Statistique Descriptive Bidimensionnelle

Séverine Affeldt

MLDS - Centre Borelli

Université de Paris

Analyse de la liaison entre deux variables quantitatives

Visualisation: Faire une représentation de type nuage de points entre les variables X et Y. la forme du nuage donne une indication sur la relation entre les deux variables.

Quantification de la liaison: Calculer un coefficient de relation entre les deux variables (e.g. coefficient de corrélation linéaire). On pourra obtenir l'intensité de la liaison et, selon les coefficient, le signe.

Liaison causale: En cas de relation linéaire, on peut faire une régression linéaire d'une variable sur l'autre. S'il existe une relation causale entre les variables, la régression nous permet alors de prédire la variable réponse en fonction de la variable causale.

Le nuage de points

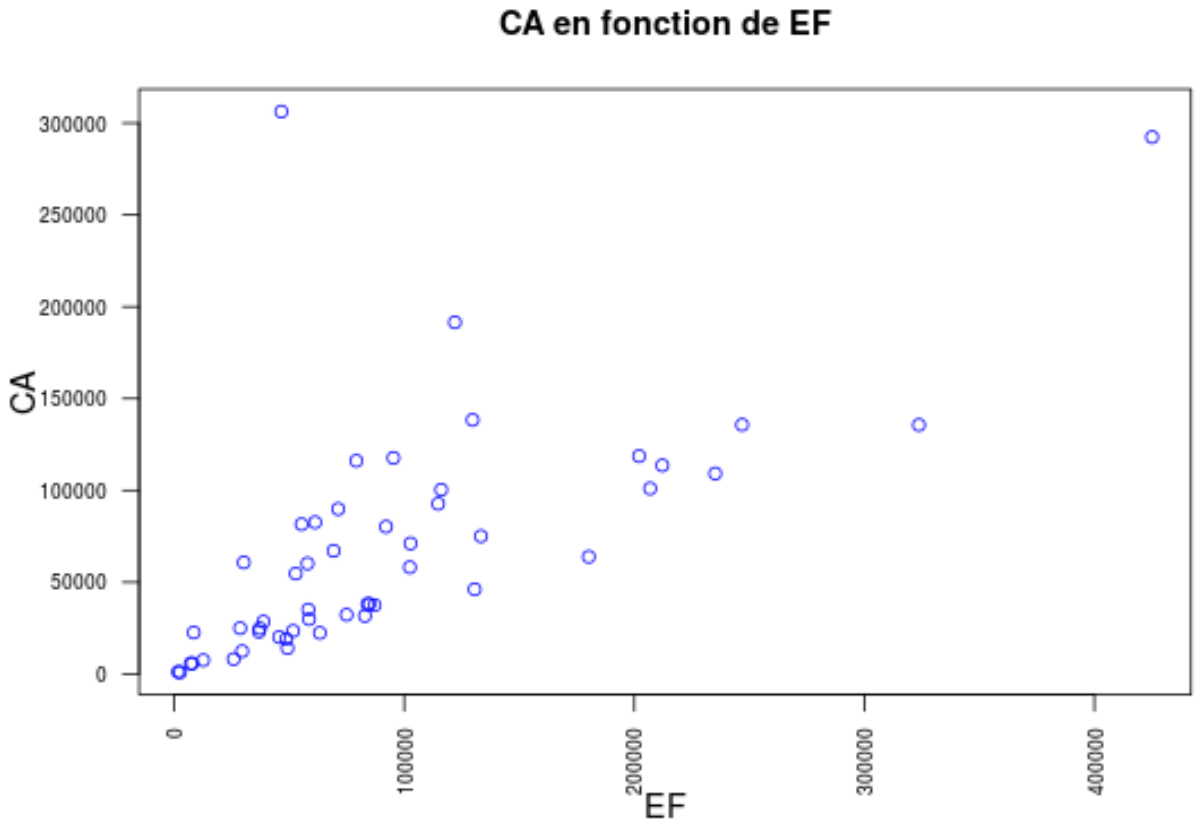
Le nuage de point permet de visualier la **variation conjointe** de deux variables X et Y. Chaque individus observé i est représenté par un point d'abscisse x_i et d'ordonnée y_i. Ce type de diagramme est également apprelé **diagramme de dispersion** (*scatter-plot*).

Les données

secteursActivite.csv

Pour 51 secteurs d'activités en France, on s'intéresse au nombre total d'entreprise (NB), à l'effectif salarié (EF) et au chiffre d'affaire hors-taxes en millions de francs (CA). Nous avons donc $n = 51$ individus et 3 variables. Le tableau ci-dessous résume les données. (Source: *Tableau de l'Economie Française*, INSEE, 1989, p.109)

Variable	Minimum	Maximum	Moyenne	Ecart-type
NB	11	41866	4135	7435
EF	1701	425082	92591	83832
CA	992	306293	68010	64532




```r
# Définition des noms de variables quantitatives à comparer
x.var = "EF"; y.var = "CA";
# Plot simple des valeurs
plot(x = data()[, x.var], y = data()[, y.var], col = "blue",
     las = 2, cex.axis = 0.7,
     main = paste(y.var, "en fonction de", x.var),
     xlab = x.var, ylab = y.var, cex.lab = 1.2)

```

Nuage de points dans une application Shiny

01_analyse_bidim_quantitative.R

Variable quantitative

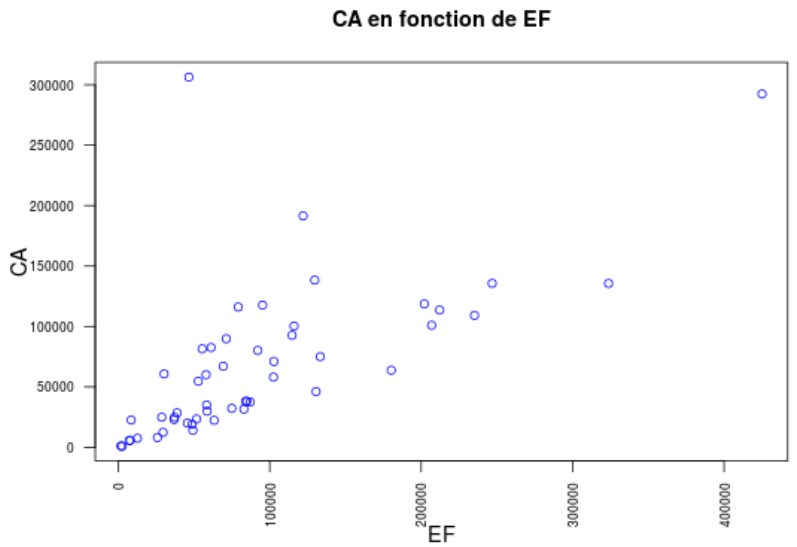
Choisir un fichier CSV

Browse... secteursActivite.csv

Upload complete

Load

Nuage de points | Caractéristiques | Table



Nuage de points dans une application Shiny

01_analyse_bidim_quantitative.R

Nuage de points | Caractéristiques | Table

	Minimum	Maximum	Moyenne	Ecart-type
NB	11	41866	4135	7435
EF	1701	425082	92591	83832
CA	992	306293	68010	64532

Nuage de points dans une application Shiny

01_analyse_bidim_quantitative.R

Nuage de points | Caractéristiques | Table

code	secteur	NB	EF	CA
4	Prod. Combustibles min.solides, cokéfaction	19	49251	14111
5	Production de pétrole et de gaz naturel	120	46594	306293
6	Production et distribution d'électricité	731	129723	138389
7	Distribution de gaz	103	30255	60767
8	Distribution d'eau et chauffage urbain	227	28658	24957
9	Extraction et préparation de minerai de fer	25	2582	992
10	Sidérurgie	67	92085	80237
11	Première transformation de l'acier	238	37204	24853
12	Extract. Et prépar. De minerais non ferreux	40	1701	1159
13	Métallurgie, 1ere transf. Des mét. Non ferreux	303	55374	81550
14	Production de minéraux divers	270	12621	7535

Voir aussi les deux nouveaux panels,

- *Histogrammes dos à dos*
- *Nuage de points et Histogrammes*

Covariance

La covariance de deux variables quantitatives X et Y est la moyenne du produit des écarts aux moyennes,

$$cov(X, Y) = s_{XY} = \frac{1}{n} \sum_{i=1}^{n} [x_i - \bar{x}][y_i - \bar{y}]$$

C'est une grandeur **symétrique** ($s_{XY} = s_{YX}$) et **réelle**.

Attention: sa valeur dépend des unités de mesure dans lesquelles sont exprimées X et Y!

Coefficient de corrélation linéaire

Le coefficient de **corrélation linéaire** (ou coefficient de Pearson) ne dépend pas des unités de mesure des deux variables. En effet, il est le **rapport** entre la **covariance** et le **produit des écarts-types**,

$$corr(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

C'est une grandeur **symétrique** ($r_{XY} = r_{YX}$) et **réelle** comprise entre -1 et $+1$.

Attention: le coefficient de corrélation linéaire ne donne aucune information de causalité.

Variable centrée et réduite

Soit X une variable quantitative de moyenne $\bar{x}$ et d'écart-type s_x.

La **variable centrée** associée à X est $X - \bar{x}$. Elle est de moyenne nulle et d'écart-type s_x.

La **variable centrée et réduite** associée à X est $\frac{X-\bar{x}}{s_X}$. Elle est de moyenne nulle et d'écart-type $s_x = 1$.

Donc, la covariance de deux variables centrées et réduites est égale à leur coefficient de corrélation linéaire (leur variance étant égale à 1)

Interprétation du coefficient de corrélation

Un signe est **positif** indique que les variables X et Y varient dans le même sens. Un signe **négatif** indique une variation opposée.

La **force** de la relation entre les variables X et Y est donnée par la **valeur absolue** du coefficient de corrélation linéaire. Une liaison forte donne un coefficient de valeur absolue proche de 1 et une liaison faible est indiquée par une coefficient proche de 0.

Si le coefficient de corrélation linéaire vaut **+1** ou **-1**, il existe une **équation affine** ente les variables X et Y. Attention: la causalité ne peut pas être identifiée avec ce coefficient.

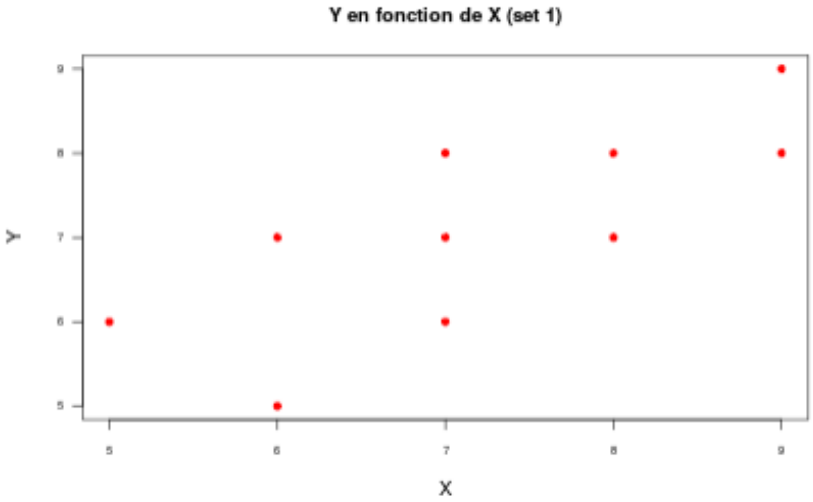
Exemple de calcul du coefficient de corrélation entre EF et CA

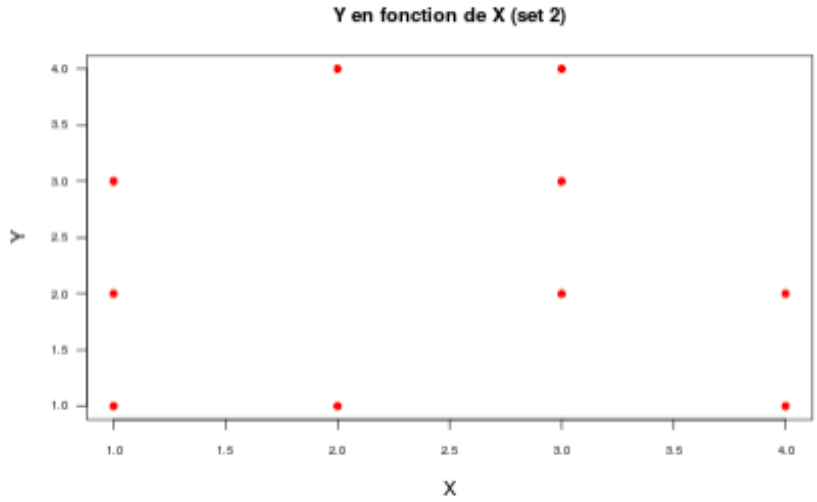
02_analyse_bidim_quantitative_corr.R

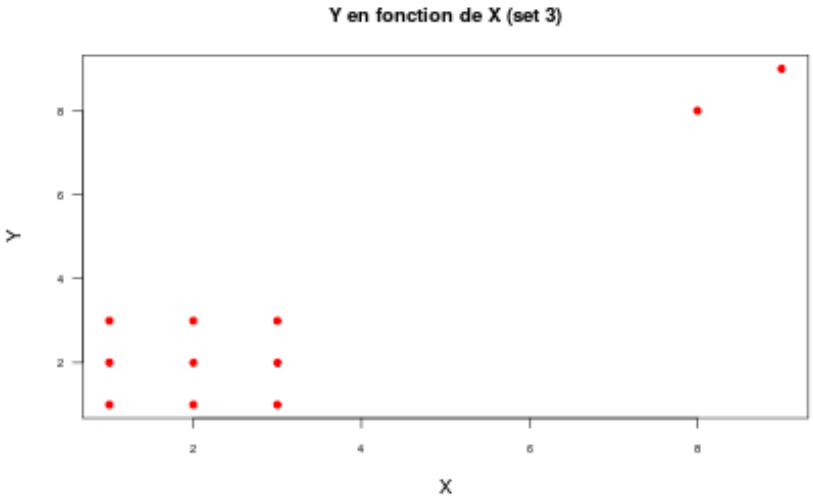
- la liaison linéaire est positive et moyenne (0.66)

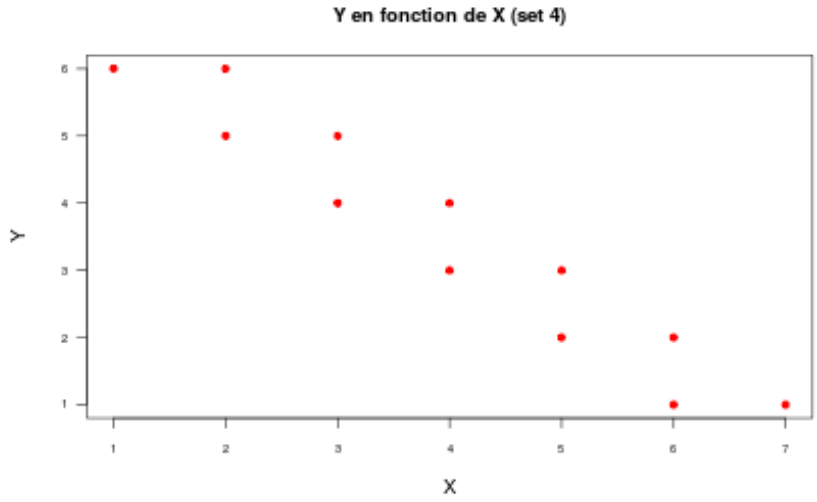
$\Rightarrow$ en moyenne, plus le nombre de salariés d'un secteur est important, plus le chifre d'affaire est grand

! la relation a un force moyenne, d'autres facteurs que EF doivent influencer CA









	moy. X	moy. Y	sd. X	sd. Y	corr(X, Y)
data1	7.20	7.10	1.25	1.14	0.76
data2	2.40	2.30	1.11	1.10	-0.02
data3	3.18	3.18	2.62	2.62	0.92
data4	4.00	3.50	1.78	1.78	-0.96

03_analyse_bidim_quantitative_corr_expNuages.R

Régression linéaire entre deux variables quantitatives

Quand? Le nuage de point indique nettement une relation, le coefficient de corrélation linéaire est fort et on suppose *a priori* une relation de causalité (e.g. X cause Y).

Comment? On utilise le critère des moindres carrés, i.e. on minimise la quantité,

$$F(a,b) = \sum_{i=1}^{n} \{y_i - [ax_i + b]\}^2$$

La solution de cette minimisation est,

$$\hat{a} = \frac{s_{XY}}{s_X^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

Exemple: Régression linéaire de la variable CA sur la variable EF,

$$\hat{a} = \frac{cov(EF, CA)}{var(EF)} =\simeq 0.509$$

$$\hat{b} = \bar{CA} - \hat{a}\bar{EF} =\simeq 20884$$

Variable quantitative

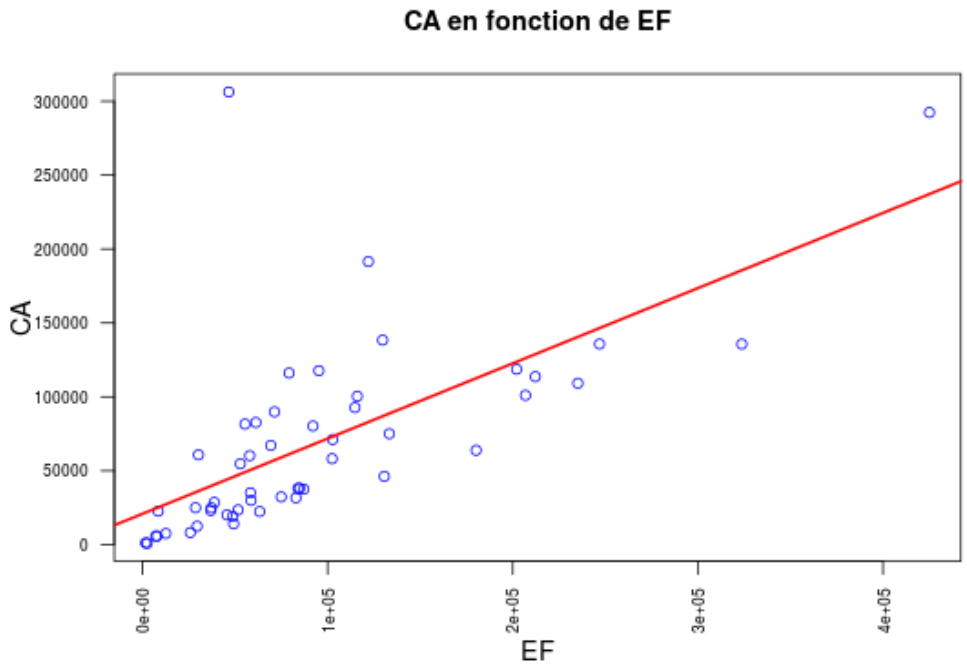
Choisir un fichier CSV

Browse... secteursActivite.csv

Upload complete

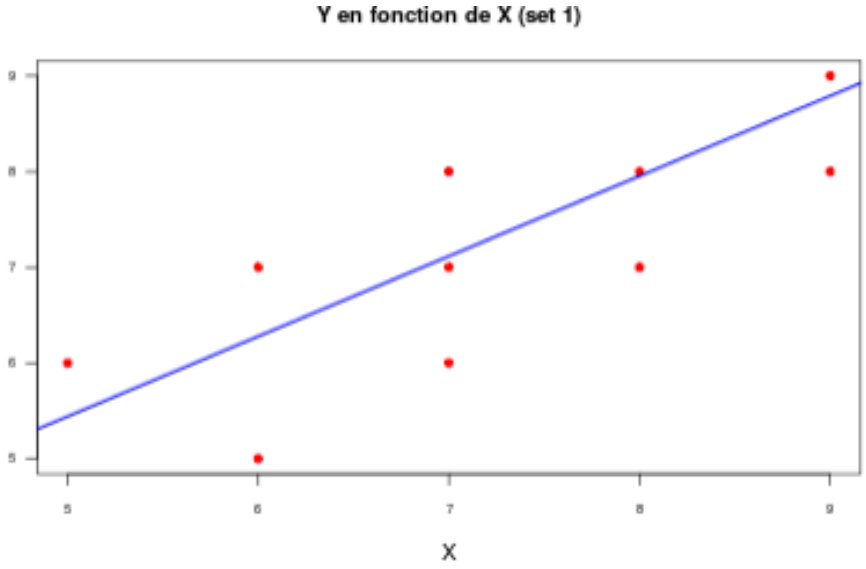
Load

Nuage de points | Caractéristiques | Table | Aide

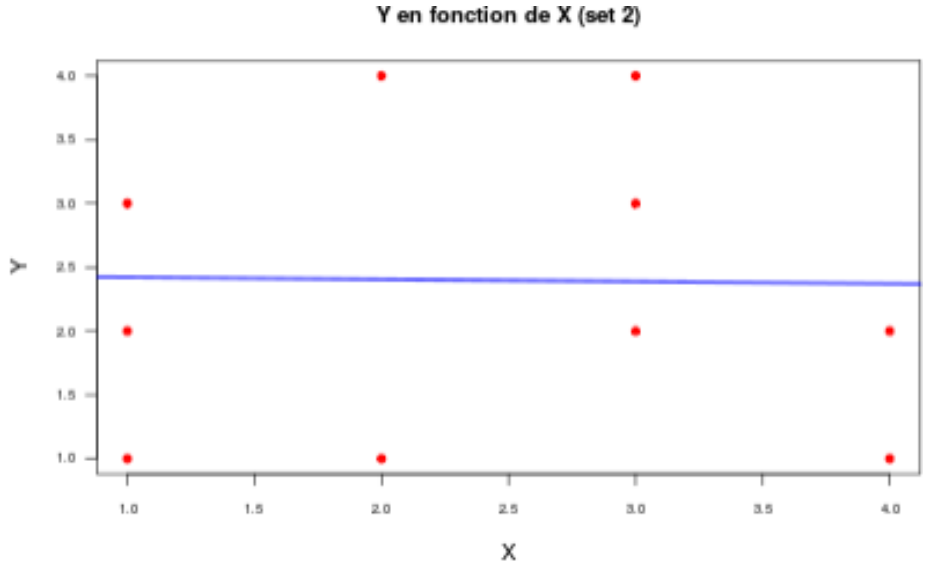


Coefficient de corrélation linéaire = 0.66

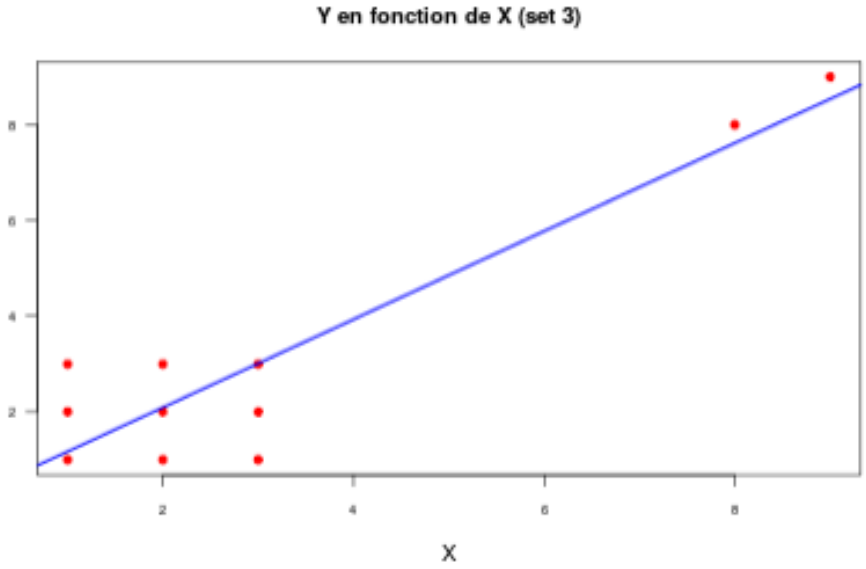
04_analyse_bidim_quantitative_corr_regression.R



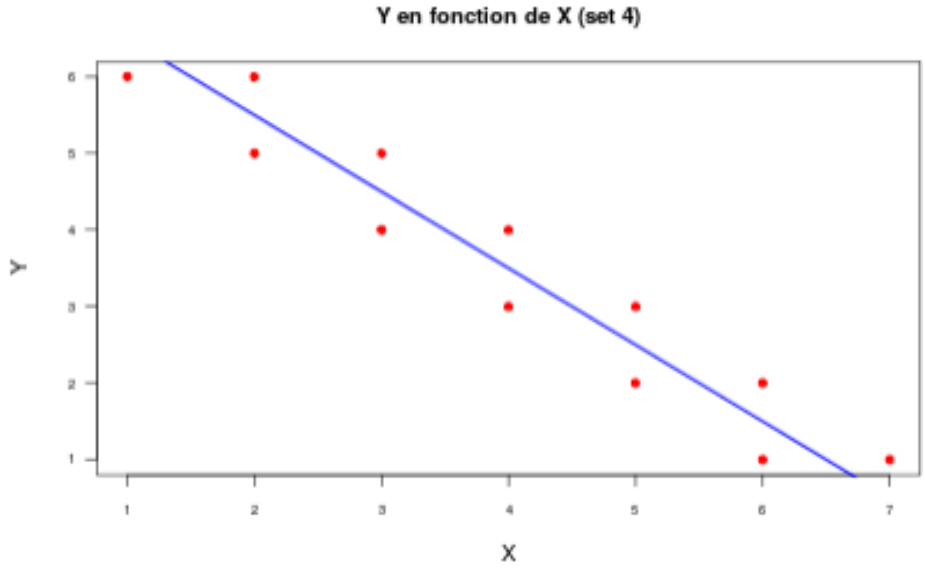
Coefficient de corrélation linéaire = 0.76 (lien fort)



Coefficient de corrélation linéaire = -0.02 (pas de lien)



Coefficient de corrélation linéaire = 0.92 (artéfact du aux outliers)



Coefficient de corrélation linéaire = -0.96 (lien fort))

04_analyse_bidim_quantitative_corr_regression.R

Variable quantitative *vs* qualitative

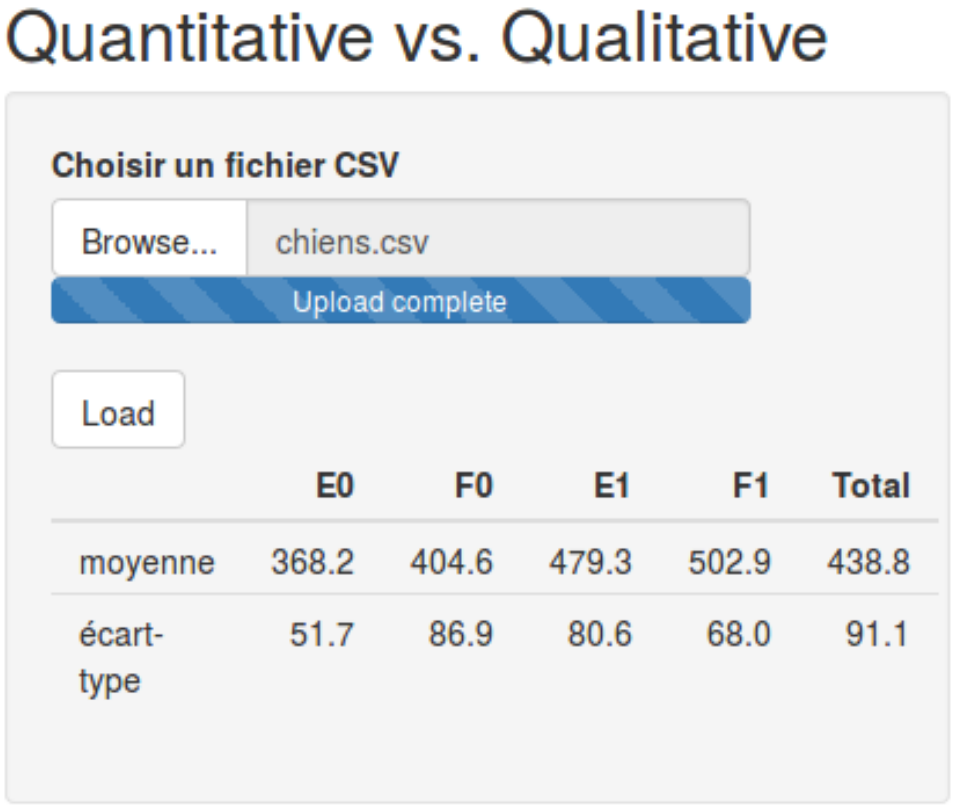
On considère une variable quantitative Y définit sur les r **modalités** d'une variable qualitative X. On peut analyser Y (e.g. moyenne, écart-type) pour chaque **classe** C_l, définie par X.

Les données[1]

chiens.csv

On considère 19 chiens ayant reçu du pentobarbital et on observe l'effet sur leur **rythme cardiaque** via une variable **quantitative Y**. Y mesure en millisecondes le temps entre deux battements. La variable **qualitative X** est la combinaison de deux facteurs: (1) la pression d'administration de CO_2 (élevé **E** ou faible **F**), et (2) la présence d'halothane (présence **1** ou absence **0**). On a donc **4 modalités**: E0, F0, E1 et F1.

[1] extrait de *Applied Multivariate Statistical Analysis*, R.A. Johnson & D.W. Wichern, 2007



	E0	F0	E1	F1	Total
moyenne	368.2	404.6	479.3	502.9	438.8
écart-type	51.7	86.9	80.6	68.0	91.1

05_analyse_bidim_quantQual.R.R

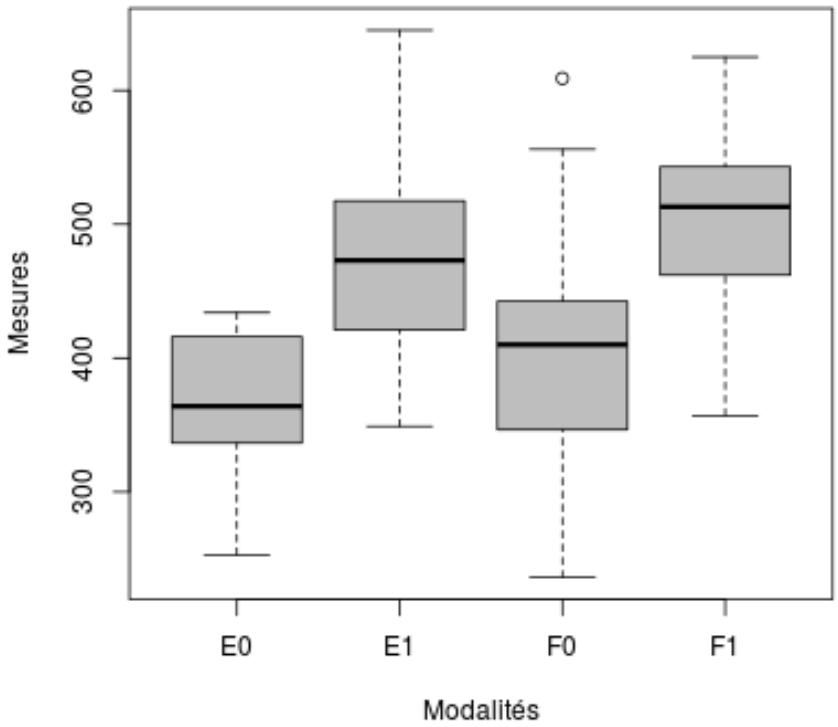
On estime la moyenne et l'écart-type de Y pour chaque classe C_l. On calcule également ces valeurs pour la totalité des observations. La variable Y est influencée par la modalité de X.

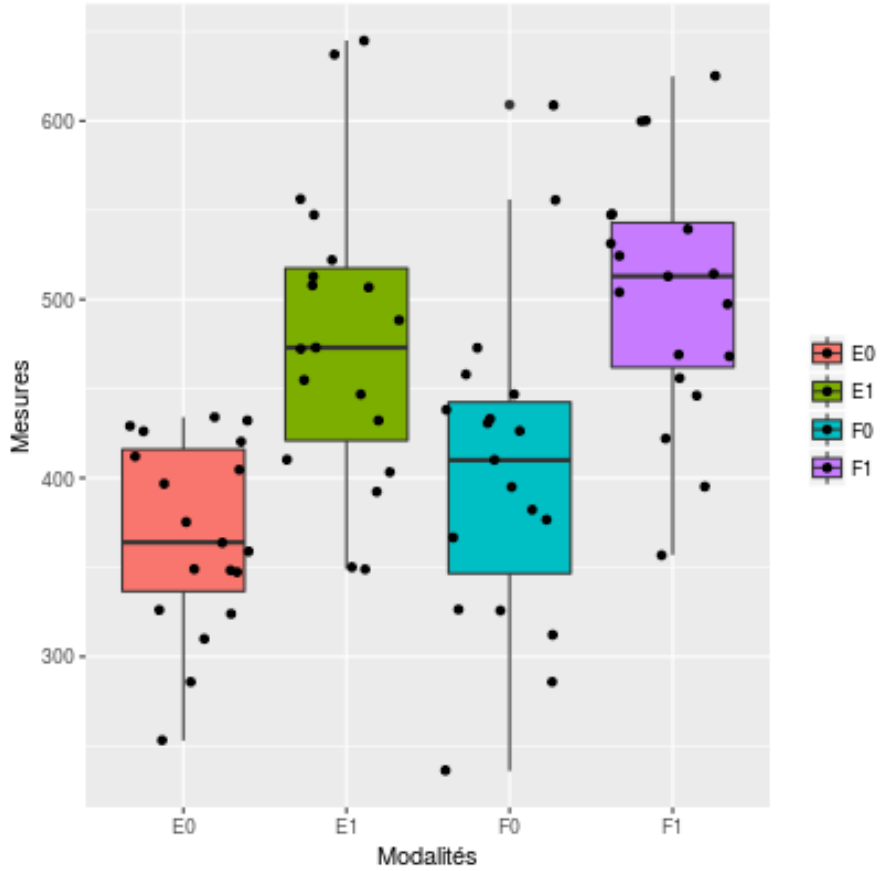
Les valeurs suggèrent une relation entre la variable quantitative Y et la variables qualitative X.

Boîtes parallèles

On représente la variable quantitative Y par une boîte (*boxplot*) pour chaque modalité r de la variable qualitative X.

Boîtes parallèles | Table





`05_analyse_bidim_quantQual.R.R`

La représentation en boîtes parallèles suggère également une relation entre la variable quantitative Y et la variables qualitative X.

Indice de liaison entre une variable quantitative et une variable qualitative

Formules de décomposition

On consière une variable **quantitative Y** et un variables **qualitative X** à r modalités. Le nombre total d'observations est n. Une modalités r comporte n_l observations.

On décompose la moyenne de y selon les modalités de X suivant la formule,

$$\bar{y} = \frac{1}{n} \sum_{l=1}^{r} n_l \bar{y}_l$$

On décompose la variance de y selon les modalités de X suivant la formule,

$$s_Y^2 = \frac{1}{n} \sum_{l=1}^{r} n_l (\bar{y}_l - \bar{y})^2 + \frac{1}{n} \sum_{l=1}^{r} n_l s_l^2 = s_E^2 + s_R^2$$

$\mathbf{s_E^2}$ variance *expliquée* par la partition ou variance *inter-classes*
$\mathbf{s_R^2}$ variance *résiduelle* ou variance *intra-classes*

$\Rightarrow$ Plus $\mathbf{s_E^2}$ est grande par rapport à $\mathbf{s_R^2}$, plus la relation entre Y et X est forte

Rapport de corrélation

Afin d'estimation la force de la liaison, on définit l'indice suivant,

$$c_{Y|X} = \sqrt{\frac{s_E^2}{s_Y^2}} = \frac{s_E}{s_Y} = \frac{s_E}{\sqrt{(s_E^2 + s_R^2)}}$$

- $c_{Y|X} = 1 \Leftrightarrow s_R^2 = 0 \Leftrightarrow \forall l = 1, ..., r, \quad s_l = 0$

  Donc, Y est constante dans chacune des classes, et la seule connaissance de la modalité de X donne la valeur de Y. C'est une liaison totale.

- $c_{Y|X} = 0 \Leftrightarrow s_E^2 = 0 \Leftrightarrow \forall l = 1, ..., r, \quad \bar{y}_l = \bar{y}$

  Donc, la modalité de X n'a aucune influence sur le valeur de Y. Il n'y a aucune liaison entre les variables.

Exemple (données `chiens.csv`)

- $s_E^2 = 2973.94$
- $s_R^2 = 5628.18$
- $\Rightarrow c_{Y|X} \simeq 0.59$

Deux variables qualitatives

On considère une variable X à r modalités, $x_1, x_2, ..., x_r$, et une variable Y à c modalités, $y_1, y_2, ..., y_c$. Les données associées peuvent être présentées dans une **table de contingence** de dimension $r \times c$ comportant les quantités $\mathbf{n_{lh}}$ qui correspondent aux **effectifs conjoints**,

	y_1	...	y_h	...	y_c	sommes
x_1	n_{11}	...	n_{1h}	...	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_l	n_{l1}	...	n_{lh}	...	n_{lc}	n_{l+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	...	n_{rh}	...	n_{rc}	n_{r+}
sommes	n_{+1}	...	n_{+h}	...	n_{+c}	n

La dernière colonne et la dernière ligne contiennent les *effectifs marginaux*. Les fréquences conjointes sont définies par $f_{lh} = \frac{n_{lh}}{n}$ et les fréquences marginales par $f_{l+} = \frac{n_{l+}}{n}$.

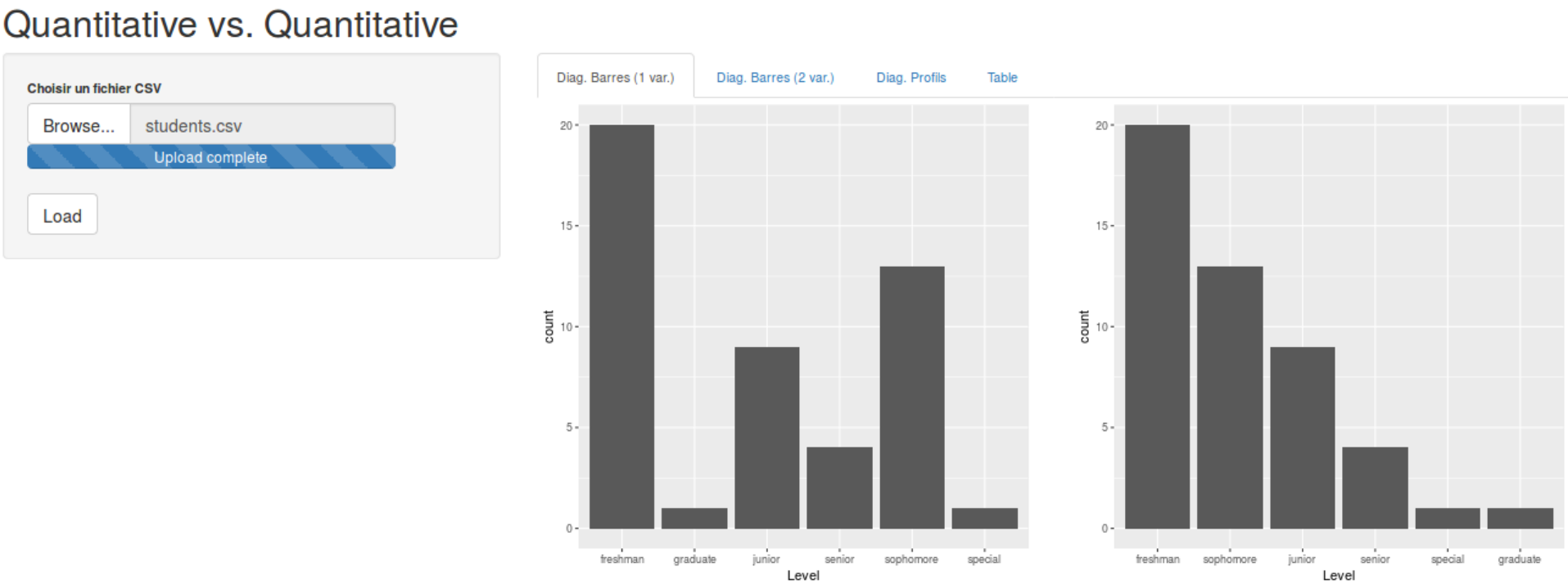
Les données

students.csv

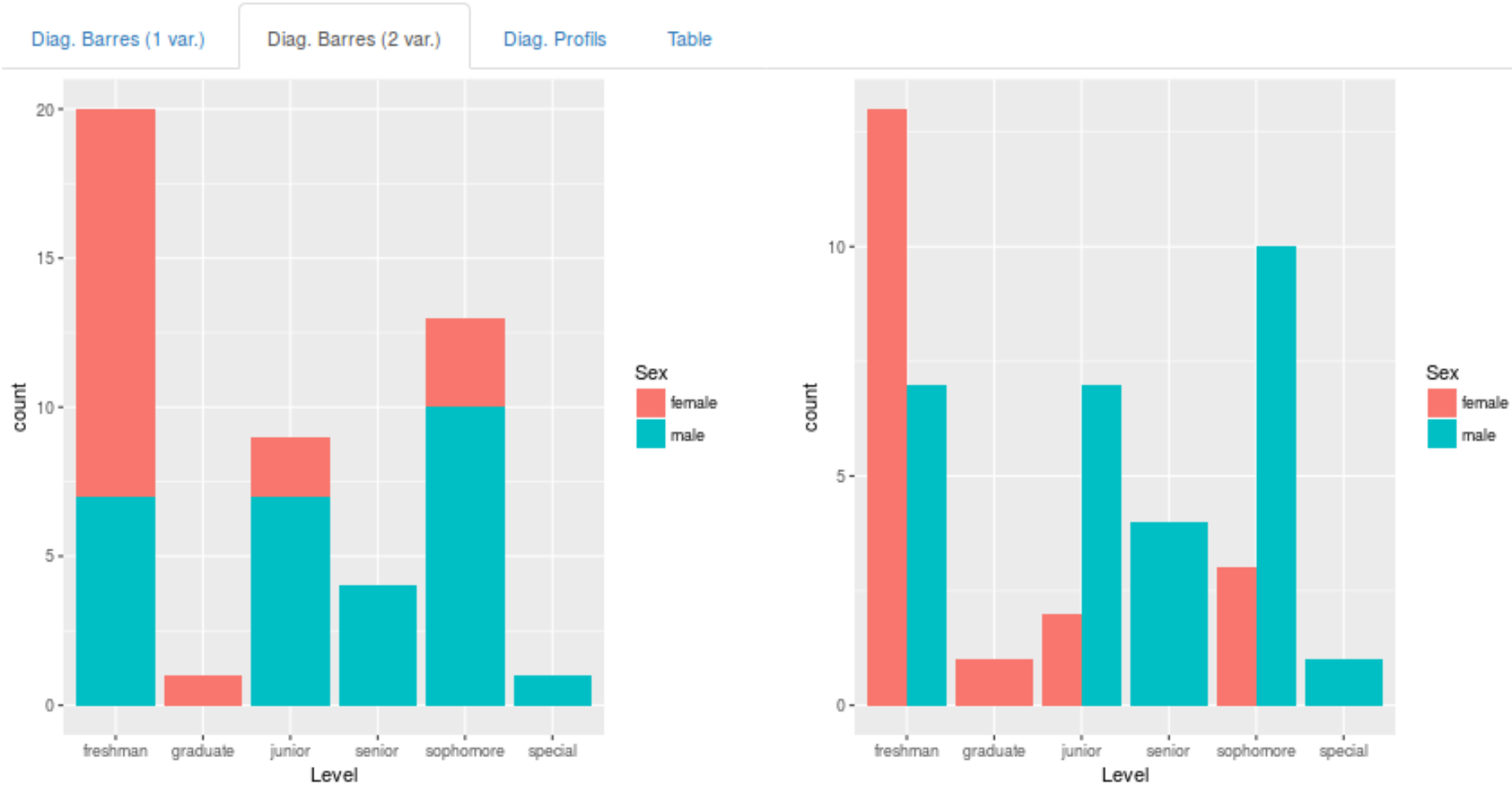
Une étude a été menée sur 48 étudiants concernant leur orientation et leur niveau scolaire ainsi que des informations personnels (eg. famille, groupe sanguin). On s'interesse à la relation entre niveau de l'étudiant et genre.

	freshman	graduate	junior	senior	sophomore	special	sommes
female	13	1	2	0	3	0	19
male	7	0	7	4	10	1	29
sommes	20	1	9	4	13	1	48

Diagramme en barres des profils-colonnes

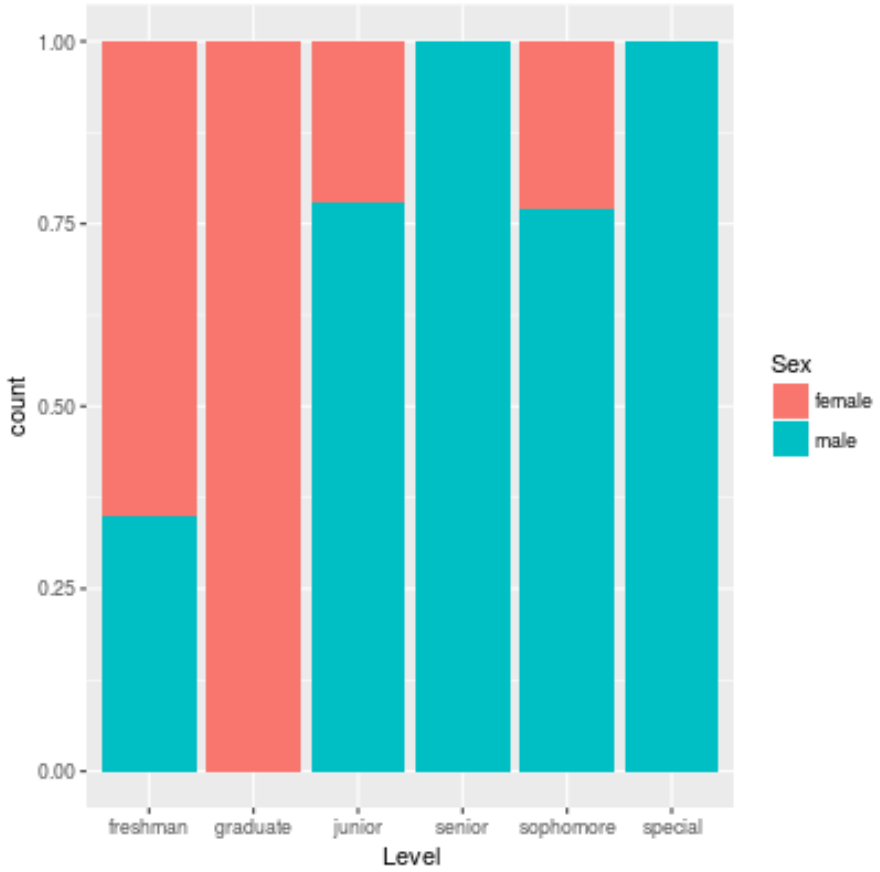


06_analyse_bidim_qualQual.R



06_analyse_bidim_qualQual.R

Diag. Barres (1 var.) | Diag. Barres (2 var.) | Diag. Profils | Table



Sex	Level	Freq
female	freshman	13
male	freshman	7
female	graduate	1
male	graduate	0
female	junior	2
male	junior	7
female	senior	0
male	senior	4
female	sophomore	3
male	sophomore	10
female	special	0
male	special	1

06_analyse_bidim_qualQual.R

Diag. Barres (1 var.) | Diag. Barres (2 var.) | Diag. Profils | Table

Show 25 entries Search:

Sex	Major	Major2	Major3	Level	Brothers	Sisters	BirthOrder	MilesHome	MilesLocal	Sleep	BloodType	Heigh
female	actuarial science			freshman	0	1	1	107.54	0.43	7.0		62.00
male	statistics	computer science		junior	3	2	3	64.70	0.20	6.0	A	72.00
male	statistics	japanese	spanish	freshman	0	1	1	155.80	0.70	9.0	B	72.00
male	mathematics	statistics		senior	0	1	1	83.90	1.20	8.0		71.50
male	statistics			sophomore	0	0	1	269.70	0.30	7.0		71.00
female	statistics			graduate	0	0	1	68.60	0.90	7.5	O	69.00
male	economics	statistics		sophomore	1	1	1	2.40	0.10	7.5	AB	72.00
male	political science	international studies	african studies	senior	1	1	1	114.50	0.80	7.0	B	70.50

06_analyse_bidim_qualQual.R

Khi-deux, χ^2

Pour quantifier la liaison entre deux variables qualitatives, l'indicateur fondamental est le χ^2. Toutefois, son usage n'est pas pratique, et on utilise plutôt des indices dérivés comme le Φ^2, le T de Tschuprow ou le C de Cramér.

Rappel de la table de contingence

On considère une variable X à r modalités, $x_1, x_2, ..., x_r$, et une variable Y à c modalités, $y_1, y_2, ..., y_c$. Les données associées peuvent être présentées dans une **table de contingence** de dimension $r \times c$ comportant les quantités $\mathbf{n_{lh}}$ qui correspondent aux **effectifs conjoints**,

	y_1	...	y_h	...	y_c	sommes
x_1	n_{11}	...	n_{1h}	...	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_l	n_{l1}	...	n_{lh}	...	n_{lc}	n_{l+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	...	n_{rh}	...	n_{rc}	n_{r+}
sommes	n_{+1}	...	n_{+h}	...	n_{+c}	n

La dernière colonne et la dernière ligne contiennent les *effectifs marginaux*. Les fréquences conjointes sont définies par $f_{lh} = \frac{n_{lh}}{n}$ et les fréquences marginales par $f_{l+} = \frac{n_{l+}}{n}$.

Définition du χ^2

Pour calculer le χ^2, on compare les **effectifs conjoints** (ou *observés*) n_{lh} aux **effectifs standards** (qui correspondent à l'*absence de liaison*) $\frac{n_{l+}n_{+h}}{n}$. On mesure l'écart à la non liaison, c'est-à-dire l'importance de la laison.

$$\chi^2 = \sum_{l=1}^{r}\sum_{h=1}^{c} \frac{\left(n_{lh} - \frac{n_{l+}n_{+h}}{n}\right)^2}{\frac{n_{l+}n_{+h}}{n}} = n\left[\left(\sum_{l=1}^{r}\sum_{h=1}^{c} \frac{n_{lh}^2}{n_{l+}n_{+h}}\right) - 1\right]$$

Le χ^2 est:

- toujours positif
- d'autant plus grand que la liaison entre les deux variables est forte
- dépendant de c, r et n
- non majoré

Phi-deux, Φ^2

$$\Phi^2 = \frac{\chi^2}{n}$$

Le Φ^2:

- ne dépend plus de n
- dépend de c et r

Cet indice est peu utilisé dans la pratique, mais joue un rôle important en Analyse Factorielle des Correspondances.

Le coefficient T de Tschuprow

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}$$

Le T de Tschuprow:

- ne dépend pas de n, c et r
- $0 \leq T \leq 1$ (proche de 1 pour une liaison forte)
- est difficile à interpréter dans l'absolu

NB:
En pratique, T est rarement supérieur à 0.5.
La valeur de T est généralement comprise entre 0.1 et 0.3

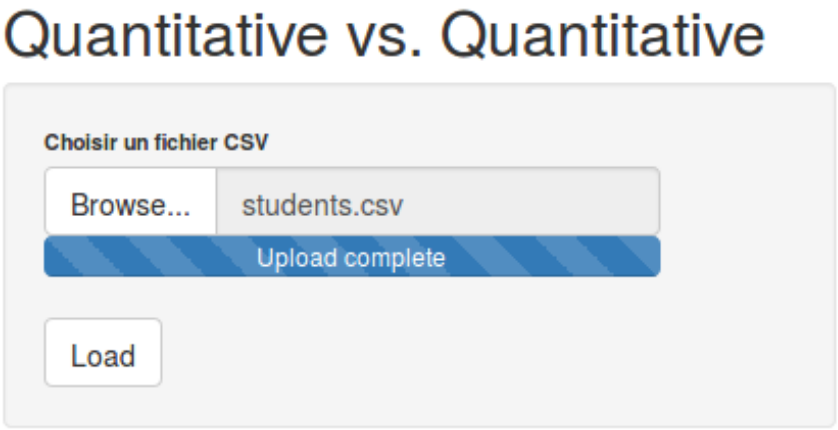
Le coefficient C de Cramér

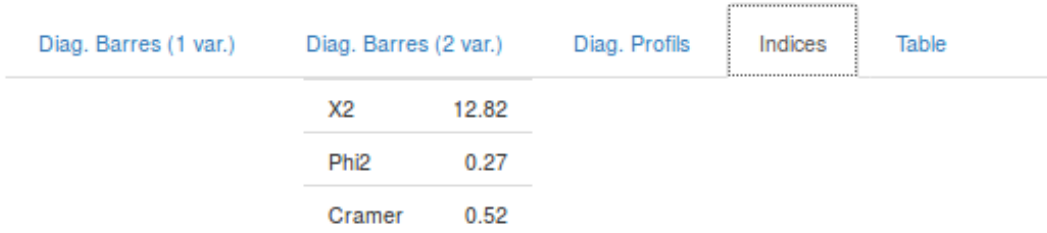
$$C = \sqrt{\frac{\Phi^2}{(d-1)}} \quad , \qquad d = inf(r, c)$$

Le C de Cramèr:

- ne dépend pas de n, c et r
- $0 \leq T \leq 1$ (proche de 1 pour une liaison forte)
- est difficile à interpréter dans l'absolu

NB:
En pratique, C est rarement supérieur à 0.5.
La valeur de C est généralement comprise entre 0.1 et 0.3





07_analyse_bidim_qualQual_chiPhiTC.R

Bibliography

- Baccini, Alain. *Statistique Descriptive Elementaire*. (2010).
- Mazerolle, Fabrice. *Statistique Descriptive*. (2009).
- Grolemund, Garrett. *Teach Yourself Shiny*.
- *The R Graph Gallery*.
- *R ggplot2 Examples*.
- *Histrogrammes avec R, S. ZIRAH*.