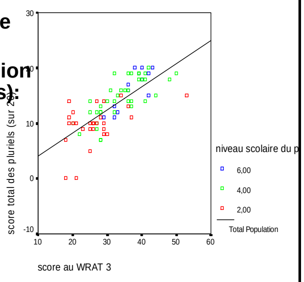
Objectifs du chapitre sur la corrélation

- Comprendre ce qu'est une corrélation
- Savoir représenter une corrélation à l'aide
  - d'un diagramme de dispersion
  - du coefficient de corrélation

2 façons de représenter une corrélation (1)

- <u>Visuellement</u>:
  - tableau de contingence (données qualitatives)
  - diagramme de dispersion (données quantitatives): coordonnées de 2 variables sur plan cartésien



2 façons de représenter une corrélation (2)

- <u>Mathématiquement</u>:
  - coefficient de Bravais-Pearson
    - Coefficient standardisé de covariance (données continues)
    - variant entre –1 et +1
    - 0 indique l'absence d'association
  - indique
    - la force et
    - la direction de la relation
  - autres:
    - φ,
    - Spearman,
    - bisériel,
    - etc.

L'équation du coefficient de corrélation de Bravais-Pearson

- La formule conceptuelle

$$r_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

- La formule de calcul

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

Mnémonique

- La formule conceptuelle …

$$r_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

- s'écrit aussi

$$r_{xy} = \frac{Co\,var_{xy}}{s_x\, s_y}$$

Un exemple: calcul et interprétation: association ou cause-effet?

# cigognes	population
125	55000
150	55500
175	65000
200	67500
248	68000
250	70000
250	75000

Un exemple: calcul

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

Cigo-gnes	Popu-lation	C^2	P^2	CxP
125	55000			
150	55500			
175	65000			
200	67500			
248	68000			
250	70000			
250	75000			

Exemple de calcul (1)

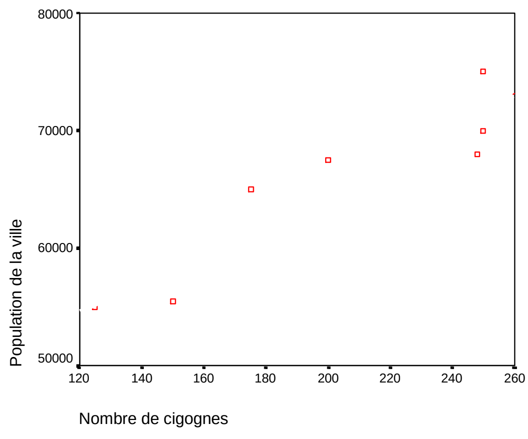
1398	456000	295254	3.0035 x 10^{10}	93189000
∑X	∑Y	∑X²	∑Y²	∑XY

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

$$r_{xy} = \frac{93189000 - \frac{(1398)(456000)}{7}}{\sqrt{295254 - \frac{1398^2}{7}}\sqrt{30035000000 - \frac{456000^2}{7}}}$$

$$r_{xy} = \frac{93189000 - \frac{637480000}{7}}{\sqrt{295254 - \frac{1954404}{7}}\sqrt{30035000000 - \frac{207930000000}{7}}}$$

$$r_{xy} = \frac{2119286}{\sqrt{16053.429}\ldots} = \frac{93189000 - 91068714}{\ldots} = \frac{2119286}{2301135.3} = .920974$$



Problèmes 9.1-9.3

bébé	fertile	% nais.
6,1	43,0	9,2
7,1	55,3	12,0
7,4	48,5	10,4
6,3	38,8	9,8
6,5	46,2	9,8
5,7	39,9	7,7
6,6	43,1	10,9
8,1	48,5	9,5
6,3	40,0	11,6
6,9	56,7	11,6

Problèmes 9.1-9.3: calculs

Y	X_1	X_2	YX_1	YX_2	Y^2	X_1^2	X_2^2	X_1X_2
6,1	43,0	9,2						
7,1	55,3	12,0						
7,4	48,5	10,4						
6,3	38,8	9,8						
6,5	46,2	9,8						
5,7	39,9	7,7						
6,6	43,1	10,9						
8,1	48,5	9,5						
6,3	40,0	11,6						
6,9	56,7	11,6						

Calcul 9.1

Y	X_1	YX_1	Y^2	X_1^2
67	460	3106,54	453,28	21516

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

$$r_{xy} = \frac{3106{,}54 - \frac{(460)(67)}{10}}{\sqrt{21516 - \frac{460^2}{10}}\sqrt{453{,}28 - \frac{67^2}{10}}}$$

$$r_{xy} = \frac{3106{,}54 - \frac{30820}{10}}{\sqrt{21516 - \frac{211600}{10}}\sqrt{453{,}28 - \frac{4489}{10}}} = \frac{3106{,}54 - 3082}{\sqrt{21516 - 21160}\sqrt{453{,}28 - 448{,}9}}$$

$$r_{xy} = \frac{24{,}54}{\sqrt{356}\sqrt{4{,}38}} = \frac{24{,}54}{18{,}87 \times 2{,}09} = \frac{24{,}54}{39{,}49} = {,}62$$

r_{xy} en tant qu'estimateur de ρ

- biais:
  - surévaluation
  - solution: réduire l'indice
- correction mathématique:

$$r_{ajusté} = \sqrt{1 - \frac{(1 - r_{xy}^2)(N-1)}{(N-2)}}$$

utilisation en régression,
alors N-2 = N-p-1

Comment savoir si r_{xy} égale 0 ou non?

- test t:

$$t_{(dl)} = \frac{r_{xy}\sqrt{N-2}}{\sqrt{1 - r_{xy}^2}}$$

avec N-2 comme dl

D'autres coefficients de corrélation (1)

- ϕ: corrélation pour 2 variables nominales

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

- r_{pb}: corrélation entre une variable nominale et une variable à intervalles égaux

$$r_{pb} = \frac{(M_p - M_q)(\sqrt{n_p n_q})}{N\sigma_{total}}$$

= r

D'autres coefficients de corrélation (2)

- rb: corrélation pour une variable nominale en pratique mais continue en réalité

$$r_b = \frac{(M_p - M_q)}{N\sigma_{total}} x \frac{(n_p n_q)}{y}$$

- ρ: corrélation entre deux variables ordinales

$$r_s = 1 - \left(\frac{6\sum_{i=1}^{N} D^2}{N(N^2-1)} \right)$$

Un exemple de calcul pour le r de Spearman (1)

# cigo-gnes	Rang cig.	Popu-lation	Rang popu.	Diff. rangs
125		55000		
150		55500		
175		65000		
200		67500		
248		68000		
250		70000		
250		75000		

Un exemple de calcul pour le r de Spearman (2)

$$r_s = 1 - \left(\frac{6\sum_{i=1}^{N} D^2}{N(N^2-1)} \right) = 1 - \left(\frac{6 \times 0{,}5}{7 \times (49-1)} \right) =$$

$$r_s = 1 - \left(\frac{3}{7(48)} \right) = 1 - \left(\frac{1}{7 \times (16)} \right) = 1 - \frac{1}{112} = 0{,}99107142857$$

Problèmes 9.1-9.3

bébé	rb	fertile	rf	Diff.	D^2
6,1		43,0			
7,1		55,3			
7,4		48,5			
6,3		38,8			
6,5		46,2			
5,7		39,9			
6,6		43,1			
8,1		48,5			
6,3		40,0			
6,9		56,7			

le r de Spearman du problème 9.1

$$r_s = 1 - \left(\frac{6\sum_{i=1}^{N} D^2}{N(N^2-1)} \right) = 1 - \left(\frac{6 \times 32}{10 \times (100-1)} \right) =$$

$$r_s = 1 - \left(\frac{192}{10(99)} \right) = 1 - \left(\frac{192}{990} \right) = 1 - 0{,}1939 = 0{,}8061$$

De la corrélation aux corrélations

- **les tableaux de corrélation**:
  - relations multiple 2 x 2
    - savoir trouver les patrons à l'œil
- **l'analyse factorielle**:
  - approche mathématique pour réduire les tableaux de corrélations en éléments principaux

Un exemple: calcul et interprétation: association ou cause-effet?

# cigognes	population
125	55000
150	55500
175	65000
200	67500
248	68000
250	70000
250	75000

Réflexion

Cote 1	Cote 2
1	11
2	12
3	13
4	14
5	15
6	16
7	17
8	18