REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'enseignement supérieur et de la recherche scientifique

Université de ABU BAKR BELKAID Tlemcen

Département de mathématiques

Mémoire de fin d'étude
Pour l'obtention du diplôme de licence en mathématique

Description bidimensionnelle et mesure de liaison entre variables

Option : probabilités et statistique

Présenté par :

Mlle HAKIKI

Sous la direction de madame **Benssadat**

Année universitaire : 2012/2013

Dédicace

- **A mes très chers parents pour leur soutien moral.**
- **A mes très chers frères.**
- **A tous mes proches et amis.**

Remerciements

Je tiens tout d'abord à remercier Dieu le tout puissant, qui m'a donné la force et la patience d'accomplir ce modeste travail.

Je remercie Mme Benssadat qui s'est penchée sur mon travail pour l'évaluer.

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience.

Enfin, je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Sommaire

1 Introduction

En statistique, étudier les phénomènes aléatoires revient parfois à étudier les liaisons entre différentes variables observées. L'étude qui met en évidence ces liens est ce qu'on appelle communément l'étude des corrélations.les méthodes et les indices de dépendance varient selon la nature (qualitative, ordinale, numérique) des variables étudiées.

On peut rendre compte de l'existence d'un lien entre deux variables numériques ou plus à l'aide du coefficient de corrélation linéaire, qui intervient dans les formules de différents indicateurs de liens statistiques.

Lorsque les variables ne sont plus uniquement quantitatives, on dispose du rapport de corrélation qui est utilisé pour caractériser l'association entre une variable quantitative et un variable qualitative, comme il peut mesurer la liaison entre deux variables numériques lorsque la relation s'écarte de la linéarité.

Lorsque les variables sont ordinales, on parle de corrélation des rangs, ou interviennent deux coefficients d'une grande importance qui sont : le coefficient de Spearman et le coefficient de Kendall.

Et finalement lorsque les variables sont toutes qualitatives, on adopte une mesure différente de toutes celles qui l'a précèdent qui est l'écart à l'indépendance.

C'est ce que nous allons tenter de préciser dans les paragraphes qui suivent.

2 Liaison entre deux variables numériques

Lorsqu'il ne s'agit que de deux variables X et$Y$, on dit alors qu'il y a corrélation s'il y a dépendance en moyenne : à $X = x$ fixé la moyenne $\bar{Y}$ est fonction de X, lorsque la nature de la liaison est linéaire on parle alors de coefficient de corrélation linéaire.

2.1 Le coefficient de corrélation linéaire

2.1.1 Définition

Ce coefficient est en effet spécialement adapté à la mesure d'une liaison «linéaire » (ou plutôt affine).

On considère le couple (X, Y) et un échantillon observé de n valeurs numériques de ce couple :

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$; On note :

- Les moyennes observées : $\bar{x}\ et\ \bar{y}$ avec :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i$$

- Les variances observées : s_x^2 et s_y^2 avec :

$$s_x^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n}\sum_{1}^{n}(y_i - \bar{y})^2$$

- La covariance observée : s_{xy} avec :

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

Alors le coefficient de corrélation linéaire est définit comme le quotient :

$$\boxed{r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \sum_1^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}}$$

- r est le coefficient de « Bravais Pearson » qui est une normalisation de la covariance par le produit des écarts-type, ce qui fait que ce nombre réel sans dimension soit compris entre -1 et 1.
- $|r| = 1$ lorsque les points de coordonnées $(x_i, y_i), i = 1 \dots n$ sont parfaitement alignés, ce qui est traduit par la relation linéaire exacte : $ax_i + by_i + c = 0 \quad \forall i$.
- En présence de n couples, on a deux vecteur de R^n :

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Considérons les deux vecteurs des variables centrées de R^n :

$$x' = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} \qquad y' = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

r est le cosinus formé par les vecteurs x' et y' :

En effet, la covariance de x et y est en fait le produit scalaire des variables centrées i.e de x' et y' : $s_{xy} = x'.y' = \|x'\|\|y'\| \cos(\widehat{x'}, \widehat{y'})$ et les écarts type sont la norme des variables centrées i.e $s_x = \|x'\|$ et $s_y = \|y'\|$, alors :

$$r = \frac{s_{xy}}{s_x\, s_y} = \cos(\widehat{x'}, \widehat{y'}).$$

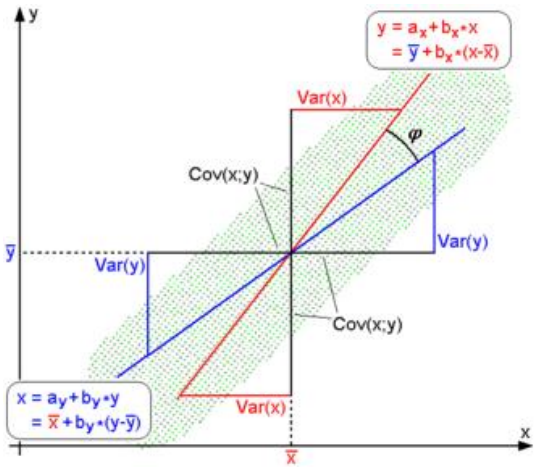


Figure 1

2.1.2 Du bon usage du coefficient de r

Déjà mentionné ci-dessus, r sert à quantifier l'intensité de la dépendance linéaire entre 2 variables, c'est-à-dire son usage doit être réservé à des nuages où les points sont répartis de part et d'autre d'une tendance linéaire, sauf qu'il est très sensible et non résistant aux données aberrantes (une observation qui se trouve « loin » des autres observations) et n'est donc pas robuste.

La figure ci-dessous montre les risques d'un usage inconsidéré du coefficient de corrélation linéaire r.

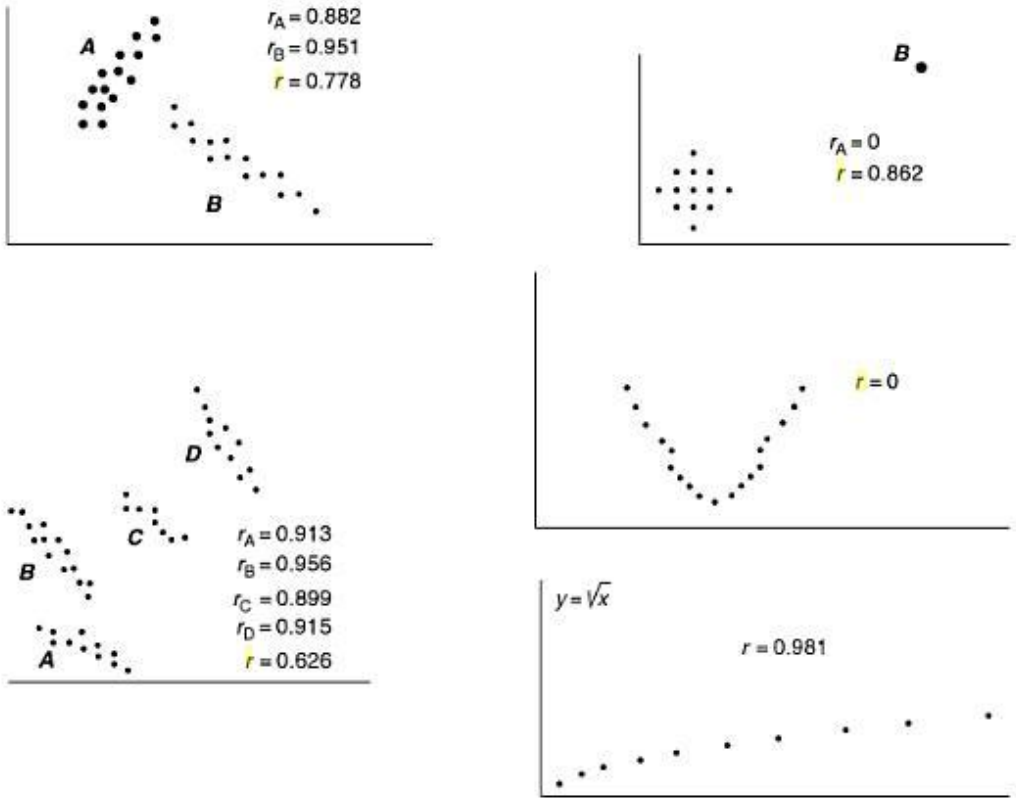


Figure 2

Les 4 nuages de la figure 2 ont mêmes moyennes, mêmes variances et même coefficient de corrélation :

$$\bar{x} = 9 \qquad \bar{y} = 7.5$$
$$s_x^2 = 10.0 \qquad s_y^2 = 3.75$$
$$r = 0.82$$

Seul le premier nuage justifie l'usage de r.

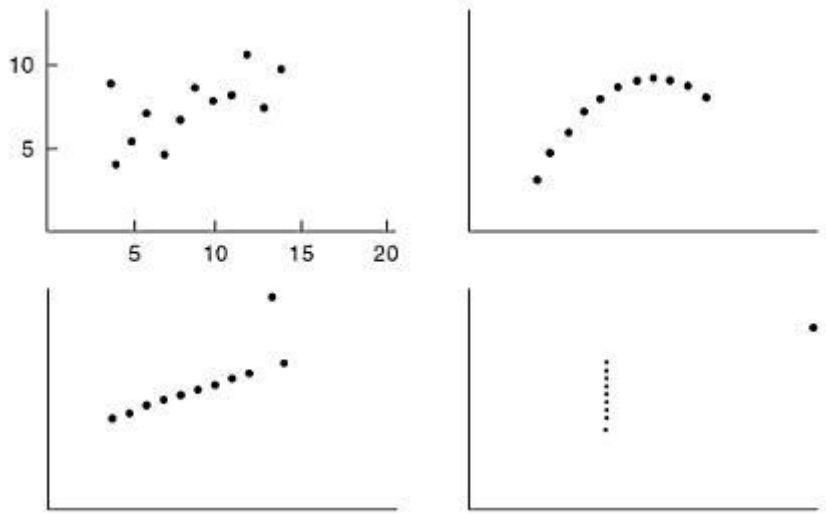


Figure 3

De plus, le coefficient de corrélation n'est pas transitif.

2.1.3 Matrice de corrélation

Lorsque l'on observe n valeurs numériques de p variables, les corrélations des variables (entre elles) sont regroupées dans une matrice appelée matrice de corrélation R, indiquant l'influence des variables les unes sur les autres.

Considérons les matrices suivantes :

_ $X = \begin{bmatrix} x_1^1 & \dots & x_1^p \\ \vdots & x_i^j & \vdots \\ x_n^1 & \dots & x_n^p \end{bmatrix}$ x_i^j : est la i-ème valeur prise par la j-ème variable.

_ $A = I - \frac{11'}{n} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}[1 \quad \dots \quad 1] = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} & \dots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \dots & \frac{1}{n} \end{bmatrix}$

Matrice carrée de taille n, appelée opérateur de centrage ;de terme général : $a_{ii} = 1 - \frac{1}{n}$ $a_{ij} = -\frac{1}{n}\ i \neq j.$

_ $Y = \begin{bmatrix} x_1^1 - \overline{x^1} & \dots & x_1^p - \overline{x^p} \\ \vdots & x_i^j - \overline{x^J} & \vdots \\ x_n^1 - \overline{x^1} & \dots & x_n^p - \overline{x^p} \end{bmatrix}$ la matrice des données centrées, et est telle que :$Y = AX$.

_ La matrice V des variances covariances des p variables :

$$V = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{12} & s_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{1p} & \dots & \dots & s_p^2 \end{bmatrix}$$

V s'obtient en utilisant la matrice Y : $V = \frac{1}{n} Y'Y$.

_ Si on pose $D_{1/s}$ la matrice diagonale suivante :

$$D_{1/s} = \begin{bmatrix} 1/s_1 & 0 & \dots & 0 \\ 0 & 1/s_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1/s_p \end{bmatrix}$$

On a alors : $R = D_{1/s} V D_{1/s}$.

2.1.4 Les propriétés de la matrice de corrélation

Si l'on désigne par ${x^{cr}}_i^j$ les valeurs centrées réduites de la j-ème variable, c'est-à-dire ${x^{cr}}_i^j = \frac{x_i^j - \overline{x^j}}{s_{x^j}}$; $\overline{x^j}$ est la moyennes de la j-ème variable, et s_{x^j} est l'écart-type de la j-ème variable.

Alors $\overline{x^{crj}} = 0$ et $s^2_{x^{crj}} = 1$ (la variance de la j-ème variable centrée réduite).

$$s_{x^{crj}\, x^{crk}} = \frac{1}{n} \sum_1^n \left({x^{cr}}_i^j\right)\left({x^{cr}}_i^k\right) = \frac{1}{n} \frac{\sum_1^n \left(x_i^j - \overline{x^j}\right)\left(x_i^k - \overline{x^k}\right)}{s_{x^j} s_{x^k}} = r_{x^j x^k}$$

$s_{x^{crj}\, x^{crk}}$ est la covariance entre la j-ème variable centrée réduite et la k-ème variable centrée réduite).

La matrice de variance covariance des données centrées réduites est :

$$V = \begin{bmatrix} 1 & s_{x^{cr1}\, x^{cr2}} & \dots & s_{x^{cr1}\, x^{crp}} \\ s_{x^{cr2}\, x^{cr1}} & 1 & & \vdots \\ \vdots & & \ddots & \\ s_{x^{crp}\, x^{cr1}} & \dots & & 1 \end{bmatrix} = \begin{bmatrix} 1 & r_{x^1 x^2} & \dots & r_{x^1 x^p} \\ r_{x^2 x^1} & 1 & & \vdots \\ \vdots & & \ddots & \\ r_{x^p x^1} & \dots & & 1 \end{bmatrix} = R$$

Donc :

- R est identique à la matrice de variance covariance des données centrées réduites.
- R est symétrique (due à la symétrie de la covariance) et semi définie positive.

Pour illustrer l'usage de cette matrice, partons de l'exemple suivant :

Exemple

On se propose d'étudier la relation existant entre les variables suivantes : cylindrée, puissance, longueur, largeur, poids et vitesse de pointe pour 18 véhicules.

	Nom	Cyl	Puis	Lon	Lar	Poids	Vitesse
1	ALFASUD-TI-1350	1350	79	393	161	870	165
2	AUDI-100-L	1588	85	468	177	1110	160
3	SIMCA-1370-GLS	1294	68	424	168	1050	152
4	CITREN-GS-CLUB	1222	59	412	161	930	151
5	FIAT-132-1600 GLS	1585	98	439	164	1105	165
6	LANCIA-BETA-1300	1297	82	429	169	1080	160
7	PEUGEOT-504	1796	79	449	169	1160	154
8	RENAULT-16-TL	1565	55	424	163	1010	140
9	RENAULT-30-TS	2664	128	452	173	1320	180
10	TOYOTA-COROLLA	1166	55	399	157	815	140
11	ALFETTA-	1570	109	428	162	1060	175
12	PRINCESS-1800-HL	1798	82	445	172	1160	158
13	DATSUN-200L	1998	115	469	169	1370	160
14	TAUNUS-2000-GL	1993	98	438	170	1080	167
15	RANCHO	1442	80	431	166	1129	144
16	MAZDA-9295	1769	83	440	165	1095	165
17	OPEL-REKORD-L	1979	100	459	173	1120	173
18	LADA-1300	1294	68	404	161	950	140

La matrice V calculé avec $n-1$ en dénominateur :

	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	139823.5294	6069.7451	5798.7059	1251.2941	40404.2941	3018.5686
PUIS	6069.7451	415.1928	288.9118	56.3922	2135.6961	208.8791
LON	5798.7059	288.9118	488.7353	99.7647	2628.3824	127.7353
LAR	1251.2941	56.3922	99.7647	28.2353	521.7059	30.5098
POIDS	40404.2941	2135.6961	2628.3824	521.7059	18757.4412	794.1078
VITESSE	3018.5686	208.8791	127.7353	30.5098	794.1078	147.3889

Matrice de variance et covariance V

La matrice R est la suivante :

	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	1.00000	0.79663	0.70146	0.62976	0.78895	0.66493
PUIS	0.79663	1.00000	0.64136	0.52083	0.76529	0.84438
LON	0.70146	0.64136	1.00000	0.84927	0.86809	0.47593
LAR	0.62976	0.52083	0.84927	1.00000	0.71687	0.47295
POIDS	0.78895	0.76529	0.86809	0.71687	1.00000	0.47760
VITESSE	0.66493	0.84438	0.47593	0.47295	0.47760	1.00000

Matrice de corrélation R

Compte tenu des deux matrices, on constate que les variables sont fortement corrélées.

La figure suivante, appelée matrice de dispersion, est très utile : elle permet en un seul graphique de juger des liaisons entre toutes les variables.

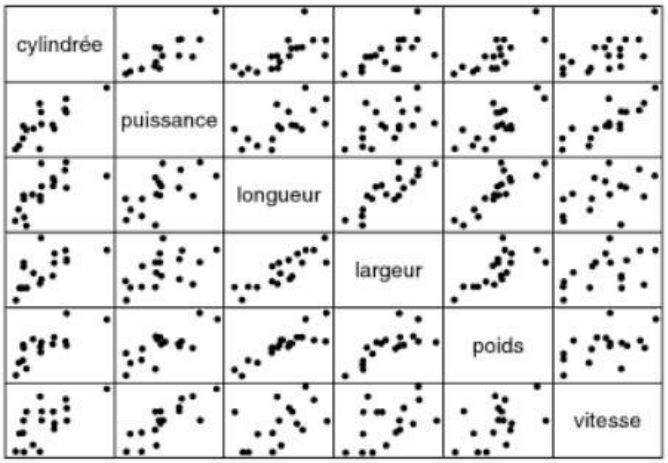


Figure 4

2.1.5 Caractère significatif d'un coefficient de corrélation : (test de significativité)

On prélève n observations au hasard dans une population où X et Y sont indépendants ; le premier test qui vient à l'esprit est la signification (significativité) de la corrélation, c'est-à-dire le coefficient de corrélation est significativement différent de zéro ?

Quelles seraient les valeurs possibles de r ? Ou plus exactement la distribution de probabilité de la variable R qui correspond à cet échantillonnage ?

On suppose à priori que le coupe(X, Y) suit une loi normale bivariée, et on formalise notre problème sous deux hypothèses entre lesquelles il faut choisir :

$$H_0 : \rho = 0 \ (hypothèse\ nulle)$$

$$H_1 : \rho \neq 0 \ (hypothèse\ alternative)$$

2.1.5.1 Statistique du test :

Sous H_0 : $t = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}}$ suit une loi de Student à (n-2) degrés de liberté.

2.1.5.2 Région critique :

La région critique (rejet de l'hypothèse nulle) du test au risque α est :

$$R.C.: |t| > t_{1-\frac{\alpha}{2}}(n-2)$$

Où $t_{1-\frac{\alpha}{2},n-2}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-2$ degrés de liberté.

Pour $n = 4$, on remarquera que R suit une loi uniforme sur $[-1, 1]$ et donc que toutes les valeurs possibles sont équiprobables.

On a :

$$E(R) = 0 \qquad et \qquad V(R) = \frac{1}{n-1}$$

2.1.5.3 Un exemple numérique :

Toujours avec les caractéristiques numériques des 18 véhicules :

Le coefficient de corrélation entre la cylindrée est la puissance vaut : $r = 0.79663$.

On souhaite tester sa significativité au risque $\alpha = 0.05$.

Nous devons calculer les éléments suivants :

- La statistique du test : $t = \frac{0.79663}{\sqrt{\frac{1-0.79663^2}{18-2}}} = 21.08660$.
- Le seuil théorique au risque α est $t_{0.975,(18-2)} = 2.120$.
- Nous rejetons donc l'hypothèse nulle car $t > t_{0.975,16}$ c'est-à-dire que les deux variables sont fortement liées.

Lorsque n est grand $(n > 100)$, $R\sqrt{n}$ suit une loi $N(0,1)$, et donc on compare $r\sqrt{n}$ à 1.96 (le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$ dans le cas $\alpha = 0.05$.

Sous H_1, c'est-à-dire $\rho \neq 0$, la loi exacte de R bien que connue est très difficilement exploitable on notera cependant que :

$$E(R) \approx \rho - \frac{\rho(1-\rho^2)}{2n} \qquad R\ est\ biaisé\ pour\ \rho$$
$$V(R) = \frac{(1-\rho^2)^2}{n-1}$$

La figure 5 donne les distributions d'échantillonnage de r pour différentes valeurs de ρ, avec $n = 10$.

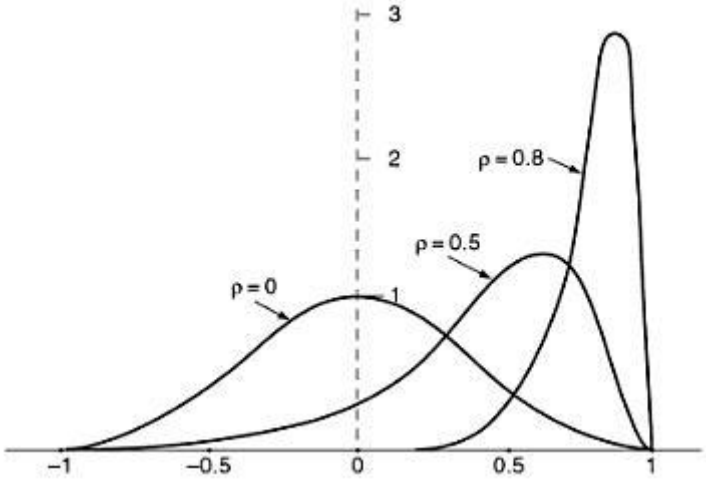


Figure 5

Lorsque n est grand ($n > 25$) :

On dispose de la transformation de Fisher définie par :

$$Z = \frac{1}{2}\log\frac{1+R}{1-R}$$

Z suit une loi normale de moyenne :

$$E(Z) = \frac{1}{2}\log\frac{1+r}{1-r}$$

Et de variance :

$$V(Z) = \frac{1}{n-3}$$

Partant du fait que$R = \frac{e^{2Z}-1}{e^{2Z}+1}$, (car $Z = argtan(R)$) on pourra retrouver $E(R) et\, V(R)$.

Nous pouvons nous appuyer sur cette statistique pour réaliser le test de significativité ci-dessus, mais plus intéressant encore, la transformation nous offre la possibilité de tester des valeurs à priori de r, et de trouver des intervalles de confiance.

Lorsque le couple (X, Y) n'est pas gaussien les résultats précédents restent utilisables à condition que n soit grand (en pratique $n > 30$), mais le fait de trouver que r n'est pas significativement différent de 0 n'entraine pas nécessairement l'indépendance.

2.2 Corrélation partielle

Il arrive fréquemment que la forte corrélation entre deux variables résulte d'une corrélation commune avec une troisième variable.

Le coefficient de corrélation partielle constitue donc un moyen de mesure de la liaison, ou la dépendance entre deux variables après élimination de l'effet d'une ou plusieurs variables (en les fixant).

2.2.1 Définition

Pour un triplet (X, Y, Z) et un échantillon observé de n valeurs numériques de ce triplet$((x_1, y_1, z_1), \dots, (x_n, y_n, z_n))$, le coefficient de corrélation partielle de X et Y avec Z noté $r_{xy.z}$ est défini à partir des corrélations brutes :

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

L'idée est de retrancher de r_{xy} le double effet des corrélations qu'entretiennent X et Y avec$Z$. Puis un terme de normalisation est introduit de manière à ce que$-1 \leq r_{xy.z} \leq 1$.

2.2.2 Démonstration de la formule

La corrélation partielle entre X et Y (étant donné) est la corrélation simple entre X et Y, étant enlevé l'effet linéaire de Z c'est-à-dire : $r_{xy.z} = r_{x.z\ y.z}$

Pour cela, soit :

$$X_{.z} = X - aZ$$

$$Y_{.z} = Y - bZ$$

Où a et b sont les coefficients obtenus de la régression :

$$a = \frac{s_{xz}}{s_z^2}$$

$$b = \frac{s_{yz}}{s_z^2}$$

La corrélation (simple) entre $X_{.z}$ et $Y_{.z}$ est : $\frac{s_{x.zy.z}}{s_{x.z}s_{y.z}}$.

On calcule :

$$s_{x.zy.z} = s_{xy} - as_{yz} - bs_{xz} + abs_z^2$$

$$s_{x.z}^2 = s_x^2 - 2as_{xz} + a^2s_z^2$$

$$s_{y.z}^2 = s_y^2 - 2bs_{yz} + b^2s_z^2$$

En substituant a et$b$, on obtient :

$$s_{x.zy.z} = s_{xy} - \frac{s_{xy}s_{yz}}{s_z^2}$$

$$s_{x.z}^2 = s_x^2 - \frac{s_{xz}^2}{s_z^2}$$

$$s^2_{y.z} = s^2_y - \frac{s^2_{yz}}{s^2_z}$$

Donc :

$$r = \frac{s_{xy} - \frac{s_{xz}s_{yz}}{s^2_z}}{\sqrt{\left(s^2_x - \frac{s^2_{xz}}{s^2_z}\right)\left(s^2_y - \frac{s^2_{yz}}{s^2_z}\right)}} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r^2_{xz})(1-r^2_{yz})}}$$

Pour le modèle normal à p dimensions, le coefficient de corrélation partielle ou conditionnelle est obtenu à partir de la matrice des covariances partielles.

2.2.3 Remarques

- On peut calculer toutes les corrélations partielles à partir de la matrice de corrélation simple.
- On pourra enlever l'effet linéaire d'une deuxième variable, puis d'une troisième…de façon récursive, ainsi si l'on veut éliminer l'effet linéaire de Wen plus de$Z$, il suffit de remplacer dans la formules précédentes les corrélations simples par les corrélations partielles :

$$r_{xy.zw} = \frac{r_{xy.z} - r_{xw.z}r_{yw.z}}{\sqrt{(1-r^2_{xw.z})(1-r^2_{yw.z})}}$$

2.2.4 Exemple numérique

Le TiO_2 et le SiO_2 sont des bons indices de la maturité magmatique des roches volcaniques. On pourrait vouloir éliminer l'effet de la différenciation magmatique sur les corrélations entre les autres variables. Lors de la différenciation magmatique, les minéraux Ferro-magmatiques cristallisent en premier. On observera donc typiquement une corrélation positive entre FeO etMgO.

Cependant, ces deux éléments se trouvent en compétition pour occuper les mêmes sites de cristallisation sur les minéraux. Ceci entraine que pour des roches de maturité magmatique comparable, on devrait observer une corrélation négative entre FeO etMgO.

On a alors mesuréSiO_2, MgO etFeO et on a obtenu, avec 30 observations les corrélations simples suivantes entre ces trois éléments :

	SiO_2	MgO	Feo
SiO_2	1	-0.86	-0.75
MgO	-0.86	1	0.50
Feo	-0.75	0.50	1

Ainsi la corrélation partielle entre MgO etFeO (étant donné l'effet de SiO_2 enlevé) est :

$$r_{MgO\ FeO.SiO_2} = \frac{0.50 - (-0.86)(-0.75)}{\sqrt{(1-(-0.86)^2(1-(-0.75)^2}} = -0.429$$

C'est loin d'être la même situation par rapport au coefficient de corrélation simple !

2.2.5 Signification d'un coefficient de corrélation partielle

Si l'hypothèse de normalité est vérifiée, nous adoptons la même démarche que pour la corrélation simple (brute). Les hypothèses à tester sont :

$$H_0 : r_{xy.z} = 0$$

$$H_1 : r_{xy.z} \neq 0$$

La statistique du test est $t = \frac{r_{xy.z}}{\sqrt{\frac{1-r_{xy.z}^2}{n-3}}}$ qui, sous H_0 suit uneT_{n-3}, (le degré de liberté est diminué de 1 qui, en général est le nombre de variables fixées).

Et la région critique du test est définie par :

$$R.C. : |t| > t_{1-\frac{\alpha}{2},(n-3)}$$

Où $t_{1-\frac{\alpha}{2},(n-3)}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-3$ degrés de liberté, avec α le risque.

Reprenons notre exemple :

Au seuil 5% la valeur critique est : 2.052 et t vaut2.468, donc la liaison est significative.

2.3 Corrélation multiple entre une variable numérique et p autres variables numériques

2.3.1 Définition

Soit une variable numérique Y et un ensemble de p variables numériques $X^1, X^2, \dots, X^p$.

Le coefficient de corrélation multiple R est alors la valeur maximale prise par le coefficient de corrélation linéaire entre Y et une combinaison linéaire des X^i:

$$R = \sup_{a_1,a_2,\dots,a_p} r(Y; \sum_{i=0}^{p} a_i X^i)$$

On a toujours $0 \leq R \leq 1$.

En d'autres termes, si on pose$Y^* = b_0 + b_1 X^1 + \cdots + b_p X^p$, on désire que Y^*soit le plus proche possible de Y.

Alors si l'espace des variablesR^n, est muni de la métrique$D$, on exigera que $\|Y - Y^*\|^2$soit minimal.

Donc$R = 1$, s'il existe une combinaison linéaire des X^i telle que :

$$Y = a_0 + \sum_{i=1}^{p} a_i X^i$$

2.3.2 Interprétation géométrique

On rappelle que le coefficient de corrélation est le cosinus de l'angle formé de R^n par des variables centrées.

Considérons le sous-espace W de R^n (de dimension au plus égale à$p+1$), engendré par les combinaisons linéaires des X^iet la constante1.

Y^*est alors la projection orthogonale sur le sous-espace W. On a aussi

$$R = \frac{cov(Y,Y^*)}{s_Y s_{Y^*}}$$

Et puis que :

$$cov(Y,Y^*) = \|Y-\bar{Y}\|\|Y^*-\bar{Y}\|\cos(Y-\bar{Y},Y^*-\bar{Y})$$

$$s_Y = \|Y-\bar{Y}\| \quad s_{Y^*} = \|Y^*-\bar{Y}\|$$

Alors

$$R = \frac{cov(Y,Y^*)}{s_Y s_{Y^*}} = \cos(Y-\bar{Y},Y^*-\bar{Y})$$

R est le cosinus de l'angle θformé par la variable centrée $Y-\bar{Y}$ et$W$, c'est-à-dire l'angle formé par $Y-\bar{Y}$ et sa projection orthogonale $Y^*-\bar{Y}$ sur W(voir figure 6).

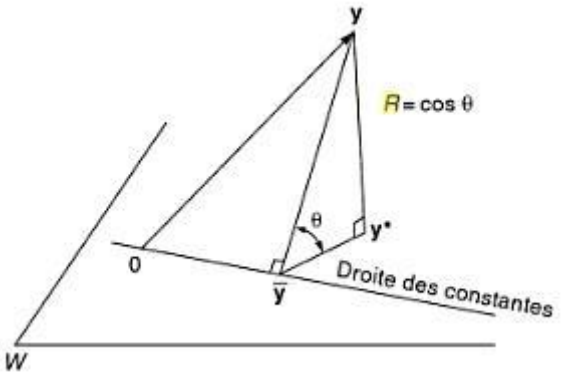


Figure 6

2.3.3 Calcul de R

Comme tout coefficient de corrélation linéaire, son carré s'interprète en terme de variance expliquée :

$$R^2 = \frac{\sum(Y_i-\bar{Y})^2 - \sum(Y_i-Y_i^*)^2}{\sum(Y_i-\bar{Y})^2} = \frac{\|Y-\bar{Y}\|^2 - \|Y-Y^*\|^2}{\|Y-\bar{Y}\|^2} = \frac{\|Y^*-\bar{Y}\|^2}{\|Y-\bar{Y}\|^2} = \frac{s_{Y^*}^2}{s_Y^2} = \cos^2\theta$$

Soit A la matrice de projection orthogonale sur$W$, alors :

$$R^2 = \frac{(Y-\bar{Y})'\,A(Y-\bar{Y})}{\|Y-\bar{Y}\|^2}$$

En effet :

$$\|Y^* - \bar{Y}\|^2 = (Y^* - \bar{Y})'(Y^* - \bar{Y}) = (AY - \bar{Y})'(AY - \bar{Y}) = Y'A'AY - Y'A'\bar{Y} - \bar{Y}'AY + \bar{Y}'\bar{Y}$$
$$= Y'AY - Y'A\bar{Y} - \bar{Y}'AY + \bar{Y}'A\bar{Y} = (Y - \bar{Y})'\, A(Y - \bar{Y})$$

Car :

$$Y^* = AY$$
$$A = A' \ (symetrique)$$
$$A^2 = A \ (idempotente)$$

En particulier si Y est centré, c'est-à-dire $\bar{Y} = 0$:

$$R^2 = \frac{Y'\, A\, Y}{\|Y\|^2}$$

2.3.4 Signification d'un coefficient de corrélation multiple :

Si les n observations étaient issues d'une population gaussienne où Y est indépendante des X^i alors on a :

$$\frac{R^2}{1 - R^2}\frac{n - p - 1}{p} = F(p, n - p - 1)$$

On retrouve comme cas particulier la loi du coefficient de corrélation linéaire simple en faisant$p = 1$.

3 Liaison entre variables ordinales : corrélation des rangs

Souvent, on ne dispose que d'un ordre sur un ensemble d'individus et non de valeurs numériques d'une variable mesurable : soir par ce qu'on a que des données du type classement (classement A, B, C, D, E…etc.), ou bien par ce que les valeurs numériques d'une variable n'apportent que peu comparé à leur ordre (notes de copies,…etc.) .

Le rang d'une observation est donc une version « réduite » de sa coordonnée.

L'idée est alors de substituer aux valeurs observées leurs rangs. Nous créons donc deux nouvelles colonnes : $r_i = \ rang(x_i)$, correspondant au rang de l'observation x_i dans la colonne de X et$s_i = rang(y_i)$, avec r_i et s_i sont des permutations différentes des n premiers entiers.

3.1 Le coefficient de Spearman

Il est ni plus ni moins que le coefficient de Pearson calculé sur les rangs, et donc il mesure l'association entre deux variables mesurées dans une échelle ordinale.

$$r_s = \frac{cov(r, s)}{s_r s_s} = \frac{\sum_{i=0}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=0}^{n}(r_i - \bar{r})^2 \sum_{i=0}^{n}(s_i - \bar{s})^2}}$$

Compte tenu du fait que les rangs soient des permutations de $[1 \ldots n]$ et de l'absence d'ex aequo on a :

$$\bar{r} = \bar{s} = \frac{1}{n}\sum_{i=0}^{n} i = \frac{1}{n}.\frac{n}{2}(n+1) = \frac{n+1}{2}$$

$$s_s^2 = s_r^2 = \frac{1}{n}\sum_{i=0}^{n}(r_i - \bar{r})^2 = \frac{n^2-1}{12}$$

D'où :

$$r_s = \frac{\frac{1}{n}\sum_{i=0}^{n} r_i s_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}}$$

En posant $d_i = r_i - s_i$: différence des rangs d'un même objet selon les deux classements, on a :

$$\sum_{i=0}^{n} r_i s_i = -\frac{1}{2}\sum_{i=0}^{n} -2r_i s_i = -\frac{1}{2}\sum_{i=0}^{n}\left[(r_i - s_i)^2 - r_i^2 - s_i^2\right] = -\frac{1}{2}\sum_{i=0}^{n}(r_i - s_i)^2$$
$$+\frac{1}{2}\sum_{i=0}^{n} r_i^2 + \frac{1}{2}\sum_{i=0}^{n} s_i^2 = -\frac{1}{2}\sum_{i=0}^{n}(r_i - s_i)^2 + \sum_{i=0}^{n} r_i^2$$

Or :

$$\sum_{i=0}^{n} r_i^2 = \frac{n(n+1)(2n+1)}{6}$$

somme des carrés des nombres entiers, d'ou :

$$r_s = \frac{-\frac{1}{2n}\sum_{i=0}^{n} d_i^2 + \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}}{\frac{(n^2-1)}{12}}$$
$$= -\frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)} + \frac{2(n+1)(2n+1) - 3(n+1)^2}{n^2-1}$$

Le deuxième terme vaut1, et on a l'expression équivalente pratique :

$$r_s = 1 - \frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)}$$

$r_s = 1 \Rightarrow \sum_{i=0}^{n} d_i^2 = 0 \Rightarrow r_i = s_i \ \forall i$: les deux classements sont identiques ;

$r_s = -1 \Rightarrow$ les deux classements sont inverses l'un à l'autre ;

$r_s = 0 \Rightarrow$ les deux classements sont indépendants.

3.1.1 Signification d'un coefficient de Pearson

En présence d'un échantillon de n couples de rangs: $(r_1, s_1),\ (r_2, s_2), \ldots, (r_n, s_n)$, obtenus, soit par observation directe d'un couple (R, S) de rangs, soit par transformation en rangs des valeurs d'un couple (X, Y) de valeurs réelles :

Lorsque $n \leq 100$, on se rapportera à la table du coefficient de corrélation de Spearman (fournie en annexe)

La région critique est $|R_s| > k$:

- Si $R_s > k$: il y a concordance de classements ;
- Si $R_s < -k$: il y a discordance de classements.

Lorsque $\mathrm{n} > 100$, on admet que$\mathrm{R_s}$, est distribué comme une normale$\mathrm{N}(0, \frac{1}{\sqrt{\mathrm{n}-1}})$, et donc il suffit de comparer $R_s\sqrt{n-1}$ à la valeur critique lue dans une table de loi normale centrée réduite.

Lorsque les observations proviennent d'un couple normal (X, Y) de corrélation ρ et que l'on calcule r_s à la place de r, si n est grand on a les relations approchées suivantes :

$$r_s = \frac{6}{\pi}\sin^{-1}(\frac{\rho}{2}) \quad ou \quad \rho = 2\sin(\frac{\pi}{6}r_s)$$

3.1.2 Exemple

Mettons en relation la taille et le poids de 15 personnes qu'on ordonnera de manière croissante comme suit :

Numéro	Taille(m)	Poids(Kg)	R_i(taille)	S_i(poids)	d_i^2
1	1.697	77.564	15	14	1
2	1.539	55.000	3	1	4
3	1.629	76.657	10	12	4
4	1.633	62.596	11	6	25
5	1.500	58.068	2	3	1
6	1.679	72.575	14	11	9
7	1.643	82.000	13	15	4
8	1.626	76.667	9	13	16
9	1.543	58.060	5	2	9
10	1.542	71.668	4	10	36
11	1.621	68.039	8	8	0
12	1.577	70.060	7	9	4
13	1.557	61.689	6	5	1
14	1.496	67.585	1	7	36
15	1.637	59.874	12	4	64
				Somme	214

A partir de la dernière formule, on a :

$$r_s = 0.6179$$

Au risque$\alpha = 5\%$, la valeur critique, d'après la table du coefficient de corrélation de Pearson, est :

$$k = 0.521$$

On a$r_s > k$, alors il y a concordance des classements.

3.2 Le coefficient de corrélation des rangs de Kendall

3.2.1 Aspect théorique :

Afin de savoir si deux variables aléatoires X et Y varient dans le meme sens ou en sens contraire, on peut considérer le signe du produit $(X_i - X_j)(Y_i - Y_j)$ ou (X_i, Y_i); (X_j, Y_j) sont des réalisations indépendantes du couple(X, Y).

Le coefficient de Kendall repose justement sur la notion de paires discordantes et concordantes.

1. On dit que les paires observations i et j sont concordantes si et seulement si $(X_i > X_j$ alors $Y_i > Y_j)$ ou $(X_i < X_j$ alors$Y_i < Y_j)$. Nous simplifions l'écriture avec$(X_i - X_j)(Y_i - Y_j) > 0$.
2. On dit que les paires sont discordantes lorsque $(X_i > X_j$ alors $Y_i < Y_j)$ ou $(X_i < X_j$ alors$Y_i > Y_j)$, en d'autres termes$(X_i - X_j)(Y_i - Y_j) < 0$.

Le τ de Kendall théorique, calculé sur la population, est défini par :

$$\tau = 2 \times P\left((X_i - X_j)(Y_i - Y_j) > 0\right) - 1$$

Ou

$$\tau = P\left((X_i - X_j)(Y_i - Y_j) > 0\right) - P\left((X_i - X_j)(Y_i - Y_j) < 0\right)$$

Et donc la seule différenciation avec le r de Spearman est le fait que τ de Kendall se lise comme une probabilité. Il est le fruit de la différence entre deux probabilités : celle d'avoir des paires concordantes et celle d'avoir des paires discordantes.

Ce coefficient est compris entre -1 et1.

Si (X, Y) est un couple gaussien de coefficient de corrélationρ, alors :

$$\tau = \frac{2}{\pi} \sin^{-1} \rho$$

On a$\tau \leq \rho$, $\tau = \rho$ n'est vrai que pour $\rho = 0$ et$\rho = \mp 1$.

Notons qu'il est possible de calculer directement τ sur des données continues (X et Y) sans qu'il soit nécessaire de les transformer en rangs. Le τ de Kendall s'applique naturellement aussi lorsque l'une des variables est continue, l'autre ordinale.

3.2.2 Calcul sur un échantillon

Pour un échantillon de taille$n$, posons P le nombre de paires concordantes, et Q le nombre de paires discordantes, alors τ peut être défini de la manière suivante :

$$\tau = \frac{P - Q}{\frac{1}{2} n(n-1)}$$

Le dominateur représente naturellement le nombre total de paires c'est-à-dire : $\frac{1}{2} n(n-1) = \binom{n}{2}$.

On note 1 si deux individus i et j sont dans le même ordre (c'est-à-dire pour les paires concordantes).

Et on note -1 si les deux classements discordent (les paires discordantes).

On somme les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples distincts, soit S cette somme, on a :

$$S = P \times 1 + Q \times (-1) = P - Q$$

$$S_{max} = 1 \times \frac{n(n-1)}{2} = \frac{n(n-1)}{2}$$

La somme maximale, lorsque toutes les paires sont concordantes.

$$S_{min} = (-1) \times \frac{n(n-1)}{2} = -\frac{n(n-1)}{2}$$

La somme minimale, lorsque toutes les paires sont discordantes.

Donc

$$S_{max} = -S_{min}$$

Alors le coefficient τ est :

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

Compte tenu de ce qui précède :

$\tau = 1$: le classement selon X concorde symétriquement avec le classement selon$Y$.

$\tau = -1$: les classements sont inversés.

3.2.3 Calcul pratique de la somme S

La manière la plus simple de calculer τ est de trier les données selon$X$, puis de comptabiliser la quantité suivante :

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v_{ij}$$

Où

$$v_{ij} = \begin{cases} +1 \; si \; Y_i < Y_j \\ -1 \; si \; Y_i > Y_j \end{cases}$$

Et

$$v_i = \sum_{j=i+1}^{n} v_{ij}$$

v_i est l'écart entre le nombre de paires concordantes et discordantes relativement à l'observation.

3.2.4 Signification du coefficient de Kendall

Dès que$n \geq 8$, nous pouvons nous appuyer sur la normalité asymptotique de τ sous l'hypothèse d'indépendance de X et Y.

On a :

$$u = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = 3\tau\sqrt{\frac{n(n-1)}{2(2n+5)}}$$

Suit une loi normale centrée réduite, sousH_0: $\tau = 0$, ou encore$\tau \sim N(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}})$.

Et la région critique du test pour un risque αs'ecrit :

$$|u| > u_{1-\frac{\alpha}{2}}$$

3.2.5 Exemple

Détaillons les calculs sur le même exemple abordé auparavant. Nous limitons l'effectif à $n = 8$, et les données sont triées selon la taille de la personne.

	Taille	Poids							
1	1.496	67.585	67.585						
2	1.500	58.068	-1	58.068					
3	1.539	55.000	-1	-1	55.000				
4	1.542	71.668	+1	+1	+1	71.668			
5	1.543	58.060	-1	-1	+1	-1	58.060		
6	1.557	61.689	-1	+1	+1	-1	+1	61.689	
7	1.577	70.060	+1	+1	+1	-1	+1	+1	70.060
8	1.621	68.039	+1	+1	+1	-1	+1	+1	-1
Somme			-1	+2	+5	-4	+3	+2	-1

$$\tau = 0.214 \qquad u = 0.741$$

Pour$\alpha = 5\%$, le seuil critique du test est $u_{0.975} = 1.96$: pas de liaison significative entre les deux classements.

A part le cas ou les variables sont ordinales, les coefficients de corrélation des rangs sont très utiles pour tester l'indépendance de deux variable **non normales** lorsque l'échantillon est **petit** : on sait en effet qu'on ne peut appliquer alors le test du coefficient de corrélation linéaire. Les tests de corrélation des rangs sont alors les seuls applicables, car ils ne dépendant pas de la distribution sous-jacente.

Ils sont robustes car insensibles à des valeurs aberrantes.

Les coefficients de corrélation des rangs sont en fait des **coefficients de dépendance monotone** car ils sont invariants pour toute transformation croissante de variables.

Les coefficients de corrélation de rang permettent de tester l'existence d'une relation monotone entre deux variables. Ainsi le nuage des points suivant où $y = \ln(x)$ donne un coefficient de corrélation linéaire $r = 0.85$ mais des coefficients de Spearman et de Kendall égaux à1.

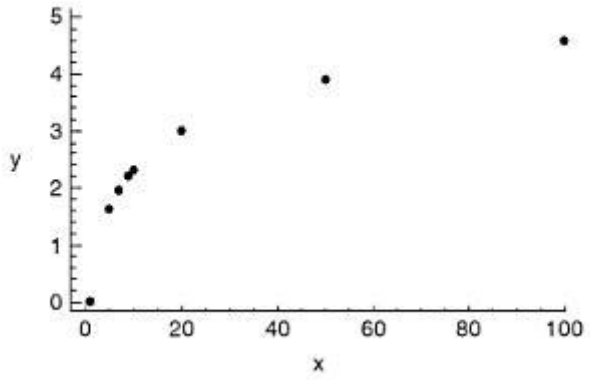


Figure 7

Lorsque les coefficients de corrélation de rang sont nettement supérieurs au coefficient de corrélation linéaire, des transformations monotones non linéaires sur certaines variables peuvent se révéler utiles.

4 Liaison entre une variable numérique et une variable qualitative

4.1 Le rapport de corrélation théorique

Lorsque la relation s'écarte de la linéarité, le coefficient de corrélation n'est plus adapté, pour cela nous utiliserons un autre indicateur, qui est le rapport de corrélation.

Le rapport de corrélation n'est pas une mesure symétrique et elle repose sur la notion d'espérance conditionnelle, en effet $y = E(Y|X = x)$ est la courbe de régression qui fournit un résumé de Y lorsque X prend la valeur$x$. Il a une portée plus large que la simple alternative pour mesurer une liaison non linéaire entre deux variables quantitatives, et donc il peut être utilisé pour caractériser l'association entre une variable quantitative Y et une variable qualitative $\mathcal{X}$, ainsi le rapport de corrélation théorique est défini comme le rapport entre la variabilité de Y expliquée par $\mathcal{X}$ et la variance totale de Y :

$$\eta^2_{Y/\mathcal{X}} = \frac{V[E(Y/\mathcal{X})]}{V(Y)}$$

Le rapport de corrélation est défini sur l'intervalle $[0, 1]$:

- Lorsqu'il est égal à0, cela veut dire que la connaissance de $\mathcal{X}$ ne donne aucune information sur Y ; la moyenne de Y est la meme quelque soit la valeur de $\mathcal{X}$.
- Lorsqu'il est égal à1, la connaissance de $\mathcal{X}$ permet de determiner avec certitude la valeur de Y.

_ Lorsque $\mathcal{X}$ et Y sont indépendants, il est nul.

4.2 Le rapport de corrélation empirique

Si $\mathcal{X}$ a k catégories (modalités), on notera $\mathrm{n}_1, \mathrm{n}_2, \dots, \mathrm{n}_\mathrm{k}$ les effectifs observés et $\bar{\mathrm{y}}_1, \bar{\mathrm{y}}_2, \dots, \bar{\mathrm{y}}_\mathrm{k}$ les moyennes de Y pour chaque catégorie et $\bar{y}$, la moyenne totale.

Si l'on note e^2 l'équivalent empirique de η^2, alors il est défini de la manière suivante :

$$e^2 = \frac{\frac{1}{n}\sum_{i=0}^{k} n_i(\bar{y}_i - \bar{y})^2}{s_Y^2}$$

Avec :

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2$$

Où les s_i^2 sont les variances de Y à l'intérieur de chaque catégorie.

En effet :

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_i^j - \bar{y})^2$$

Avec :

y_i^j : la j-ème valeur de Y pour la catégorie i.

Et :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_i \qquad et \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_i^j$$

En écrivant :

$$y_i^j - \bar{y} = y_i^j - \bar{y}_i + \bar{y}_i - \bar{y}$$

On obtient :

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_i^j - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_i^j - \bar{y}_i\right)^2 + \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 + \frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_i^j - \bar{y}\right)(\bar{y}_i - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i \frac{1}{n_i}\sum_{j=1}^{n_i}(y_i^j - \bar{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2 + \frac{1}{n}\sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{y})^2$$

Le troisième terme est nul car :

$$\frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_i^j-\bar{y})(\bar{y}_i-\bar{y}) = \frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_i^j(\bar{y}_i-\bar{y}) - \frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\bar{y}(\bar{y}_i-\bar{y})$$
$$= \frac{2}{n}\sum_{i=1}^{k}n_i\bar{y}_i(\bar{y}_i-\bar{y}) - \frac{2}{n}\sum_{i=1}^{k}n_i\bar{y}_i(\bar{y}_i-\bar{y}) = 0$$

Donc :

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{k}n_i s_i^2 + \frac{1}{n}\sum_{i=1}^{k}n_i\,(\bar{y}_i-\bar{y})^2$$

Avec :

- $\frac{1}{n}\sum_{i=1}^{k}n_i\,(\bar{y}_i-\bar{y})^2$ est appelée variance intercatégories.
- $\frac{1}{n}\sum_{i=1}^{k}n_i s_i^2$ est appelée variance intracatégorie.

Si $\bar{y}_1=\bar{y}_2=\cdots=\bar{y}_k$, c'est-à-dire$y_1^j=y_2^j=\cdots\ y_k^j\ \forall j$.

Alors :

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{y}_i=\bar{y}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{y}_1=\bar{y}_1=\bar{y}_2=\cdots=\bar{y}_k$$

D'où :

$$e^2=\frac{\frac{1}{n}\sum_{i=0}^{k}n_i(\bar{y}_i-\bar{y})^2}{s_Y^2}=\frac{(\bar{y}_1-\bar{y}_1)^2}{s_Y^2}=0$$

Absence de dépendance en moyenne.

Si tous les individus d'une catégorie de $\mathcal{X}$ ont même valeur de Y et ceci pour chaque catégorie, c'est-à-dire : $y_i^1=y_i^2=\cdots=y_i^{n_i}$.

Alors :

$$\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_i^j\Rightarrow\bar{y}_i=y_i^j\ \forall j\ \Rightarrow\ \frac{1}{n}\sum_{i=1}^{k}n_i s_i^2=\frac{1}{n}\sum_{i=1}^{k}n_i\frac{1}{n_i}\sum_{j=1}^{n_i}(y_i^j-\bar{y}_i)^2=0$$

D'où :

$$e^2=\frac{\sum_{i=0}^{k}n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=0}^{k}n_i(\bar{y}_i-\bar{y})^2}=1$$

Si on transforme $\mathcal{X}$ en une variable numérique $\tilde{X}$ à k valeurs, en attribuant à chaque catégorie i de $\mathcal{X}$ une valeur numérique égale à$\bar{y}_i$, alors e^2 sera égal à r^2(le coefficient de corrélation).

Lorsqu'il n'y a que deux classes de moyennes $\bar{y}_1$et $\bar{y}_2$:

$$e^2 = \frac{\frac{n_1 n_2}{n^2}(\bar{y}_1 - \bar{y}_2)^2}{s_Y^2}$$

En effet :

Pour deux classes :

$$n = n_1 + n_2 \qquad \text{et} \qquad \bar{y} = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_2)$$

Alors :

$$\begin{aligned}
e^2 &= \frac{1}{s_Y^2}\frac{1}{n}[n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2] \\
&= \frac{1}{s_Y^2}\frac{1}{n}\left[n_1\left(\bar{y}_1 - \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_2)\right)^2 + n_2\left(\bar{y}_2 - \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_2)\right)^2\right] \\
&= \frac{1}{s_Y^2}\frac{1}{n}\left[n_1(\frac{n_2}{n}(\bar{y}_1 - \bar{y}_2))^2 + n_2(\frac{n_1}{n}(\bar{y}_1 - \bar{y}_2))^2\right] \\
&= \frac{1}{s_Y^2}\frac{1}{n}\left[\left(\frac{n_1 n_2^2}{n^2} + \frac{n_2 n_1^2}{n^2}\right)(\bar{y}_1 - \bar{y}_2)^2\right] \\
&= \frac{1}{s_Y^2}\frac{1}{n}\left[\frac{n_1 n_2}{n^2}(n_1 + n_2)(\bar{y}_1 - \bar{y}_2)^2\right] \\
&= \frac{1}{s_Y^2}\frac{n_1 n_2}{n^2}(\bar{y}_1 - \bar{y}_2)^2
\end{aligned}$$

4.2.1 Signification du rapport de corrélation

Sous l'hypothèse nulle $\eta^2 = 0$ et sous condition que les distributions conditionnelles de Y pour chaque catégorie soient gaussiennes, de moyenne et de variance identiques (hypothèse d'homoscédacité), la statistique :

$$F = \frac{\frac{e^2}{k-1}}{\frac{1-e^2}{n-k}} = \frac{n-k}{k-1}\frac{e^2}{1-e^2}$$

Suit une loi de Fisher à $(k-1, n-k)$ degrés de liberté, avec k le nombre de classes.

4.2.2 Exemple

Le magazine **CAPITAL** a donné pour 100 villes françaises les valeurs du taux de la taxe d'habitation. On traite la variation du taux de taxe d'habitation Y selon la zone géographique X .

Ville	Taux taxe d'habitation	Zone géographique	Ville	Taux taxe d'habitation	Zone géographique
Aix-en-Provence	18.94	Sud-est	Limoge	17.24	Centre

Ajaccio	22.06	Sud-est	Lorient	16.74	Ouest
Amiens	17.97	Nord	Lyon	19.09	Sud-est
Angers	18.86	Ouest	Maisons-Alfort	10.30	Ile-de-France
Annecy	14.97	Sud-est	Marseille	21.93	Sud-est
Antibes	14.30	Sud-est	Mérignac	19.39	Sud-ouest
Antony	11.07	Ile-de-France	Metz	16.62	Est
Argenteuil	16.90	Ile-de-France	Montauban	12.72	Sud-ouest
Arles	24.49	Sud-est	Montpellier	21.40	Sud-ouest
Asnières-sur-Seine	10.13	Ile-de-France	Montreuil	13.67	Ile-de-France
Aubervilliers	12.45	Ile-de-France	Mulhouse	16.05	Est
Aulnay-Sous-Bois	15.59	Ile-de-France	Nancy	18.21	Est
Avignon	22.41	Sud-est	Nanterre	6.13	Ile-de-France
Beauvais	15.37	Nord	Nantes	21.13	Ouest
Belfort	16.20	Est	Neuilly-sur-Seine	3.68	Ile-de-France
Besançon	20 .20	Est	Nice	19.75	Sud-est
Béziers	22.14	Sud-ouest	Nîmes	30.23	Sud-ouest
Blois	17.07	Centre	Niort	19.19	Centre
Bordeaux	22.11	Sud-ouest	Noisy-le-Grand	16.91	Ile-de-France
Boulogne-Billancourt	9.46	Ile-de-France	Orléans	20.05	Centre
Bourges	15.77	Centre	Paris	9.15	Ile-de-France
Brest	25.99	Ouest	Daum	21.31	Sud-ouest
Brive-la-Gaillarde	15.82	Centre	Perpignan	15.87	Sud-ouest
Caen	16.12	Ouest	Pessac	20.71	Sud-ouest
Calais	23.36	Nord	Poitiers	21.55	Centre
Cannes	19.72	Sud-est	Quimper	16.67	Ouest
Chalon-sur-Saône	17.30	Centre	Reims	14.98	Est
Chambéry	18.71	Sud-est	Rennes	21.75	Ouest
Champigny/Marne	15.09	Ile-de-France	Roubaix	27.97	Nord
Charleville-Mézières	17.30	Est	Rouen	20.97	Ouest
Châteauroux	17.37	Centre	Rueil-Malmaison	14.93	Ile-de-France
Cholet	14	Ouest	Saint-Denis	9.17	Ile-de-France
Clermont-Ferrand	15.85	Centre	Saint-Etienne	19.90	
Colmar	16.31	Est	St-Maur-des-fossés	10.82	Ile-de-France
Colombes	14.16	Ile-de-France	Saint-Nazaire	16.36	Ouest
Courbevoie	4.86	Ile-de-France	Saint-Quentin	20.46	Nord
Créteil	17.58	Ile-de-France	Sarcelles	19.32	Ile-de-France
Dijon	18.75	Centre	Sartrouville	12.38	Ile-de-France
Drancy	10.42	Ile-de-France	Strasbourg	22.04	Est
Dunkerque	28.69	Nord	Toulon	19.37	Sud-est
Evreux	21.27	Ouest	Toulouse	19.23	Sud-ouest
Fontenay-sous-Bois	12.10	Ile-de-France	Tourcoing	33.61	Nord
Grenoble	19.43	Sud-est	Tours	20.79	Centre

Ivry-sur-Seine	9.16	Ile-de-France	Troyes	18.11	Est
La rochelle	18.75	Centre	Valence	16.25	Sud-est
La Seyne-sur-mer	25.98	Sud-est	Vénissieux	18.70	Sud-est
Laval	19.48	Ouest	Versailles	8.95	Ile-de-France
Le havre	17.67	Ouest	Villeneuve-d'Ascq	29.96	Nord
Le mans	17.54	Ouest	Villeurbanne	19.85	Sud-est
Lille	36.17	Nord	Vitry-sur-Seine	11.50	Ile-de-France

Le rapport de corrélation est tel que :

$\eta^2_{Y/X} = 0.56$ et correspond à $F = 20.05$

5 Liaison entre deux variables qualitatives

On veut à présent étudier l'association entre deux variables qualitatives.

5.1 Tableau de contingence, marges et profils

Soit $\mathcal{X}$ et $\mathcal{Y}$ deux variables qualitatives catégorielles possédant respectivement $r$et s catégories décrivant un ensemble de n individus, les données de l'échantillon observé sont répertoriées dans un tableau croisé à $r$lignes et $s$colonnes appelé tableau de contingence regroupant les effectifs n_{ij}, n_{ij} étant le nombre d'individus de la population qui possèdent à la fois la modalité y_j pour la variable $\mathcal{Y}$ et la modalité x_i pour la variable $\mathcal{X}$, comme suit :

$\mathcal{X}$ \ $\mathcal{Y}$	y_1	y_2	……	y_s	
x_1	n_{11}	n_{12}		n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2s}	$n_{2.}$
⋮					
x_r	n_{r1}	n_{r2}		n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$		$n_{.s}$	n

Avec des notations standards on a :

$$n_{i.} = \sum_j n_{ij} \qquad n_{.j} = \sum_i n_{ij}$$

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement marges en lignes (on a sommé les colonnes d'une seule ligne) et marges en colonnes (on a sommé les lignes d'une seule colonne).

Deux lectures différentes d'un même tableau de contingences sont possibles selon que l'on privilégie l'une de l'autre des deux variables : lecture en ligne ou lecture en colonne.

On appelle tableau des profils-lignes le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$ (la somme de chaque ligne est ramenée à 100%), et tableau des profils-colonnes, le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{.j}}$ (la somme de chaque colonne est ramenée à 100%).

Il est plus intéressant de ramener les effectifs par rapport aux tableaux marginaux en ligne ou en colonne, en effet, chaque ligne (colonne) définit un sous-ensemble de la population, nous pouvons calculer les proportions pour chaque groupe, les comparer entre elles et les comparer avec les proportions dans la population globale. On parle alors de profils c'est-à-dire fréquences conditionnelles que l'on oppose aux fréquences marginales lues dans la dernière ligne ou colonne du tableau.

5.1.1 Exemple

Comme illustration, nous utilisons le fichier GERMAN CREDIT qui recense les caractéristiques de 1000 demandeurs de crédits. Il comporte 23 variables avec entre autre l'objet de la demande de crédit (achat de voiture, équipements,…etc.), le statut de la personne (mariée, divorcée,…, etc.), son emploi (qualifié, non qualifié,…etc.)…etc. Mais nous nous intéressons aux croisements entre la variable " housing" (logement-$\mathcal{Y}$) qui peut prendre trois valeurs possibles (trois modalités) :

- For free (pas de charges à payer : soit la personne habite avec sa famille, soit elle a un logement de fonction, etc.).
- Own (propriétaire).
- Rent (locataire) ; et la variable "job" qu'elle peut prendre quatre modalités différentes :
- High qualif/mgm/self emp (en d'autres termes : le management et les professions libérales…).
- Skilled (travail qualifié).
- Unemp/ unskilled/ non res (sans emploi, emploi non qualifié et non résident…).
- Unskilled resident (les résidents avec un travail non qualifié).

La taille de l'échantillon est1000.

Le tableau de contingence est le suivant :

$\mathcal{Y}$ $housing$ / $\mathcal{X}$ job	High qualif/ self emp/ mgm	skilled	Unemp/ unskilled/ non res	unskilled	Total
For free	$n_{11} = 33$	$n_{12} = 63$	$n_{13} = 4$	$n_{14} = 8$	$n_{1.} = 108$
Own	$n_{21} = 94$	$n_{22} = 452$	$n_{23} = 13$	$n_{24} = 154$	$n_{2.} = 713$
rent	$n_{31} = 21$	$n_{32} = 115$	$n_{33} = 5$	$n_{34} = 38$	$n_{3.} = 179$
total	$n_{.1} = 148$	$n_{.2} = 630$	$n_{.3} = 22$	$n_{.4} = 200$	1000

- 713 personnes possèdent leur logement.
- 148 personnes occupent une fonction managériale, ou exercent une profession libérale.
- Etc …

De ce tableau, on déduit le tableau des profils-colonnes :

$\mathcal{Y}$ $housing$ $\mathcal{X}$ job	High qualif/ self emp/ mgm	skilled	Unemp/ unskilled/ non res	unskilled	$n_{i.}/n$
For free	22%	10%	18%	4%	11%
Own	4%	72%	59%	77%	71%
rent	14%	18%	23%	19%	18%
	100%	100%	100%	100%	100%

- 71% des personnes possédent leur logement, cette proportion passe à 4% chez les personnes occupant un poste hautement qualifié.
- Elle est plus élevée chez les personnes sous emploi, ces mêmes personnes ne sont souvent pas propriétaires de leur logement (rent).

Et le tableau des profils-lignes :

$\mathcal{Y}$ $housing$ $\mathcal{X}$ job	High qualif/ self emp/ mgm	skilled	Unemp/ unskilled/ non res	unskilled	
For free	31%	58%	4%	7%	100%
Own	13%	63%	2%	22%	100%
rent	12%	64%	3%	21%	100%
$n_{.j}/n$	15%	63%	2%	20%	100%

- On a 15% des demandeurs de crédit occupent un emploi hautement qualifié. Chez les personnes occupant gratuitement un logement, cette proportion passe à 31%.

On remarquera que la moyenne des profils-lignes (avec des poids correspondant aux effectifs marginaux des lignes) n'est autre que le profil marginal des colonnes :

$$\sum_{i=1}^{r} \frac{n_{ij}}{n_{i.}} \left(\frac{n_{i.}}{n}\right) = \frac{n_{.j}}{n}$$

Et que l'on a de même :

$$\sum_{j=1}^{s} \frac{n_{ij}}{n_{.j}} \left(\frac{n_{.j}}{n}\right) = \frac{n_{i.}}{n}$$

5.2 L'écart à l'indépendance

Lorsque la connaissance de $\mathcal{X}$ ne change pas les distributions conditionnelles de $\mathcal{Y}$, on parle d'indépendance entre $\mathcal{X}$ et $\mathcal{Y}$, et l'indépendance empirique se traduit par :

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

En effet :

On peut parler d'indépendance entre $\mathcal{X}$ et $\mathcal{Y}$, lorsque tous les profils-lignes sont identiques.

On a donc $\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{rj}}{n_{r.}}$ $\forall j$, alors :

$$\frac{n_{.j}}{n} = \sum_{i=1}^{r} \frac{n_{ij}}{n_{i.}}\left(\frac{n_{i.}}{n}\right) = \sum_{i=1}^{r} \frac{n_{1j}}{n_{1.}}\left(\frac{n_{i.}}{n}\right) = \frac{n_{1j}}{n_{1.}}\frac{1}{n}\sum_{i=1}^{r} n_{i.} = \frac{n_{1j}}{n_{1.}} = \frac{n_{ij}}{n_{i.}} \Rightarrow n_{ij} = \frac{n_{i.}n_{.j}}{n}$$

5.2.1 La statistique d^2

L'idée est de comparer les effectifs observés avec les effectifs théoriques que l'on obtiendrait si les variables étaient indépendants.

On adopte généralement la mesure d^2 notée aussi X^2 ou χ^2, définie comme :

$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Et donc cette statistique quantifie l'écart (la distance) entre les effectifs et les effectifs théoriques.

En reprenant l'exemple des demandeurs de crédits :

$\mathcal{Y}$ $housing$ / $\mathcal{X}$ job	High qualif/ self emp/ mgm	skilled	Unemp/ unskilled/ non res	unskilled	
For free	15.98	68.0	2.4	21.6	108
Own	105.5	449.2	15.7	142.6	713
rent	26.5	112.8	3.9	35.6	179
	148	630	22	200	1000

$$d^2 = 18.11 + 0.37 + 1.11 + 8.56 + 1.26 + 0.62 + 0.46 + 0.91 + 1.14 + 0.04 + 0.29 + 0.14 = 32.41$$

Nous savons qu'en situation d'indépendance le χ^2 vaut zéro. En revanche il peut prendre une valeur strictement positive sans que cela soit le reflet d'une liaison significative entre $\mathcal{X}$ et $\mathcal{Y}$, il faut donc définir une valeur seuil, ou une borne supérieure, et dans quel cas elle est atteinte ?

Pour répondre à cette question, il faudra utiliser le résultat suivant :

$$d^2 = n\left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

En effet :

$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left[\sum_i \sum_j \frac{\left(n_{ij}^2 - \left(\frac{n_{i.}n_{.j}}{n}\right)^2 - 2n_{ij}(n_{i.}n_{.j})\right)}{n_{i.}n_{.j}}\right]$$

$$= n\left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} + \sum_i \sum_j \frac{n_{i.}n_{.j}}{n^2} - 2\sum_i \sum_j \frac{n_{ij}}{n}\right] = n\left[\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

Comme $\frac{n_{ij}}{n_{.j}} \leq 1 \Rightarrow \frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \frac{n_{ij}}{n_{i.}}$ $d'o\grave{u}$ $\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \sum_i \sum_j \frac{n_{ij}}{n_{i.}}$

Or $\sum_i \sum_j \frac{n_{ij}}{n_{i.}} = r$ d'où$d^2 \leq n(r-1)$, on pourrait montrer de meme que $d^2 \leq n(s-1)$ donc $d^2 \leq \inf(s-1, r-1)$

Cette borne n'est atteinte que lorsqu'il y a dépendance fonctionnelle, en effet pour que,

$d^2 = n(r-1)$, il faut que$\frac{n_{ij}}{n_{i.}} = 1\ \forall i$, c'est-à-dire s'il existe qu'une case non nulle dans chaque ligne, ce cas est celui ou $\mathcal{Y}$ est fonctionnellement lié à$\mathcal{X}$, sans pour autant la réciproque comme le montre la figure ci-dessous :

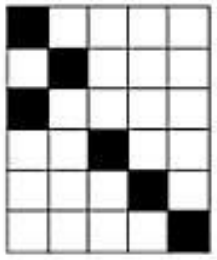


Figure 8

La dépendance fonctionnelle réciproque nécessite l'égalité des modalités : $r = s$, dans ce cas le tableau de contingence reste sous forme diagonale après permutation des lignes ou des colonnes.

Divers coefficients liés au d^2ont été proposés tels que :

_ Le coefficient de contingence K de Pearson, défini de la manière suivante :

$C = \sqrt{\frac{d^2}{d^2+n}}$ commen$n > 0$, forcément cette mesure est comprise entre 0 et 1.

_ Le coefficient T de Tschuprow, défini par :

$T = \sqrt{\frac{d^2}{n\sqrt{(r-1)(s-1)}}}$, aussi compris entre 0et 1, on peut traduire ce coefficient comme un pourcentage d'informations expliquées par la liaison.

_ Le coefficient de V de Cramer, sa formule est la suivante :

$V = \sqrt{\frac{d^2}{n\ inf(s-1,r-1)}}$, il varie entre 0 et1.

Et dernièrement le coefficient $\phi^2 = \frac{d^2}{n}$ qui permet d'éliminer l'effet taille en normalisant le d^2par$n$.

Pour notre exemple “Housing“×“job“ :

- $\phi^2 = \frac{32.41}{1000} = 0.0324$
- $V = \sqrt{\frac{32.41}{1000 \times \min(3-1,4-1)}} = 0.1273$
- $T = \sqrt{\frac{32.41}{1000 \times \sqrt{(3-1)(4-1)}}} = 0.1150$
- Et $C = \sqrt{\frac{32.41}{32.41+1000}} = 0.1772$

5.3 Contribution du χ^2

Détecter une liaison significative, c'est bien ; comprendre la nature de la liaison, c’est mieux. La différence entre le tableau observé (tableau desn_{ij}) et le tableau théorique (tableau des$\frac{n_{i.}n_{.j}}{n}$) permet de construire un indicateur, le résidu, pour chaque case le résidu est égal à$n_{ij} - \frac{n_{i.}n_{.j}}{n}$.

Le plus intéressant est sans doute le signe du résidu qui indique le sens de l’association entre les catégories i de $\mathcal{X}$ et j de $\mathcal{Y}$:

- Positive : attraction entre les caractères.
- Négative : répulsion des caractères.

Pour notre exemple : Housing × Job, le tableau des résidus est le suivant :

$\mathcal{Y}$ $housing$ $\mathcal{X}$ job	High qualif/ self emp/ mgm	skilled	Unemp/ unskilled/ non res	unskilled
For free	$17.02 = 33 - 15.98$	-5.0	1.6	-13.6
Own	-11.5	2.8	-2.7	11.4
rent	-5.5	2.2	1.1	2.2

Pour mesurer l’importance relative d’une case du tableau dans la caractérisation de la liaison, nous lui associons une valeur dite contribution au χ^2égale à : $\frac{\left(n_{ij} - (n_{i.}n_{.j})\right)^2}{\frac{n_{i.}n_{.j}}{n}} \frac{1}{d^2}$, elle indique la fraction d’information qu’apporte la case dans la caractérisation de la liaison entre les variables.

Plus forte sera la contribution, plus la case apporte de l’information.

Pour compléter l’analyse des écarts, il est d’usage d’associer à la contribution le signe du résidu, afin que l’on identifie s’il s’agit d’une attraction ou une répulsion des modalités.

5.4 Cas des tableaux 2×2

Si $\mathcal{X}$ et $\mathcal{Y}$ n’ont que deux modalités chacune le tableau de contingence n’a alors que 4 cases d’effectifs a, b, c, et d.

$\mathcal{Y}$ $\mathcal{X}$	1	2
1	a	b
2	c	d

Avec :

$$n_{11} = a$$
$$n_{12} = b$$
$$n_{21} = c$$
$$n_{22} = d$$

d^2 peut alors s'exprimer par la formule :

$$d^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

En effet :

$$d^2 = n\left[\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$
$$= n\left[\frac{a^2}{(a+b)(a+c)} + \frac{b^2}{(a+b)(b+d)} + \frac{c^2}{(a+c)(c+d)} + \frac{d^2}{(b+d)(c+d)} - 1\right]$$
$$= n\left[\frac{a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c\,)}{(a+b)(c+d)(a+c)(b+d)} - 1\right]$$
$$= n\left[\frac{a^2d^2 + b^2c^2 - 2abcd}{(a+b)(c+d)(a+c)(b+d)}\right] = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Remarquons que ϕ est équivalent à la valeur absolue du coefficient de corrélation entre les variables, si nous attribuons des valeurs arbitraires à leurs catégories.

5.5 Caractère significatif de l'écart à l'indépendance

La statistique du test est l'indicateurd^2, sous l'hypothèse nulle, c'est-à-dire sous l'hypothèse d'indépendance des deux caractères ; elle suit approximativement une loi $\chi^2_{(r-1)(s-1)}$: le nombre de degré de liberté, s'agit du nombre total des cases moins le nombre de cas que nous pouvons déduire des autres lorsque les marges sont fixées.

Dans notre exemple, le nombre de degré de liberté est donc égal à $2 \times 3 = 6$; la formule générale est $ddl = (r-1)(s-1)$.

En se fixant un risque d'erreurα, c'est-à-dire une valeur qui, s'il y avait indépendance, n'aurait qu'une probabilité faible d'être dépassé (usuellement, on prend$\alpha = 5\%$). On rejette l'hypothèse d'indépendance si le chi-carré observé est plus grand que le chi-carré théorique, c'est-à-dire $d^2 > \chi^2_{1-\alpha}(r-1)(s-1)$où :

$\chi^2_{1-\alpha}(r-1)(s-1)$ Est le quantile d'ordre $1-\alpha$ de la loi du χ^2à $(r-1)(s-1)$ ddl. Ainsi sur l'exemple : pour un risque $\alpha = 5\%$, la valeur seuil est$\chi^2_{0.95}(6) = 12.56$, et puis que$d^2 = 32.41 > 12.59$, l'hypothèse d'indépendance est rejetée et donc au risque$\alpha = 5\%$, le mode d'occupation du logement et le type d'emploi sont liés chez les demandeurs de crédit.

Beaucoup de choses sont dite concernant la piètre qualité de l'approximation à l'aide de loi du χ^2 lorsque les effectifs sont faibles. Certains affirment que le test d'indépendance n'est pas valide s'il existe au moins une case ou l'effectif théorique est inférieur à5. D'autres assouplissent cette condition en indiquant que l'approximation est acceptable dès que 80% des cases aient l'effectif théorique supérieur à 5. Enfin lorsque nous manipulons des tableaux 2×2 (ddl=1) l'approximation est invalidée s'il existe au moins une case avec un effectif théorique inférieur à 10 ; on peut néanmoins s'en sortir en introduisant une modification de la statistique dud^2, dite correction de Yates qui est :

$$d^2 = \frac{n\left[|ad - bc| - \frac{n}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

L'information importante qu'il faut retenir est que les faibles valeurs de l'effectif théorique ont tendance à "gonfler" la valeur de la statistique, indiquant à tort une liaison significative.

5.5.1 Exemple

Nous croisons deux variables « own telephone ($\mathcal{Y}$) », qui indique si le client demandeur de crédit possède un numéro de téléphone enregistré à son nom, et « foreign worker ($\mathcal{X}$) », qui indique si le demandeur de crédit est travailleur étranger ou non.

On a le tableau de contingence suivant :

$\mathcal{X}$ / $\mathcal{Y}$	NO	YES	TOTAL
NONE	32	564	596
YES	5	399	404
TOTAL	37	963	1000

Alors :

$$d^2 = \frac{1000(32 \times 399 - 564 \times 5)^2}{(32+564)(32+5)(399+564)(399+5)} = 11.534955$$

On adopte le codage suivant :

$$\mathcal{X} = \begin{cases} 1 \; si \; "foreign \; worker" = "yes" \\ 0 \; si \; "foreign \; worker" = "no" \end{cases}$$

$$\mathcal{Y} = \begin{cases} 1 \; si \; "own \; telephone" = "yes" \\ 0 \; si \; "own \; telephone" = "none" \end{cases}$$

Alors :

$$\bar{x} = 0.963 \qquad et \qquad \bar{y} = 0.404$$

Et :

$$r = \frac{\sum_1^{1000}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^{1000}(x_i - \bar{x})^2 \sum_1^{1000}(y_i - \bar{y})^2}} = 0.10740090$$

$$\phi = \sqrt{\frac{d^2}{1000}} = 010740090$$

Nous observons effectivement que r est exactement identique à ϕ.

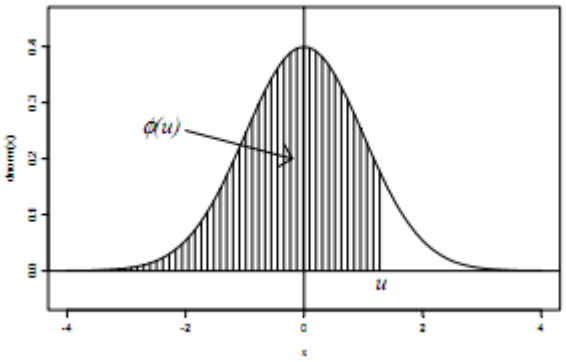
6 Conclusion

L'étude des corrélations est aujourd'hui nécessaire dans presque tous les secteurs de l'activité humaine, les méthodes statistiques qu'elle propose devraient faire partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du biologiste et de l'informaticien, car en s'appuyant sur l'existence des liens statistiques entre deux phénomènes ou plus elle permet de confirmer une théorie ou de la contredire.

Annexes

Table 1 : la table de la loi normale centrée réduite

U une variable aléatoire de loi $N(0,1)$, la table donne la valeur de $\phi(u) = P(U \leq u)$



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

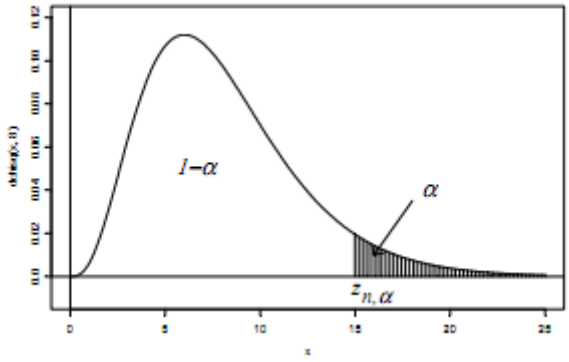
Grandes valeurs de u

u	3.0	3.5	4.0	4.5
$\Phi(u)$	0.9987	0.99977	0.999968	0.999997

Table 2 : table de la loi du χ^2

X une variable aléatoire de loi du χ^2 à n degrés de liberté, et α un réel de $[0,1]$,

La table donne la valeur $z_{n,\alpha} = F_{\chi^2}^{-1}(1-\alpha)$, telle que $P(X > z_{n,\alpha}) = \alpha$.

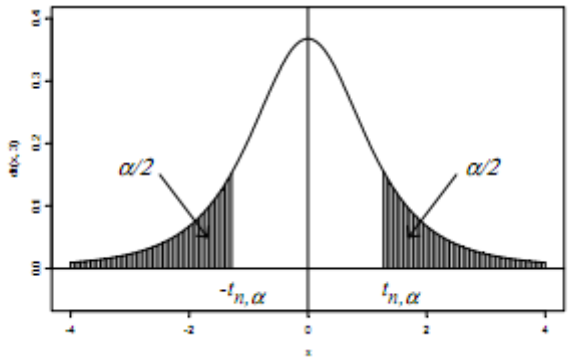


n \ α	0.995	0.990	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00004	0.0002	0.001	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.80
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.60	13.82
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70

Table 3 : table de la loi de Student

X une variable de loi $St(n)$ et α un réel de $[0,1]$,

La table donne la valeur $t_{n,\alpha} = F^{-1}_{St(n)}(1-\frac{\alpha}{2})$ telle que $(|X| > t_{n,\alpha}) = \alpha$.

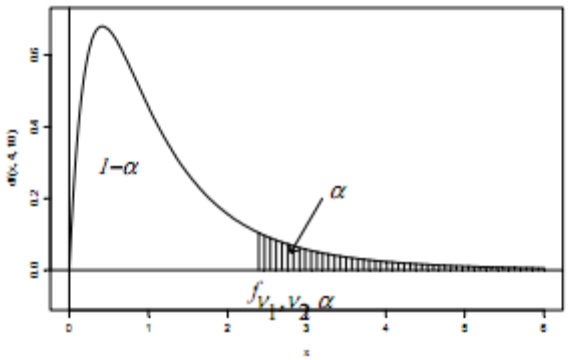


n \ α	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.62
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.527	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$+\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Table 4 : table de la loi de Fisher-Snedecor

X une variable aléatoire de la loi de (v_1, v_2) , la table donne la valeur

$$f_{v_1,v_2,\alpha} = F^{-1}_{F(v_1,v_2)}(1-\alpha)$$ **telle que** $$P(X > f_{v_1,v_2,\alpha}) = \alpha$$, **pour** $\alpha = 5\%$.



v_2 \ v_1	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	+∞
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	241.9	243.9	246.5	248.0	249.1	251.1	252.2	253.0	254.2
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.40	19.41	19.43	19.45	19.45	19.47	19.48	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.69	8.66	8.64	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.84	5.80	5.77	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.60	4.56	4.53	4.46	4.43	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.92	3.87	3.84	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.49	3.44	3.41	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.20	3.15	3.12	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.99	2.94	2.90	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.83	2.77	2.74	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.70	2.65	2.61	2.53	2.49	2.46	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.60	2.54	2.51	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.51	2.46	2.42	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.44	2.39	2.35	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.38	2.33	2.29	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.33	2.28	2.24	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.29	2.23	2.19	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.25	2.19	2.15	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.21	2.16	2.11	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.18	2.12	2.08	1.99	1.95	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.16	2.10	2.05	1.96	1.92	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.13	2.07	2.03	1.94	1.89	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.11	2.05	2.01	1.91	1.86	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.09	2.03	1.98	1.89	1.84	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.07	2.01	1.96	1.87	1.82	1.78	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.99	1.93	1.89	1.79	1.74	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.90	1.84	1.79	1.69	1.64	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.03	1.95	1.85	1.78	1.74	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.82	1.75	1.70	1.59	1.53	1.48	1.39
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.77	1.70	1.65	1.54	1.48	1.43	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.75	1.68	1.63	1.52	1.45	1.39	1.28
+∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.64	1.57	1.52	1.39	1.32	1.24	1.00

Table 5 : table du coefficient de corrélation des rangs de Spearman entre deux variables indépendantes.

Valeurs r de R, ayant une probabilité α d'être dépassée en valeur absolue.

n \ α	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
4	0.600	1.000	1.000						
5	0.500	0.800	0.900	1.000	1.000				
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.224	0.406	0.503	0.587	0.671	0.727	0.776	0.825	0.860
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.802	0.835
14	0.200	0.367	0.464	0.538	0.622	0.675	0.723	0.776	0.811
15	0.189	0.354	0.443	0.521	0.604	0.654	0.700	0.754	0.786
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.732	0.765
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507

n \ α	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456
52	0.095	0.180	0.231	0.274	0.323	0.356	0.386	0.422	0.447
54	0.094	0.177	0.226	0.268	0.317	0.349	0.379	0.414	0.439
56	0.092	0.174	0.222	0.264	0.311	0.343	0.372	0.407	0.432
58	0.090	0.171	0.218	0.259	0.306	0.337	0.366	0.400	0.424
60	0.089	0.168	0.214	0.255	0.300	0.331	0.360	0.394	0.418
62	0.087	0.165	0.211	0.250	0.296	0.326	0.354	0.388	0.411
64	0.086	0.162	0.207	0.246	0.291	0.321	0.348	0.382	0.405
66	0.084	0.160	0.204	0.243	0.287	0.316	0.343	0.376	0.399
68	0.083	0.157	0.201	0.239	0.282	0.311	0.338	0.370	0.393
70	0.082	0.155	0.198	0.235	0.278	0.307	0.333	0.365	0.388
72	0.081	0.153	0.195	0.232	0.274	0.303	0.329	0.360	0.382
74	0.080	0.151	0.193	0.229	0.271	0.299	0.324	0.355	0.377
76	0.078	0.149	0.190	0.226	0.267	0.295	0.320	0.351	0.372
78	0.077	0.147	0.188	0.223	0.264	0.291	0.316	0.346	0.368
80	0.076	0.145	0.185	0.220	0.260	0.287	0.312	0.342	0.363
82	0.075	0.143	0.183	0.217	0.257	0.284	0.308	0.338	0.359
84	0.074	0.141	0.181	0.215	0.254	0.280	0.305	0.334	0.355
86	0.074	0.139	0.179	0.212	0.251	0.277	0.301	0.330	0.351
88	0.073	0.138	0.176	0.210	0.248	0.274	0.298	0.327	0.347
90	0.072	0.136	0.174	0.207	0.245	0.271	0.294	0.323	0.343
92	0.071	0.135	0.173	0.205	0.243	0.268	0.291	0.319	0.339
94	0.070	0.133	0.171	0.203	0.240	0.265	0.288	0.316	0.336
96	0.070	0.132	0.169	0.201	0.238	0.262	0.285	0.313	0.332
98	0.069	0.130	0.167	0.199	0.235	0.260	0.282	0.310	0.329
100	0.068	0.129	0.165	0.197	0.233	0.257	0.279	0.307	0.326

Pour $n > 100$ on admet que R est distribuée comme une $N(0; \frac{1}{\sqrt{n-1}})$.

Bibliographie

[1] YADOLAH DODGE, statistique dictionnaire encyclopédique, 2ème édition, Edition Springer Verlag, France 2007.

[2] DRESS François, les probabilités et la statistique de A à Z, Edition Dunod, France 2007.

[3] GILBERT SAPORTA, probabilités et analyse des données statistiques, 3ème edition , Edition Technip, 2006.

Documents en ligne :

_ RICCO RAKOTOMALALA, Analyse de corrélation : Etude des dépendances, variables quantitatives.

_ RICCO RAKOTOMALALA, Etude des dépendances, variables qualitatives : tableaux de contingence et mesures d'association.

Résumé

L'étude des corrélations est aujourd'hui nécessaire dans tous les secteurs de l'activité humaine. Elle permet de mettre en évidence la relation entre différents phénomènes statistiques étudiés, pour confirmer une théorie ou de la contredire.

Les méthodes et les indices de liaison varient selon la nature des variables étudiées (qualitative, ordinale, ou numérique).

Abstract

The study of correlations is now required in almost all human activity sectors. It can highlight the relation between various studied statistical phenomena, in order to confirm a theory or to contradict it.

The methods and bond index vary depending on the nature of the studied variables (qualitative, ordinal, or numerical).

ملخص

دراسة الارتباطات هي اليوم مطلوبة، تقريبا في جميع قطاعات النشاط البشرى فهي تسمح بإبراز العلاقة بين مختلف الظواهر الإحصائية المدروسة لتأكيد نظرية أو مناقضتها.

الأساليب و مؤشرات الترابط تختلف اعتمادا على طبيعة المتغيرات المدروسة.