Chapitre 4. Mesures d'association

Introduction

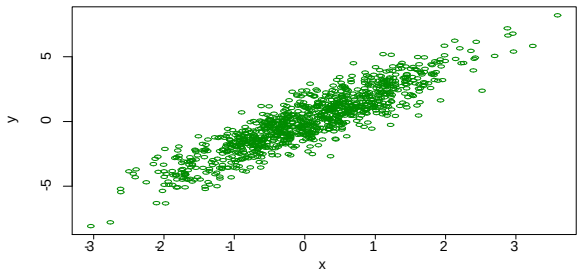
La corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance. Cette corrélation est très souvent réduite à la corrélation linéaire entre variables quantitatives. Par exemple, en épidémiologie, on cherche à savoir si l'exposition à un facteur de risque entraîne l'apparition d'une maladie, en sociologie, on cherche à savoir s'il y a un lien entre la profession du père et la filière choisie par un étudiant à l'université...etc. En particulier on essaie de savoir l'existence d'une liaison entre deux ou plusieurs variables.

Lorsque deux variables X et Y ne sont pas indépendantes entre elles (ou bien elles sont liées l'une à l'autre), on s'intéresse souvent à l'étude de cette liaison par l'analyse graphique à travers le nuage de points ou par des indicateurs numériques pour essayer de quantifier ce que l'on voit.

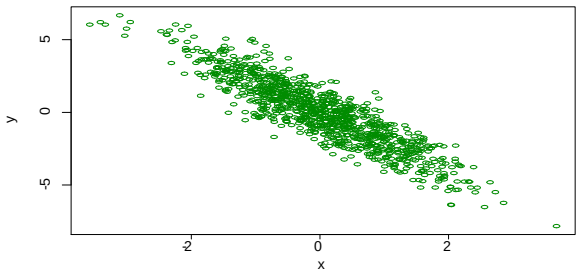
1. Analyse graphique d'une liaison

Les distributions statistiques à deux dimensions peuvent être représentées graphiquement sous forme de nuage de points dans un plan. Le principe est que chaque couple (X, Y) est représenté par un point (en abscisse la variable X, en ordonnée la variable Y), l'ensemble de points forme un nuage de points dont la forme permette de caractériser le type de liaison. Les cas possibles sont les suivants :

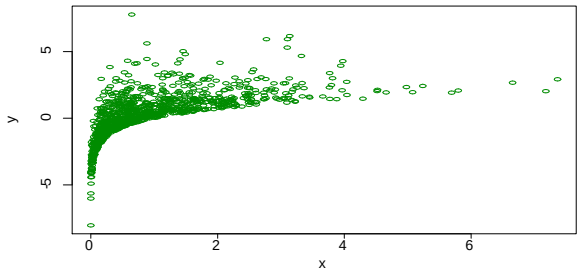
- **Liaison linéaire positive :** les deux variables X et Y varient dans le même sens.
- **Liaison linéaire négative:** X et Y varient en sens inverse.
- **Liaison non-linéaire monotone positive:** X et Y varient dans le même sens mais la pente est différente selon le niveau de X.
- **Liaison non-linéaire non-monotone:** il y a une relation fonctionnelle entre X et Y, mais la liaison n'est pas monotone.
- **Absence de liaison:** la valeur de X ne donne plus d'indication sur la valeur de Y.



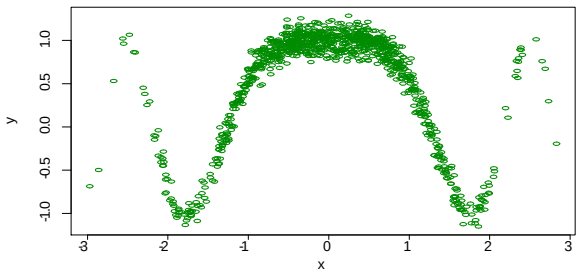
Liaison linaire positive



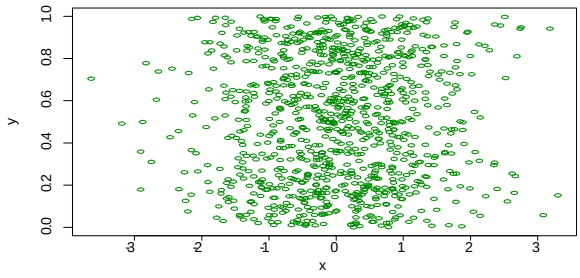
Liaison linaire ngative



Liaison non-linaire monotone positive



Liaison non-linaire non-monotone

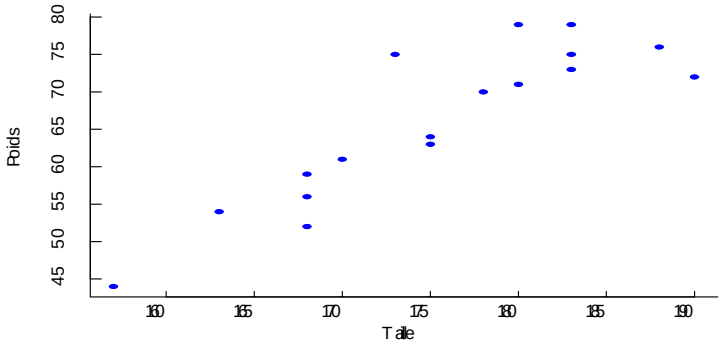


Absence de liaison

Exemple 1 : *On souhaite étudier l'existence d'une liaison entre la taille et le poids de* 17 *individus:*

$$\begin{aligned} Taille &= c(188, 173, 178, 183, 168, 168, 163, 180, 183, 175, 175, 183, 157, 190, 170, 180, 165) \\ Poids &= c(76, 75, 70, 75, 56, 52, 54, 71, 79, 64, 63, 73, 44, 72, 61, 79, 59) \end{aligned}$$

Les données sont représentées graphiquement dans la figure suivante :



Reprsentation simultane du taille et poids des individus

Cette figure montre que la taille et le poids ont tendance à varier dans le même sens. Ils'agit donc d'une liaison linéaire positive.

2. Indicateurs de liaison linéaire

Pour avoir un accès pratique à la quantification du sens et d'intensité de liaison peuvent exister entre les variables, les indicateurs le plus utilisée sont la covariance et le coefficient de corrélation de Pearson.

2.1. Covariance et corrélation

Définition 1: *Soient X et Y deux variables, la covariance entre X et Y est la quantité:*

$$COV(X, Y) := E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

Propriété 1: *Lorsque X et Y sont indépendantes, $COV(X, Y) = 0$, mais la réciproque est généralement fausse. On peut donner le contre exemple suivantes : Soit $X \sim N(0, 1)$ et $Y = X^2$. Alors ,*

$$COV(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0.$$

Donc $COV(X, Y) = 0$, mais au contraire X et Y ne sont pas indépendantes.

Interprétation:

La valeur de covariance est interprétée comme suit :

- $COV(X,Y) > 0$: la liaison est positive.
- $COV(X,Y) = 0$: absence de liaison monotone.
- $COV(X,Y) < 0$: la liaison est négative.

Exemple 2: *Reprenons l'exemple des poids et tailles, la valeur de covariance entre la taille et le poids est égale à $+83.67$. Ce qui confirme quantitativement ce que le nuage de points nous suggère visuellement, à savoir une liaison positive entre la taille et le poids d'une personne.*

Définition 2: *Le coefficient de corrélation linéaire de Pearson entre deux variables X et Y est une normalisation de leur covariance par le produit de leur écarts-types, sa formule est :*

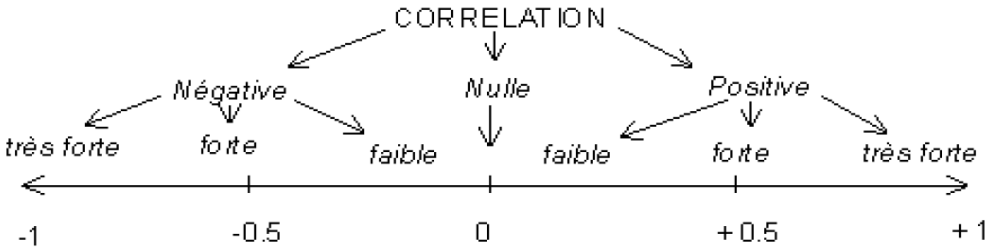
$$R(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,1]\,.$$

Remarques 1: *Le coefficient R est indépendant des unités de mesure. La corrélation d'une variable avec elle même est $R(X,X) = 1$. Si X et Y sont indépendants, alors $R = 0$. La réciproque est fausse, sauf lorsque le couple de variables (X,Y) suit une loi normale bivariée, nous avons l'équivalence.*

Interprétation:

Selon la valeur du coefficient R, nous avons les interprétations suivantes :

- Si R est proche de 1, il existe une forte liaison linéaire positive entre X et Y.
- Si R est proche de 0, il n'y a pas de liaison linéaire entre X et Y.
- Si R est proche de -1, il existe une forte liaison linéaire négative entre X et Y.



Interprtation du coefficient de corrlation

Remarques 2: *Le signe de R indique donc le sens de la liaison tandis que la valeur absolue de R indique l'intensité de la liaison.*

Exemple 3: *On reprend l'exemple de la liaison entre la taille et le poids. La valeur de corrélation est égale à $+0.88$, ce qui indique qu'il y a une liaison positive très forte reliant la taille et le poids d'une personne.*

Définition 3: *Soit un échantillon de n observations d'un couple (X,Y), la corrélation empirique est définie par :*

$$\hat{R}_{XY} := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Il s'agit d'un estimateur asymptotiquement baisé et convergent, avec :

$$E\left(\hat{R}_{XY}\right) = R - \frac{R\left(1-R^2\right)}{2n} \quad \text{et} \quad Var\left(\hat{R}_{XY}\right) = \frac{\left(1-R^2\right)^2}{n}.$$

2.2. Test paramétrique sur le coefficient de corrélation linéaire

Après le calcul du coefficient de corrélation $\hat{R}$ estimé sur un échantillon de taille n, il faut déterminer si le coefficient de corrélation R est significativement différent de 0. Le test est de l'hypothèse:

$$H_0 : R = 0, \text{ ‹‹absence de liaison linéaire entre } X \text{ et } Y\text{››},$$

$$H_1 : R = 0, \text{ ‹‹existence d'une liaison entre } X \text{ et } Y\text{››}.$$

La statistique de test sous H_0 est :

$$T = \frac{\hat{R}\sqrt{n-2}}{\sqrt{1-\hat{R}^2}}$$

qui suit une loi de Student à $(n-2)$ degrés de liberté. Au risque α, H_0 est rejetée si

$$|T| > t_{1-\frac{\alpha}{2}}(n-2)$$

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.

Exemple 4: *Nous revenons à l'exemple précédent du même échantillon, mais nous voulons savoir maintenant s'il existe une liaison entre le poids et l'âge des individus, les données sont dans le tableau suivant :*

poids	76	75	70	75	56	52	54	71	79	64	63	73	44	72	61	79	59
Âge	41	42	32	39	30	33	26	30	53	32	47	34	23	36	31	29	28

On pose les hypothèses de test comme suit :

$$H_0 : \textit{absence de liaison entre l'âge et le poids.}$$

$$H_1 : \textit{existence d'une liaison entre l'âge et le poids.}$$

Pour un risque d'erreur $\alpha = 5\%$, la statistique de test sous l'hypothèse nulle est $T = 2.84$, le quantile d'ordre $1-\frac{\alpha}{2}$ correspondant à 15 degré de liberté est $t(0.975, 15) = 2.13$. Puisque $2.84 > 2.13$, on peut rejeter donc H_0 et affirme qu'il existe une liaison entre le poids et l'âge des personnes à 95%.

3. Autres mesures d'association

D'une part, on peut voir le coefficient de corrélation de Pearson de différentes manières pour obtenir des informations supplémentaire comme le phi de Pearson. D'autre part, il ne caractérise qu'une liaison linéaire pour l'étude de liaison non linéaire, plusieurs mesures on été proposées, certaines sont basées sur les rangs des observations comme le rho de Spearman, d'autres sont basées sur la notion de paires concordances et discordances comme le tau de Kendall.

3.1. Phi de Pearson

Le coefficient de corrélation ϕ de Pearson permet de mesurer l'intensité de la liaison entre deux variables binaires (codées 0 ou 1). Le calcul est réalisé à travers le coefficient de Pearson sur les variables binaires ou sur un tableau composé de deux lignes et deux colonnes comme suit :

$Y \quad vs \quad X$	1	0
1	a	b
0	c	d

Sa formule est :

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Le coefficient phi est similaire au coefficient de corrélation de Pearson dans son interprétation. Alors il varie entre -1 et 1. Plus il est proche de ces bornes plus la liaison est forte entre les deux variables. $\phi = 0$ indique une situation d'indépendance.

Remarque 3 : *On utilise souvent le codage* 1 *de la modalité qui nous intéresse et* 0 *à la second. De plus, ce codage détermine le signe de* ϕ*, mais il n'a pas d'incidence sur la valeur absolue du coefficient.*

Exemple 5: *On veut étudier la liaison entre les caractères : ‹‹être fumeur›› (plus de* 20 *cigarettes par jour, pendant* 10 *ans) et ‹‹avoir un cancer de la gorge››, sur une population de* 140 *personnes. Les résultats sont dans le tableau suivant:*

Observé	*cancer*	*non cancer*
fumeur	*42*	*27*
non fumeur	*13*	*58*

Nous obtenons,

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{42 \times 58 - 27 \times 13}{\sqrt{69 \times 85 \times 71 \times 55}} = 0.44$$

Il s'agit donc, d'une faible dépendance positive entre ‹‹être fumeur›› et ‹‹avoir un cancer de la gorge››.

3.2. Mesures d'association en Epidémiologie

Un facteur F est un facteur de risque pour une Maladie M si l'exposition au facteur F modifie la probabilité d'apparition de la maladie.

Observé	Malades	non Malades
Exposés	a	b
non Exposés	c	d

L'effet d'un facteur de risque est mesuré par le Risque Relatif (RR) et Odds Ratio (OR). La mesure d'association la plus simple entre une exposition et une maladie est le **Risque Relatif** (RR) donné par:

$$RR = \frac{\text{Risque de développer la maladie si exposé au facteur}}{\text{Risque de développer la maladie si non exposé}} = \frac{R_1}{R_0},$$

avec

$$R_1 = a/(a+b) \qquad \text{et} \quad R_0 = c/(c+d)$$

d'où

$$RR = \frac{a(c+d)}{c(a+b)}.$$

Ainsi,
$RR = 1$: ce n'est pas un facteur de risque . Il n'y a pas de relation démontrée entre la maladie et le facteur de risque étudié
$RR > 1$: le facteur étudié est considéré comme facteur de risque
$RR < 1$: facteur protecteur
$RR = 13$ signifie que le risque de devenir malade est 13 fois plus important chez les exposés que chez les non-exposés.

Si on ne peut pas calculer le risque relatif (études des cas-témoins ou contrôle de la proportion de malades/non malades), dont l'estimation RR est impossible, on calcule l'odds ratio (OR) . C'est le rapport des côtes d'exposition chez les cas et chez les Témoins:

$$OR = (a/c)/(b/d) = \frac{ad}{cb} = \frac{R_1}{(1-R_1)} \frac{(1-R_0)}{R_0}$$

Même interprétation que le RR.

Exemple 6: *On reprond l'exemple de la liaison entre les caractères : ‹‹être fumeur›› et ‹‹avoir un cancer de la gorge››:*

Observé	*cancer*	*non cancer*
fumeur	*42*	*27*
non fumeur	*13*	*58*

On a

$$RR = \frac{R_1}{R_0} = \frac{a/(a+b)}{c/(c+d)} = \frac{42/(42+27)}{13/(13+58)} = \frac{0.6087}{0.1831} = 3.32$$

Donc le risque d'avoir un cancer de la gorge est 3.32 *fois plus important chez les fumeurs que chez les non-fumeurs.*

Exemple 7: *Considérons l'exemple d'étude de l'effet de la consommation de viande nsuffisamment cuite et la contamination par le toxoplasma gondii pour un échantillon de* 160 *femmes:*

Observé	*toxoplasmose*	*pas de toxoplasmose*
Viande insuffisamment cuite	*44*	*15*
Pas de consommation de viande insuf. cuite	*36*	*65*

$OR = (44 \times 65)/(36 \times 15) = 5.3$. *Donc, les femmes ayant consommé de la viande insuffisamment cuite sont plus fréquemment contaminées par le toxoplasma gondii que les femmes n'en ayant pas consommé.*

3.3. Concept de concordance

Soit $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ un échantillon de n observations d'un couple (X, Y). Il existe $\complement_n^2 = \frac{n!}{2(n-2)!}$ pairs de distributions distinctes de couples (x_i, y_i) et (x_j, y_j) qui sont dites concordantes où discordantes selon :

1. **Concordantes**: $(x_i - x_j)(y_i - y_j) > 0$ i.e. $(x_i < x_j$ et $y_i < y_j)$ où $(x_i > x_j$ et $y_i > y_j)$.

2. **Discordantes**: $(x_i - x_j)(y_i - y_j) < 0$ i.e. $(x_i < x_j$ et $y_i > y_j)$ où $(x_i > x_j$ et $y_i < y_j)$.

Définition 4: *La fonction de concordance est la différence entre la probabilité de concordance et celle de discordance entre deux couples* (X_1, Y_1) *et* (X_2, Y_2) *. Elle est donnée par :*

$$Q := P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right].$$

Comme les variables aléatoires sont continues, alors

$$P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right] = 1 - P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right]$$

donc,

$$Q = 2P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - 1,$$

où

$$P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] = P\left[X_1 < X_2, Y_1 < Y_2\right] + P\left[X_1 > X_2, Y_1 > Y_2\right].$$

3.4. Rho de Spearman

Le coefficient de corrélation rho de Spearman permet d'analyser les liaisons non linéaires entre les rangs des observations des variables. La valeur de ce coefficient notée ρ est équivalente au coefficient de corrélation de Pearson. Il a été développé par Spearman. Autrement dit, le coefficient rho de Spearman est la corrélation de Pearson appliquée sur les rangs.

Définition 5 : *Soit une série de* n *observations* $\{(x_i, y_i)\}_{1 \leq i \leq n}$ *d'un couple* (X, Y)*, on note les rangs observés de* X *et de* Y *par*

$$r_i = rang(x_i), \qquad s_i = rang(y_i), \quad i = 1, ..., n$$

et les moyennes et les variances des rangs observés :

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i, \quad \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i, \quad S_r^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2 \quad S_s^2 = \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})^2?$$

La covariance entre les rangs observés :

$$S_{rs} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s}).$$

Alors, le coefficient rho de Spearman noté ρ, devient :

$$\rho := \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2 \sum_{i=1}^{n}(s_i - \bar{s})^2}} = \frac{S_{rs}}{S_r S_s}.$$

Compte tenu de certaines propriétés des rangs, nous pouvons déduire une expression simplifiée

$$\rho := 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} D_i^2, \qquad avec \quad D_i = r_i - s_i.$$

Comme la valeur de ρ est comprise entre -1 et 1 nous avons les interprétations suivantes :

1. Si tous les classements des paires sont concordantes, $\rho = 1$.
2. Si tous les classements des paires sont totalement indépendants, $\rho = 0$.
3. Si tous les classements des paires sont discordantes, $\rho = -1$.

3.5. Tau de Kendall

Le tau de Kendall est défini pour mesurer la liaison non linéaire entre deux variables. Il donne une mesure de la corrélation entre les rangs des observations. On peut exprimer le tau de Kendall de deux manières différentes, soit en fonction des observations, ou en fonction de la concordance.

Définition 6: *Soit (X_1, Y_1) un vecteur aléatoire et (X_2, Y_2) un vecteur indépendant mais de même loi que (X_1, Y_1). Le tau de Kendall noté τ est définit par :*

$$\tau_{XY} := P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Définition 7: *Soit une série de n observations $\{(x_i, y_i)\}_{1\leq i\leq n}$ d'un couple (X, Y). Le tau de Kendall est définit par :*

$$\tau := \frac{(n_c - n_d)}{N} = \frac{(n_c - n_d)}{\frac{1}{2}n(n-1)}; \quad \tau\epsilon[-1, 1]$$

où

n_c : *nombre de paires concordantes,*

n_d : *nombre de paires discordantes,*

N : *nombre total de paires.*

Comme la valeur de τ est comprise entre -1 et 1 nous avons les interprétations suivantes :

4. Si tous les paires sont concordantes, alors $\tau = 1$.
5. Si les deux classements de paires sont totalement indépendants, alors $\tau = 0$.
6. Si tous les paires sont discordantes, alors $\tau = -1$.

Remarque 4: *Le tau de Kendall et le rho de Spearman sont des mesures utilisées pour la caractérisation d'une liaison non linéaire, mais la seule différenciation entre les deux coefficients est que le tau de Kendall peut considérer comme une probabilité et le rho de Spearman s'interpète comme une proportion de variance expliquée. Il y'a cependant une relation entre les valeurs de ces deux coefficients. En effet,*

$$-1 \leq 3\tau - 2\rho \leq 1,$$

et lorsque n est assez grand, et τ, ρ pas trop proches de 1 :

$$\rho = \frac{3}{2}\tau.$$

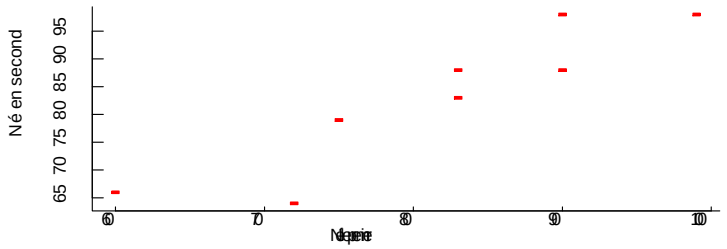
Finalement, si (X, Y) suit une loi normale bivariée, alors

$$\tau = \frac{2}{\pi} \arcsin \rho.$$

Exemple 8 : *Dans cet exemple, on fait passer des tests d'intelligence à 8 couples de vrais jumeaux. Le but est devoir s'il y a une liaison entre les tests de celui qui est né en premier et ceux de celui qui est né en second. Les scores plus élevés correspondant à de meilleurs résultats aux tests:*

couple de jumeaux	1	2	3	4	5	6	7	8
Né le premier	90	75	99	60	72	83	83	90
Né en second	88	79	98	66	64	83	88	98

Les données sont représentées graphiquement dans la figure suivante :



Reprsentation simultane des tests d'intelligence de vrais jumeaux

A partir la figure on peut voir une liaison entre les résultats du test d'intelligence d'un couple de jumeaux. De plus, la valeur de rho de Spearman $\rho = 0.93$ et tau de Kendall $\tau = 0.78$ affirme numériquement ce que l'on voit. En effit,

r_i	6.5	3	8	1	2	4.5	4.5	6.5
s_i	5.5	3	7.5	2	1	4	5.5	7.5
D_i	1	0	0.5	-1	1	0.5	-1	-1

$$\rho = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} D_i^2 = 1 - \frac{6}{8\,(63)} 5.5 = 0.93$$

De plus, on $N = \complement_n^2 = 28$ nombre total de paires et

$$n_c = 25 : \textit{nombre de paires concordantes},$$
$$n_d = 3 : \textit{nombre de paires discordantes},$$

donc,

$$\tau = \frac{n_c - n_d}{N} = \frac{22}{28} = 0.78.$$

3.6. Fonction R de calcul du coefficient de corrélation

La fonction **cor**() de **R** peut être utilisée pour calculer le coeficient de corrélation entre deux variables, X et Y. Un format simplifié de la fonction est:

```
# x et y sont des vecteurs de type numéric
  cor(x, y, method = c("pearson", "kendall", "spearman"))
```

Exemple 9:

```
  x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1)
  y <- c( 2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 3.8)
  cor(x,y, method="pearson")
  [1] 0.5712
```

La fonction **cor.test**() peut être utilisée pour calculer le niveau de significativité de la corrélation. Elle teste l'association entre deux variables en utilisant les méthodes de Pearson, Kendall ou de Spearman. Le format simplifié de la fonction :

```
  cor.test(x, y, method=c("pearson", "kendall", "spearman"))
```

* * * * * * * *K. Pearson* * * * * * *

Karl Pearson (27 mars 1857–27 avril 1936), mathématicien britannique, est un des fondateurs de la statistique moderne appliquée à la biomédecine (biométrie et biostatistique). Il est principalement connu pour avoir développé le coefficient de corrélation et le test du khi − 2. Il est aussi l'un des fondateurs de la revue Biometrika, dont il a été rédacteur en chef pendant 36 ans et qu'il a hissée au rang des meilleures revues de statistique mathématique. Il reçoit la médaille Darwin en 1898. Il reçoit la médaille Rudolf Virchow de la Société d'anthropologie de Berlin en 1932. L'International Statistical Institute décerne depuis 2013 le prix Karl-Pearson en son honneur.

* * * * * * * *C.E. Spearman* * * * * * *

Charles Edward Spearman, né le 10 septembre 1863 à Londres où il meurt le 17 septembre 1945, est un psychologue anglais connu pour son travail sur l'intelligence et facteur g. Il fait des recherches en statistique et analyse factorielle. La corrélation de Spearman est nommée d'après lui. Président de la British Psychological Society (1923-1926) et Président de la British Association for the Advancement of Science (1925). La médaille Spearman, créée en 1965 par la British Psychological Society pour récompenser les travaux de jeunes chercheurs en psychologie du Royaume-Uni, est nommée en son honneur.

* * * * * * * *M.G. Kendall* * * * * * *

Maurice George Kendall, né le 6 septembre 1907 à Kettering et mort le 29 mars 1983, est un statisticien britannique. Après des études de mathématiques au St John's College de Cambridge, Kendall entame sa carrière au ministère de l'Agriculture du Royaume-Uni. De 1949 à 1961, il occupe la chaire de statistique à la London School of Economics. À partir de 1972, Kendall est directeur de l'enquête mondiale sur la fécondité menée par l'Institut international de statistique en collaboration avec l'ONU. Membre de la Royal Statistical Society en 1943, il en a été le président à deux reprises. Maurice Kendall a été fait chevalier en 1974.