Analyse bivariée (partie 1)

Vincent Audigier
vincent.audigier@lecnam.net

CNAM, Paris

STA101

Plan

Rappels

Introduction

Coefficient de corrélation linéaire

Caractère significatif

Corrélation sur les rangs

Causalité

Terminologie

- Population : groupe d'individus soumis à une étude
- Individu (statistique) : élément issu de la population
- Echantillon : partie d'une population

Pour chaque individu, on observe un ensemble de caractères $X_1, X_2, \ldots, X_j, \ldots, X_p$ appelés **variables**

La valeur de la j variable observée sur le i-ème individu est notée x_{ij}

Typologie des variables

- variable qualitative : variable à valeurs non-numériques (où la moyenne n'a pas de sens). Ses valeurs sont appelées **modalités**
  - nominale (ou catégorielle) : absence d'ordre entre les modalités
  - ordinale : existence d'un ordre total
- variable quantitative : variable à valeurs numériques (où la moyenne a un sens)
  - continue : à valeurs dans un intervalle réel
  - discrète : dans le cas contraire

Données *ozone*

- Données climatiques et de pollution à l'ozone mesurées durant l'été 2001 à Rennes (112 individus)

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
87	15.60	18.50	18.40	4	4	8	0.69	-1.71	-0.69	84	Nord	Sec
	17.00	18.40	17.70	5	5	7	-4.33	-4.00	-3.00	87	Nord	Sec
92	15.30	17.60	19.50	2	5	4	2.95	1.88	0.52	82	Est	
114	16.20	19.70	22.50	1		0	0.98			92		Sec
94	17.40	20.50	20.40	8	8	7	-0.50	-2.95	-4.33	114	Ouest	Sec
80	17.70	19.80	18.30	6	6	7	-5.64	-5.00	-6.00		Ouest	Pluie
...	...	...	...	...	...	...	...	...	...	...	...	

- maxO3, maxO3v : maximum d'ozone journalier et maximum de la veille
- T9, T12, T15 : température à 9h, 12h, 15h
- Ne9, Ne12, Ne15 : nébulosité à 9h, 12h, 15h
- Vx9 , Vx12, Vx15 : force du vent à 9h, 12h, 15h
- vent : direction du vent
- pluie : présence de pluie

Analyse univariée, bivariée, multivariée

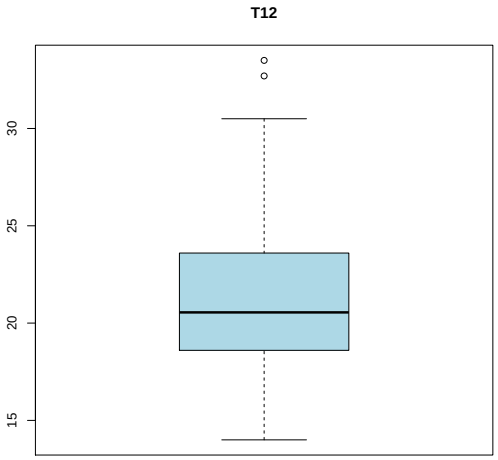
- Analyse univariée : la description porte sur chacune des variables
- Analyse bivariée : la description porte sur des couples de variables
- Analyse multivariée : la description porte sur l'ensemble des variables du jeu de données

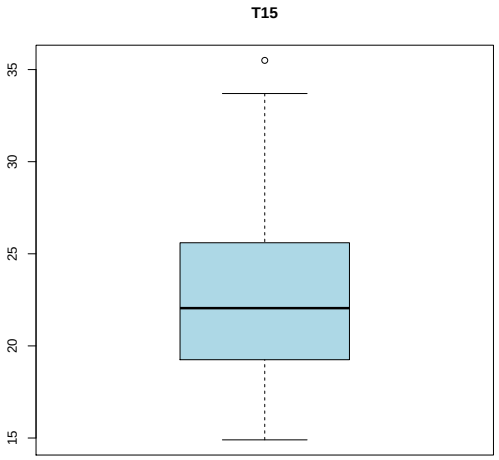
Analyse univariée

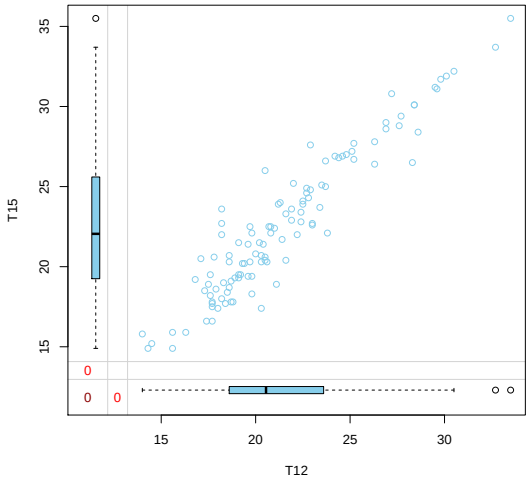
- L'analyse univariée d'une variable s'effectue différemment selon que celle-ci soit de nature qualitative (ordonnée ou non) ou quantitative (discrète ou continue)
- Ce type d'analyse est indispensable pour avoir une première idée de la distribution des variables, ainsi que pour identifier de potentielles "anomalies" dans les données
- Elle s'effectue par la présentation de tableaux ou de graphiques spécifiques

Limites de l'analyse univariée

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
20010601	87	15.60	18.50	18.40	4	4	8	0.69	-1.71	-0.69	84	Nord	Sec
20010602		17.00	18.40	17.70	5	5	7	-4.33	-4.00	-3.00	87	Nord	Sec
20010603	92	15.30	17.60	19.50	2	5	4	2.95	1.88	0.52	82	Est	
20010604	114	16.20	19.70	22.50	1		0	0.98			92		Sec
20010605	94	17.40	20.50	20.40	8	8	7	-0.50	-2.95	-4.33	114	Ouest	Sec
20010606	80	17.70	19.80	18.30	6	6	7	-5.64	-5.00	-6.00		Ouest	Pluie
...	...	...	...	...	...	...	...	...	...	...	...	...	







Aucune information sur le lien entre les deux variables.

$\Rightarrow$ **Analyse bivariée** : résumer le lien entre les deux variables

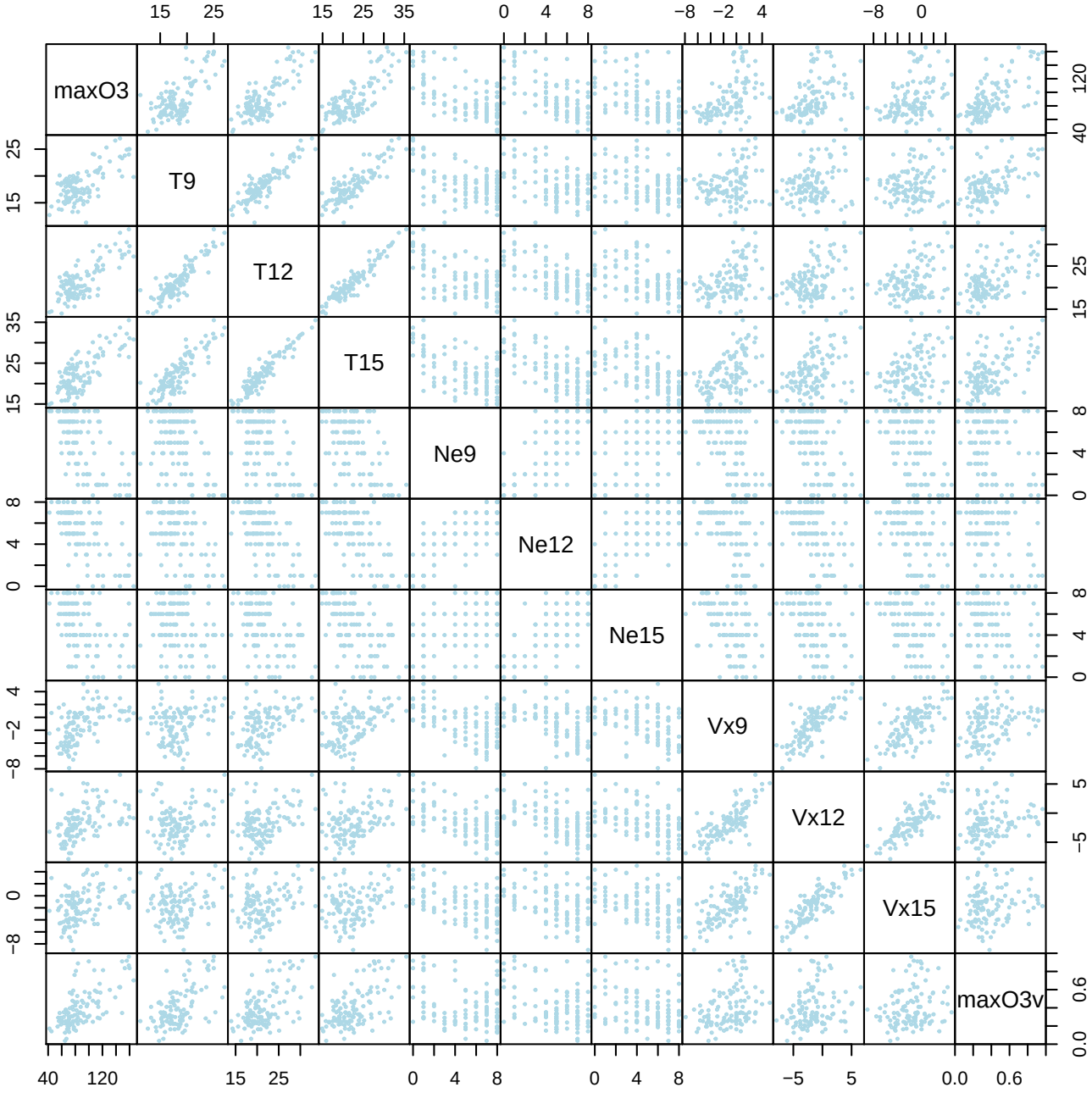
Analyse bivariée

- Soit un échantillon de n individus avec deux mesures chacun : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- L'analyse de la liaison entre deux variables sera fonction la nature de X et Y
  - cas X et Y quantitatives
  - cas X quantitative et Y qualitative
  - cas X et Y qualitatives
- **Question:** Peut-on quantifier le lien entre X et Y ? Cette liaison est-elle significative ?

Analyse bivariée

- Soit un échantillon de n individus avec deux mesures chacun : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- L'analyse de la liaison entre deux variables sera fonction la nature de X et Y
  - cas X et Y quantitatives
  - cas X quantitative et Y qualitative
  - cas X et Y qualitatives
- **Question:** Peut-on quantifier le lien entre X et Y ? Cette liaison est-elle significative ?

Exemples



Autres exemples

- Pourcentage de masse grasse et âge
- Note donnée à un jeu vidéo et nombre de ventes en France
- Nombre de personnes vaccinées et nombre de malades
- Quantité d'alcool consommée et espérance de vie
- Nombre d'heures de travail et note en première session
- etc

Deux cas de figure

- X et Y sont deux variables aléatoires
  Elles ont un rôle symétrique, on ne cherche a priori pas à expliquer/prédire l'une par l'autre.

  $\Rightarrow$ corrélation

- Y est aléatoire, mais X est fixe
  Elles ont un rôle asymétrique, on pense pouvoir expliquer/prédire Y à partir de X.

  $\Rightarrow$ régression

Plan

Rappels

Introduction

Coefficient de corrélation linéaire

Caractère significatif

Corrélation sur les rangs

Causalité

Coefficient de corrélation linéaire

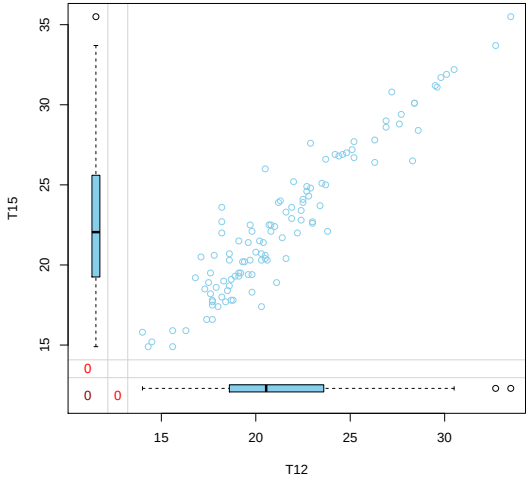
$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\times\frac{(y_i-\bar{y})}{s_y}$$

Propriétés

- $-1 \leq r \leq 1$
- $|r| = 1$ traduit un lien linéaire parfait entre X et Y
- $r = 0$ traduit une absence de lien linéaire
- si X et Y normales, alors $r = 0$ traduit l'indépendance
- symétrique
- non-transitif (ex : X et Y non liées et $Z = X + Y$ X$\sim$Z, Z$\sim$Y mais X$\nsim$Y)

Coefficient de corrélation linéaire

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

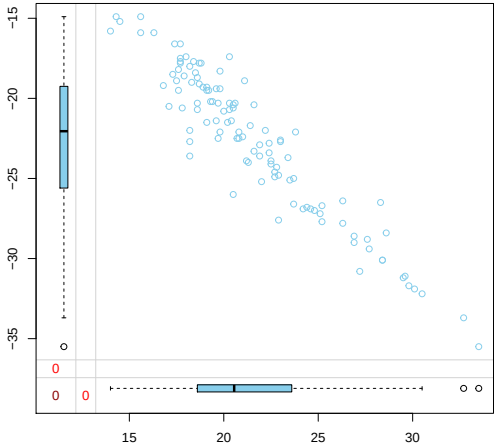


Propriétés

- $-1 \leq r \leq 1$
- $|r| = 1$ traduit un lien linéaire parfait entre X et Y
- $r = 0$ traduit une absence de lien linéaire
- si X et Y normales, alors $r = 0$ traduit l'indépendance
- symétrique
- non-transitif (ex : X et Y non liées et $Z = X + Y$ X$\sim$Z, Z$\sim$Y mais X$\nsim$Y)

Coefficient de corrélation linéaire

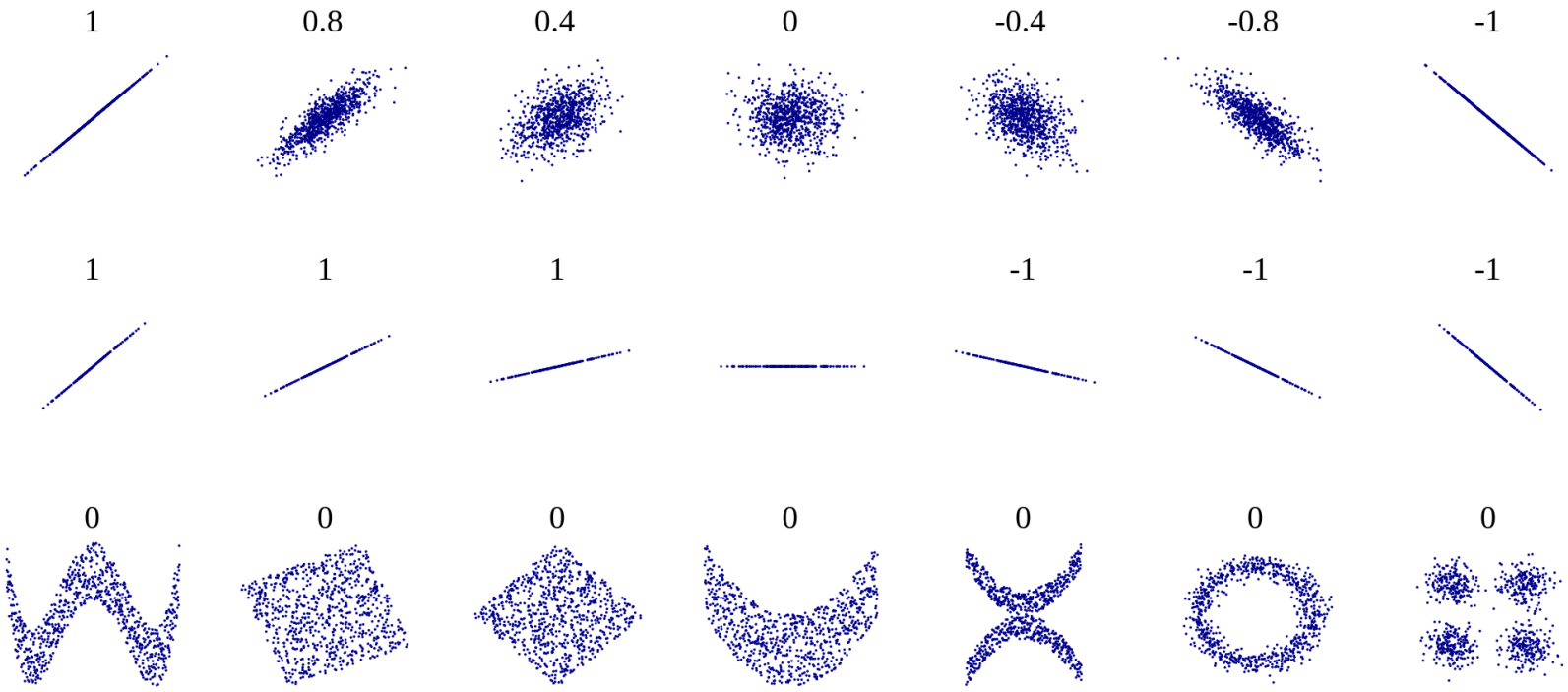
$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\times\frac{(y_i-\bar{y})}{s_y}$$



Propriétés

- $-1 \leq r \leq 1$
- $|r| = 1$ traduit un lien linéaire parfait entre X et Y
- $r = 0$ traduit une absence de lien linéaire
- si X et Y normales, alors $r = 0$ traduit l'indépendance
- symétrique
- non-transitif (ex : X et Y non liées et $Z = X + Y$ X$\sim$Z, Z$\sim$Y mais X$\nsim$Y)

Exemples de coefficients de corrélation



Attention ! Un coefficient de correlation nul ne signifie pas qu'il n'y a pas de lien entre les deux variables

Données ozone

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
maxO3	1.00	0.70	0.78	0.77	-0.62	-0.64	-0.48	0.53	0.43	0.39	0.68
T9	0.70	1.00	0.88	0.85	-0.48	-0.47	-0.33	0.25	0.22	0.17	0.58
T12	0.78	0.88	1.00	0.95	-0.58	-0.66	-0.46	0.43	0.31	0.27	0.56
T15	0.77	0.85	0.95	1.00	-0.59	-0.65	-0.57	0.45	0.34	0.29	0.57
Ne9	-0.62	-0.48	-0.58	-0.59	1.00	0.79	0.55	-0.50	-0.53	-0.49	-0.28
Ne12	-0.64	-0.47	-0.66	-0.65	0.79	1.00	0.71	-0.49	-0.51	-0.43	-0.36
Ne15	-0.48	-0.33	-0.46	-0.57	0.55	0.71	1.00	-0.40	-0.43	-0.38	-0.31
Vx9	0.53	0.25	0.43	0.45	-0.50	-0.49	-0.40	1.00	0.75	0.68	0.34
Vx12	0.43	0.22	0.31	0.34	-0.53	-0.51	-0.43	0.75	1.00	0.84	0.22
Vx15	0.39	0.17	0.27	0.29	-0.49	-0.43	-0.38	0.68	0.84	1.00	0.19
maxO3v	0.68	0.58	0.56	0.57	-0.28	-0.36	-0.31	0.34	0.22	0.19	1.00

Table: matrice des corrélations

Plan

Rappels

Introduction

Coefficient de corrélation linéaire

Caractère significatif

Corrélation sur les rangs

Causalité

Coefficient ρ

- r varie selon l'échantillon, mais la liaison entre deux variables ne varie que par les variables considérées
- r est une version empirique du coefficient de corrélation ρ

$$\begin{aligned}\rho(X,Y) &= E\left[\frac{X-E[X]}{\sqrt{Var[X]}} \times \frac{Y-E[Y]}{\sqrt{Var[Y]}}\right] \\ &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}\end{aligned}$$

- r est une estimation de ρ
- Quelles sont les valeurs de ρ compatibles, avec un certain degré de confiance, avec nos données ? Dans quelle mesure peut-on dire que $\rho \neq 0$?

Intervalle de confiance (1)

- Si X et Y sont normalement distribuées, R ne suit pas pour autant une loi normale (NB : sa loi est connue, mais en général complexe à utiliser)
- Mais $Z = f(R) = \frac{1}{2} \ln \left(\frac{1+R}{1-R} \right)$ suit **approximativement** une loi normale ($n > 25$)
  - d'espérance $f(\rho) = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$
  - de variance $\dfrac{1}{n-3}$
- Principe de l'IC pour ρ :
  - construire un IC $[b_{inf}; b_{sup}]$ pour $\mu_Z = f(\rho)$
  - transformer les bornes selon f^{-1} pour trouver un IC de ρ

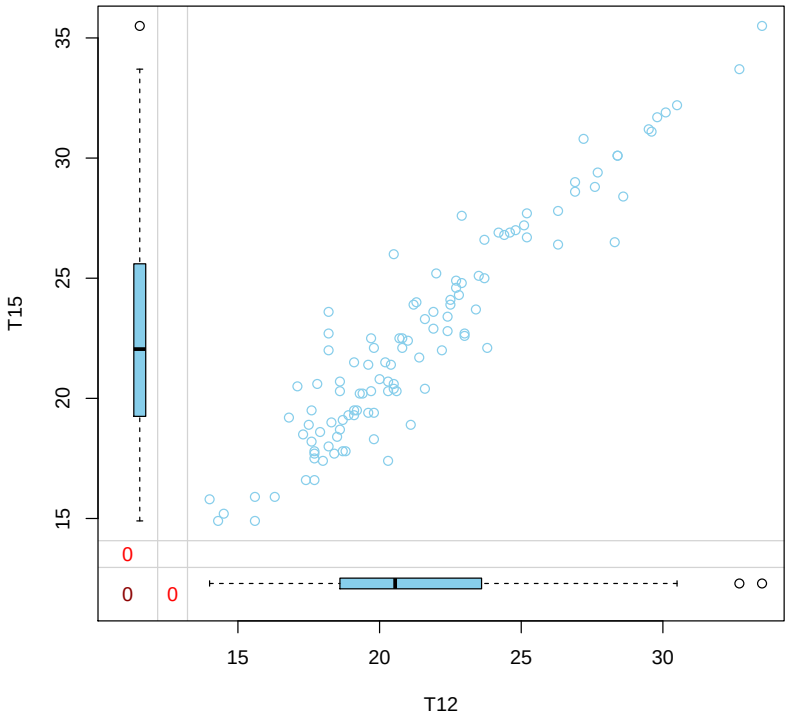
Intervalle de confiance (2)

- On définit b_{inf} et b_{sup} tels que $P(b_{inf} \leq \mu_Z \leq b_{sup}) = 1 - \alpha$ avec $\mu_Z = f(\rho)$
- $b_{inf} = z - \dfrac{q_{1-\alpha/2}}{\sqrt{n-3}}$ et $b_{sup} = z + \dfrac{q_{1-\alpha/2}}{\sqrt{n-3}}$ avec $z = f(r)$
- La fonction réciproque de f est la tangente hyperbolique $f^{-1}(z) = \frac{e^{2z}-1}{e^{2z}+1}$
- En appliquant f^{-1} sur les bornes, on obtient l'intervalle de confiance pour ρ

$$\mathrm{IC}_{(1-\alpha)}(\rho) = \left[\frac{e^{2b_{inf}} - 1}{e^{2b_{inf}} + 1}; \frac{e^{2b_{sup}} - 1}{e^{2b_{sup}} + 1}\right]$$

Exemple

- Données ozone
- Température à 12 et 15 heures



$$r = 0.946193$$
$$z = 1.794114$$
$$b_{inf} = 1.606383$$
$$b_{sup} = 1.981844$$
$$IC_{95\%}(\rho) = [0.92; 0.96]$$

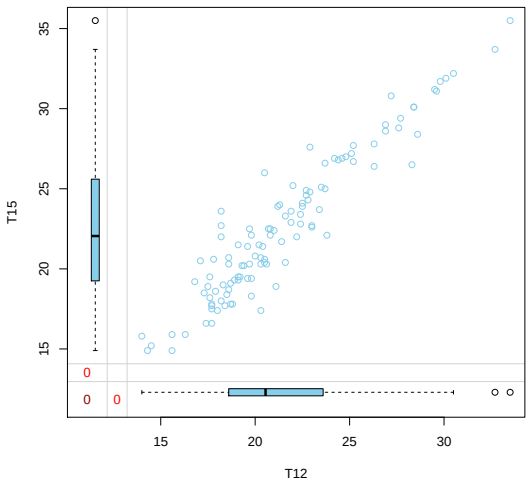
Test d'association

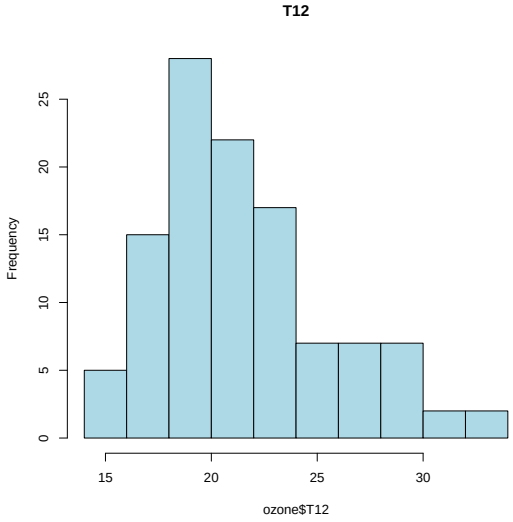
- H_0: $\rho = 0$ contre H_1: $\rho \neq 0$
- Sous l'hypothèse que X et Y sont Gaussiens

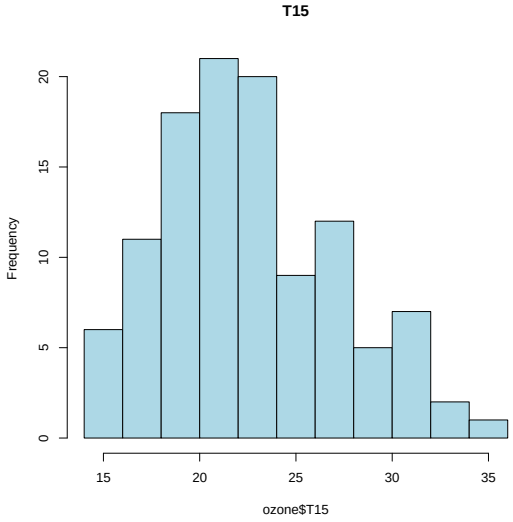
$$T_c = R\sqrt{\frac{n-2}{1-R^2}} \sim_{H_0} t_{(n-2)}$$

- Test bilatéral
  - on calcule r et t_c que l'on compare à $t_{1-\alpha/2,n-2}$
  - si $|t_c| < t_{1-\alpha/2,n-2}$, on ne rejette pas H_0 (au risque α)
    On ne peut pas dire qu'il y ait une liaison entre X et Y
  - si $|t_c| \geq t_{1-\alpha/2,n-2}$, on rejette H_0 (au risque α)
    On conclut que X et Y sont liées
- **Attention !** Un test significatif ne signifie pas une association forte.

Exemple







$$
\begin{aligned}
r &= 0.946193 \\
t_c &= 0.946 \times \sqrt{\frac{110}{1 - 0.946^2}} \\
&\simeq 30.67 \\
t_{0.975,110} &= 1.98
\end{aligned}
$$

$t_{0.975,110} < |t_c| \Rightarrow$ il existe bien une liaison au risque 5%

Remarques

- Le test d'indépendance à l'aide du coefficient de corrélation de Pearson nécessite l'hypothèse de normalité des distributions des deux variables.
- Le test est relativement robuste (pourvu que n soit assez grand), mais il ne permet que de tester un lien linéaire
- Dans le cas non-linéaire, on peut utiliser le test de corrélation des rangs de Spearman

Coefficient de corrélation de Spearman (1)

- En pratique, on évalue le coefficient de corrélation de Pearson sur le couple des rangs des observations

- Exemple

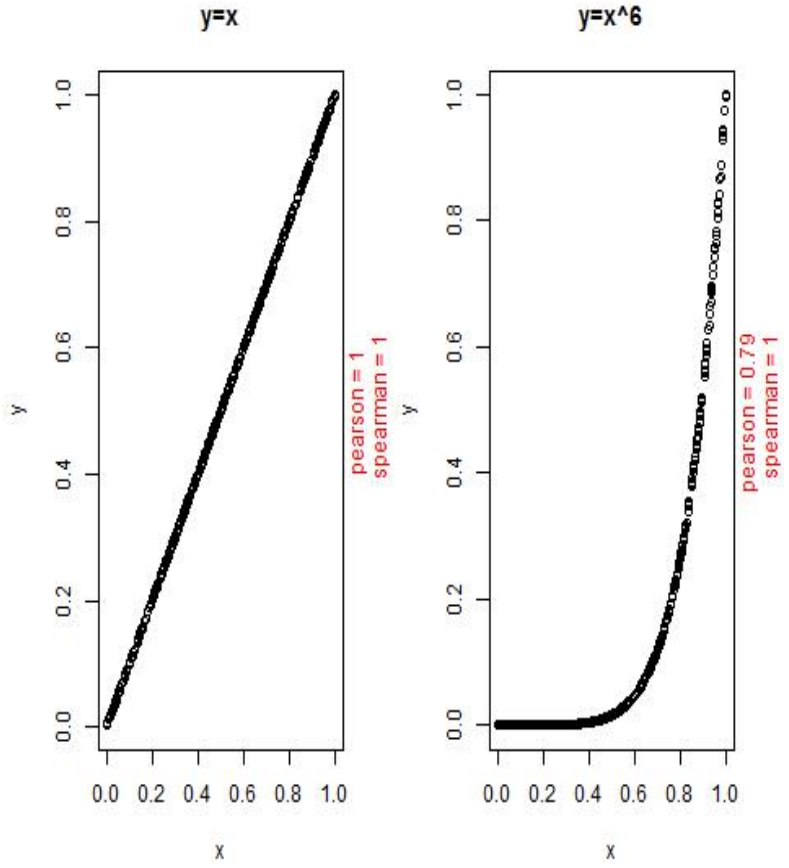
	T12	T15	Rang T12	Rang T15
20010601	18.5	18.4	27	20
20010602	18.4	17.7	26	12
20010603	17.6	19.5	12	34
20010604	19.7	22.5	44	62
20010605	20.5	20.4	55	42
20010606	19.8	18.3	46	19
...	...	...	...	...

$$\begin{aligned} r_{Pearson}(T12, T15) &= 0.946 \\ r_{Spearman}(T12, T15) &= r_{Pearson}(Rang\ T12, Rang\ T15) \\ &= 0.911 \end{aligned}$$

- NB : en cas d'ex aequo, on considère les rangs moyens

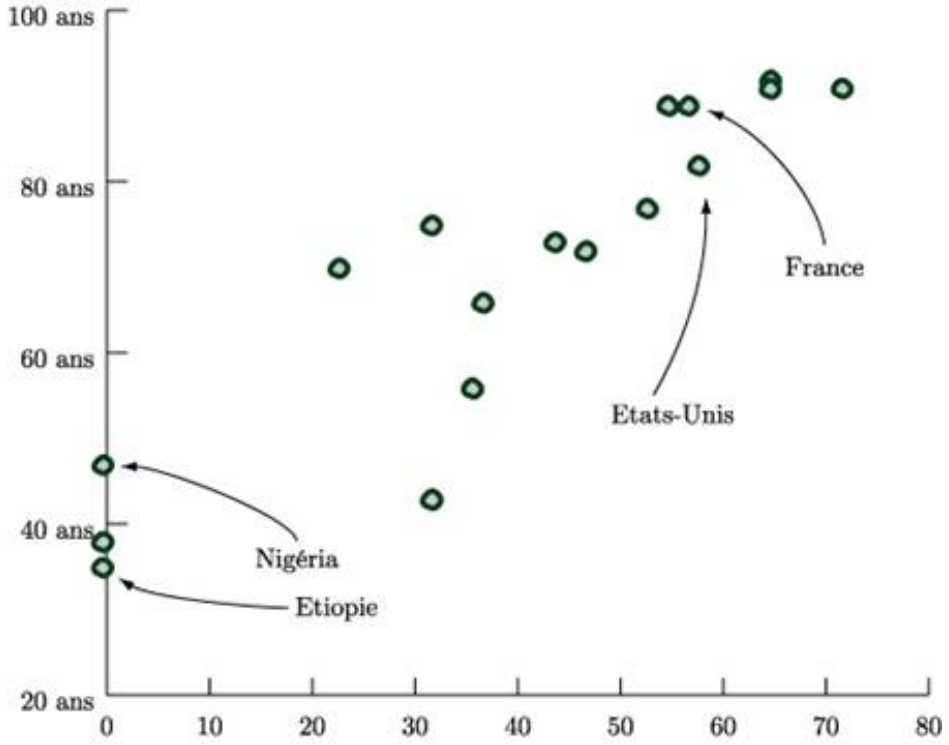
Coefficient de corrélation de Spearman (2)

- Utilisé lorsque les relations ne sont plus linéaires
- L'interprétation est identique à celle de la corrélation par rangs de Pearson
- Le test de $\rho_{Spearman} = 0$ est effectué à l'aide de tables donnant la distribution sous H_0



Causalité

Figure: Espérance de vie en fonction de la consommation d'alcool par pays. Source : Joseph Klatzman, Attention statistiques !, La Découverte 1996



Démontrer une causalité nécessite de parfaitement controler les variables explicatives (essais contrôlés)

Conclusion

- Lien linéaire entre deux variables quantitatives est mesuré par la corrélation linéaire
- Dans le cas non-linéaire, on peut (parfois) utiliser les rangs des observations
- Un coefficient de correlation linéaire empirique non nul ne signifie pas un lien significatif
- Une corrélation significative ne signifie pas un lien fort
- Montrer qu'il y a un lien ne signifie pas que ce lien est nécessairement un lien causal