Chapitre 3
Etude de la liaison entre deux variables
Analyse descriptive des données et tests d'indépendance

I Introduction

Etude simultanée de deux variables X et Y définies sur une même population $\mathcal{P}$: mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables.
Exemples
Etude de la liaison entre
– le QI du père et le QI du fils (quantitatives) ;
– le salaire et le sexe (quantitative / qualitative) ;
– la couleur des yeux et la couleur des cheveux (qualitatives).

1) Notions de dépendance et d'indépendance

Variables liées : les variations de l'une dépendent des variations de l'autre.

Variables indépendantes : les deux variables varient indépendamment l'une de l'autre. Dans ce cas :
– la connaissance de la valeur prise par l'une des deux variables sur un individu n'apporte aucune information sur la valeur prise par l'autre variable sur cet individu ;
– *Exemple* : si le salaire et le sexe sont deux variables indépendantes, connaître le sexe d'un employé n'apporte aucune information sur son salaire.

Rôle des variables dans la relation : dans certains cas, une variable peut en expliquer une autre, dans d'autres cas, les variables jouent des rôles symétriques.

Vocabulaire : Pour des variables qualitatives : association. Pour des variables quantitatives : corrélation.

2) Observations

Pour étudier la relation entre deux variables, on fait des observations sur un échantillon de n individus tirés au sort dans la population.
On note x_i et y_i les valeurs de X et Y observées sur le i^{e} individu tiré au sort.
On dispose ainsi de deux échantillons appariés de mesures.

individu $n°i$	1	$\cdots$	i	$\cdots$	n
variable X	x_1	$\cdots$	x_i	$\cdots$	x_n
variable Y	y_1	$\cdots$	y_i	$\cdots$	y_n

Les méthodes utilisées pour étudier la relation dépendent du type des variables étudiées.

II Etude de la liaison entre deux variables quantitatives

Deux variables quantitatives sont corrélées si elles tendent à varier l'une en fonction de l'autre.
On parle de *corrélation positive* si elles tendent à varier dans le même sens, de *corrélation négative* si elles tendent à varier en sens contraire.
Exemple 1
On veut étudier la relation entre le QI du père (X) et le QI du fils (Y).

Sur un échantillon aléatoire de 12 couples (père, fils), on a relevé le QI du père et le QI du fils. On dispose de 2 échantillons appariés de mesures (x_i, y_i) :

couple (père, fils) n^o i	1	2	3	4	5	6	7	8	9	10	11	12
QI du père x_i	123	144	105	110	98	138	131	90	119	109	125	100
QI du fils y_i	102	138	126	133	95	146	115	100	142	105	130	120

1) Analyse descriptive des données

Comment détecter une corrélation, quelle en est la forme, le sens (les variables varient-elles dans le même sens ou bien en sens contraire), l'intensité ?
– Outil graphique : le nuage de points.
– Indicateur numérique de sens et d'intensité : coefficient de corrélation.

a) Graphique : nuage de points (*diagramme de dispersion, scatter-plot*)

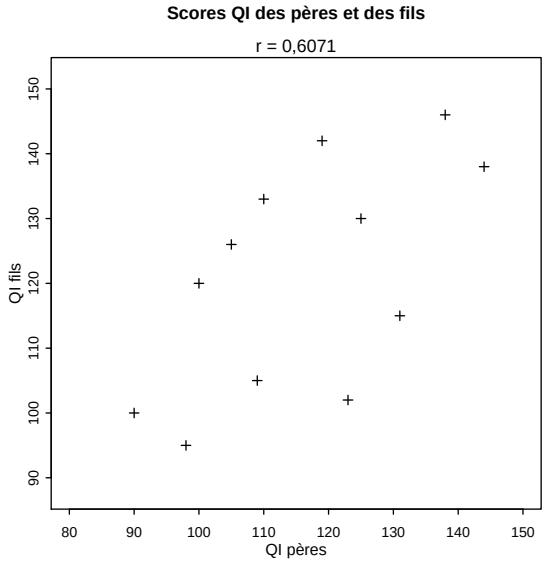
1. *Exemple 1*

   Le nuage de points, graphique 1 :

   Chaque couple (père, fils) est représenté par un point : l'abscisse est le `QI du père` et l'ordonnée le `Qi du fils`. L'ensemble forme un nuage de points.

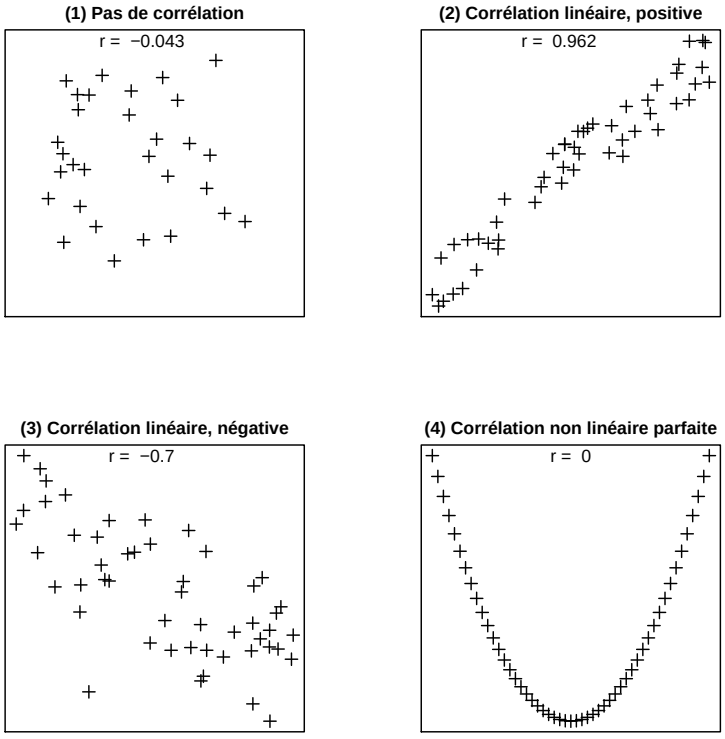
   La forme du nuage

   Le nuage est allongé , étiré du bas à gauche vers le haut à droite. Les QI ont tendance à varier dans le même sens. La corrélation observée est positive. La forme est allongée mais l'étirement est modéré.



Graphique 1

2. *Exemples-type, graphique 2 :*



Graphique 2

- (1) nuage très arrondi ; pas de relation apparente ;
- (2) nuage très étiré : le nuage a une forme linéaire très marquée. On observe sur l'échantillon une tendance de X et Y à varier dans le même sens. La corrélation observée est positive.
- (3) nuage modérément étiré (forme linéaire moins marquée) , du haut à gauche vers le bas à droite : on observe sur l'échantillon une tendance de X et Y à varier dans des sens opposés ; la corrélation observée est négative.
- (4) les points sont sur une courbe (parabole) non linéaire. La corrélation observée est parfaite, de type non linéaire. Il n'y a pas monotonicité : la courbe est d'abord décroissante pluis croissante.

3. *Forme de référence la plus simple : la droite*

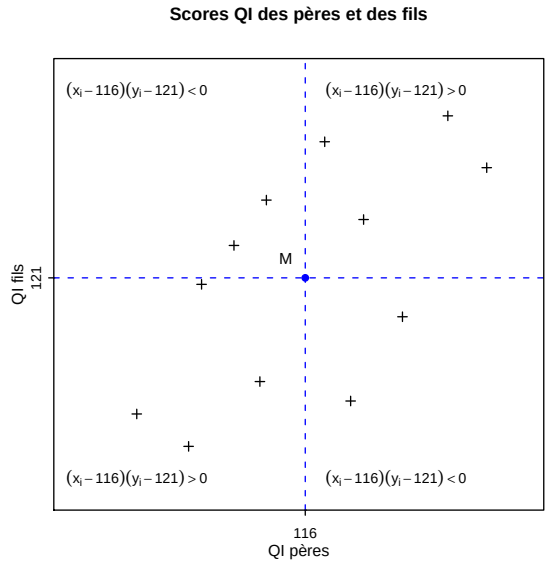
   La droite exprime une relation entre X et Y du type $Y = aX + b$.

   Si la forme du nuage s'apparente à une droite, on parle alors de corrélation linéaire entre les variables. Plus le nuage est étiré et plus la corrélation linéaire observée est forte.

   Pour mesurer la force et le sens de la corrélation linéaire, on calcule un indicateur numérique : le coefficient de corrélation linéaire.

   Remarque : dans le chapitre sur la régression linéaire, on déterminera la droite qui s'ajuste le mieux aux données.

b) Coefficient de corrélation linéaire



Graphique 3

1. *Covariance*

   Le centre du nuage (centre de gravité ou barycentre) est défini par le point M qui a pour coordonnées les moyennes $\overline{x}$ et $\overline{y}$: 116 pour `QI du père` et 121 pour `Qi du fils`. Il est représenté par un cercle plein. Les points du nuage se répartissent autour de leur centre avec une certaine dispersion.

   Cette dispersion est mesurée par la `covariance` définie par la formule (échantillon)

$$cov^*(x,y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

   Chaque individu contribue à la covariance par la quantité $(x_i - \overline{x})(y_i - \overline{y})$ qui mesure son écart au couple de moyennes.

   - Cette quantité est positive pour les couples (père, fils) dont les deux QI sont, soit tous les deux supérieurs à leur moyenne, soit tous les deux inférieurs. Elle est négative pour les couples qui ont un QI en dessous de la moyenne et l'autre QI au-dessus.
   - La covariance peut prendre n'importe quelle valeur réelle.
   - Son signe renseigne sur le sens de variation des variables. Elle est positive si les variables ont tendance à varier dans le même sens et négative en sens contraire.
   - Elle est sensible aux unités de mesure.

   Exemple 1. Résumés des données et calcul de la covariance :

$$\text{QI père :} \quad \sum x_i = 1392 \qquad \sum x_i^2 = 164566 \quad \bar{x} = 116 \quad s_x^* = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = 16,7712$$

$$\text{QI fils :} \quad \sum y_i = 1452 \qquad \sum y_i^2 = 179068 \quad \bar{y} = 121 \quad s_y^* = \sqrt{\frac{\sum y_i^2 - n\bar{y}^2}{n-1}} = 17,5188$$

$$\sum x_i\, y_i = 170394$$

$$cov^*(x,y) \underset{\text{(formule de calcul)}}{=} \frac{\sum x_i\, y_i - n\bar{x}\bar{y}}{n-1} = \frac{170394 - 12 \times 116 \times 121}{11} = 178,3636.$$

2. *Coefficient de corrélation linéaire*
   Le coefficient mesure le sens et l'intensité de la corrélation linéaire.
   Il est noté ρ dans la population et r sur l'échantillon.
   Formule de calcul sur un échantillon :

$$r(x,y) = \frac{cov^*(x,y)}{s_x^* \, s_y^*}.$$

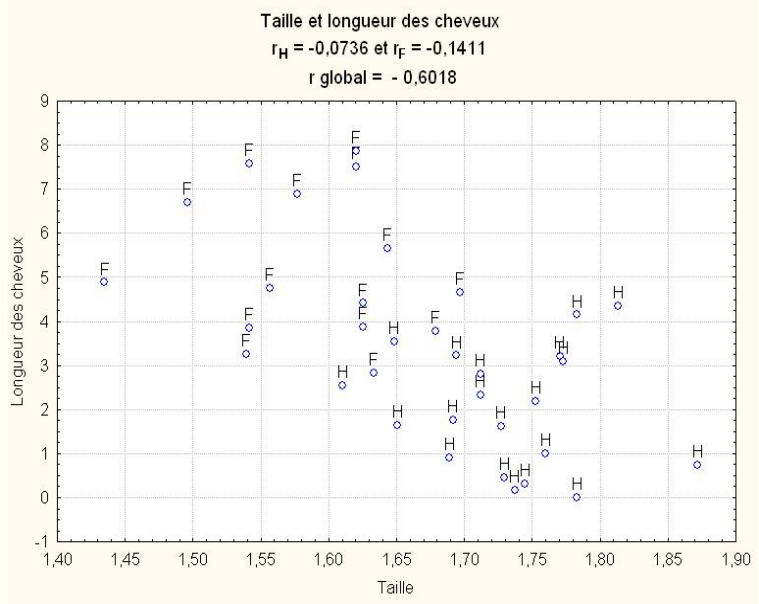
   - Le coefficient est indépendant des unités de mesure.
   - Il est compris entre -1 et 1.
   - $r = 0$ si la corrélation linéaire observée est nulle.
   - $r = \pm 1$ si la corrélation linéaire observée est parfaite. Les points sont alignés.
   - Le coefficient est positif si la liaison est positive.
   - Le coefficient est négatif si la liaison est négative.

   Exemple 1. Calcul du coefficient :

$$r(x,y) = \frac{178,3636}{16,7712 \times 17,5188} = 0,6071.$$

   La corrélation linéaire observée est positive et relativement forte.
   Pour les exemples-type précédents :
   - (1) r quasiment nul ; le nuage est rond, il ne montre pas de relation linéaire apparente entre les deux variables.
   - (2) $r = 0,962$: la corrélation linéaire observée est positive et très forte. Le nuage est très étiré.
   - (3) $r = -0,7$: la corrélation linéaire observée est négative et forte. Le nuage est modérément étiré.
   - (4) le nuage présente une corrélation non linéaire parfaite : les points sont sur une courbe $y = f(x)$. La corrélation linéaire observée est nulle : $r = 0$.

3. *Attention aux sous-groupes définis par une variable qualitative*



Graphique 4

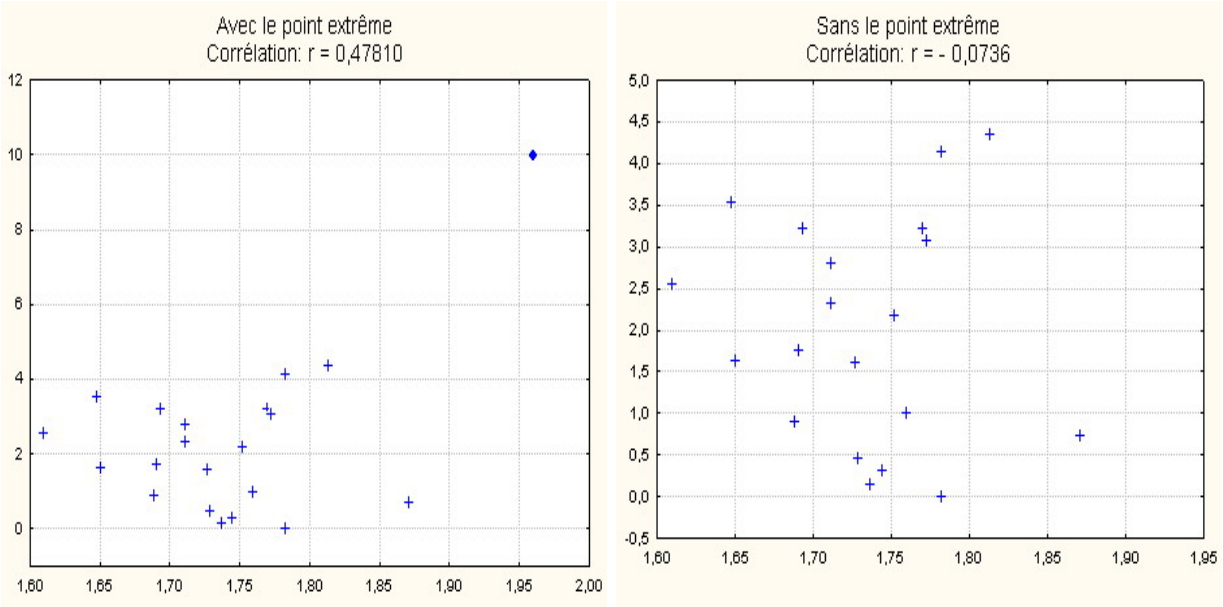
Y-a-t-il une relation entre la taille d'un individu et la longueur de ses cheveux [1] ? Le nuage donne

1. Exemple emprunté à R. rakotomalala...

l'impression d'une corrélation linéaire négative assez forte entre les deux variables avec un coefficient $r = -0,6018$.

Si l'on différencie les hommes et les femmes, on voit que pour chaque groupe le nuage est arrondi, la corrélation linéaire est quasiment nulle ($r_H = -0,0736$ et$r_F = -0,1411$).

4. *Attention aux valeurs extrêmes*



Graphique 5

2) Tests d'indépendance pour deux variables quantitatives

a) Cadre général

On étudie deux variables quantitatives X et Y définies sur une population $\mathcal{P}$.

On veut tester l'existence d'une liaison entre les deux variables.

Hypothèses du test et niveau α :

H_0 : les variables sont indépendantes

H_1 : les variables sont liées (positivement, négativement)

Le test est effectué pour un niveau α fixé.

Observations : Pour réaliser le test, on a tiré au sort un échantillon d'individus de taille n dans la population.

On dispose de deux échantillons appariés de mesures (x_i, y_i).

Exemple 1 : Existe-t-il une liaison entre le QI du père (X) et le QI du fils (Y) ?

$\mathcal{P}$: couples (père, fils)

Variable X : QI du père

Variable Y : QI du fils.

On teste :

H_0 : les variables QI sont indépendantes

H_1 : les variables QI sont liées

Le test est effectué au niveau $\alpha = 5\%$.

On dispose de deux échantillons appariés de scores observés sur $n = 12$ couples (père, fils) :

couple (père, fils) n^o i	1	2	3	4	5	6	7	8	9	10	11	12
QI du père x_i	123	144	105	110	98	138	131	90	119	109	125	100
QI du fils y_i	102	138	126	133	95	146	115	100	142	105	130	120

Tests d'indépendance : On présente dans ce document deux tests alternatifs, utilisables pour des variables continues.

Pour déterminer le test à utiliser, on doit considérer la loi du couple de variables (X, Y).

1. La loi du couple est **binormale** : test **paramétrique** sur le coefficient de corrélation linéaire (Bravais-Pearson) ; *binormalité à vérifier, ce qui est difficile (voire impossible) pour de petits échantillons. Test utilisable sans la normalité pour des échantillons assez grands.*
2. La loi du couple n'est pas **binormale** : test **non paramétrique** basé sur le coefficient de corrélation empirique de Spearman.

b) Test paramétrique sur le coefficient de corrélation linéaire

On note ρ le coefficient de corrélation linéaire de X et Y défini sur la population.

1. *Postulat requis*

   Le couple de variables (X, Y) doit se distribuer suivant une loi « binormale ».

   Généralisation de la loi normale à un couple de variables. Toute combinaison linéaire des deux variables est normale. En particulier, X et Y sont des variables normales.

   On doit vérifier ce postulat avant d'appliquer le test, voir le point 3. ci-dessous.

   Exemple 1 : on admet que le couple de QI se distribue selon une loi binormale.

2. *Déroulement du test*

   (a) *Hypothèses du test*

   Sous le postulat de binormalité, l'indépendance équivaut à $\rho = 0$. On teste alors

   $H_0 : \rho = 0$ (indépendance)

   $H_1 : \rho \neq 0$; ou $\rho > 0$; ou $\rho < 0$ (liaison, liaison positive, liaison négative).

   (b) *Observations - Statistique du test*

   La statistique est une variable aléatoire calculée sur les données de l'échantillon tiré au sort. Sa valeur observée sur l'échantillon est un résumé des données permettant de choisir entre H_0 et H_1.

   La statistique utilisée pour le test est le *coefficient de corrélation linéaire empirique*, noté R.

   Exemple 1 : Sa valeur observée sur l'échantillon tiré au sort est $r(x, y) = 0,6071$.

   La statistique R est un estimateur du coefficient ρ de la population. Sa valeur observée $0,6071$ est la valeur estimée de ρ fournie par l'échantillon.

   (c) *Loi sous H_0 de la statistique*

   R prend des valeurs entre -1 et 1.

   Sa loi sous H_0 est symétrique en 0 et dépend de n.

   Remarque : les calculs sont basés sur la statistique équivalente $T = \dfrac{\sqrt{n-2}R}{\sqrt{1-R^2}}$ qui suit sous H_0 la loi de Student à $n-2$ ddl.

(d) *Intervalle d'acceptation, région critique*

Pour le test d'une liaison ($H_1 : \rho \neq 0$) :

Sous H_0, on s'attend à observer une valeur de R proche de 0. *Les valeurs de R les plus proches de 0 sont les plus conformes à H_0.*

Sous H_1, on attend une valeur de R plus proche de -1 ou 1. *Les valeurs de R les plus extrêmes sont les plus significatives de H_1.*

➡ La RC est située aux deux extrémités du domaine.

(e) *Décision* [2]

Pour la décision, on calcule la p-valeur

$$\alpha = 2\, P_{H_0}(R \geq |r(x,y)|).$$

Règle basée sur la p-valeur : si $\alpha_{obs} \leq \alpha$, on rejette H_0 au risque d'erreur α. Sinon, on conserve H_0 avec un risque d'erreur β inconnu.

Exemple 1 :

Pour $\alpha = 5\%$, $\alpha_{obs} = 2\, P_{H_0}(R \geq 0,6071) = 0,03632 < \alpha$. On rejette H_0 et on conclut à l'existence d'un lien entre les deux QI, au risque $\alpha = 5\%$. *Le résultat du test est significatif au niveau* 5%.

Calcul réalisé avec Statistica : $t = \sqrt{10} \times \frac{0,6071}{\sqrt{1-0,6071^2}} = 2,4158$, *et p-valeur associée à* t *:* $p = 0,03662$.

(f) *Remarques*

- Pour le test d'une liaison positive, la RC est située à droite du domaine et dans ce cas, on a $\alpha_{obs} = P_{H_0}(R \geq r(x,y))$.
  Exemple 1 modifié : H_1 : les QI sont liés positivement. $n = 12$; $\alpha = 5\,\%$.
  On a $\alpha_{obs} = P_{H_0}(R \geq 0,6071) = 0,01816...$
  La p-valeur est multipliée par 2 pour le test bilatéral (test plus « conservatif »).
- Pour le test d'une liaison négative, la RC est située à gauche du domaine et dans ce cas, on a $\alpha_{obs} = P_{H_0}(R \leq r(x,y))$.

3. *Vérification du postulat*

La binormalité est difficile à vérifier, surtout pour de petits échantillons.

Pour de grands échantillons : le nuage a la forme d'ellipses concentriques.

Pour n assez grand, ici on prendra $n \geq 75$: des résultats approximatifs sur la loi de R sous H_0 permettent d'utiliser le test sans se soucier de la normalité.

Pour n petit : alternativement, on peut utiliser le test de Spearman.

c) Test non paramétrique basé sur le coefficient de corrélation empirique de Spearman

On reprend l'exemple 1.

1. *Postulat requis*

On utilise le test pour des variables X et Y continues (ce qui à priori exclut les ex-aequo). La loi du couple est quelconque.

La présence de nombreux ex-aequo, surtout sur des échantillons petits affecte le résultat du test, voir le point 3. ci-dessous.

2. Calcul de la p-valeur avec Statistica. Dans la sortie détaillée du module `matrice des corrélations`, on trouve $t = 2,4158$, $df = 10$ et la p-valeur $p = 0,03662$ associée à t . Dans le module `Calculateur de probabilités / Corrélations` de Statistica, on peut obtenir la p-valeur p à partir de r.

2. *Déroulement du test*

(a) *Hypothèses du test*

On teste ici l'existence d'une liaison qui n'est pas nécessairement linéaire.

H_0 : les variables sont indépendantes

H_1 : les variables sont liées (positivement, négativement)

Niveau α fixé.

(b) *Statistique du test*

Le test est basé sur le coefficient de corrélation empirique de Spearman, noté R_S, calculé sur les données appariées de l'échantillon. On note $r_S(x, y)$ sa valeur observée sur l'échantillon.

i. *Définition et formule de calcul sur l'échantillon*

Sur l'exemple 1

On ordonne séparément les x_i et les y_i pour transformer les données en rangs. Le coefficient de corrélation de Spearman est le coefficient de corrélation linéaire entre les deux séries appariées de rangs :

- On classe les n mesures x_i par ordre croissant. On attribue le rang 1 à la plus petite valeur et le rang n à la plus grande. On note x'_i le rang de la mesure x_i.
- On classe les n mesures y_i par ordre croissant. On note y'_i le rang de la mesure y_i.
- S'il y a des ex-aequo, on calcule les rangs moyens. Après les deux classements, on dispose de deux séries appariées de rangs.
- Le coefficient de corrélation de Spearman $r_S(x, y)$ est défini par $r_S(x, y) = r(x', y')$.
  Dans le cas où il n'y a pas d'ex-aequo, on dispose de la formule de calcul suivante :

$$r_S(x, y) = 1 - \frac{6 \sum (x'_i - y'_i)^2}{n(n^2 - 1)}.$$

Cette formule reste utilisable quand il y a peu d'ex-aequo.

Calcul de $r_S(x, y)$ pour les données de l'exemple 1 :

$n = 12$

x_i	123	144	105	110	98	138	131	90	119	109	125	100
y_i	102	138	126	133	95	146	115	100	142	105	130	120
rang x'_i	8	12	4	6	2	11	10	1	7	5	9	3
rang y'_i	3	10	7	9	1	12	5	2	11	4	8	6
$x'_i - y'_i$	5	2	-3	-3	1	-1	5	-1	-4	1	1	-3
$(x'_i - y'_i)^2$	25	4	9	9	1	1	25	1	16	1	1	9

$\sum (x'_i - y'_i)^2 = 102$ et $r_S(x, y) = 1 - \frac{6 \times 102}{12(12^2 - 1)} = 0,643$.

ii. *Propriétés*

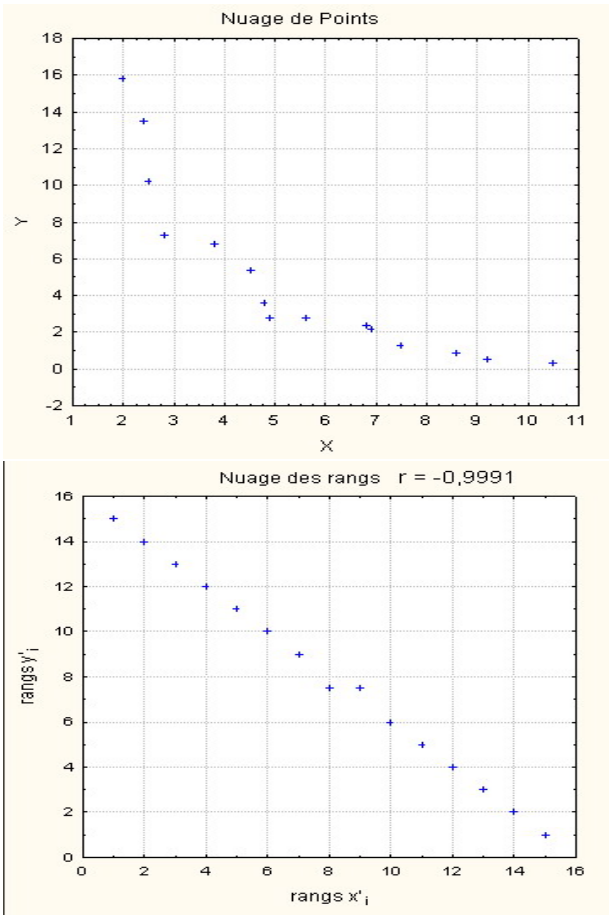
- Le coefficient de Spearman est toujours compris entre -1 et 1.
- Pour deux classements identiques, on a $r_S = +1$.
- Pour deux classements en opposition parfaite, on a $r_S = -1$.
- Lorsqu'il n'y a aucune relation entre les rangs, on a $r_S = 0$.

Pour la corrélation entre les deux variables :

Le coefficient de Spearman n'apporte aucune information sur la forme de la relation entre les variables. Il renseigne sur une relation monotone « croissante » ou « décroissante ».

– Le signe renseigne sur le sens de la corrélation : positive ou négative. Les variables ont tendance à varier dans le même sens ou bien en sens opposés.
Les valeurs ± 1 correspondent à une relation croissante ou décroissante parfaite entre les deux variables.

iii. *Exemple 2.*



graphique 6

Le nuage de points montre une relation décroissante entre X et Y. Il y a deux ex-aequo dans les valeurs y_i. Ces deux valeurs occupent les rangs 7 et 8. On leur attribue le rang moyen $7,5$. La présence des ex-aequo fait que la relation n'est pas parfaitement décroissante. Cela se traduit par des couples de rangs quasiment alignés (aux ex-aequo près) et par un coefficient de Spearman de $-0,9991$.

(c) *Loi de la statistique R_S sous H_0*

R_S prend ses valeurs entre -1 et 1. La loi est symétrique par rapport au centre 0. Elle dépend de n.

Pour n petit, $n \leq 30$, la loi exacte est tabulée.

Pour $n > 30$, on utilise une approximation normale de la loi.[3]

(d) *IA, RC et règle de décision basée sur la p-valeur*

3. *Approximation normale* Pour $n > 30$, sous H_0, on peut faire l'approximation normale : $\sqrt{n-1}R_S \underset{\text{approxt}}{\sim} N(0,1)$. Statistica utilise pour le test la statistique de Student.

- Si H_0 est vraie, les variables sont indépendantes et il n'y a donc aucune relation entre les rangs. On s'attend donc à observer sur l'échantillon un coefficient proche de 0. Les valeurs de R_S les plus proches de 0 sont les plus conformes à H_0.
- Si H_1 est vraie, on s'attend au contraire à observer un coefficient plus proche de ± 1. Les valeurs extrêmes de R_S sont les plus significatives de H_1.
  La RC est aux deux extrémités du domaine de R_S.
- *p-valeur*
  Pour le test d'une liaison, $\alpha_{obs} = 2\,P_{H_0}(R_S \geq |r_s(x,y)|)$.
  Exemple : $n = 12$; $\alpha = 5\,\%$. $\alpha_{obs} = 2\,P_{H_0}(R_S \geq 0,643) = 0,02692$.
  Décision : $\alpha_{obs} < \alpha$.
  On rejette H_0 et on conclut à l'existence d'une liaison entre les deux QI au risque d'erreur $\alpha = 5\,\%$...

(e) *Remarques*

- Si l'on veut tester une liaison positive, la RC est située à droite du domaine et dans ce cas, on a $\alpha_{obs} = P_{H_0}(R_S \geq r_s(x,y))$.
  Exemple 1 modifié : H_1 : les QI sont liés positivement. $n = 12$; $\alpha = 5\,\%$.
  On a $\alpha_{obs} = P_{H_0}(R_S \geq 0,643) = 0,01346$.
- Si l'on veut tester une liaison négative, la RC est située à gauche du domaine et dans ce cas, on a $\alpha_{obs} = P_{H_0}(R_S \leq r_s(x,y))$.

3. Problème des ex-aequo

   En présence d'ex-aequo, on calcule les rangs moyens. Il est préférable de calculer le coefficient avec la formule du coefficient de corrélation linéaire.

   Pour $n < 30$, on considère dans ce cours que l'on peut appliquer le test s'il y a très peu d'ex-aequo (2 ou 3).

   Pour $n \geq 30$, on peut l'appliquer sans problème (calcul approximatif de la p-valeur).

III Relation entre une variable qualitative et une variable quantitative

Les deux variables ne jouent pas un rôle symétrique. On veut étudier l'influence d'une variable qualitative (par exemple le sexe) sur une variable quantitative (par exemple le salaire).
L'étude est faite dans le cadre d'une **anova** : analyse de la variance à un facteur.

IV Liaison entre deux variables qualitatives

Cette dernière partie n'est pas au programme de ce cours. Le test du khi-deux d'indépendance qui est rappelé ci-dessous a été traité dans le cours de statistique de 3$^{\text{e}}$ année de licence.
Exemple 3 :
On utilise les données relevées sur un échantillon de 50 enfants de 2 à 16 ans souffrant d'un TSPT (*trouble de stress post-traumatique*) à la suite d'un accident domestique ou de circulation.
$\mathcal{P}$: enfants de 2 à 16 ans souffrant d'un TSPT à la suite d'un accident domestique ou de circulation.
Variable X : `sexe`, qualitative à $L = 2$ modalités (A_i) .
Variable Y : `type d'accident`, qualitative à $C = 2$ modalités (B_j).
On veut étudier la relation entre le sexe et le type d'accident survenu.
Pour les 50 enfants de l'échantillon, on a relevé le sexe et le type d'accident survenu.

1) Etude descriptive des données

La relation est étudiée à partir du tableau de contingence construit à partir des données. Elle est mesurée par différents coefficients d'association.

a) Tableau de contingence

On croise les deux variables et pour chaque couple de modalités (A_i, B_j), on relève l'effectif observé n_{ij} : nombre d'individus prenant simultanément les deux valeurs.
Les LC effectifs observés sont données dans un tableau de contingence.

Tableau 1. Effectifs observés n_{ij}

Sexe \ Accident	circulation	domestique	Totaux lignes L_i
fille	10	9	19
garçon	19	12	31
Totaux colonne C_j	29	21	$n = 50$

On a calculé les marges du tableau : totaux lignes L_i et totaux colonne C_j.

b) Tableau des effectifs théoriques

A partir des effectifs marginaux L_i et C_j, on peut calculer les effectifs attendus lorsque X et Y sont indépendantes.
Ces effectifs théoriques, notés e_{ij}, sont donnés par la formule

$$e_{ij} = \frac{L_i\, C_j}{n}.$$

Tableau 2. Effectifs théoriques e_{ij}

Sexe \Accident	circulation	domestique	Totaux lignes L_i
fille	$\frac{19 \times 29}{50} = 11,02$	$\frac{19 \times 21}{50} = 7,98$	19
garçon	$\frac{31 \times 29}{50} = 17,98$	$\frac{31 \times 21}{50} = 13,02$	31
Totaux colonne C_j	29	21	50

Ex. Les filles représentent $L_1/n = 19/50 = 38\%$ *des enfants de l'échantillon. S'il n'y pas de lien entre les variables, on s'attend à trouver* 38% *de filles chez les enfants ayant subi un accident domestique (soit* $29 \times 19/50 = 11,02$ *filles) et* 38% *de filles chez les enfants ayant subi un accident domestique (soit* $21 \times 19/50 = 7,58$ *filles).*

c) Coefficients d'association

La liste n'est pas exhaustive. Le coefficient le plus important est celui du khi-deux qui est utilisé pour tester l'indépendance des deux variables à partir d'un échantillon d'individus.

– *Coefficient du khi-deux* q^2
  Le coefficient est une mesure de « l'écart à la situation d'indépendance ».
  Il mesure la distance globale entre les effectifs n_{ij} relevés et les effectifs théoriques e_{ij} attendus lorsque X et Y sont indépendantes :

$$q^2 = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 0,363.$$

- le coefficient est positif ou nul ; la valeur 0 correspond à l'indépendance des variables.
  La valeur du coefficient augmente avec l'intensité de la relation mais aussi avec n et avec L et C, ce qui le rend difficilement interprétable.
- *Coefficient Phi-deux* $\Phi^2 = \dfrac{\chi^2}{n}$ ou Phi $\Phi = \sqrt{\dfrac{\chi^2}{n}}$.
  On élimine l'effet de la taille n mais la valeur augmente encore avec L et C. On l'utilise surtout pour des tables 2x2 car alors il est compris entre 0 et 1.
- *Coefficient de Cramer*
  $V = \sqrt{\dfrac{\Phi^2}{d-1}}$ où $d = \inf(L, C)$. Le coefficient est compris entre 0 et 1.

2) Test d'indépendance pour deux variables qualitatives

On reprend l'exemple 3.
Pour tester l'existence d'une liaison entre le sexe et le type d'accident survenu, on utilise le test du khi-deux d'indépendance.

a) Les différentes étapes du test

1. *Les hypothèses et le niveau du test*
   H_0 : les variables sont indépendantes
   H_1 : les variables sont liées
   $\alpha = 5\%$
2. *Les observations*
   On dispose d'un échantillon de taille $n = 50$. Les 2 échantillons appariés de mesures sont résumés par la distribution des effectifs joints observés n_{ij} donnée dans le tableau 1.
3. *La statistique du test : statistique du khi-deux*
   On utilise le coefficient empirique du khi-deux, noté Q^2 dans le cours et défini par la formule
   $$Q^2 = \sum \frac{(N_{ij} - e_{ij})^2}{e_{ij}}.$$
   - Les valeurs de Q^2 sont positives ou nulles. Plus la valeur est grande et plus l'écart à l'indépendance observé sur l'échantillon est important.
   - Sa valeur observée est $q^2_{obs} = 0,363$.
4. *loi de Q^2 sous H_0*
   Sous les conditions $n \geq 30$ et tous les $e_{ij} \geq 5$, la statistique Q^2 suit approximativement la loi du khi-deux à $(L-1)(C-1) = 1$ ddl.
5. *Région de rejet de H_0 associée à $\alpha = 5\%$*
   Sous H_0, on s'attend à observer une valeur de Q^2 proche de 0. Plus la valeur de Q^2 est grande et plus elle est en faveur de H_1.
   La région de rejet est située à l'extrémité droite du domaine. Elle contient les 5% de valeurs les plus grandes de Q^2.
6. *p-valeur α_{obs}*
   C'est la probabilité sous H_0 d'observer une valeur de Q^2 au moins aussi grande que 0,363 :
   $$\alpha_{obs} = P_{H_0}(Q^2 \geq 0,363) = 0,547.$$

7. *Décision*

   Règle basée sur la p-valeur : si $\alpha_{obs} \leq \alpha$, on rejette H_0 au risque d'erreur α.

   La p-valeur dépasse (largement) le niveau 5% choisi. On ne peut pas rejeter H_0. Au niveau 5% et avec un risque d'erreur β inconnu, on ne peut pas conclure qu'il existe un lien entre le sexe et le type d'accident survenu.