Ch2 : Statistique descriptive bivariée

Introduction: La statistique descriptive à deux dimensions a essentiellement pour but d'examiner s'il existe une certaine forme d'association entre deux variables. Par exemple, étudier l'association

- entre le revenu annuel et le nombre d'années de scolarité.
- entre l'âge et la maladie.

Ces observations peuvent être de différentes natures.

1 Variables quantitatives

Dans ce paragraphe, nous traitons une forme particulière d'association appelée corrélation linéaire.

- Diagramme de dispersion

  Pour visualiser graphiquement une corrélation, on prélève un échantillon de taille n et nous observons ensuite, sur chaque unité de l'échantillon, les valeurs de deux variables statistiques X et Y. On dispose alors de n couples d'observations (x_i, y_i) que l'on reporte sur un graphique en prenant pour abscisse la variable X et Y pour ordonnée. Le diagramme qui résulte est appelé diagramme de dispersion

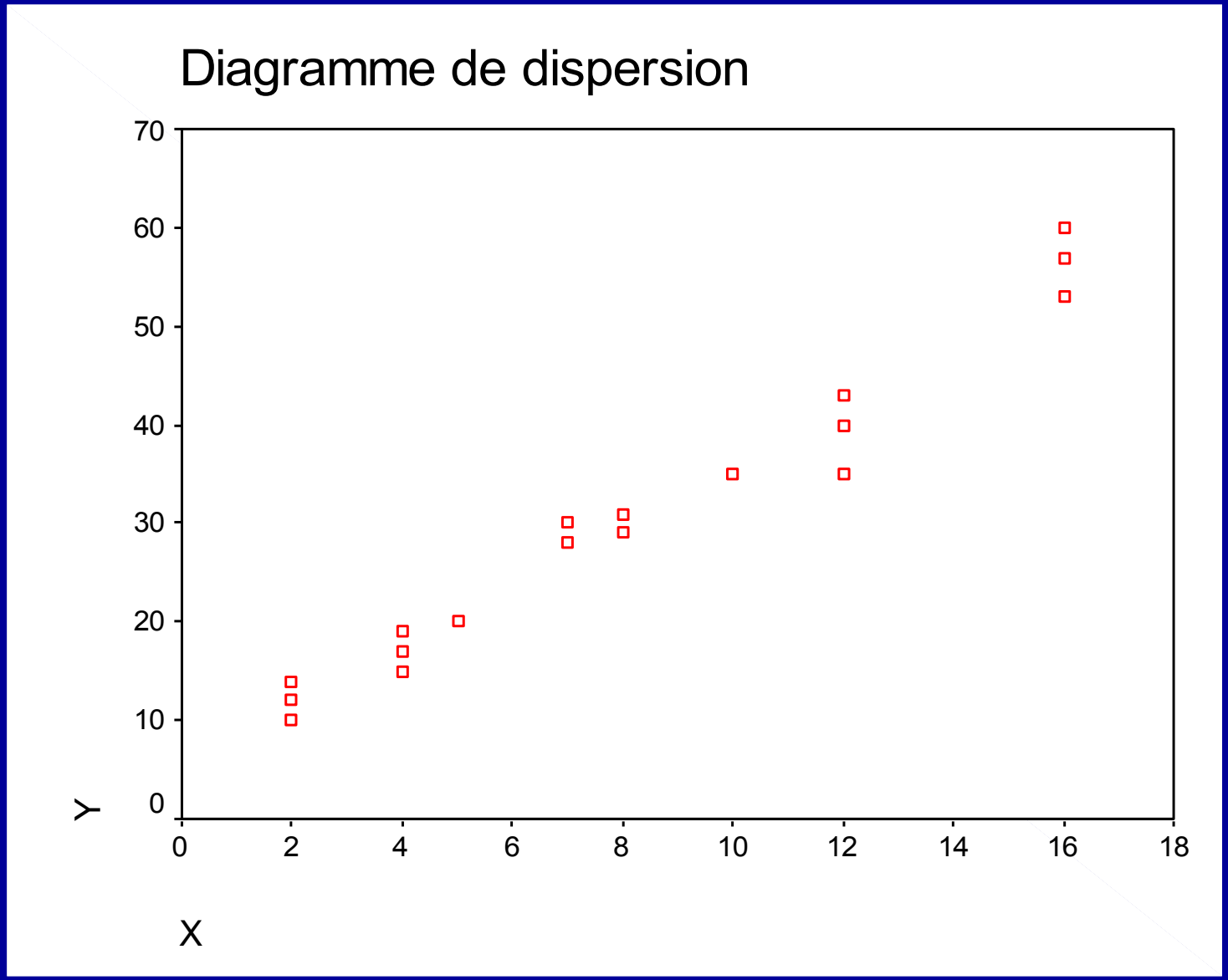


Diagramme de dispersion
70
60
50
40
30
20
10
0
Y
0
2
4
6
8
10
12
14
16
18
X

- Mesure de dépendance : Coefficient de corrélation

Le coefficient de corrélation ou Coefficient de Bavais Pearson est un indice qui permet de mesurer l'intensité de l'association linéaire entre deux variables. Il se calcule à partir des observations (x_i, y_i) de la manière suivante :

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

où

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

Covariance de (x,y)

$$S_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

sont les variances de x et y

Remarques :

- r est réservé uniquement à des nuages à répartition plus ou mois linéaire.
- Une corrélation r positive indique une dépendance linéaire positive, tandis qu'une corrélation négative r correspond à une dépendance linéaire négative.
- Un coefficient de corrélation est un nombre sans unité compris entre -1 et 1.

Définition:

- Les deux variables sont dites « fortement corrélées» lorsque la coefficient de corrélation est proche de 1 en valeur absolue

- Et «faiblement corrélées lorsque le coefficient de corrélation est proche de 0.

La Régression linéaire simple :

Lorsque le nuage de points du diagramme de dispersion est disposé avec une tendance linéaire et que les deux variables quantitatives sont fortement corrélées, on cherche à établir, dans un but explicatif et prévisionnel, une relation (qu'on suppose) linéaire reliant les deux variables de la forme $y = ax + b$, ou a et b sont deux réels à estimer. Le critère utilisé pour déterminer a et b est le critère des «moindres carré».

Critère des moindres carrés

L'idée de la méthode des moindres carrés est de choisir a et b de façon à minimiser la somme quadratique des erreurs $e_i = y_i - (ax_i + b)$, c'est à dire minimiser

$$\sum e_i^2 = L(a,b)$$

par rapport à a et b.

Les estimateurs des moindres carrés de a et b sont

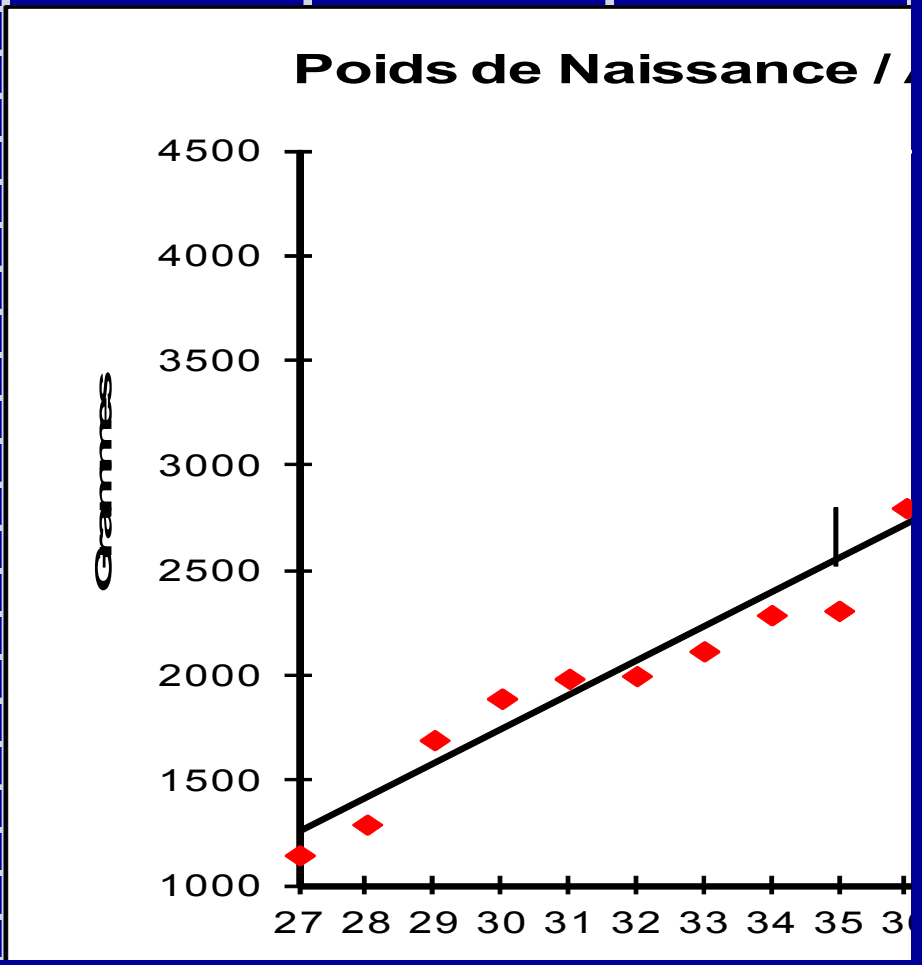
$$\hat{a} = \frac{S_{xy}}{S_{xx}} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

L'équation $y = \hat{a}x + \hat{b}$

sert alors à prédire les valeurs de la variable Y sachant les valeurs de la variable X.

Exemple: Poids de naissance en fonction du terme de la grossesse

semaines	Poids
27	1146
28	1292
29	1694
30	1892
31	1986
32	2000
33	2118
34	2290
35	2310
36	2800
37	3019
38	3210
39	3450
40	3475
41	3553
42	3582
43	3604



2 Variables qualitatives ordinales

Lorsque les deux variables sont qualitatives ordinales, on utilise la corrélation non-paramétrique par le biais du coefficient de Spearman défini par

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

- D représente, pour chaque observation, les différences de rang obtenues sur les deux variables.

- Lorsque r_s est proche de 1 en valeur absolue, on dira qu'il y a une forte liaison entre les deux caractères. Sinon, i.e si r_s est proche de 0, il n'y a pas de liaison ou de corrélation entre les deux variables.
- Mais attention, on ne saura déterminer la nature de cette liaison quand elle existe.

Exemple:

Ech.	Be	Zr	Rang Be	Rang Zr	D*D
1	1,71	62,04	5	5	0
2	1,91	71,50	10	13	9
3	1,98	68,40	12	11	1
4	1,74	61,25	7	4	9
5	1,87	64,16	9	7	4
6	1,38	58,49	3	3	0
7	0,99	30,33	1	1	0
8	1,13	39,55	2	2	0
9	1,65	64,71	4	8	16
10	2,26	71,47	16	12	16
11	1,72	63,14	6	6	0
12	1,77	67,09	8	9	1
13	2,31	85,68	17	15	4
14	2,09	88,52	15	17	4
15	2,03	88,30	14	16	4
16	2,02	77,45	13	14	1
17	1,91	68,20	11	10	1
				Somme D	70

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

$$r_s = 1 - \frac{6 \times 70}{17 \times (17^2-1)}$$

$$r_s = 0.914$$

3 Variables qualitatives nominales (ou ordinales):
Exemple 1: Cas Rola-Cola

Question :

Existe-t-il une liaison entre la boisson préférée (X)
et le goût pour le sucre (Y)?

Les données :

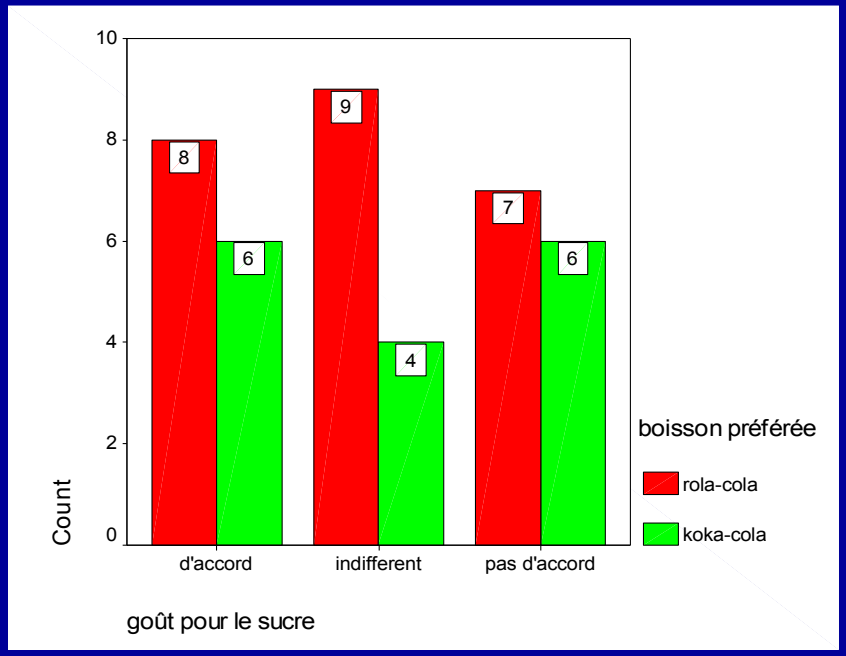
boisson préférée * goût pour le sucre Crosstabulation

Count

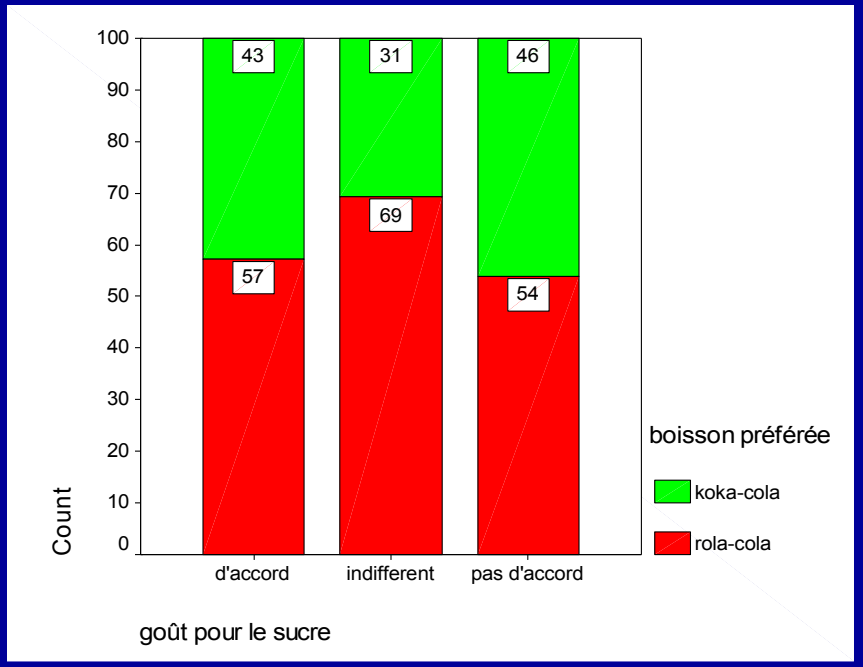
		goût pour le sucre			
		d'accord	indifferent	pas d'accord	Total
boisson préférée	rola-cola	8	9	7	24
	koka-cola	6	4	6	16
Total		14	13	13	40

Représentations graphiques du tableau de contingence

En effectif



En pourcentage



Il n'y a pas de relation entre la boisson préférée et le goût pour le sucre

On peut Construire le tableau *Boisson préférée*Goût pour le sucre* ayant les mêmes marges que le tableau d'origine et correspondant à l'indépendance

	Goût pour le sucre				
Boisson préférée	D'accord	indifférent	Pas d'accord	Total (effectif)	Total (pourcentage)
Rola-Cola Koka-Cola				24 16	60 % 40 %
Total	14	13	13	40	100 %

Distance entre le tableau observé et le tableau correspondant à l'indépendance

boisson préférée * goût pour le sucre Crosstabulation

			goût pour le sucre			
			d'accord	indifferent	pas d'accord	Total
boisson préférée	rola-cola	Count	8	9	7	24
		Expected Count	8.4	7.8	7.8	24.0
	koka-cola	Count	6	4	6	16
		Expected Count	5.6	5.2	5.2	16.0
Total		Count	14	13	13	40
		Expected Count	14.0	13.0	13.0	40.0

$$\chi^2 = \frac{(8-8.4)^2}{8.4} + \frac{(9-7.8)^2}{7.8} + \ldots + \frac{(6-5.2)^2}{5.2}$$

Exemple ***2:*** On désire vérifier si les intentions de vote aux prochaines élections provinciales varient d'une région à l'autre. Le tableau suivant donne les résultats d'un sondage effectué auprès de 1000 personnes choisies au hasard dans trois régions différentes.

	Région 1	Région 2	Région 3	Total
PLQ	205	155	100	460
PQ	125	200	80	405
Autre	70	45	20	135
Total	400	400	200	1000

On considère le tableau suivant :

Classes	A_1	A_2	.....	A_k	Total des lignes
B_1	n_{11}	n_{12}	.....	n_{1k}	$n_{1.}$
B_2	n_{21}	n_{22}	.....	n_{2k}	$n_{2.}$
B_r	n_{r1}	n_{r2}	.....	n_{rk}	$n_{r.}$
Total des colonnes	$n_{.1}$	$n_{.2}$	.....	$n_{.k}$	n

où n_{ij} représente la fréquence observée relative à la classe $A_i \times B_j$

Pour quantifier l'indépendance entre les deux caractères, on utilise le coefficient Khi-deux définie par:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - n_{eij}\right)^2}{n_{eij}}$$

où

$$n_{eij} = \frac{n_{i.} n_{.j}}{n}$$

On constate

- que χ^2 est toujours positif.
- Plus χ^2 est faible et plus on est proche de l'indépendance et plus il est grand est plus on est proche de l'association (voir table de la loi de χ^2).
- χ^2 n'est pas majoré.
- χ^2 est sensible à la taille de l'échantillon.

Pour y remédier, on introduit d'autres coefficients, le coefficient phi et de V Cramer par exemple…

Le coefficient V de Cramer

- Contrairement à χ^2 Le coefficient V Crammer est stable quand n est grand et est défini par

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n(\min(l,c)-1)}}$$

Plus V est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Il vaut 1 en cas de complète dépendance.

4. Une variable quantitative et une variable qualitative :

Etudier la relation entre la variable quantitative «âge» et variable qualitative «situation familiale» dans une population de 30 personnes (masculins).

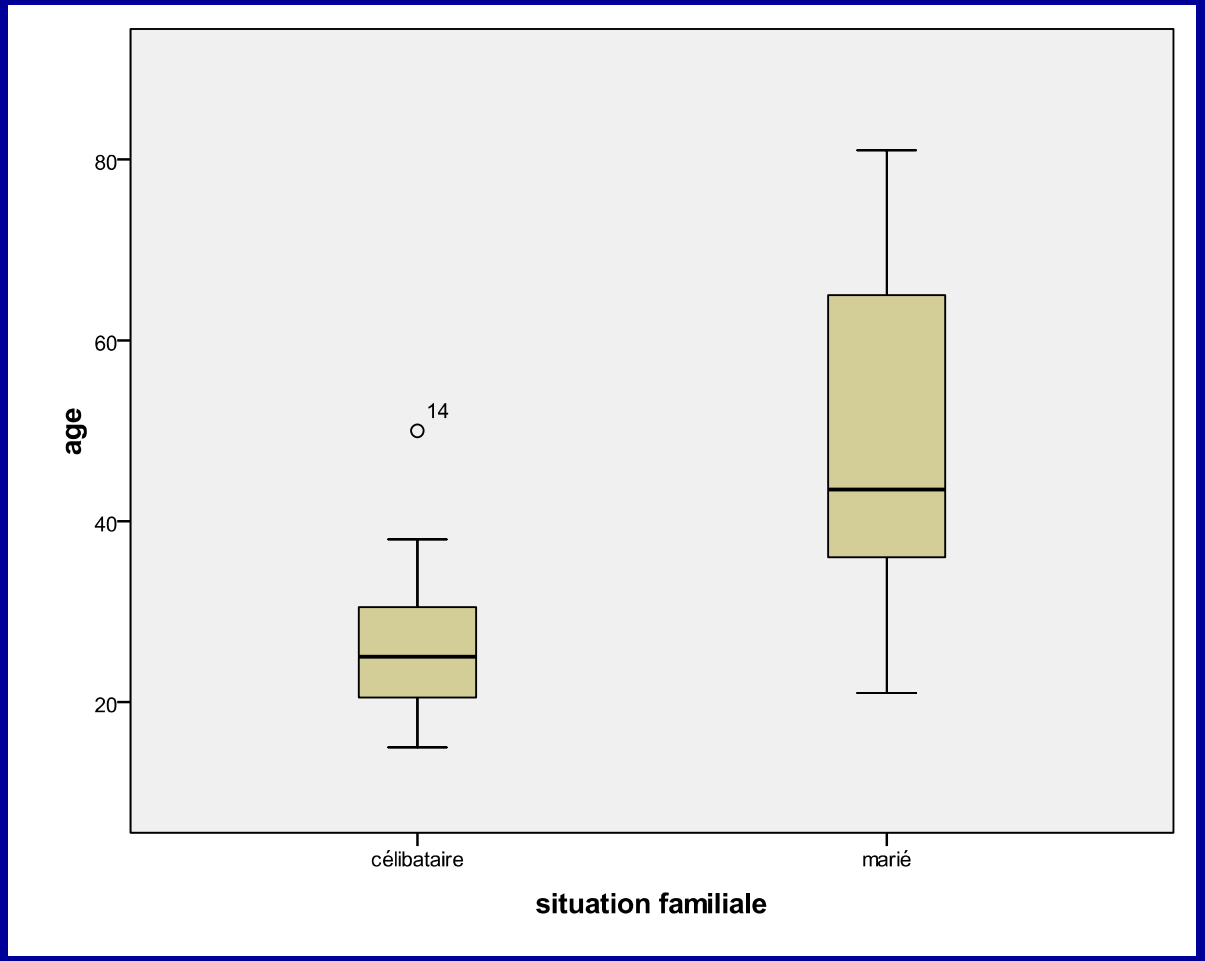
Données brutes

Age	Situation familiale
22	Célibataire
37	Célibataire
19	Célibataire
50	Célibataire
38	Célibataire
24	Célibataire
17	Célibataire
28	Célibataire
29	Célibataire
31	Célibataire
23	Célibataire
24	Célibataire
26	Célibataire
18	Célibataire

La suite du tableaux

30	Célibataire
15	Célibataire
75	Marié
37	Marié
45	Marié
81	Marié
33	Marié
21	Marié
42	Marié
29	Marié
65	Marié
53	Marié
36	Marié
49	Marié
40	Marié
68	Marié

Le diagramme ayant la forme suivante :



Commentaire:

La forme du box-plot varie en fonction de la situation familiale. Cela veut dire qu'on pourrait dire que la situation familiale dépend de l'âge.