Analyse de données L3 2024-2025

Cours n°4- La corrélation

Florian Bayer

Rappels et objectifs du cours

Lors du précédent cours, nous avons vu comment décrire une série de données.

L'analyse de données commence cependant à prendre tout sons sens lorsque l'on regarde comment se comporte une série par rapport à une autre, voir plusieurs.

Pour mesurer **l'intensité de la relation** entre deux caractères quantitatifs continus, on utilise le **coefficient de corrélation**.

Il est complémentaire à la **régression linéaire** et à la régression multiple qui visent à résumer et/ou **modéliser** un phénomène par une ou plusieurs variables.

En géographie, **identifier** puis **modéliser** des **relations** permet de comprendre un phénomène sur un espace donné, de prévoir la survenue de ce phénomène ou encore de déterminer les variables qui manquent à notre explication.

1- La corrélation

Relation et dépendance

Une relation entre deux caractères quantitatifs x et y peut-être mesurée si l'attribution des valeurs de y *dépendent* des valeurs de x ou inversement.

Par exemple, lorsque x augmente de 1, y augmente aussi de 1. Autrement dit, il y a une **relation** si les valeurs de x ne sont font pas au hasard par rapport au valeurs de y.

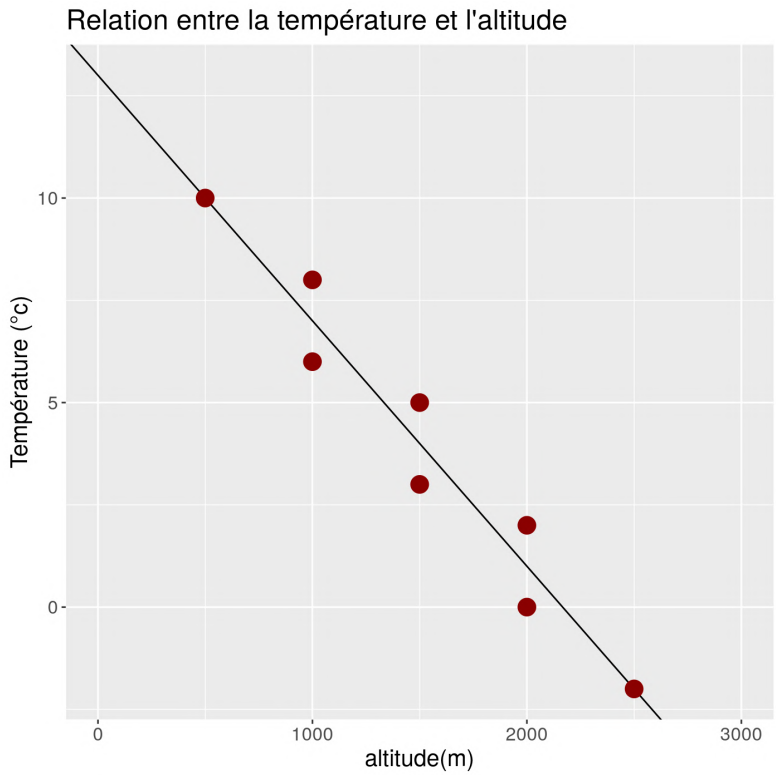
Si y dépend de x, on peut prédire avec une certaine marge d'erreur les valeurs de y en connaissant les valeurs de x à l'aide d'une fonction $y = f(x)$:

Exemple :

Il existe une relation entre la température et l'altitude, exprimée par l'équation : $T_a = -0.006a + T_0$

- T_a : température à l'altitude a
- a : altitude en mètre
- T_0 : température au niveau de la mer.

Tous les **1 m**, la température **baisse de 0,006 °C** (0.6 °C tous les 100 m)



Relation et dépendance

La notation des variables est importante. Si vous écrivez $y = f(x)$, vous postulez:

- que y est la variable à **expliquer**. On parle de variable **dépendante**.
- que x est la variable **explicative**. On parle de variable **indépendante**.

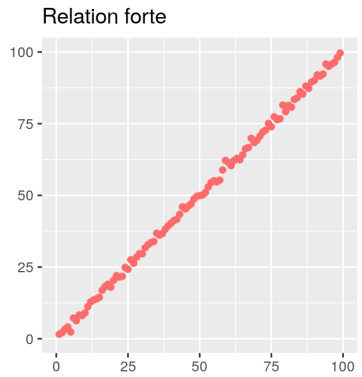
S'il existe une relation, les valeurs de x permettront de prédire les valeurs de y alors que la réciproque n'est pas toujours vrai. Il faut donc être rigoureux et précis lors de l'énoncé de votre hypothèse et **réfléchir au sens de la dépendance**.

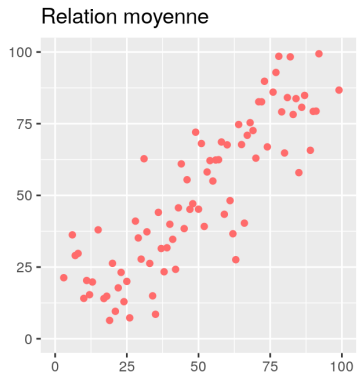
Afin de mesurer cette éventuelle relation, il est nécessaire :

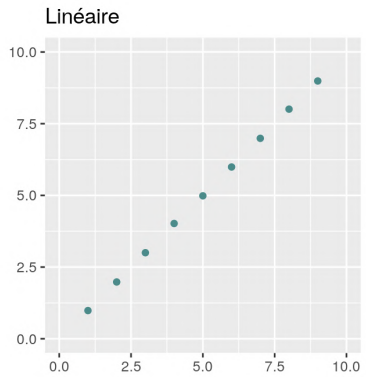
- de la visualiser sa **forme** à l'aide d'un graphique : le **diagramme de corrélation**.
- de mesurer son **intensité** et son **signe**, à l'aide d'un **coefficient de corrélation**, qu'il faudra ensuite tester significativement.
- dans certains cas, de **modéliser** la relation à l'aide d'une droite d'équation : la **régression linéaire**.

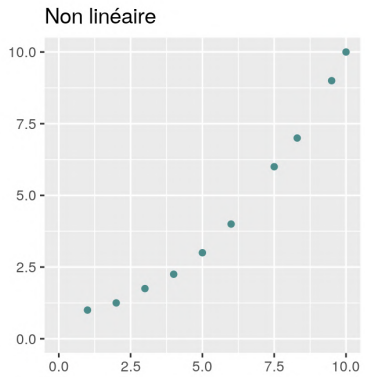
Interprétations graphiques

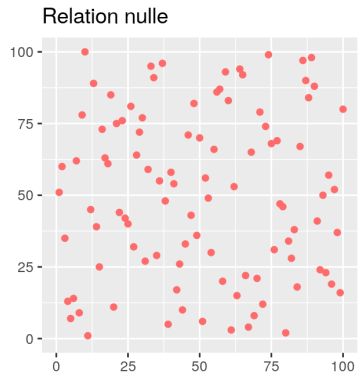
En croisant les valeurs de x et de y sur un graphique, on forme un nuage de points dont la forme permet de caractériser la relation à via son intensité, sa forme et son signe.

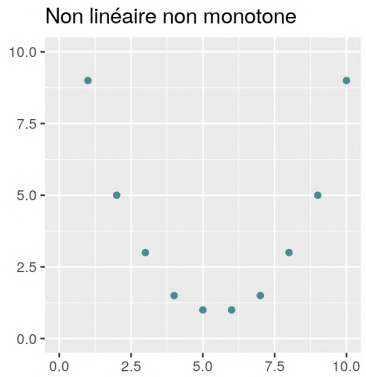




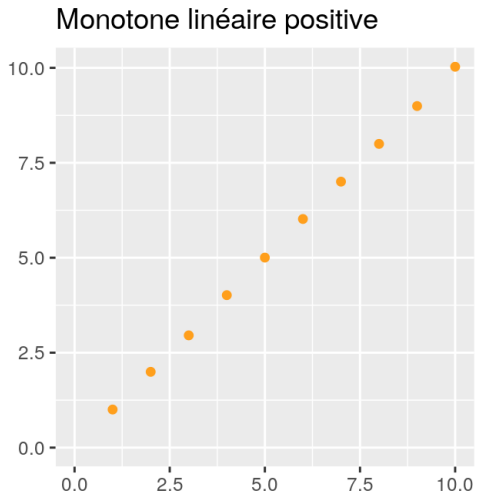


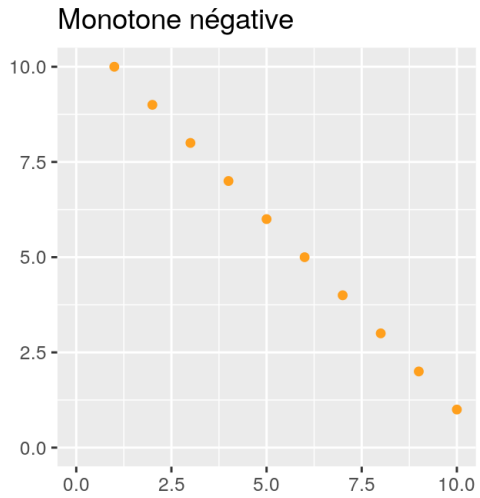


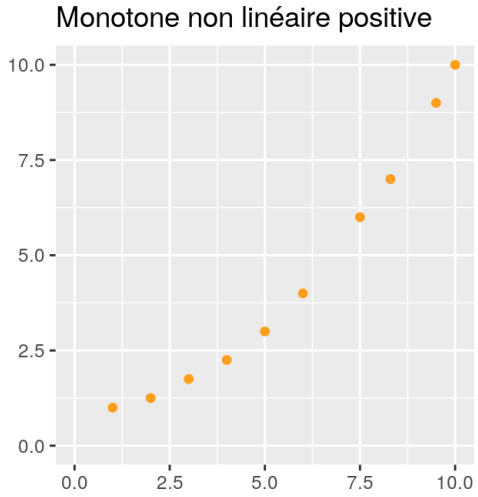


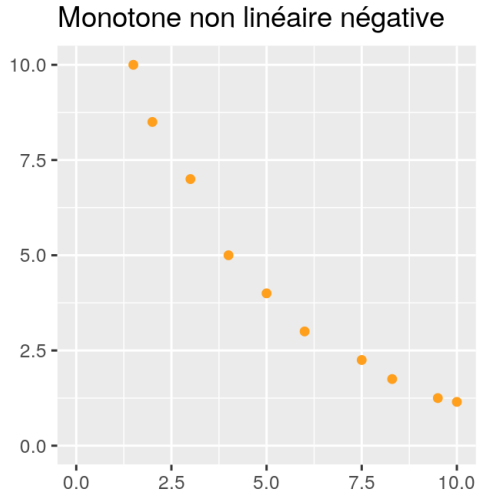


Interprétations graphiques









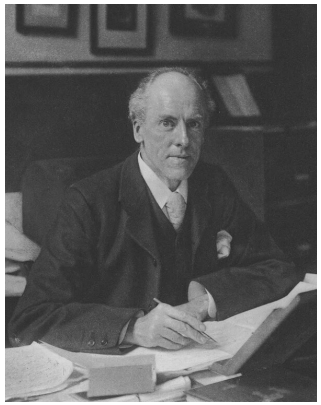
Relation et dépendance

Une fois que la relation entre x et y est entrevue graphiquement, il est possible de mesurer **l'intensité** de la relation à l'aide du **coefficient de corrélation** noté R.

Outre l'intensité d'une relation monotone, il renseigne également sur son **signe.**

Il existe plusieurs coefficients de corrélation. Les plus utilisées sont:

- Le coefficient de corrélation de **Pearson** qui permet d'analyser les **relations linéaires**. Il est en lien avec la **régression linéaire**.
- Le coefficient de corrélation de **Spearman** qui permet d'analyser les **relations non-linéaires monotones**. Il est aussi appelé coefficient de corrélation de rang.



2- Le coefficient de corrélation linéaire de Bravais-Pearson

La covariance

Le coefficient de corrélation linéaire de Bravais-Pearson permet de détecter la présence ou l'absence d'une **relation monotone linéaire** entre deux caractères quantitatifs continus.

Il est mal adapté aux relations non-linéaires.

Pour calculer ce coefficient il faut tout d'abord calculer la **covariance** : une mesure de la liaison linéaire entre deux variables quantitatives.

- Une covariance proche de zéro correspond à l'indépendance (absence de relation).
- Une covariance négative indique une relation inverse.
- Une covariance positive indique une relation de X et Y dans le même sens.

La covariance est égale à la moyenne du produit des valeurs de deux variables moins le produit des deux moyennes Cov(X,Y) = moyenne(x . y) – [moyenne(x) . moyenne(y)]

$$cov_{x,y} = \frac{\sum_{i=1}^{N} x_i . y_i}{N} - (\bar{x} . \bar{y}) \text{ ou } cov_{x,y} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

R de Bravais-Pearson

La covariance est un bon indicateur de mesure de relation, mais n'est pas standardisée, ce qui ne permet pas de comparer facilement deux covariances.

On utilise donc le **coefficient de corrélation linéaire** de deux caractères x et y qui est égal à la covariance de x et y divisée par le produit des écarts-types σ de x et y. Pour des raisons qui ne seront pas détaillées ici, l'écart-type utilisé est celui utilisé pour une population (fonction ecartypep sous Excel).

$$R_{x,y} = \frac{cov_{x,y}}{\sigma_x.\sigma_y}$$

Le coefficient de corrélation est noté R. Comme il est **standardisé**, il varie entre **-1 et +1**

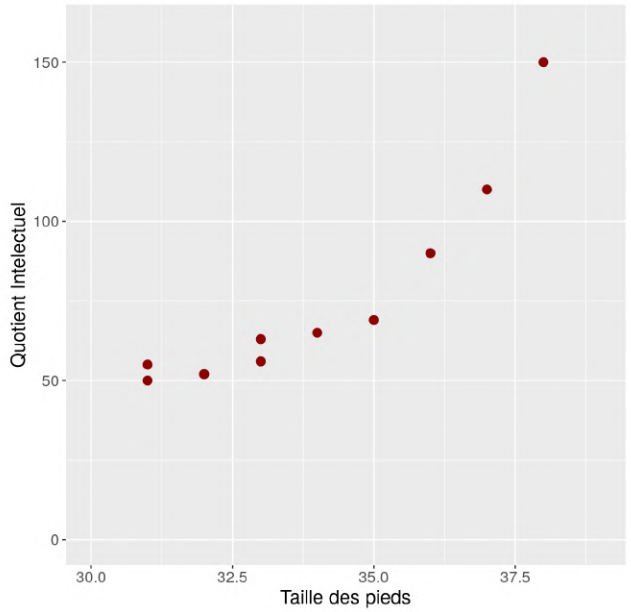
- si R est proche de -1, il existe une **forte relation linéaire négative** entre x et y
- si R est proche de 0, il n'y a pas de relation linéaire entre x et y
- si R est proche de $+1$, il existe une **forte relation linéaire positive** entre x et y, sa **forme**

Le **signe** de R indique le **sens** de la relation, sa **valeur absolue l'intensité** de la relation.

R de Bravais-Pearson : exemple

On propose d'examiner s'il existe une relation entre la capacité à épeler, mesurer par le QI y et la taille des pieds x de 10 enfants.

Enfant	xi	yi
A	31.00	50.00
B	31.00	55.00
C	32.00	52.00
D	33.00	56.00
E	33.00	63.00
F	34.00	65.00
G	35.00	69.00
H	36.00	90.00
I	37.00	110.00
J	38.00	150.00
Moyenne	**34.00**	**76.00**
Ect	**2.32**	**30.43**



Le nuage de point montre une **relation monotone positive** qui semble **non linéaire**. On décide tout de même de calculer le R de Bravais-Pearson.

R de Bravais-Pearson : exemple

Enfant	xi	yi	xi*yi
A	31	50	31*50=1 550
B	31	55	1 705
C	32	52	1 664
D	33	56	1 848
E	33	63	2 079
F	34	65	2 210
G	35	69	2 415
H	36	90	3 240
I	37	110	4 070
J	38	150	5 700
Moyenne	**34**	**76**	**2 648,1**
Ect	**2,32**	**30,43**	

La covariance est égale à la moyenne du produit des valeurs de deux variables moins le produit des deux moyennes :

$$cov_{x,y} = 2648,1 - (34 \times 76) = 64,1$$

La covariance de x et y est donc égal à **64,1**

On obtient le coefficient de corrélation de Bravais Pearson entre x et de y en divisant la covariance par le produit de l'écart-type de x et de l'écart-type de y :

$$R = \frac{64,1}{2,32.30,43} = +0,9$$

R de Bravais-Pearson : exemple

Avec $R = +0,90$, la corrélation est **positive et forte**. Cela semble indiquer qu'il existe une relation reliant le quotient intellectuel des enfants et la taille de leurs pieds.

Toutefois, le coefficient de corrélation ne nous indique pas :

- si la relation observée est **significative** (fruit du hasard ou non).
- si elle correspond à une **relation de cause à effet** entre les deux facteurs x et y étudiés.

De plus, le nuage de point observé ne montre pas un ajustement parfait des points sur une droite, mais plutôt **sur une courbe**.

On peut donc calculer le **R de Spearman** pour mesurer un éventuel meilleur ajustement **non-linéaire**.

R de Bravais-Pearson : limites

En principe, le coefficient de Bravais-Pearson ne peut s'appliquer que

- pour des distributions **gaussiennes**.
- sans valeurs **exceptionnelles** min ou max (outliers).

Il arrive très souvent que ces conditions ne soient pas vérifiées. Elles conduisent alors à des interprétations faussées. C'est pourtant le coefficient le plus largement répandu.

De plus, ne pas montrer une relation linéaire ne signifie pas l'absence d'une autre relation. Dans l'exemple précédent, le coefficient de corrélation de Bravais-Pearson indiquait un bon ajustement, alors que le nuage de point montre que la relation n'est sans doute pas linéaire.

3- Le coefficient de corrélation de rang de Spearman

Le coefficient de corrélation de rang de Spearman

Le coefficient de corrélation de **Spearman** ne se base pas sur les valeurs des individus x_i et y_i mais sur leur **rang** $r(x_i)$ et $r(y_i)$.

Il permet de déterminer l'existence d'une relation entre le rang des observations pour deux caractères x et y. Cette propriété permet de démontrer l'existence de **relations monotones linéaires ou non**.

On peut donc l'utiliser pour des distributions **non gaussiennes** ou sur des données avec des **valeurs extrêmes**.

En contrepartie, il est plus difficile à calculer manuellement, il est moins efficaces sur des rangs ex-æquo et il n'intervient pas dans la modélisation par régression linéaire.

$$R = 1 - \frac{6 \sum [r(x_i) - r(y_i)]^2}{N^3 - N}$$

Avec $r(x_i)$ et $r(y_i)$ le rang de x et y dans la distribution et N le nombre d'individus

R de Spearman : exemple

Enfant	xi	yi	Rxi	Ryi	Rxi-Ryi	(Rxi-Ryi)²
A	31	50	1.5	1	0.5	0.25
B	31	55	1.5	3	-1.5	2.25
C	32	52	3.0	2	1.0	1.00
D	33	56	4.5	4	0.5	0.25
E	33	63	4.5	5	-0.5	0.25
F	34	65	6.0	6	0.0	0.00
G	35	69	7.0	7	0.0	0.00
H	36	90	8.0	8	0.0	0.00
I	37	110	9.0	9	0.0	0.00
J	38	150	10.0	10	0.0	0.00
Somme	**340**	**760**	**55.0**	**55**	**0.0**	**4.00**

Pour les rangs $r(x_i)$ et $r(y_i)$ ex-æquo, on calcule la moyenne ou la médiane.

- La somme du carré des différences de rang étant égale à **+4**
- le nombre d'individus étudiés est égal à **10**

On en déduit la valeur du coefficient de corrélation de Spearman :

$$R = 1 - \frac{6.4}{10^3 - 10} = +0,98$$

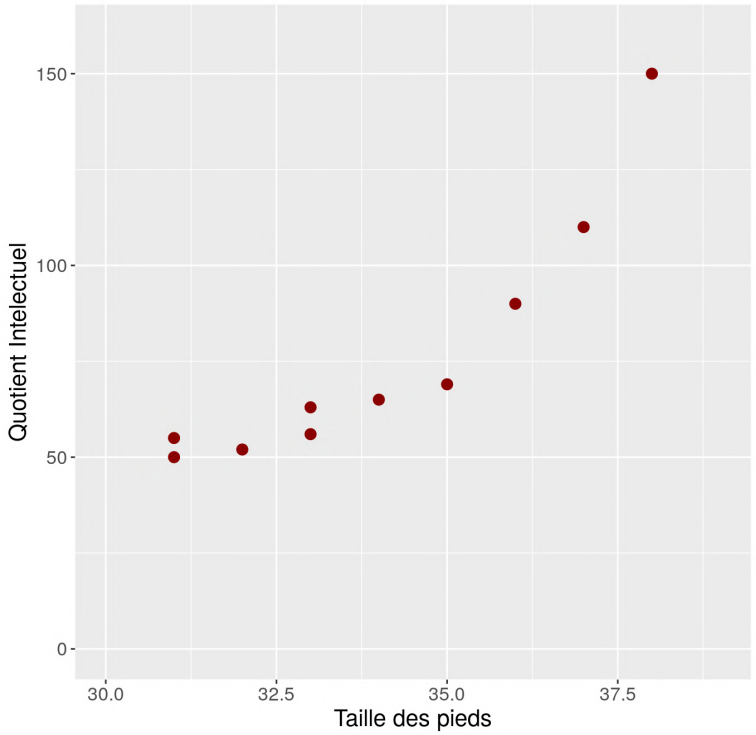
R de Spearman : exemple

La relation mise en évidence avec le coefficient de corrélation de Bravais-Pearson est **confirmée** avec le R de Spearman.

Elle est cependant **plus forte avec ce dernier**, ce qui peut laisser supposer une **relation non-linéaire** entre x et y.

Le nuage de points confirme cette hypothèse. Cependant, le calcul du coefficient de corrélation n'est pas suffisant.

Comme pour le $\chi 2$, il faut **tester** la relation afin de déterminer si elle est liée au hasard ou non.



Test de la significativité de la corrélation

Le test se déroule de la même manière que celui du $\chi 2$

- On pose H0 : il n'y a pas de relation entre les deux caractères x et y
- On fixe un risque d'erreur α pour le rejet de H0 (5%)
- On calcul le degré de liberté z.
  - Pour Bravais-Pearson : le nombre de couples Xi,Yi - le nombre de variables explicatives – 1 (sur 10 individus : $z = 10 - 1 - 1 = 8$)
  - Pour Spearman : le nombre de couples (z=10)
- On calcule la valeur absolue du coefficient de corrélation R(X,Y) dans la table correspondante (Pearson ou Spearman)
- On calcule la valeur théorique R(α, z) de ce coefficient
- Si R théorique > R calculé, l'hypothèse H0 ne peut pas être rejetée.

Si R théorique < R calculé, l'hypothèse H0 est rejetée au risque α

Test du R de Bravais-Pearson

Dans notre exemple, il y a 10 individus :

- $z = 10 - 1 - 1 = 8$
- On choisi un risque α de rejeter H0 à tort de 5%
- La valeur du R de Bravais-Pearson pour $z = 8$ et $\alpha = 0.05$ est de 0,6319
- R théorique (0,63) < R calculé (0,90)
- On peut rejeter H0 et accepter H1 avec un risque de 5% de rejeter H0 à tort.
- Avec un risque de 2%, la relation est toujours significative (R théorique = 0,7155)

N.B. : la plupart des logiciels de statistiques donnent la p-value

z / α	0.10	0.05	0.02	z / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	16	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	17	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	18	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	19	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	20	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	25	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	30	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	35	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	40	0.2573	0.3044	0.3578
10	0.4973	0.5750	0.6581	45	0.2428	0.2875	0.3384
11	0.4762	0.5529	0.6339	50	0.2306	0.2732	0.3218
12	0.4575	0.5324	0.6120	60	0.2108	0.2500	0.2948
13	0.4409	0.5139	0.5923	70	0.1954	0.2319	0.2737
14	0.4259	0.4973	0.5742	80	0.1829	0.2172	0.2565
15	0.4124	0.4821	0.5577	90	0.1726	0.2050	0.2422
				100	0.1638	0.1946	0.2301

Test du R de Spearman

Dans notre exemple, il y a 10 individus :

- $z = 10$
- On choisi un risque α de rejeter H0 à tort de 5%
- La valeur du R de Bravais-Pearson pour $z = 10$ et $\alpha = 0.05$ est de 0,56
- R théorique (0,56) < R calculé (0,98)
- On peut rejeter H0 et accepter H1 avec un risque de 5% de rejeter H0 à tort.
- Avec un risque de 1%, la relation est toujours significative (R théorique = 0,75)

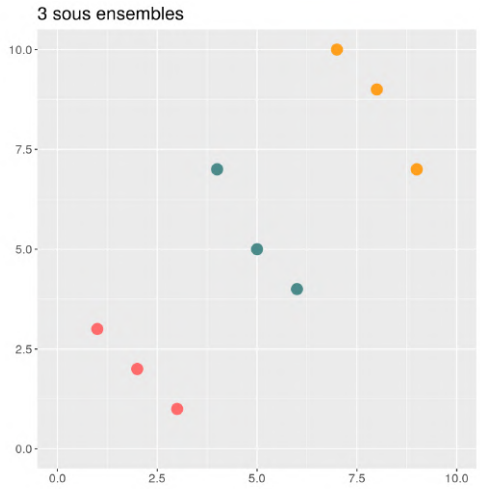
z / α	0.05	0.01	z / α	0.05	0.01
4	1.00	-	24	0.34	0.49
5	0.90	1.00	26	0.33	0.47
6	0.83	0.94	28	0.32	0.45
7	0.71	0.89	30	0.31	0.43
8	0.64	0.83	35	0.28	0.40
9	0.60	0.78	40	0.26	0.37
10	0.56	0.75	45	0.25	0.35
12	0.51	0.71	50	0.24	0.33
14	0.46	0.64	55	0.22	0.32
16	0.42	0.60	60	0.21	0.30
18	0.40	0.56	70	0.20	0.28
20	0.38	0.53	80	0.19	0.26
22	0.36	0.51	100	0.17	0.23

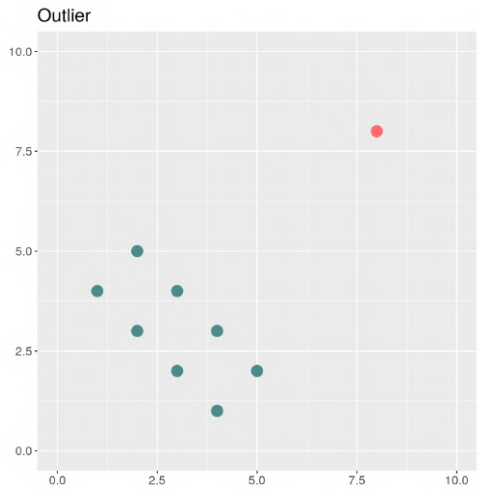
Vérification des résultats

Vérifiez **toujours** la forme du nuage de points. Des sous ensembles ou des outliers peuvent radicalement changer les résultats.

Vérifiez si les deux R significatifs sont proches :

- Si R(Pearson) > R(Spearman) = présence de valeurs exceptionnelles ?
- Si R(Spearman) > R(Pearson) = non-linéarité ?





Conclusions

Concepts clés

La corrélation est uns outil trè puissant permettant la mise en relation des caractères quantitatifs. Il est néanmoins nécessaire :

- De prendre du recul par rapport aux données étudiées (Qu'elles sont les données les plus susceptibles d'expliquer un phénomène observé)
- De faire attention aux éventuels biais de confusion (le % de consommateurs de café est lié au % de cancer des bronches, mais parce que le % de consommateurs de café est aussi lié au % de fumeurs)
- De poser les hypothèses adéquats avant de lancer vos analyses
- De vérifier les prérequis à l'utilisation du coefficient de corrélation
- De vérifier la forme du nuage de points

Concepts clés

Le prochain cours s'intéressera à un autre aspect de la mise en relation des caractères quantitatif : la modélisation

Nous verrons comment expliquer une variable par une autre dans le cadre de la régression linéaire :

- Peut-on expliquer le taux d'abstention par l'âge ?

et plus généralement comment expliquer une variable par plusieurs autres avec la régression multiple :

- Peut-on expliquer le taux d'abstention par l'âge, le niveau de scolarisation, le niveau de revenu ?