Analyse bivariée (partie 1)

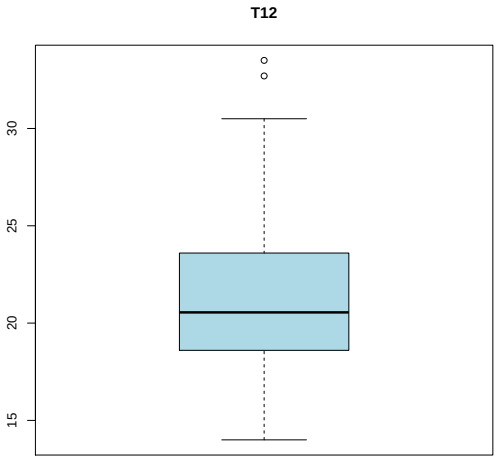
Vincent Audigier
vincent.audigier@lecnam.net

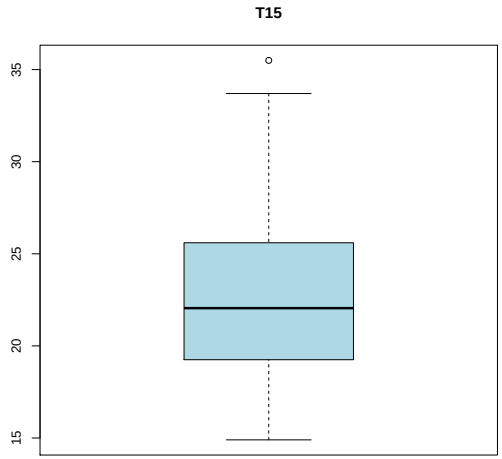
CNAM, Paris

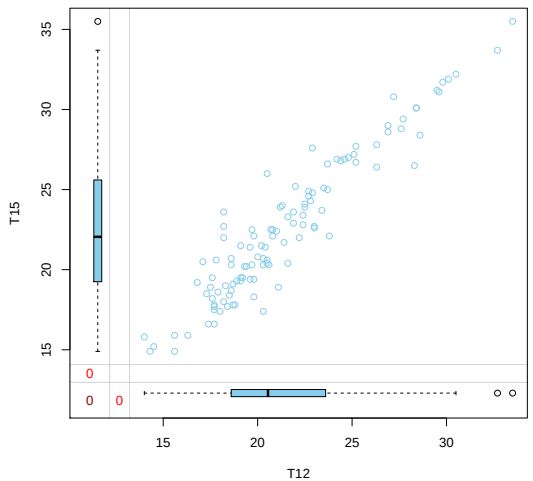
STA101 2019-2020

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

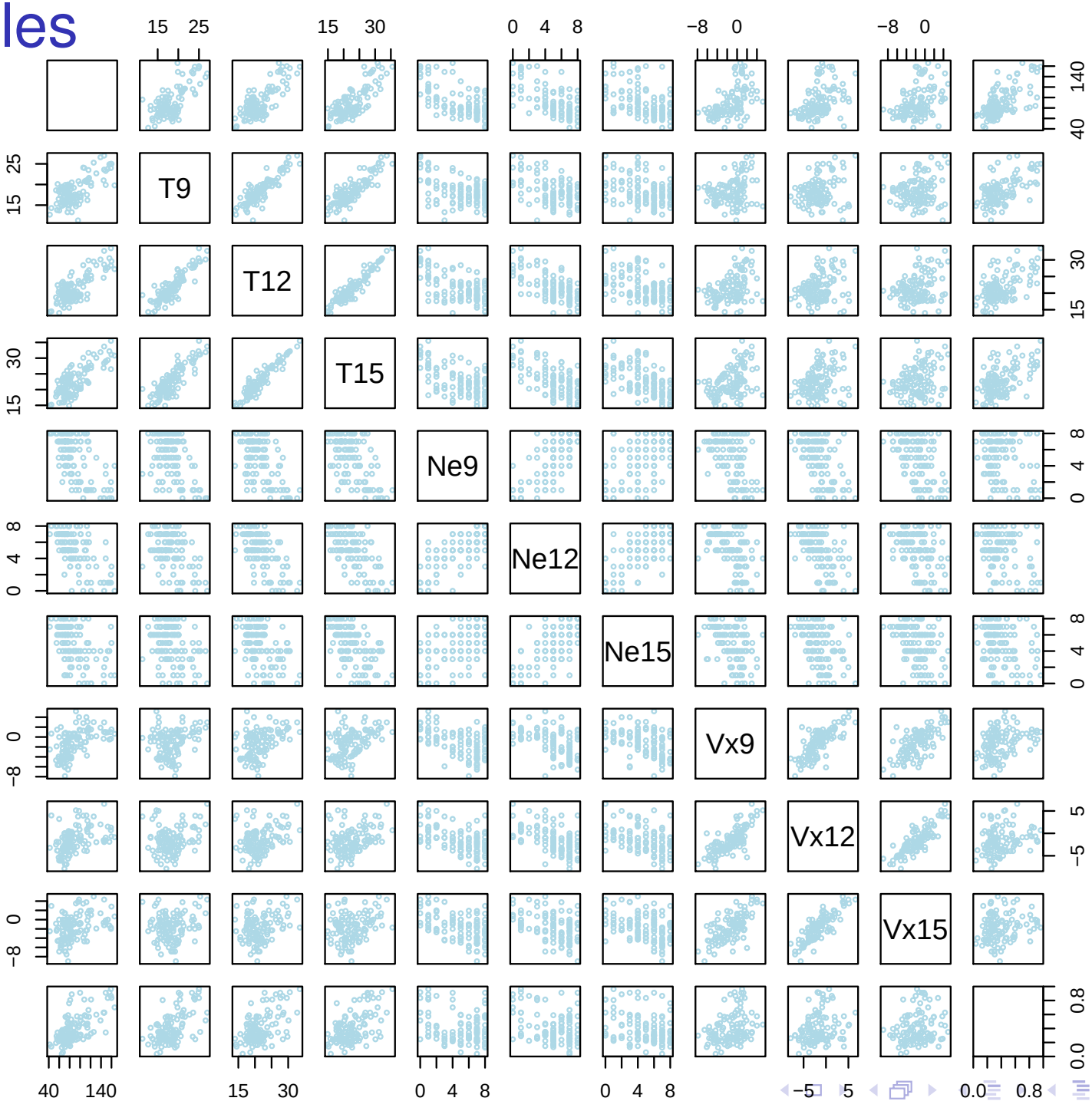
Analyse bivariée

- 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 gilets jaunes et prix du baril de pétrole
- Quantité d'alcool consommé et espérance de vie

Deux cas de figure

- X et Y sont deux variables aléatoires
  Elles ont un rôle symétrique, on ne cherche a priori pas à prédire l'une par l'autre.
  $$\Rightarrow \text{corrélation}$$

- Y est aléatoire, mais X est fixe
  Elles ont un rôle asymétrique, on pense pouvoir prédire Y à partir de X.
  $$\Rightarrow \text{régression}$$

Plan

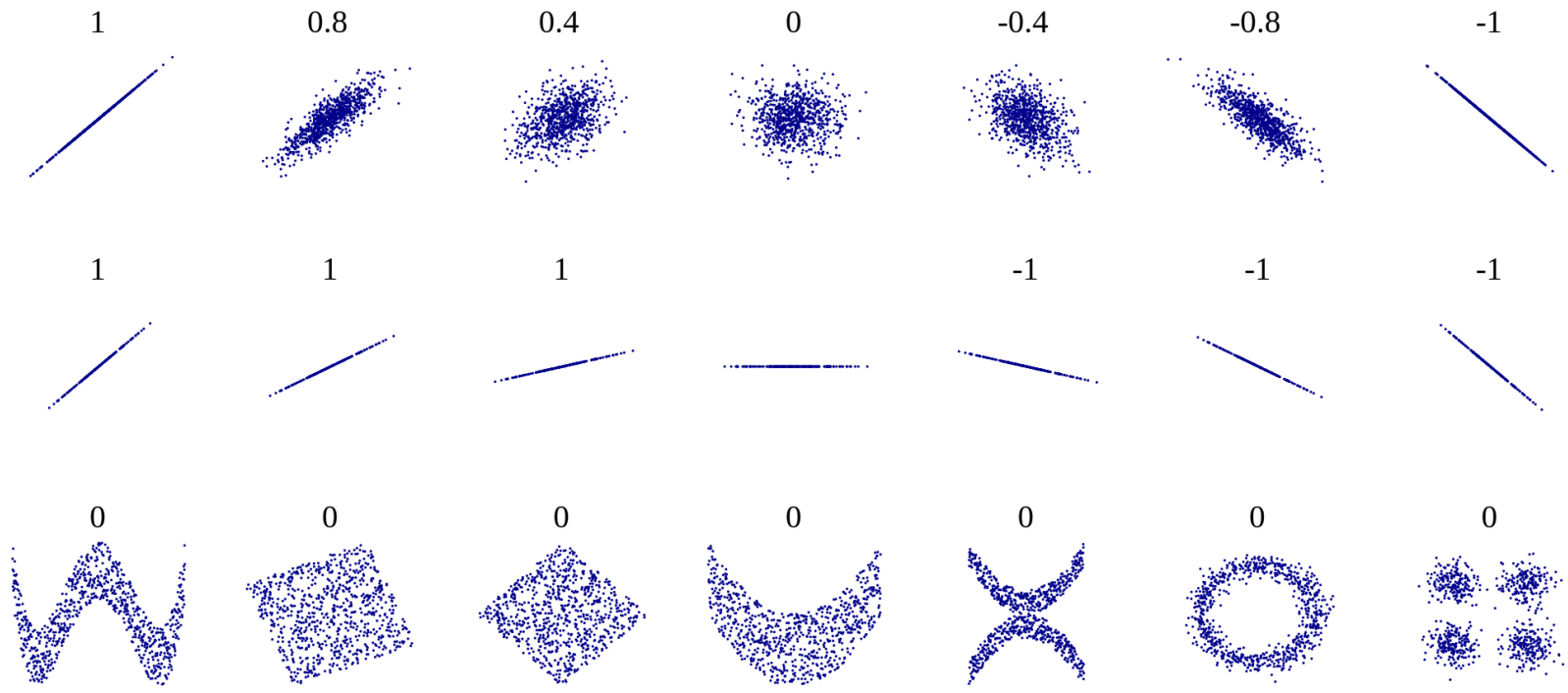
Coefficient de corrélation linéaire

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

Propriétés :

- symétrique
- non-transitif (ex : X et Y non liées et $Z = X + Y$ X$\sim$Z, Z$\sim$Y mais X$\nsim$Y)
- $-1 \leq r \leq 1$
- $|r| = 1$ 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

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

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

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

- 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 distribués, R ne suit pas pour autant une loi normale (NB : sa loi est connue, mais en général complexe à utiliser)

- Mais $Z = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right)$ suit **approximativement** une loi normale ($n > 25$)
  - d'espérance $\frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$
  - de variance $\frac{1}{n-3}$
- Principe de l'IC pour ρ :
  - calculer r, le transformer selon $f(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$
  - définir deux bornes z_1, z_2 telles que $P(z_1 \leq Z \leq z_2) = 1 - \alpha$ ($\alpha = 5\%$)
  - transformer les bornes selon f^{-1}

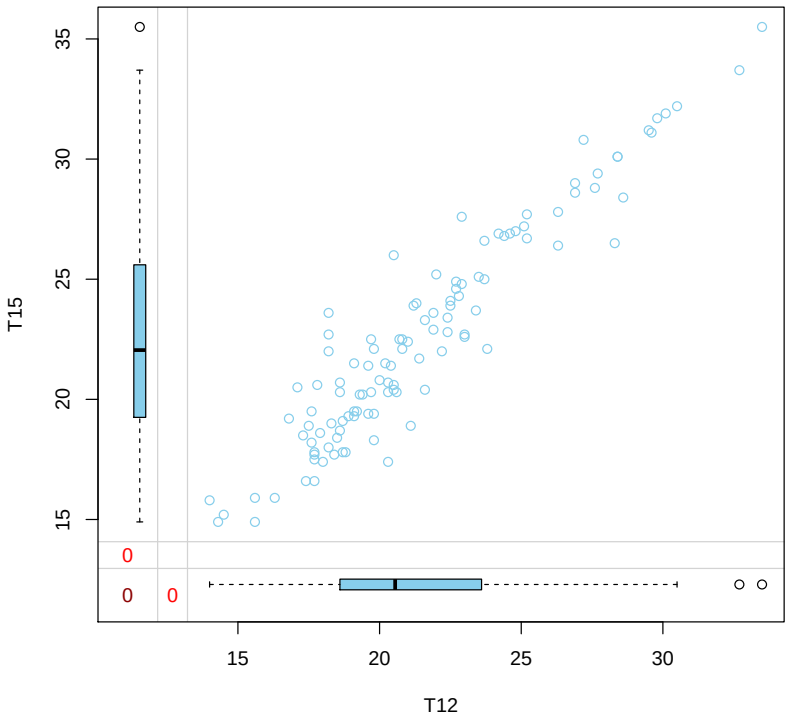
Intervalle de confiance (2)

- On définit z_1 et z_2 tels que $P(z_1 \leq Z \leq z_2) = 1 - \alpha$ avec $Z = f(R)$
- $z_1 = z - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$ et $z_2 = z + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$
- 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

$$\text{IC}_{(1-\alpha)}(\rho) = \left[\frac{e^{2z_1}-1}{e^{2z_1}+1}; \frac{e^{2z_2}-1}{e^{2z_2}+1}\right]$$

Exemple

- Données ozone
- Température à 12 et 15 heures



$$r = 0.946193$$
$$z = 1.794114$$
$$z_1 = 1.606383$$
$$z_2 = 1.981844$$
$$IC_{95\%}(\rho) = [0.92; 0.96]$$

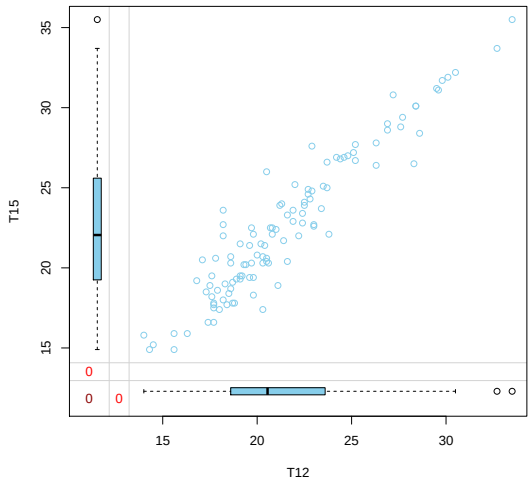
Test d'association

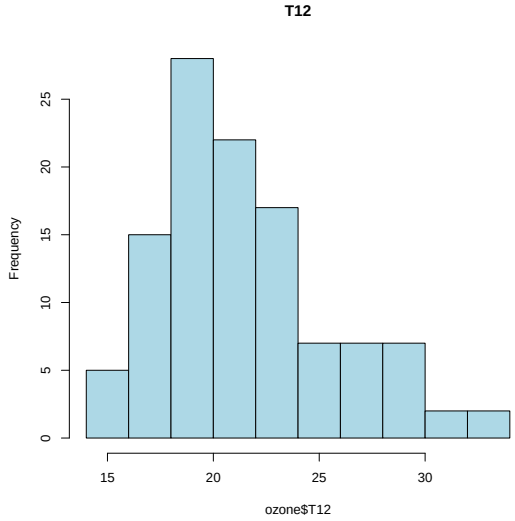
- H_0: $\rho = 0$ contre H_1: $\rho \neq 0$
- Sous l'hypothèse que X et Y sont Gaussiens

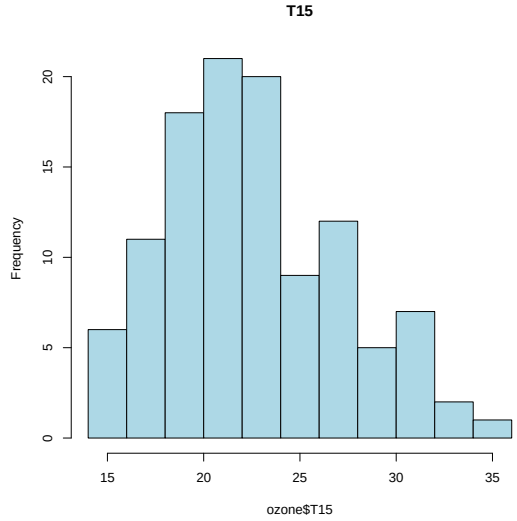
$$T_c = R\sqrt{\frac{n-2}{1-R^2}} \sim_{H_0} t_{(n-2)}$$

- 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%

Coefficient de corrélation de Spearman

- 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 nécessite toujours un **lien linéaire**
- Utilisation du **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

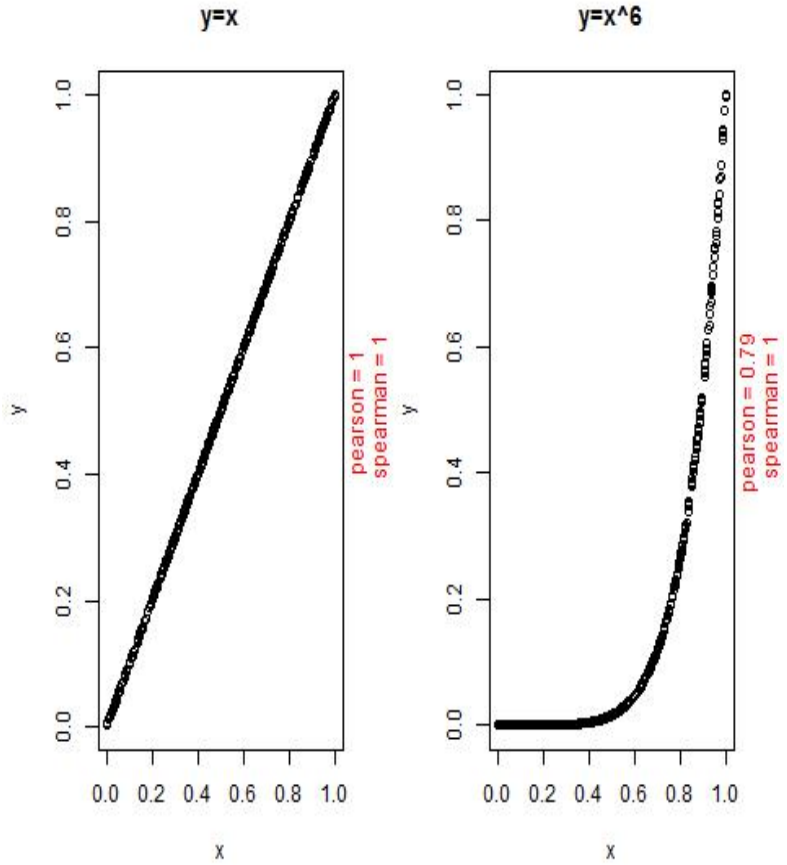
	T12	T15	Rang T12	Rang T15
20010601	18.5	18.4	27	20
20010602	18.4	17.7	26	12
20010603	17.6	19.5	12	34
20010604	19.7	22.5	44	62
20010605	20.5	20.4	55	42
20010606	19.8	18.3	46	19
...	...	...	...	...

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

- 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



Plan

Régression

- Le coefficient de corrélation résume le lien entre deux variables aléatoires
- Parfois, X n'est pas une variable aléatoire (ex: temps de mesure fixé dans une expérience)
- On peut aussi vouloir s'intéresser simplement à l'effet de X sur Y
- Objectif : expliquer comment varie Y en fonction de X et prédire Y à partir de valeurs de X
  - X et Y ne jouent plus un rôle symétrique
  - X est appelée variable **explicative**
  - Y est appelée variable à **expliquée**

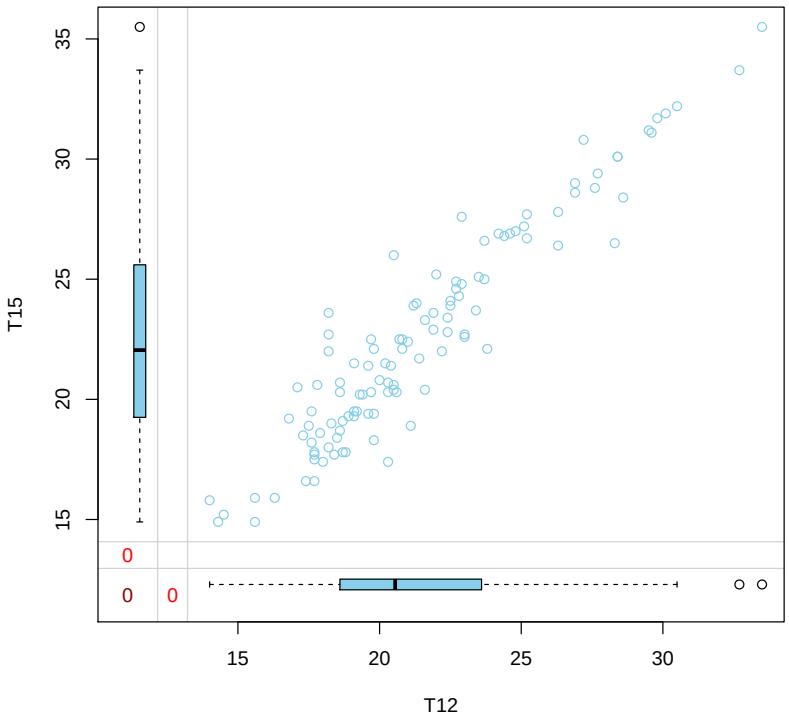
Modèle de régression linéaire

- On cherche à estimer E($Y|X = x$) sous la forme d'une fonction
- Modèle général: E$[Y|X = x] = f(x)$ ou $Y = f(X) + \varepsilon$
  où ε est une variable aléatoire d'espérance nulle qui représente l'erreur résiduelle du modèle
- Modèle linéaire = le plus simple

$$E[Y|X = x] = \alpha + \beta X$$
$$Y = \alpha + \beta X + \varepsilon \qquad \text{avec } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Exemple

Y : température à 15h
X : température à 12h



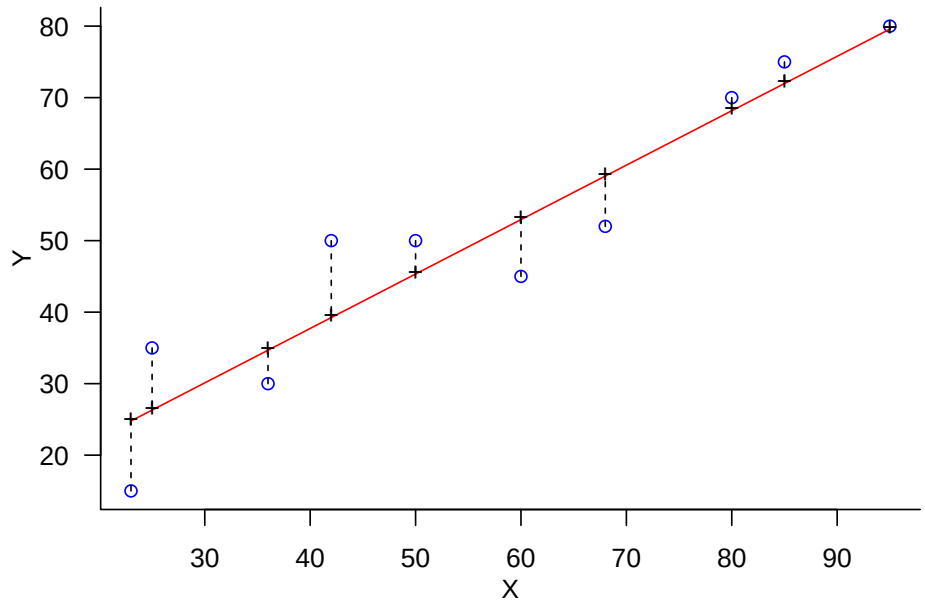
- $E[Y|X = x]$ est la moyenne de la température à 15h quand il fait la température x à 12h
- $E[Y|X = x] = \alpha + \beta x$: la température moyenne à 15h en fonction de la température à 12h s'obtient par la formule $\alpha + \beta x$ (lien linéaire)
- De façon équivalente, on écrit $Y = \alpha + \beta X + \varepsilon$: la température à 15h pour une observation, c'est une fonction linéaire de la température à 12h, plus une erreur (gaussienne centrée)

Estimation

- modèle estimé :

$$y_i = a + bx_i + \varepsilon_i$$

- a et b sont calculés de façon à ce que la distance entre les observations et les prédictions (distance parallèle à l'axe de ordonnées) soit minimale



- Critère des **moindres carrés**: minimiser

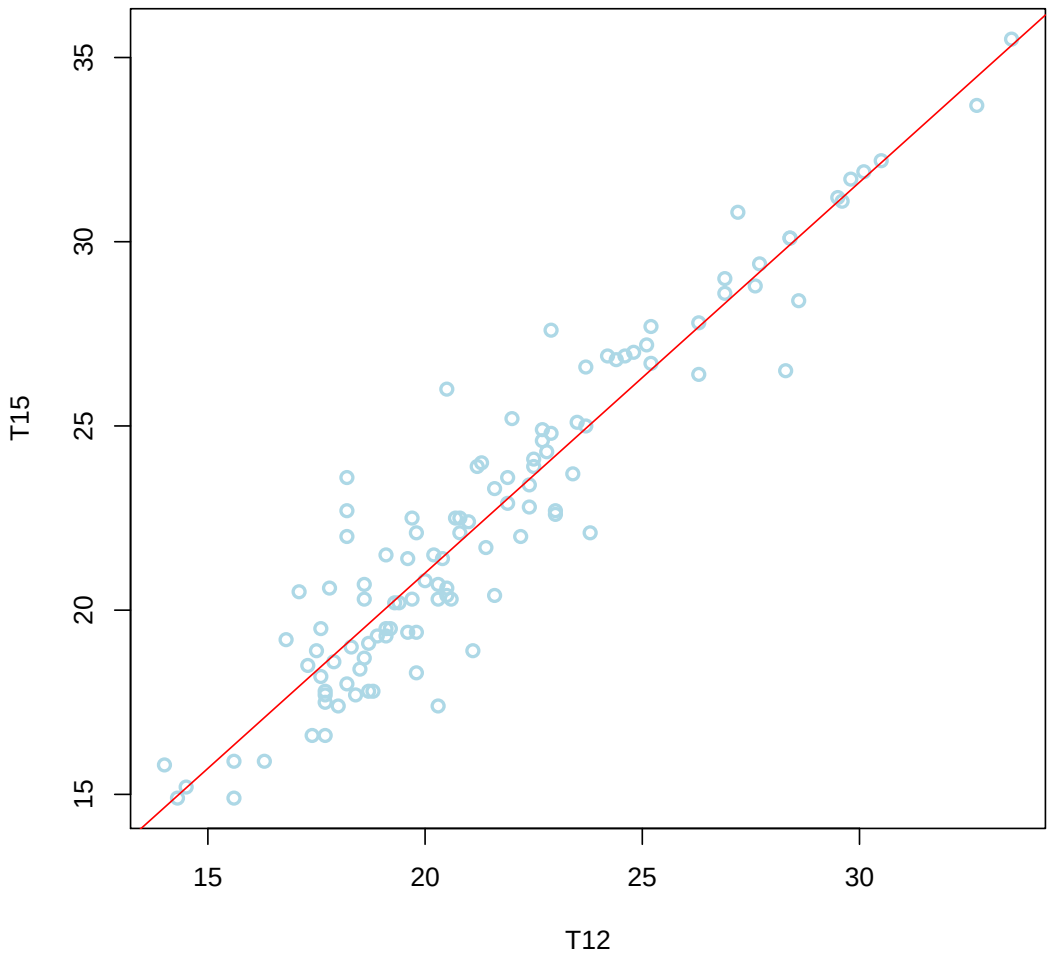
$$C_{MC} = \sum \varepsilon_i^2 = \sum (y_i - a - bx_i)^2$$

Solutions

$$
\begin{aligned}
b &= \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
a &= \bar{y} - b\bar{x}
\end{aligned}
$$

Comme il s'agit de coefficients calculées à partir des données, ce sont des estimations de β et α. Pour cette raison, on les note aussi $\hat{\beta}$ et $\hat{\alpha}$

Exemple



a=-0.2025
b=1.0605

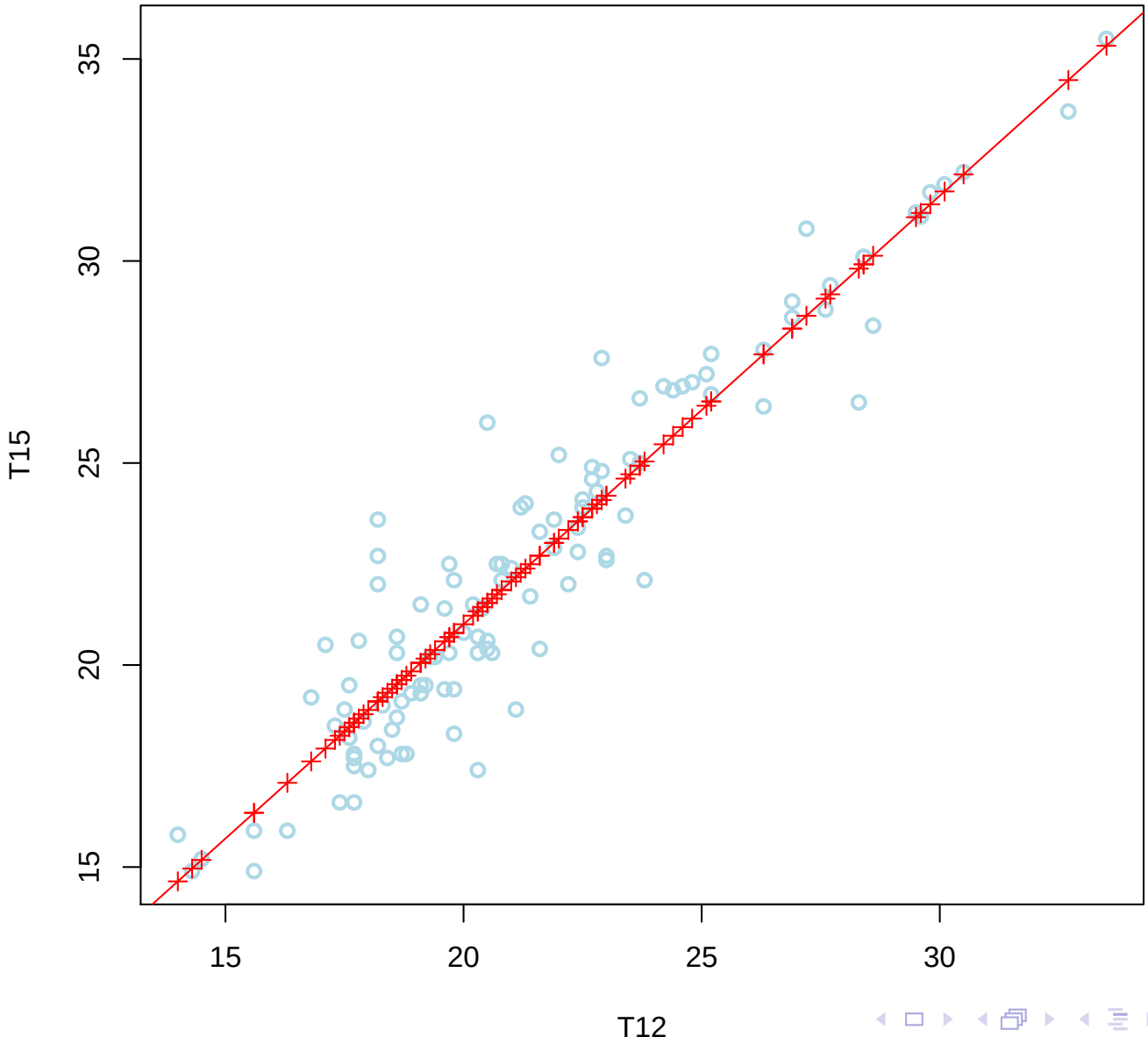
Interprétation

- la valeur des coefficients du modèle a un sens
  - *b* nous indique que quand la variable explicative augmente d'une unité, en moyenne, la variable réponse augmente de *b* unité
  - *a* nous indique la moyenne de la variable réponse quand la variable explicative est nulle
- il est également possible de construire des intervalles de confiance pour α et β
- on peut aussi tester la nullité de ces coefficients

Remarque

- La droite de régression nous fournit une approximation de la moyenne de Y pour les différentes valeurs de x
- C'est naturellement également une estimation de y_i en fonction de x_i
- On note cette estimation $\hat{y}_i$
- $\hat{y}_i = a + bx_i$

Exemple



Coefficient de détermination

Décomposition de la variance

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \\ \textit{variabilite totale} &= \textit{variabilite expliquee} + \textit{variabilite residuelle} \end{aligned}$$

On mesure la qualité de l'ajustement du modèle par

$$\begin{aligned} R^2 &= \frac{\textit{variabilite expliquee}}{\textit{variabilite totale}} \\ &= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \end{aligned}$$

Propriétés

- $0 \leq R^2 \leq 1$
- $R^2 = 0 \Leftrightarrow \textit{variabilite expliquee} = 0$
- $R^2 = 1 \Leftrightarrow \textit{variabilite residuelle} = 0$

Prédiction

En pratique, on utilise souvent le modèle de régression pour la prédiction d'une valeur future

Exemple : on souhaite prédire la température à 15h à partir de la température à 12h pour un nouvel individu. On note x_0 la température à 12h.

$$E[Y_0] = \alpha + \beta x_0$$

Il est naturel d'estimer $E[Y_0]$ par $a + bx_0$

Corrélation vs. régression ?

- contexte d'utilisation différents
- mais mathématiquement les deux sont liés

$$r = b\frac{s_x}{s_y} \Leftrightarrow b = r\frac{s_y}{s_x}$$

- le coefficient de corrélation est symétrique
- alors que la pente change si on permutent X et Y