

$$\begin{array}{ccc} P & \rightarrow & M = M_1 \times M_2 \\ \omega & & c(\omega) = (c_1(\omega), c_2(\omega)) \end{array}$$ 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.

$$\begin{array}{ccc} \mathrm{M}^{\mathrm{card}} & \rightarrow & \mathrm{R} \\ \mathrm{v} & & \mathrm{l}_{\mathrm{card}}(\mathrm{v}) \end{array}$$ où *c*ard est n la taille de l'échantillo n ou N la taille 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 $$\begin{array}{ccc} P & \rightarrow & R \\ \omega & & c_2(\omega) = r(c_1(\omega)) + \varepsilon(c_1(\omega)) \end{array}$$

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

Croisement qualitatif × qualitatif (1)

- 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.
- On va définir un tableau dit de contingence.

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

$n_{i,j}$ est le nombre d'individus ω tels que $C_1(\omega) = x_i$ et $C_2(\omega) = y_j$

- On note k le nombre de modalités du caractère C_1 et l le nombre de modalités du caractère C_2.
- On note x_1, …., x_k les valeurs de C_1 et y_1, …., y_l les valeurs de C_2.

Croisement qualitatif × qualitatif (2)

- 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)
- On enrichit donc le tableau dit de contingence avec les effectifs marginaux.

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

effectifs marginaux de C_1.

$$\boldsymbol{n}_{i,.} = \sum_{j=1}^{l} \boldsymbol{n}_{i,j}$$

effectifs marginaux de C_2.

$$\boldsymbol{n}_{.,j} = \sum_{i=1}^{k} \boldsymbol{n}_{i,j}$$

$$\boldsymbol{card} = \sum_{j=1}^{l} \boldsymbol{n}_{.,j} = \sum_{i=1}^{k} \boldsymbol{n}_{.i,.} = \sum_{i=1}^{k} \sum_{j=1}^{l} \boldsymbol{n}_{i,j}$$

- Pour faire des comparaisons pertinentes, il faudra compléter par des calculs de fréquences comme l'explique les slides suivants.

Croisement qualitatif × qualitatif (3)

- Des effectifs ne sont pas directement comparables tandis que des fréquences sont toujours comparables
- On définit donc le tableau de contingence avec des fréquences.

	y_1		y_j		y_l	
x_1	$f_{1,1}$				$f_{1,l}$	$f_{1,.}$
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

fréquences marginales de C_1.

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

fréquences marginales de C_2.

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

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

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

Croisement qualitatif × qualitatif (4)

- L'analyse croisée consiste à chercher des correspondances entre des modalités de C_1 et des modalités de C_2.
- On définit donc deux nouvelles notions : profils lignes et profils colonnes.
- Un profil ligne est la répartition en fréquences du caractère C_2 dans une sous population définie par $P_{i,.} = \{\omega / C_1(\omega) = x_i\}$
- Un profil colonne est la répartition en fréquences du caractère C_1 dans une sous population définie par $P_{.,j} = \{\omega / C_2(\omega) = y_j\}$

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	$f_{1/k}$		$f_{j/k}$		$f_{l/k}$	$f_{k,.}$
	$f_{.,1}$		$f_{.,j}$		$f_{.,l}$	

$f_{j/i}$ et $f_{.j}$ sont directement comparables. Elles donnent une information sur le même phénomène mais dans deux populations différentes.

La ligne des fréquences marginales de C_2 est appelé profil moyen.

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

Croisement qualitatif × qualitatif (5)

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	1/6	2/6	3/6
x_2	1/6	2/6	3/6
x_3	1/6	2/6	3/6
Fréq. marginales	1/6	2/6	3/6

Ex 1 : Profils colonnes	Y_1	y_2	y_3	Fréq marginales
x_1	1/111	1/111	1/111	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 (6)

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 (7)

- C_1 et C_2 ne sont pas liés ⇔ les profils lignes sont égaux ⇔ les profils colonnes sont égaux
- On nous donne la répartition de C_1 et C_2 . A quoi sont égales les fréquences $f_{i,j}$ si C_1 et C_2 ne sont pas liés ?
- Théorème $\forall (i,j) \in \{1,...,k\} \times \{1,...,l\} \quad f_{i,j} = f_{i,.} \cdot f_{.,j} \Rightarrow C_1 \text{ et } C_2$ sont indépendants :

Preuve : Calculons les profils lignes : $\forall (i,j) \in \{1,...,k\} \times \{1,...,l\} \quad f_{j/i} = \frac{f_{i,j}}{f_{i,.}} = \frac{f_{i,.} \cdot f_{.,j}}{f_{i,.}} = f_{.,j}$

Tous les profils lignes sont égaux au profil ligne moyen. CQFD

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

$f_{i,}.f_{.,j}$ est la proportion théorique de la case (i,j) si C_1 et C_2 sont indépendants

- On peut démontrer que $f_{i,j} = f_{i,.} \cdot f_{.,j}$ est la seule configuration possible pour que C_1 et C_2 soient indépendants

Croisement qualitatif × qualitatif (8)

- Comment mesurer le lien de dépendance entre les C_1 et C_2 ?

Tableau de contingence théorique si C_1 et C_2 sont indépendants

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

Tableau de contingence observé

	y_1		y_j		y_l	
x_1	$f_{1,1}$				$f_{1,l}$	$f_{1,.}$
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

- On notera $t_{i,j}$ l'effectif théorique de la case (i,j). $t_{i,j}$ = card . $f_{i,.}f_{.,j}$
- Rappel : $n_{i,j}$ = card . $f_{i,j}$
- Intuitivement, il faudrait trouver une formule de distance entre ces deux matrices.
- Mr Pearson a créé la formule suivante : $$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(n_{i,j} - t_{i,j}\right)^2}{t_{i,j}}$$

Croisement qualitatif × qualitatif (9)

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

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 (10)

- Coefficients normalisés
  - Coefficient de contingence : $CC = \sqrt{\frac{\chi^2}{\chi^2 + card}}$

    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 : $V = \sqrt{\frac{\chi^2}{card.[\min(k,l)-1]}}$

    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 pas 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

Deux caractères quantitatifs (1)

On considère C_1 et C_2 deux caractères quantitatifs

P	ω_1		ω_i		ω_{card}
C_1	$x_1=c_1(\omega_1)$		$x_i=c_1(\omega_i)$		$x_{card}=c_1(\omega_{card})$
C_2	$y_1=c_2(\omega_1)$		$y_i=c_2(\omega_i)$		$y_{card}=c_2(\omega_{card})$

On considère f une fonction de R dans R. On cherche à approximer le caractère C_2 en fonction du caractère C_1 à l'aide de f. On calcule l'erreur quadratique *EQ(f)*.

$$EQ(f)=\frac{1}{card}\sum_{i=1}^{card}(y_i-f(x_i))^2$$

L'ensemble des fonctions est infini. On se restreint aux fonctions affines *f(x) = a.x + b*. On cherche a et b qui minimisent l'erreur quadratique :

$$\min_{(a,b)\in R\times R} EQ(a,b)=\frac{1}{card}\sum_{i=1}^{card}(y_i-a.x_i-b)^2$$

Pour a fixé, on cherche b qui minimise EQ. $EQ_q : \begin{matrix} R & \rightarrow & R \\ b & & EQ_a=\mathrm{EQ(a,b)} \end{matrix}$

EQ est une fonction quadratique convexe. Il suffit donc d'annuler la dérivée en b.

$$\frac{\mathrm{dEQ_a(b)}}{\mathrm{db}}=\frac{\partial EQ(a,b)}{\partial b}=\frac{-2}{card}\sum_{i=1}^{card}(y_i-a.x_i-b)=0\Leftrightarrow \bar{y}-a\bar{x}-b=0\Leftrightarrow \hat{b}(a)=\bar{y}-a\bar{x}$$

où $\bar{y}=\overline{C_2}$ et $\bar{x}=\overline{C_1}$

Pour conclure, on cherche donc à minimiser EQ par rapport à a la fonction suivante :

$$EQ(a,\hat{b}(a))=\frac{1}{card}\sum_{i=1}^{card}(y_i-a.x_i-\bar{y}+a.\bar{x})^2=\frac{1}{card}\sum_{i=1}^{card}((y_i-\bar{y})-a.(x_i-\bar{x}))^2$$

Deux caractères quantitatifs (2)

(1) $\dfrac{dEQ(a,b(a))}{da} = \dfrac{-2}{card}\sum_{i=1}^{card}(x_i-\bar{x}).((y_i-\bar{y})-a.(x_i-\bar{x})) = 0 \Leftrightarrow \dfrac{\sum_{i=1}^{card}(x_i-\bar{x}).((y_i-\bar{y})}{card} = a.\dfrac{\sum_{i=1}^{card}(x_i-\bar{x})^2}{card}$

(2) *le terme* $\dfrac{\sum_{i=1}^{card}(x_i-\bar{x}).(y_i-\bar{y})}{card}$ est appelé covariance de C_1 et C_2 et est noté $\text{cov}(C_1,C_2)$

(3) Le terme $\dfrac{\sum_{i=1}^{card}(x_i-\bar{x})^2}{card}$ est la variance du caractère C_1 et est notée $\text{var}(C_1)$.

(4) Si C_1 n'est pas constant alors $var(C_1)$ est strictement positif.

en conclusion le couple optimal est $(\hat{a},\hat{b}) = \left(\dfrac{\text{cov}(C_1,C_2)}{\text{var}(C_1)}, \overline{C}_2 - \dfrac{\text{cov}(C_1,C_2)}{\text{var}(C_1)}.\overline{C}_1\right)$.

(5) Rappel l'écart - type de C_1 noté $\sigma(C_1)$ vaut $\sqrt{\text{var}(C_1)}$.

On définit le coefficient de corrélation linéaire de Pearson : $r(C_1,C_2) = \dfrac{\text{cov}(C_1,C_2)}{\sigma(C_1).\sigma(C_2)}$

(6) Une autre version du couple optimal est $(\hat{a},\hat{b}) = \left(r(C_1,C_2).\dfrac{\sigma(C_2)}{\sigma(C_1)}, \overline{C}_2 - r(C_1,C_2).\dfrac{\sigma(C_2)}{\sigma(C_1)}.\overline{C}_1\right)$.

Deux caractères quantitatifs (3)

(7) Propriété : $|\text{cov}(C_1, C_2)| \leq \sigma(C_1).\sigma(C2)$

Preuve : l'inégalité de Cauchy - Schwartz nous donne $\left|\sum_{i=1}^{card}(x_i - \bar{x}).(y_i - \bar{y})\right| \leq \sqrt{\sum_{i=1}^{card}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{card}(y_i - \bar{y})^2}$

on en déduit $\frac{\left|\sum_{i=1}^{card}(x_i - \bar{x}).(y_i - \bar{y})\right|}{card} \leq \frac{\sqrt{\sum_{i=1}^{card}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{card}(y_i - \bar{y})^2}}{card} \Leftrightarrow \frac{\left|\sum_{i=1}^{card}(x_i - \bar{x}).(y_i - \bar{y})\right|}{card} \leq \sqrt{\frac{\sum_{i=1}^{card}(x_i - \bar{x})^2}{card}}.\sqrt{\frac{\sum_{i=1}^{card}(y_i - \bar{y})^2}{card}}$

en conclusion $|\text{cov}(C_1, C_2)| \leq \sigma(C_1).\sigma(C_2)$

(8) On déduit de la propriété (7) que $-1 \leq r(C_1, C_2) = \frac{\text{cov}(C_1, C_2)}{\sigma(C_1).\sigma(C_2)} \leq 1$

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| ≅ 0 : on ne peut rien dire sur un lien éventuel entre C_1 et C_2.

(9) Calculons l'erreur globale $EQ(\hat{a},\hat{b}) = \frac{1}{card}\sum_{i=1}^{card}(y_i - \hat{a}x_i - (\bar{y} - \hat{a}\bar{x}))^2 = \frac{1}{card}\sum_{i=1}^{card}((y_i - \bar{y}) - \hat{a}(x_i - \bar{x}))^2 = \sigma^2(C_2)$

$EQ(\hat{a},\hat{b}) = \sigma^2(C_2) + \hat{a}^2.\sigma^2(C_1) - 2.\hat{a}.\text{cov}(C_1, C_2)$

On a les deux égalités suivantes : $\text{cov}(C_1, C_2) = r(C_1, C_2).\sigma(C_1).\sigma(C_2)$ et $\hat{a} = \frac{r(C_1, C_2).\sigma(C_2)}{\sigma(C_1)}$

On en déduit : $EQ(\hat{a},\hat{b}) = \sigma^2(C_2) + r^2(C1, C2).\sigma^2(C_2) - 2.r^2(C1, C2).\sigma^2(C_2) = \sigma^2(C_2)(1 - r^2(C1, C2))$

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.

Deux caractères quantitatifs (4)

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

(10) Propriété : $\overline{\hat{a}C_1 + \hat{b}} = \overline{C}_2$

Preuve : $\overline{\hat{a}C_1 + \hat{b}} = \dfrac{\sum_{i=1}^{card} \hat{a}x_i + \bar{y} - \hat{a}.\bar{x}}{card} = \dfrac{card.\bar{y}}{card} + \dfrac{\sum_{i=1}^{card} \hat{a}x_i}{card} - \dfrac{card.\hat{a}.\bar{x}}{card} = \bar{y} + \hat{a}.\bar{x} - \hat{a}.\bar{x} = \bar{y} = \overline{C}_2$

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

(11) Propriété : $\text{var}(\hat{a}C_1 + \hat{b}) = \text{var}(C_2).r^2(C1, C_2)$

Preuve : $\text{var}(\hat{a}C_1 + \hat{b}) = \dfrac{\sum_{i=1}^{card} (\hat{a}x_i + \hat{b} - \bar{y})^2}{card} = \dfrac{\sum_{i=1}^{card} (\hat{a}x_i + \bar{y} - \hat{a}.\bar{x} - \bar{y})^2}{card} = \hat{a}^2 . \dfrac{\sum_{i=1}^{card} (x_i - \bar{x})^2}{card}$

$\text{var}(\hat{a}C_1 + \hat{b}) = \dfrac{r^2(C1, C_2).\text{var}(C_2).\text{var}(C_1)}{\text{var}(C_1)} = r^2(C1, C_2).\text{var}(C_2)$

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.

Deux caractères quantitatifs (5)

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

Les valeurs communes aux deux régression s

$\overline{C}_1$	$\overline{C}_2$	$\sigma^2(C_1)$	$\sigma^2(C_2)$	$r(C_1,C_2)$
113.2	20.3	34.76	7.61	0.90013

C_2 par rapport à C_1 : $\hat{a} = r(c_1,c_2)\dfrac{\sigma(C_2)}{\sigma(C_1)} = 0.90013\dfrac{\sqrt{7.61}}{\sqrt{34.76}} = 0.45$ et $\hat{b} = 20.3 - 0.45.113.2 = -30.64$

La variance de C_2 vaut 8.46, la variance de C_2 expliquée par C_1 vaut $8.46.(.90013)^2 = 6.85$ et la variance résiduelle de C_2 non expliquée par C_1 vaut $8.46.(1-(.90013)^2) = 1.6054$. L'écart - type résiduel vaut $\sqrt{1.6054} = 1.27\,Kg$

En moyenne si on estime le poids avec la droite de régression on fera une erreur de 1.27 kg

C_1 par rapport à C_2 : $\hat{a}' = r(c_1,c_2)\dfrac{\sigma(C_1)}{\sigma(C_2)} = 0.90013\dfrac{\sqrt{34.76}}{\sqrt{7.61}} = 1.92$ et $\hat{b}' = 113.2 - 1.92 * 20.3 = 74.22$

La variance de C_1 vaut 34.76, la variance de C_1 expliquée par C_2 vaut $34.76 * (.90013)^2 = 28.16$ et la variance résiduelle de C_1 non expliquée par C_2 vaut $34.76.(1-(.90013)^2) = 6.60$. L'écart - type résiduel vaut $\sqrt{6.6} = 2.57\,cm$.

En moyenne si on estime la taille avec la droite de régression on fera une erreur de 2.57cm

L'équation de la régression de C_2 en C_1 : $y = 0.45.x - 30.64$

L'équation de la régression de C_1 en C_2 : $x = 1.92.y + 74.22 \Leftrightarrow y = \dfrac{x - 74.22}{1.92} = 0.52.x - 38.65$

Deux caractères quantitatifs (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	$y = a + \frac{b}{x}$	$x' = \frac{1}{x}$	$y' = a + b.x'$
logistique	$y = \frac{1}{1 + e^{-(a.x+b)}}$	$y' = \log\left(\frac{y}{1-y}\right)$	$y' = a.x + b$

Croisement qualitatif × quantitatif (1)

- On croise C_1 un caractère qualitatif avec C_2 un caractère quantitatif. On note k le nombre de modalités du caractère C_1 .et x_i la $i^{ème}$ modalité de C_1.
- Question : Est-ce que les variations de C_2 sont différentes d'une modalité à une autre modalité de C_1 ?
- Le caractère C_1 partitionne la population en k sous populations. On note k_i la valeur de la $i^{ème}$ modalité de C_1 et k_i l'effectif de la $i^{ème}$ sous population.

- On définit k+1 populations : $\boldsymbol{P_i} = \{\, \boldsymbol{\omega} \in \boldsymbol{P} / \boldsymbol{C_1}(\boldsymbol{\omega}) = \boldsymbol{x_i} \,\}$ $\quad \Omega = \{\, \boldsymbol{P_i} / \boldsymbol{i} \in \{1,\ldots,\boldsymbol{k}\} \,\}$
- On peut étudier le caractère C_2 à travers 2+k populations : P, Ω et $P_{i\ i\in\{1,\ldots,k\}}$

$$\begin{array}{ccc} \boldsymbol{P} & \rightarrow & \boldsymbol{R} \\ \boldsymbol{\omega} & & \boldsymbol{C_2}(\boldsymbol{\omega}) \end{array}$$

Les variations de C_2 dans la population

$$\begin{array}{ccc} \boldsymbol{P_i} & \rightarrow & \boldsymbol{R} \\ \boldsymbol{\omega} & & \boldsymbol{C}_{2,\boldsymbol{i}}^{\boldsymbol{intra}}(\boldsymbol{\omega}) = \boldsymbol{C_2}(\boldsymbol{\omega}) \end{array}$$

Les variations de C_2 dans les sous populations définies par C_1

$$\begin{array}{ccc} \Omega & \rightarrow & \boldsymbol{R} \\ \boldsymbol{P_i} & & \boldsymbol{C_2}^{\boldsymbol{inter}}(\boldsymbol{P_i}) = \overline{\boldsymbol{C}}_{2,\boldsymbol{i}}^{\boldsymbol{intra}} \end{array}$$

Les variations de C_2 en réduisant chaque sous population à un représentant

Croisement qualitatif × quantitatif (2)

$$(1) : \quad \overline{C}_2 = \frac{1}{card} \sum_{i=1}^{k} k_i . \overline{C}_{2,i}^{intra}$$

Preuve :

$$\frac{1}{card} \sum_{i=1}^{k} k_i . \overline{C}_{2,i}^{intra} = \frac{1}{card} \sum_{i=1}^{k} \left(k_i . \frac{1}{k_i} \sum_{\omega \in P_i} C_2(\omega) \right) = \frac{1}{card} \sum_{i=1}^{k} \sum_{\omega \in P_i} C_2(\omega) = \overline{C}_2$$

En d'autres termes, la moyenne du caractère C_2 sur la population P est la moyenne des moyennes de C_2 sur les sous populations P_i pondérées par les effectifs k_i de ces sous populations

On définit les trois variances suivantes :

$$Var^{inter}(C_2) = Var(C_2^{inter}) = \frac{1}{card} \sum_{i \in \{1,..,k\}} k_i . (C_2^{inter}(P_i) - \overline{C}_2)^2$$

$$Var^{totale}(C_2) = Var(C_2) = \frac{1}{card} \sum_{\omega \in P} (C_2(\omega) - \overline{C}_2)^2$$

$$Var^{intra}(C_2) = \frac{1}{card} \sum_{i=1}^{k} k_i . Var(C_{2,i}^{intra})$$

Croisement qualitatif × quantitatif (3)

Théorème : $Var^{totale}(C_2) = Var^{inter}(C_2) + Var^{intra}(C_2)$

$$Var^{totale}(C_2) = \frac{1}{card}.\sum_{\omega\in P}(C_2(\omega)-\overline{C}_2)^2 = \frac{1}{card}.\sum_{i=1}^{k}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_2)^2$$

$$Var^{totale}(C_2) = \frac{1}{card}.\sum_{i=1}^{k}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra}+\overline{C}_{2,i}^{intra}-\overline{C}_2)^2$$

$$Var^{totale}(C_2) = \frac{1}{card}.\left(\begin{array}{l}\sum_{i=1}^{k}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra})^2+\sum_{i=1}^{k}\sum_{\omega\in P_i}(\overline{C}_{2,i}^{intra}-\overline{C}_2)^2\\ -2.\sum_{i=1}^{k}\left((\overline{C}_{2,i}^{intra}-\overline{C}_2).\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra})\right)\end{array}\right)$$

$$Var^{totale}(C_2) = \frac{1}{card}.\left(\sum_{i=1}^{k}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra})^2+\sum_{i=1}^{k}\sum_{\omega\in P_i}(\overline{C}_{2,i}^{intra}-\overline{C}_2)^2-2.\sum_{i=1}^{k}\left((\overline{C}_{2,i}^{intra}-\overline{C}_2).0\right)\right)$$

$$Var^{totale}(C_2) = \frac{1}{card}.\left(\sum_{i=1}^{k}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra})^2+\sum_{i=1}^{k}\sum_{\omega\in P_i}(\overline{C}_{2,i}^{intra}-\overline{C}_2)^2\right)$$

$$Var^{totale}(C_2) = \frac{1}{card}.\left(\sum_{i=1}^{k}k_i.\frac{1}{k_i}\sum_{\omega\in P_i}(C_2(\omega)-\overline{C}_{2,i}^{intra})^2+\sum_{i=1}^{k}k_i.(\overline{C}_{2,i}^{intra}-\overline{C}_2)^2\right)$$

$$Var^{totale}(C_2) = \frac{1}{card}.\sum_{i=1}^{k}k_i.Var(C_{2,i}^{intra})+\frac{1}{card}.\sum_{i=1}^{k}k_i.(\overline{C}_{2,i}^{intra}-\overline{C}_2)^2$$

$$Var^{totale}(C_2) = Var^{inter}(C_2) + Var^{intra}(C_2)$$

Croisement qualitatif × quantitatif (4)

- Pour étudier le lien entre un caractère qualitatif 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.
- 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 k_i où est l'effectif de chaque sous population.
- 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.
- Nous avons l'égalité suivante : $Var^{totale}(C_2) = Var^{inter}(C_2) + Var^{intra}(C_2)$
- On en déduit une mesure du lien entre C_1 et C_2 avec l'expression $\frac{Var^{inter}(C_2)}{Var^{totale}(C_2)}$
- Cette expression varie entre 0 et 1. Plus sa valeur est proche de 1 plus les deux caractères sont liés