Chapitre 5
Statistiques descriptives bivariées

Serge COHEN (1) Sandrine CABUT (2)

(1) IPANEMA UAR 3461, CNRS, MHNH, UVSQ, Ministère de la Culture, Gif-sur-Yvette
(2) Aix-Marseille Univ., CNRS, Minist. Culture, LAMPEA, Aix-en-Provence

Après les descriptions unidimensionnelles, on s'intéresse généralement à l'étude conjointe de deux variables*. Les méthodes et les représentations employées dépendront alors de la nature (type) des données en jeu (qualitative et/ou quantitative).

1 Deux variables qualitatives

Soit X et Y deux variables qualitatives* à respectivement r et s modalités* décrivant un échantillon* de n individus* statistiques. On réalise un **tableau de contingence** (ou tableau/tri croisé), tel que représenté par le tableau 1, croisant les modalités de X en lignes et celles de Y en colonnes et contenant les effectifs* $n_{i\,j}$, appelés effectifs joints ou conjoints, tels que X présente la modalité i et Y la modalité j.

Y / X	y_1	y_2	...	y_j	...	y_s	Total : marge ligne
x_1	n_{11}	n_{12}					$n_{1.}$
x_2	n_{21}	n_{22}					$n_{2.}$
...							
x_i				n_{ij}			$n_{i.}$
...							
x_r							$n_{r.}$
Total : marge colonne	$n_{.1}$	$n_{.2}$				$n_{.s}$	n

Tab. 1 : Tableau de contingence des variables qualitatives X (r modalités) et Y (s modalités).

Les marges lignes et les marges colonnes sont respectivement égales à :

$$n_{i\,.} = \sum_{j=1}^{s} n_{i\,j} \text{ et } n_{.\,j} = \sum_{i=1}^{r} n_{i\,j}$$

Ainsi la valeur i de la marge ligne correspond à la somme des individus présentant la modalité i de la variable X (ligne i de la colonne total" du tableau 1) ; et la valeur j de la marge colonne correspond à la somme des individus présentant la modalité j de la variable Y (colonne j de la ligne "total" du tableau 1). Le nombre total d'individus est noté indifféremment $n = n_{..} = \sum_{i\,j} n_{i\,j}$

Selon la variable intéressant l'utilisateur, il calcule ensuite les profils-lignes $\mathcal{X}_i$ ou les profils-colonnes $\mathcal{Y}^j$. Ceux-ci correspondent aux fréquences f (ou pourcentages) se ramenant respectivement au total d'une ligne, soit $\frac{n_{i\,j}}{n_{i\,.}}$ (d'où l'appelation fréquence conditionnelle), ou au total d'une colonne, soit $\frac{n_{i\,j}}{n_{.\,j}}$. Les diagrammes en barre étudiés au chapitre précédent (*cf.* 4 2.1) seront déclinés de sorte à afficher les profils-lignes ou les profils-colonnes, préférentiellement en fréquence ou pourcentage pour une meilleure lisibilité et comparaison.

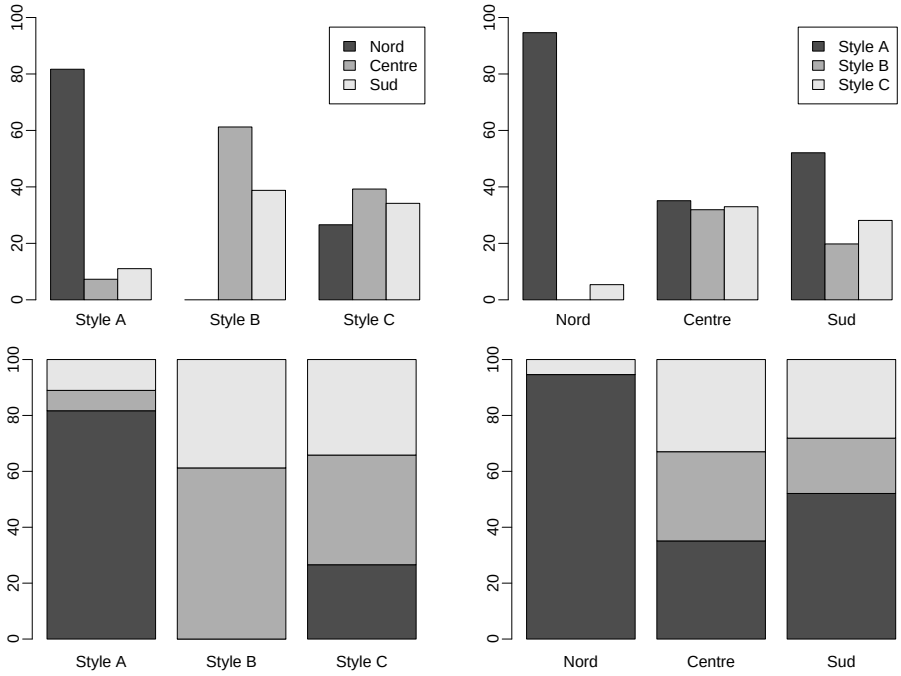
Exemple

Une étude porte sur 581 vases italiens, chaque vase étant alors un individu statistique du point de vue de cette étude. Chaque vase est caractérisé par deux variables qualitatives : sa région de provenance, avec trois modalités (Nord, Centre ou Sud) et son style avec également trois modalités (A, B ou C). Le jeu de données peut alors être résumé par son tableau de contingence ainsi que les profils-lignes (profil par provenance) et profil-colonne (profil par style) représentés au tableau 2. Les profils-lignes permettent de comparer la provenance des différents styles de vases tandis que les profils-colonnes permettent de comparer la répartition des différents styles entre chaque région. Ainsi selon la question d'intérêt, l'utilisateur pourra calculer l'un ou l'autre et réaliser le diagramme en barres correspondant pour visualiser graphiquement ce résultat.

	Nord	Centre	Sud	Total
Style A	*370*	*33*	*50*	**453**
Style B	*0*	*30*	*19*	**49**
Style C	*21*	*31*	*27*	**79**
Total	**391**	**94**	**96**	581

	Nord	Centre	Sud	Total
Style A	81.7	7.3	11.0	100.0
Style B	0.0	61.2	38.8	100.0
Style C	26.6	39.2	34.2	100.0
Total	67.3	16.2	16.5	100.0

	Nord	Centre	Sud	Total
Style A	94.6	35.1	52.1	78.0
Style B	0.0	31.9	19.8	8.4
Style C	5.4	33.0	28.1	13.6
Total	100.0	100.0	100.0	100.0



Tab. 2 : Tableau de contingence d'un ensemble de vases italiens.
Les nombres en italique correspondent aux effectifs conjoints, ceux en gras aux effectifs marginaux. À partir de ce tableau de contingence, on peut calculer, au choix, les profils lignes (à gauche) et les profils colonnes (à droite) (un exemple de chaque en fond grisé), le plus souvent exprimés en pourcentages.

On peut également représenter les effectifs à travers l'utilisation d'une présentation en mosaïque (Figure 1).

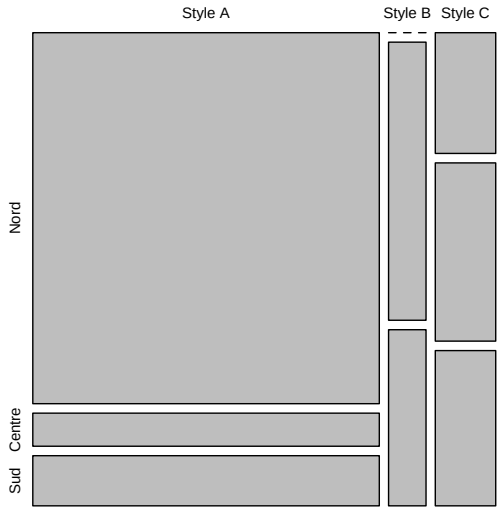


FIG. 1 : Représentation en mosaïque du jeu de données exposé au tableau 2

Face à ce type de jeu de données, il est souvent pertinent de vouloir caractériser s'il y a indépendance entre les variables observées. Dans ce cas, tous les profils-lignes $\mathcal{X}_i$ sont identiques[1] : la connaissance de X ne change pas les distributons conditionnelles de Y. Il s'ensuit :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \tag{1.1}$$

Pour répondre à cette question sur l'indépendance, il faut utiliser un test d'inférence (*cf.* 8). Ces tests sont fondés sur des critères qui sont déjà informatifs, en dehors de la formalisation du test lui-même. Nous en présentons trois, basés sur l'écart à l'indépendance : **critère du χ^2**, **coefficient de contingence**, **V de Cramer**.

1.1 Critère du χ^2

Le critère le plus souvent utilisé est le χ^2, plus précisément le χ^2 de Pearson (Everitt *et al.*, 2001). Ce critère compare les effectifs* conjoints observés aux effectifs conjoints qui aboutiraient aux mêmes effectifs marginaux tout en étant exactement équirépartis[2], illustrés par la figure 2. Ces derniers effectifs sont nommés effectifs théoriques ou idéaux. Ce critère doit impérativement être calculé sur les effectifs et non sur des fréquences ou pourcentages, de plus son utilisation dans un test d'inférence nécessite qu'aucun des effectifs conjoints théoriques ne soit trop faible (en général on considère que chaque effectif doit être supérieur à 5). Finalement, son utilisation en dehors du cadre du test statistique est rendue difficile, car sa valeur dépend du nombre de modalités pour chacune des deux variables.

[1] Réciproquement, tous les profils-colonnes $\mathcal{Y}_j$ sont également identiques.
[2] À noter que ces effectifs *idéaux* ou *théoriques*, par construction, ne sont pas entiers.

	Nord	Centre	Sud	Total
Style A	370 / 304.86	33 / 73.29	50 / 74.85	453
Style B	0 / 32.98	30 / 7.93	19 / 8.10	49
Style C	21 / 53.17	31 / 12.78	27 / 13.05	79
Total	391	94	96	581

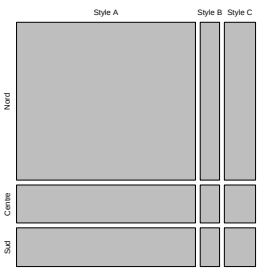


FIG. 2 : Le critère du χ^2 est calculé sur la base de la différence entre les effectifs conjoints observés ($n_{i\,j}$), en haut à gauche de chaque case, et des effectifs conjoints *idéaux* ($\widehat{n_{i\,j}}$) calculés sur la seule base des effectifs marginaux, en bas à droite de chaque case (eq. 1.2). À droite est représentée la mosaïque obtenue à partir des effectifs *théoriques* qui permet de visualiser en quoi ces effectifs sont *idéaux*. Dans cet exemple, en utilisant la définition donnée par l'équation 1.3, on obtient une valeur de $\chi^2 = 213,7624$

Les effectifs *idéaux*, notés $\widehat{n_{i\,j}}$, découlant de la relation (1.1) en cas d'indépendance, peuvent être calculés avec la formule :

$$\widehat{n_{i\,j}} = \frac{n_{i\,.} n_{.\,j}}{n_{..}} \tag{1.2}$$

La valeur du χ^2 de Pearson est alors la distance quadratique (écart élevé au carré) pondérée entre les effectifs conjoints observés et les effectifs conjoints *idéaux* :

$$\chi^2 = \sum_{i\,j} \frac{(n_{i\,j} - \widehat{n_{i\,j}})^2}{\widehat{n_{i\,j}}} \tag{1.3}$$

La valeur de χ^2 est donc positive, et elle n'est nulle que si les effectifs observés sont exactement égaux aux effectifs *idéaux*. Plus l'écart à zéro est élevé, plus les effectifs observés s'éloignent des effectifs correspondant au cas idéal d'indépendance entre les deux variables. Toutefois, un tel écart peut éventuellement être expliqué par un effet aléatoire, et c'est l'objet d'un test statistique d'évaluer la probabilité qu'une différence observée soit due au simple hasard ou, alternativement, soit significative d'une liaison entre les deux variables.

1.2 Coefficient de contingence

Comme indiqué *supra*, l'échelle de lecture du χ^2 dépend du nombre de modalités de chacune des variables, ce qui rend complexe son utilisation quand il s'agit de comparer des couples de variables avec des nombres de modalités différents. Pour remédier à cette difficulté, on utilise souvent le **coefficient de contingence**, noté CC qui est calculé sur la base du χ^2 mais avec une échelle allant de 0 à (presque) 1[3]. Ainsi, les valeurs proches de 0 indiquent l'indépendance entre les deux variables, et les valeurs proches de 1 indiquent un degré d'association élevé entre elles.

[3] le CC est strictement inférieur à 1, et il ne peut tendre vers 1 que quand $n_{..}$ tend vers l'infini.

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n..}} \tag{1.4}$$

Dans l'exemple utilisé, nous avons donc $CC = 0,519$

1.3 V de Cramer

Un peu plus compliqué à calculer que le coefficient de contingence, le critère V de Cramer a l'avantage de pouvoir prendre toutes les valeurs entre 0 et 1. En ce sens, une valeur de $V = 0$ correspond à une indépendance idéale entre les deux variables, alors qu'une valeur $V = 1$ signifie qu'il y a une dépendance *complète* entre les deux variables. Ce critère compare le χ^2 observé avec la valeur maximale que pourrait prendre le χ^2 en conservant les mêmes effectifs marginaux, valeur qu'on note $\chi^2_{\max}$ dans l'équation 1.5. La valeur de $\chi^2_{\max}$ dépend de l'effectif total et du nombre de modalités de chacune des variables, donc le nombre de lignes, r et le nombre de colonnes s du tableau des effectifs conjoints.

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n..(\min(r,s) - 1)}} \tag{1.5}$$

Pour notre exemple, nous avons donc $r = s = 3$ et $V = 0,607$.

Les indices dérivés du χ^2 ne sont pas les seules mesures de dépendance utilisables, il en existe bien d'autres, adaptées aux questionnements posés. On citera par exemple le **kappa de Cohen**, destiné à mesurer la concordance entre deux variables qualitatives ayant les mêmes modalités. Cette situation se rencontre dans le cadre d'observateurs différents chargés d'attribuer les p modalités d'une caractéristique aux mêmes n individus statistiques.

2 Une variable quantitative croisée à une variable qualitative

L'ensemble des paramètres [indicateurs] statistiques (tendance centrale, dispersion, forme…) et des représentations univariées (histogramme, boite à moustaches…) vus au chapitre 4, seront recalculés pour chaque sous-échantillon de la variable quantitative* correspondant aux modalités de la variable qualitative*. Par exemple, les paramètres décrivant les valeurs isotopiques (variable quantitative) seront recalculées pour chaque espèce (variable qualitative) séparemment (*cf.* 3.3, Tableau 2).

Exemple

Dans cet exemple, nous décrivons la concentration (en fait le *log-ratio* de la concentration[4]) d'un élément chimique comme caractéristique participant à la signature* chimique d'une ressource minérale exploitable pour une production matérielle. Ici,

[4]Il s'agit ici d'un processus de transformation des données élémentaires utile pour les études de provenance (*cf.* 2.1.2. et, en particulier, pour l'exemple du fer 2.1.5).

pour caractériser chaque signature, on s'intéresse à un ensemble d'échantillons (78 au total) correspondants à 9 classes* (une classe par signature), étiquetées de A à I. La classe d'un échantillon est donc une variable qualitative. Pour chaque échantillon, nous avons également une variable quantitative qui représente, sous la forme d'un *log-ratio*, la concentration en Europium (un élément chimique). Le tableau 3 propose un résumé chiffré de l'ensemble des données. Le résumé graphique de ce même jeu de données est présenté sous forme d'un *boxplot* (boite à moustaches*) (Tukey, 1977) à la figure 3.

	n	Moyenne	σ	Min.	Q1	Médiane	Q3	Max.	MAD
A	2	-0,903	0,247	-1,078	-0,991	-0,903	-0,816	-0,729	0,259
B	4	-0,961	0,069	-1,052	-0,984	-0,955	-0,932	-0,884	0,058
C	5	-0,797	0,035	-0,845	-0,821	-0,779	-0,773	-0,764	0,022
D	15	-0,574	0,103	-0,650	-0,631	-0,600	-0,571	-0,228	0,046
E	6	-1,091	0,147	-1,328	-1,154	-1,074	-0,985	-0,933	0,135
F	8	-1,113	0,040	-1,183	-1,128	-1,113	-1,095	-1,052	0,029
G	15	-0,721	0,157	-1,064	-0,786	-0,744	-0,659	-0,383	0,085
H	14	-0,820	0,212	-1,089	-0,990	-0,794	-0,735	-0,250	0,152
I	9	-0,722	0,315	-0,898	-0,839	-0,812	-0,797	0,112	0,040

TAB. 3 : Résumé numérique du croisement entre une variable qualitative représentant la classe d'un échantillon, de A à I, et d'une variable quantitative caractérisant la concentration en *Europium*. Pour chaque classe sont indiqués le nombre d'échantillons (n), la moyenne*, l'écart-type* (σ), puis les quantiles* : minimum, 1er quartile (Q1), médiane, 3e quartile (Q3) et maximum, et finalement l'estimation de l'écart-type à partir de la *MAD*

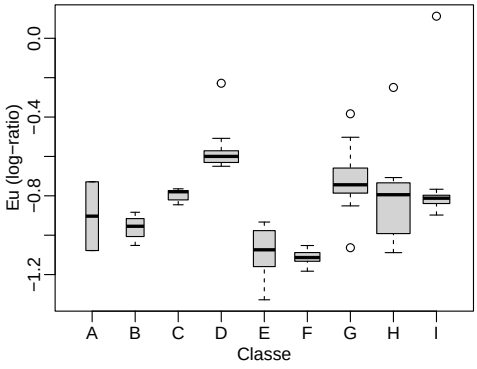


FIG. 3 : Résumé graphique. La représentation préférée sera le *boxplot* (boite à moustaches). Pour chaque classe, la ligne noire représente la médiane*, le rectangle gris s'étend du 1er au 3e quartile. Les *moustaches* s'étendent du minimum au maximum, à l'exclusion d'éventuelles valeurs aberrantes qui sont alors représentées par des cercles. Ici, la largeur de chaque boite est proportionnelle à la racine carrée du nombre d'échantillons dans la classe. Une alternative est de représenter les boites avec la même largeur.

En ce qui concerne la tendance centrale, la médiane sera préférée à la moyenne en cas de faibles effectifs, présence de valeurs extrêmes ou dissymétrie de la distribution (*cf.* 4). Pour ce qui est de la caractérisation de la dispersion*, l'écart-type souffre de la même sensibilité aux valeurs extrêmes que la moyenne. Pour réduire cette sensibili-

té, et avoir une caractérisation plus robuste, on peut estimer l'écart-type à partir de l'écart absolu médian par rapport à la médiane (Median Absolute Deviation (MAD) en anglais), c'est-à-dire MAD = $mediane(|x_i - M|)$, ou encore calculer l'écart inter-quartile (c'est-à-dire la différence entre le 1er et le 3e quartile).

S'affranchissant de l'ordre de grandeur de la variable quantitative, contrairement à l'écart-type ($\sigma = 2$ n'ayant pas la même signification selon que la variable étudiée a un ordre de grandeur de 10 ou de 100 par exemple), le coefficient de variation permettra plus aisément de comparer les dispersions de ces sous-échantillons.

Ainsi le *boxplot* sera le résumé graphique privilégié pour visuellement comparer les valeurs, tendances centrales, dispersions, de chaque sous-échantillon ainsi formé (*cf.* 2.2.7 Figures 4 et 7, *cf.* 3.6 Figure 4). Dans le cas où les formules $Q_1 - 1,5(Q_3 - Q_1)$ et $Q_3 + 1,5(Q_3 - Q_1)$ sont utilisées pour les moustaches (respectivement inférieure et supérieure) du boxplot (cf. 4), les valeurs en dehors de ces moustaches, nommées valeurs extrêmes [aberrantes, ou outliers en anglais] sont représentées par des points isolés. La longueur de la moustache inférieure correspond alors à la valeur observée de la variable étudiée qui est directement supérieure à $Q_1 - 1,5(Q_3 - Q_1)$, et celle de la moustache supérieure à la valeur observée directement inférieure à $Q_3 + 1,5(Q_3 - Q_1)$ (Tukey, 1977).

3 Deux variables quantitatives

Soit 2 variables quantitatives X et Y observées sur n individus. On dispose donc de n couples (x_i, y_i) qui seront utilisés comme coordonnées de l'individu i dans un repère cartésien (*cf.* 2.2.7 Figure 13). La forme du nuage de points ainsi tracé donne alors des indications sur la liaison entre ces 2 variables, comme l'expose la figure 4. Ce type de graphique est appelé le **diagramme de dispersion** de X et Y.

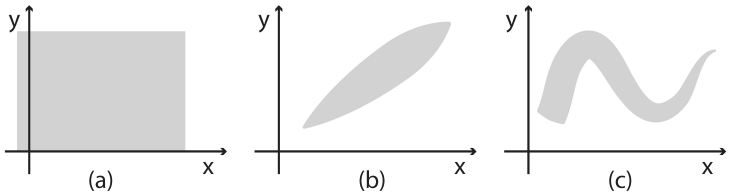


FIG. 4 : Représentation graphique d'un jeu de donnée bivarié pour lequel les variables sont quantitatives. Suivant la zone occupée par les points ainsi tracés, on met en évidence (a) l'absence de liaison entre X et Y, (b) l'existence d'une corrélation linéaire, ici positive, entre X et Y ou encore (c) l'existence d'une corrélation non linéaire entre les variables.

La première caractérisation numérique de la façon dont les variables X et Y varient de façon conjointe est la **covariance** (empirique) de ces deux variables, notée $\widehat{\sigma}_{XY}$ (pour rappel, $\widehat{x}$ et $\widehat{y}$ sont respectivement les moyennes empiriques* des variables X et Y) :

$$\widehat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x})(y_i - \widehat{y}) \tag{1.6}$$

D'autres indicateurs très utiles sont les coefficients de corrélation. Dans le cas d'une corrélation linéaire, celle-ci peut être caractérisée par le coefficient de corrélation linéaire, également appelé **coefficient de corrélation** de Pearson, noté avec la lettre r et défini par l'équation 1.7 au sein de laquelle $\widehat{\sigma}_X$ et $\widehat{\sigma}_Y$ sont respectivement les écarts-types empiriques des variables X et Y.

$$r = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X \widehat{\sigma}_Y} \tag{1.7}$$

Ce coefficient de corrélation, r, varie de -1 à 1. Sa valeur est positive pour une corrélation linéaire positive et sa valeur est négative quand la corrélation linéaire est négative. Enfin, plus sa valeur est proche de zéro, moins la corrélation linéaire est marquée. Pour autant, une valeur de zéro n'indique pas qu'il n'y a pas de corrélation entre les variables : en effet, une corrélation non linéaire peut tout à fait mener à un coefficient de corrélation de Pearson de 0. La figure 5[5] illustre de nombreux cas de corrélation, certaines linéaires d'autres non linéaires. Autre point d'attention, r est très sensible aux outliers et n'est donc pas robuste.

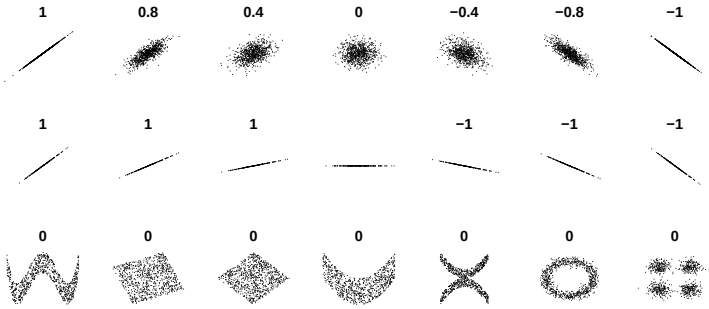


FIG. 5 : Quelques exemples de jeux de données bivariés montrant différents types de corrélation. Pour chacun de ces jeux de données, le coefficient de corrélation linéaire est indiqué juste au-dessus de chaque diagramme de dispersion. La première ligne montre que le coefficient de corrélation indique à quel point le nuage de point ressemble à une droite. La seconde ligne met en évidence que r ne renseigne pas sur la valeur de la pente de la corrélation, seulement son signe. Notons que pour l'exemple central de cette ligne, r ne peut pas être calculé car $\widehat{\sigma}_Y = 0$. Enfin, la troisième ligne montre plusieurs couples de variables X et Y corrélées non linéairement et pour lesquels r est nul.

Exemple

Reprenons l'exemple de la section précédente (Tableau 3), mais cette fois-ci en s'intéressant aux liens existants entre les concentrations en Europium (Eu), Cérium (Ce), Lanthane (La) et Césium (Cs). Dans un premier temps, on visualise les liaisons 2 à 2 de ces variables d'un seul coup d'œil en affichant les diagrammes de dispersion pour chaque couple (Figure 6).

[5] Adaptée à partir du code de Denis Boigelot dans le domaine public, https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg?uselang=fr

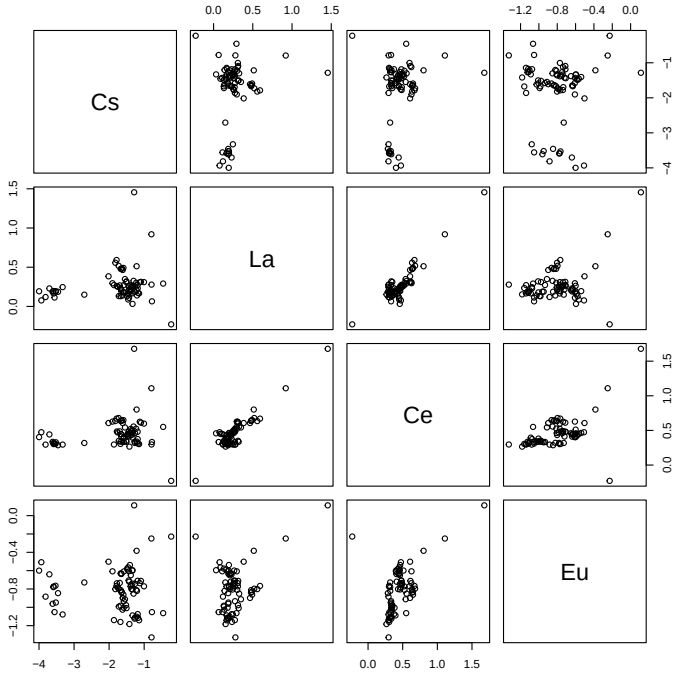


FIG. 6 : On représente, pour chaque couple de variables parmi Cs, La, Ce et Eu, le *scatter-plot* correspondant, ce qui permet de mettre en évidence d'éventuelle relation de chacune de ces variables avec l'une des autres.

L'analyse de cette figure révèle, par exemple, que les teneurs en La et Ce semblent fortement linéairement corrélées alors que celles en Cs et Eu n'ont pas l'air d'être corrélées. Cela peut être caractérisé plus précisément à travers le calcul des coefficients de corrélation de Pearson des variables deux à deux. Ceux-ci ont été calculés et sont reportés dans le tableau 4, appelé **matrice des corrélations**. Avec des coefficients de corrélation de $0,917$ et $0,020$ respectivement pour les couples (La, Ce) et (Cs, Eu), les observations effectuées sur le diagramme de dispersion se confirment.

	Cs	La	Ce	Eu
Cs	-	0,150	0,183	0,020
La	0,150	-	0,917	0,360
Ce	0,183	0,917	-	0,522
Eu	0,020	0,360	0,522	-

TAB. 4 : Matrice des corrélations, basée sur le coefficient de corrélation de Pearson.

Il est à noter, que pour conclure sur la corrélation au niveau de la population issue de l'échantillon étudié (*cf.* 4), un test de significativité de la corrélation est nécessaire (*cf.* 8). La question est alors de savoir si le coefficient de corrélation est significativement différent de zéro ou non : hypothèse nulle $H_0 : r = 0$, hypothèse alternative $H_1 : r \neq 0$.

Lorsque les variables X et Y ne suivent pas une loi normale, les résultats sur le coefficient de corrélation de Pearson restent utilisables à condition que n soit grand, dans la pratique, pour $n > 30$.

Dans le cas de variables X et Y ordinales, ou de liaison monotone[6], l'intensité de la corrélation pourra être quantifiée grâce au coefficient de corrélation de Spearman, également appelé rho de Spearman et noté ρ. Il consiste à calculer le coefficient de corrélation des rangs : les valeurs x_i et y_i de l'individu i sont remplacées par son classement (rang) pour la variable X et pour la variable Y. On obtient ainsi deux nouvelles variables composées de valeurs allant de 1 à n. Le fait que les rangs soient des permutations de $[1 \cdots n]$ permet de simplifier les calculs sur la base de l'équation (1.7), pour obtenir en l'absence d'ex-aequo :

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \tag{1.8}$$

avec d_i la différence des rangs d'un même individu selon les deux classements.

Ainsi, de même que r, ρ varie de -1 à $+1$ et s'interprète en termes de relation monotone négative ou positive ; ou encore en termes de classements inverses l'un de l'autre ou identiques. La valeur nulle signifiant toujours l'absence de lien. Ce coefficient est robuste car insensible à la présence d'outliers et son test de significativité reste valable en cas de non normalité des variables. Aussi, lorsque l'échantillon est petit, et/ou que la normalité ne peut être vérifiée, le ρ de Spearman sera préféré au coefficient de corrélation de Pearson, dont le test n'est alors pas applicable.

Le dernier coefficient de corrélation que nous citerons est également basé sur les rangs, il s'agit du τ de Kendall. En considérant tous les couples d'individus (i, j), on note :

- 1 si les deux individus sont dans le même ordre pour les deux variables : rang de $i <$ rang de j pour X et Y
- -1 si les deux classements discordent : rang de $i <$ rang de j (ou $>$) pour X et inversement pour Y

On note S la somme des valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples distincts.

Le coefficient τ vaut alors :

$$\tau = \frac{2S}{n(n-1)} \tag{1.9}$$

Comme tout coefficient de corrélation, le τ de Kendall varie entre -1 et 1, -1 indiquant des classements inversés et 1 des classements identiques. Pour savoir si la valeur

[6] croissante ou décroissante, et donc non forcément linéaire

obtenue est significative, un test statistique est également possible. Ce coefficient de corrélation sera privilégié au ρ de Spearman lorsqu'on ne dispose que de peu de données et en cas d'ex aequo sur les rangs.

Peu importe le coefficient de corrélation utilisé, il faut garder à l'esprit que leur interprétation ne doit pas être confondue avec une relation de cause à effet, comme c'est malheureusement parfois le cas.