Régression linéaire et logistique

M1 Ingénierie Statistique
Université de Nantes

Frédéric Lavancier

Objectifs d'un modèle de régression : Expliquer une grandeur Y en fonction de p grandeurs $X_1, \ldots, X_p$. Pour cela on dispose de n observations de Y et des X_j.

Exemples :

- Y : la consommation électrique quotidienne en France

X : température moyenne journalière.
Les données sont un historique de Y et X sur n jours
Question : a-t-on $Y \approx f(X)$ pour une certaine fonction f ?
En simplifiant : a-t-on $Y \approx aX + b$ pour certaines valeurs a et b ?
Si oui, $a =?$, $b =?$

- $Y = 0$ *ou* 1 : qualité d'un client (1 : bon; 0 : pas bon)

X_1 : revenu du client
X_2 : catégorie socio professionnelle (6-7 possibilités)
X_3 : âge
On modélise dans ce cas $p = \mathbb{P}(Y = 1)$. A-t-on $p \approx f(X_1, \ldots, X_p)$ pour une fonction f à valeurs dans $[0, 1]$?
On peut simplifier à des f particulières comme la fonction logistique.

La relation "approximative" $Y \approx f(X_1, \ldots, X_p)$ est un **modèle**.

Pourquoi chercher à établir un tel modèle ? Deux raisons principales :

- Objectif descriptif : quantifier l'effet marginal de chaque variable.
  Par exemple, si X_1 augmente de 10%, comment évolue Y ?
- Objectif prédictif : étant donné des nouvelles valeurs pour $X_1, \ldots, X_p$, on peut en déduire le Y (approximatif) associé.

Plan du cours :

1. Analyse bivariée
   $\rightarrow$ lien entre 2 variables
2. Régression linéaire
   $\rightarrow$ Y quantitative en fonction de $X_1, \ldots, X_p$ quantitatives
3. Analyse de la variance et de la covariance
   $\rightarrow$ Y quanti en fonction de $X_1, \ldots, X_p$ qualitatives et/ou quantitatives
4. Régression logistique
   $\rightarrow$ Y qualitative en fonction de $X_1, \ldots, X_p$ qualitatives et/ou quantitatives

1 Analyse bivariée

Type de variables

On s'intéresse au lien entre 2 variables X et Y.

On distingue deux grandes catégories, chacune déclinées en deux types.

• **Variable quantitative** : son observation est une quantité mesurée.
Exemples : âge, salaire, nombre d'infractions,...

On distingue les variables quantitatives **discrètes** dont les valeurs possibles sont finies ou dénombrables (*Exemples : nombre d'enfants, nombre d'infractions,...*) et les variables quantitatives **continues** qui peuvent prendre toutes les valeurs possibles d'un intervalle (*Exemples : taille, salaire,...*)

• **Variable qualitative** (ou **facteur**): son observation se traduit par une catégorie ou un code. Les observations possibles sont appelées les **modalités** de la variable qualitative.
Exemples : sexe, CSP, nationalité, mention au BAC,...

Lorsqu'un ordre naturel apparaît dans les modalités, on parle de variable qualitative **ordinale** (*Exemples : mention au BAC,...*). Dans le cas contraire on parle de variable qualitative **nominale** (*Exemples : sexe, CSP,...*).

Exemple du jeu de données "Pottery" : Composition chimique de poteries trouvées sur différents sites archéologiques au Royaume Uni.

	Site	Al	Fe	Mg	Ca	Na
1	Llanedyrn	14.4	7.00	4.30	0.15	0.51
2	Llanedyrn	13.8	7.08	3.43	0.12	0.17
3	Llanedyrn	14.6	7.09	3.88	0.13	0.20
4	Llanedyrn	10.9	6.26	3.47	0.17	0.22
5	Caldicot	11.8	5.44	3.94	0.30	0.04
6	Caldicot	11.6	5.39	3.77	0.29	0.06
7	IsleThorns	18.3	1.28	0.67	0.03	0.03
8	IsleThorns	15.8	2.39	0.63	0.01	0.04
9	IsleThorns	18	1.88	0.68	0.01	0.04
10	IsleThorns	20.8	1.51	0.72	0.07	0.10
11	AshleyRails	17.7	1.12	0.56	0.06	0.06
12	AshleyRails	18.3	1.14	0.67	0.06	0.05
13	AshleyRails	16.7	0.92	0.53	0.01	0.05

Les individus : les poteries numérotées de 1 à 13
Les variables : le site archéologique (facteur à 4 modalités) et différents composés chimiques (quantitatives).

Exemple du jeu de données "NO2trafic" : Concentration en NO2 mesurée à l'intérieur de voitures circulant en région parisienne, selon le type de voie empruntée (5 possibilités) et la fluidité du trafic (de A à D)

	NO2	type	fluidite
1	378.94	P	A
2	806.67	T	D
3	634.58	A	D
4	673.35	T	C
5	589.75	P	A
⋮	⋮	⋮	⋮
283	184.16	P	B
284	121.88	V	D
285	152.39	U	A
286	129.12	U	C

Les individus : les véhicules numérotées de 1 à 286
Les variables : NO2 (quantitative), type (facteur à 5 modalités) et fluidite (facteur ordinal à 4 modalités)

1. Analyse bivariée
   - Variable quantitative/ Variable quantitative
   - Variable qualitative/ Variable qualitative
   - Variable quantitative/ Variable qualitative

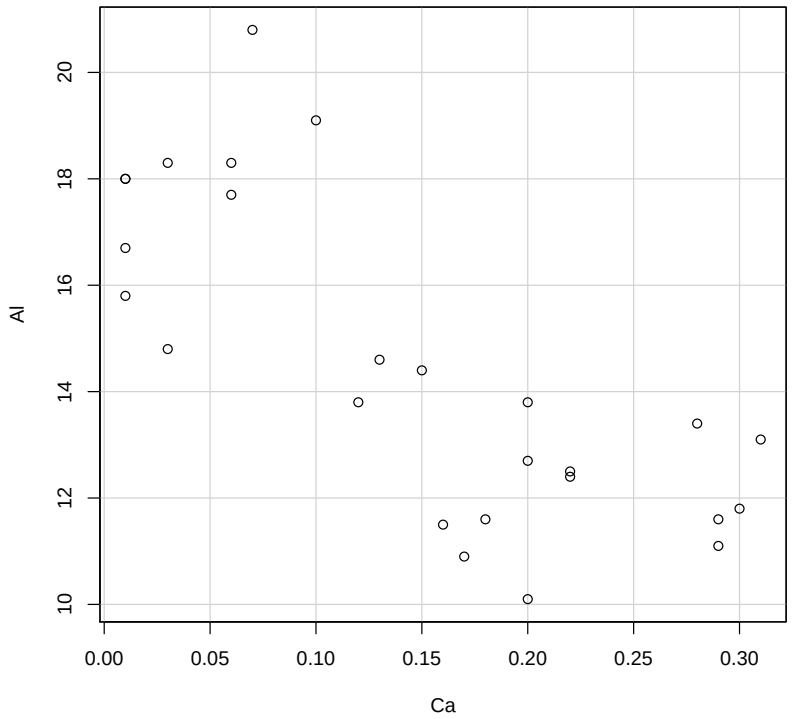
1. Analyse bivariée
   - Variable quantitative/ Variable quantitative
   - Variable qualitative/ Variable qualitative
   - Variable quantitative/ Variable qualitative

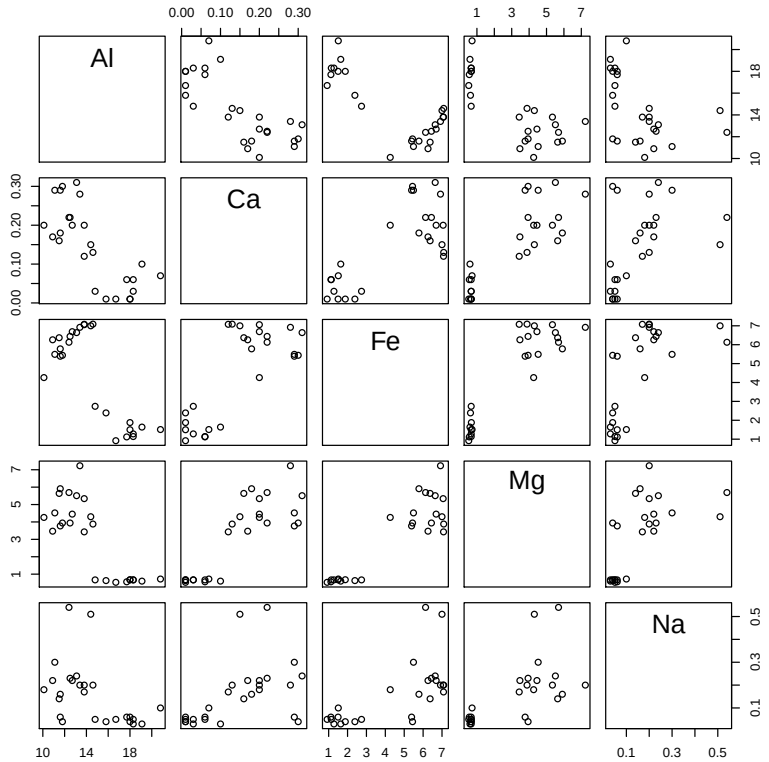
Représentation graphique : nuage de points

Soit $x_1, \ldots, x_n$ les valeurs observées de la première variable quantitative X.
Soit $y_1, \ldots, y_n$ les valeurs observées de la seconde variable quantitative Y.

On visualise le lien entre X et Y grâce au nuage des points (x_i, y_i).

Exemple : nuage de points entre "Al" et "Ca" des données "Pottery" et matrice des nuages de points entre toutes les variables.





Quantification du lien : corrélation linéaire

Le lien linéaire est quantifié par la **corrélation linéaire** de Pearson :

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

où $\bar{x}_n$ (resp. $\bar{y}_n$) désigne la moyenne empirique de X (resp. Y).

Propriétés : On déduit de l'inégalité de Cauchy Schwartz que

- La corrélation $\hat{\rho}$ est toujours comprise entre -1 et 1:
- si $\hat{\rho} = 1$, il y a un lien linéaire "parfait" positif, i.e. :

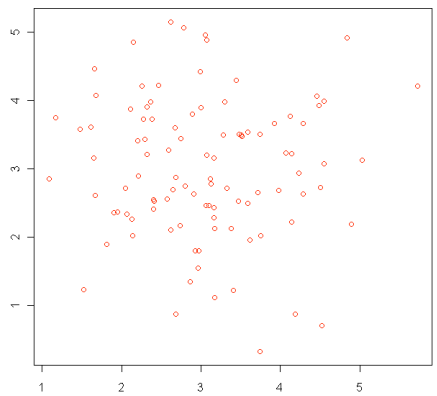
  $\hat{\rho} = 1 \quad$ ssi $\quad$ il existe $\alpha \geq 0$ et β tel que $y_i = \alpha x_i + \beta$ pour tout $i = 1, \ldots, n$

- si $\hat{\rho} = -1$, il y a un lien linéaire "parfait" négatif, i.e. :

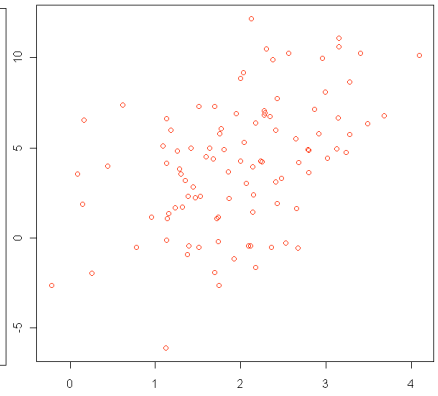
  $\hat{\rho} = -1 \quad$ ssi $\quad$ il existe $\alpha \leq 0$ et β tel que $y_i = \alpha x_i + \beta$ pour tout $i = 1, \ldots, n$

- si $\hat{\rho} = 0$, il n'y a aucun lien linéaire (mais il peut exister un lien non-linéaire).

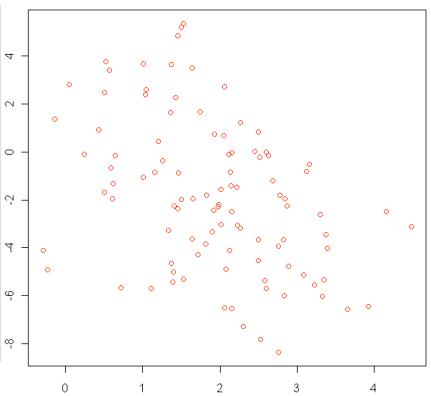
Quelques exemples de nuages de points avec la corrélation correspondante.



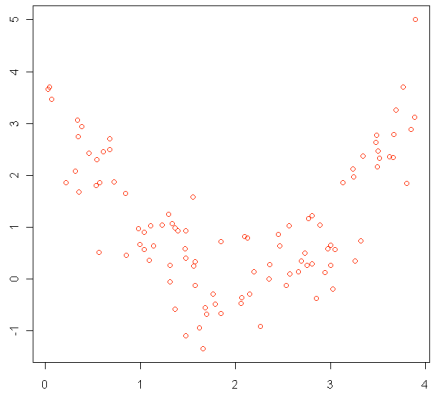
Aucun lien ($r \approx 0$)



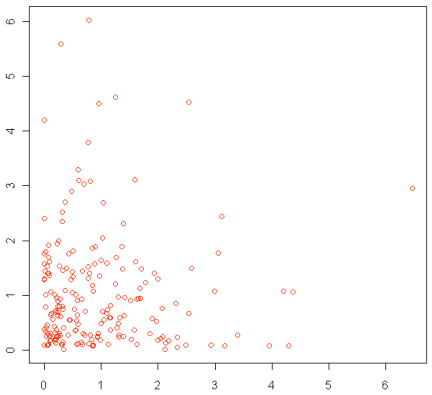
Lien linéaire ($r \approx 0.4$)



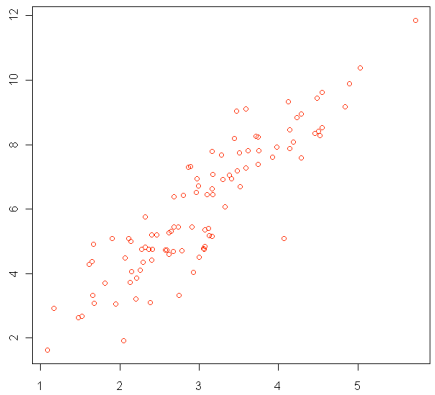
Lien linéaire ($r \approx -0.4$)



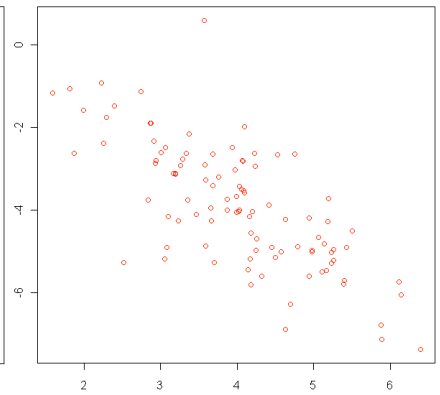
Lien non-linéaire($r \approx 0$)



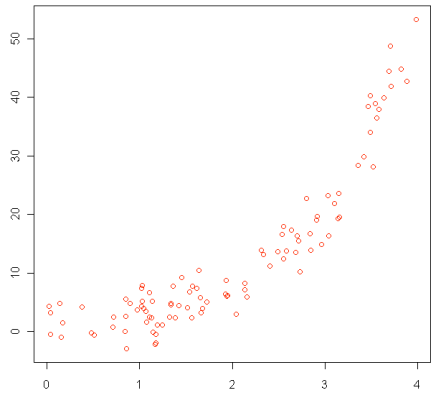
Aucun lien ($r \approx 0$)



Lien linéaire ($r \approx 0.9$)



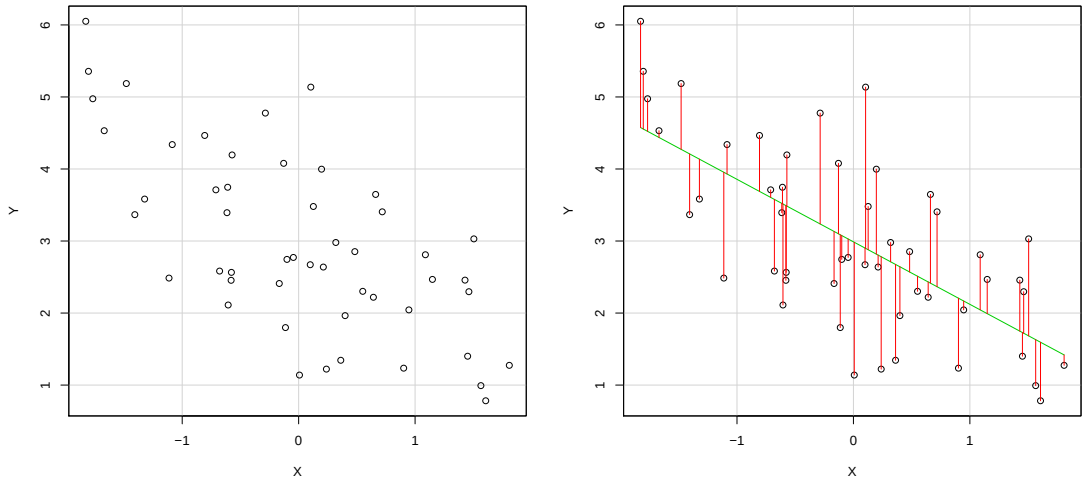
Lien linéaire ($r \approx -0.8$)



Lien non-linéaire ($r \approx 0.8$)

Ajuster la droite des moindres carrés

Droite des moindres carrés : Il s'agit de la droite qui passe "le mieux" au milieu des points (x_i, y_i), au sens où la somme des distances en rouge prises au carré est minimale. Il s'agit de la **régression linéaire** de Y sur X.



L'équation de la droite recherchée est donc $y = \hat{a}x + \hat{b}$ où $\hat{a}$ et $\hat{b}$ vérifient :

$$(\hat{a}, \hat{b}) = \underset{(a,b)}{argmin} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

On trouve (vérifiez-le !), si $var(X) \neq 0$:

$$\hat{a} = \frac{cov(X,Y)}{var(X)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

en notant *var* et *cov* la variance et la covariance empirique.

Tester si la corrélation est significative

$\hat{\rho}$ est un estimateur de la corrélation théorique ρ entre X et Y défini par

$$\rho = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)]}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

On peut vouloir tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$

Si (X, Y) est Gaussien, on peut montrer que $T \sim St(n-2)$ sous H_0 où

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

et $St(n-2)$ désigne la loi de Student à $n-2$ degrés de liberté. On en déduit

$$RC_\alpha = \{|T| > t_{n-2}(1-\alpha/2)\}.$$

Sous R : fonction cor.test

1. Analyse bivariée
   - Variable quantitative/ Variable quantitative
   - Variable qualitative/ Variable qualitative
   - Variable quantitative/ Variable qualitative

X : premier facteur à I modalités
Y : second facteur à J modalités.
n_{ij} : nombre d'individus ayant la modalité i pour X et j pour Y.
$n_{i.}$: nombre d'individus ayant la modalité i pour X
$n_{.j}$: nombre d'individus ayant la modalité j pour Y

$$n_{i.} = \sum_{j=1}^{J} n_{ij}, \quad n_{.j} = \sum_{i=1}^{I} n_{ij}, \quad n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

Les effectifs n_{ij} sont résumés dans un **tableau de contingence**.

Exemple : Pour les variables "type" et "fluidite" du jeu de données NO2trafic, le tableau de contingence est :

	type				
fluidite	P	U	A	T	V
A	21	21	19	9	9
B	20	17	16	8	7
C	17	17	16	8	7
D	20	20	18	8	8

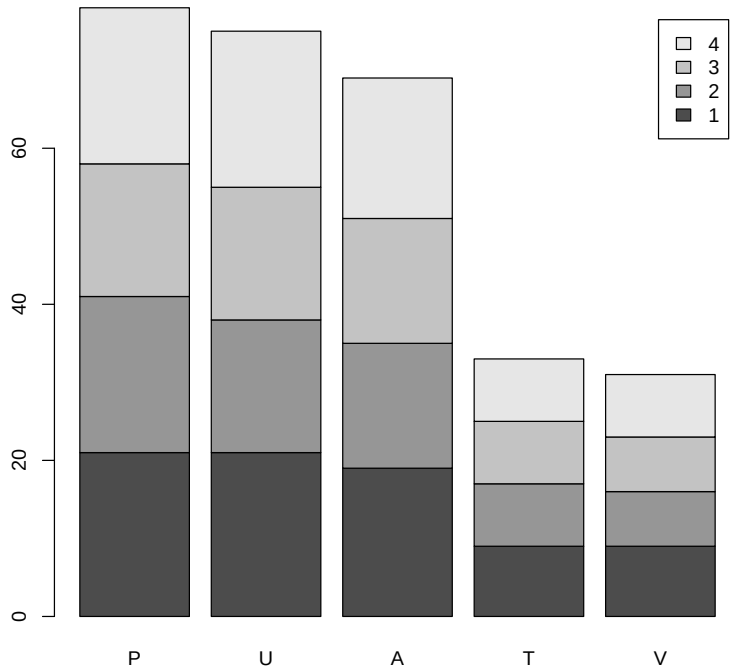
Sous R : table(X,Y)

On résume le tableau de contingence par des diagrammes en batons "croisés", soit par empilement (à gauche), soit côte à côte (à droite).

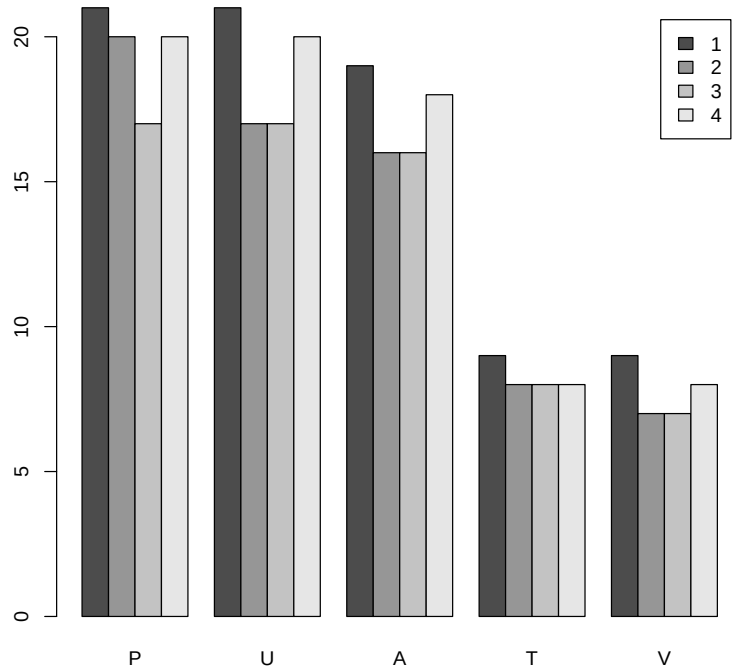
Sous R : si le tableau de contingence se nomme `tab`, barplot(tab) ou barplot(tab,beside=TRUE)

Exemple : pour le graphe croisant les variables "type" et "fluidite",

barplot(tab,legend.text=TRUE)



barplot(tab,beside=TRUE,legend.text=TRUE)



Remarque : si on souhaite représenter les fréquences et non les effectifs, il suffit de diviser `tab` par l'effectif total n, barplot(tab/n).

Pour quantifier le lien entre les deux facteurs, on calcule la distance du χ^2 (khi-deux)

$$d^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cette distance mesure la différence entre les effectifs observés n_{ij} et les effectifs théoriques s'il y avait indépendance : dans ce cas la fréquence observée dans i et j, $\frac{n_{ij}}{n}$, vaudrait le produit des fréquences marginales $\frac{n_{i.}}{n}\frac{n_{.j}}{n}$.

Test du χ^2 : H_0 : X et Y indépendants contre H_1 : le contraire
Sous H_0, $d^2 \sim \chi^2((I-1)(J-1))$ lorsque $n \to \infty$ d'où

$$RC_\alpha = \{d^2 > \chi^2_{(I-1)(J-1)}(1-\alpha)\}$$

est une région critique au niveau asymptotique α, avec $\chi^2_{(I-1)(J-1)}(1-\alpha)$ le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

Sous R : fonction chisq.test

$\to$ Pour aller plus loin dans la compréhension du lien : AFC.

1. Analyse bivariée
   - Variable quantitative/ Variable quantitative
   - Variable qualitative/ Variable qualitative
   - Variable quantitative/ Variable qualitative

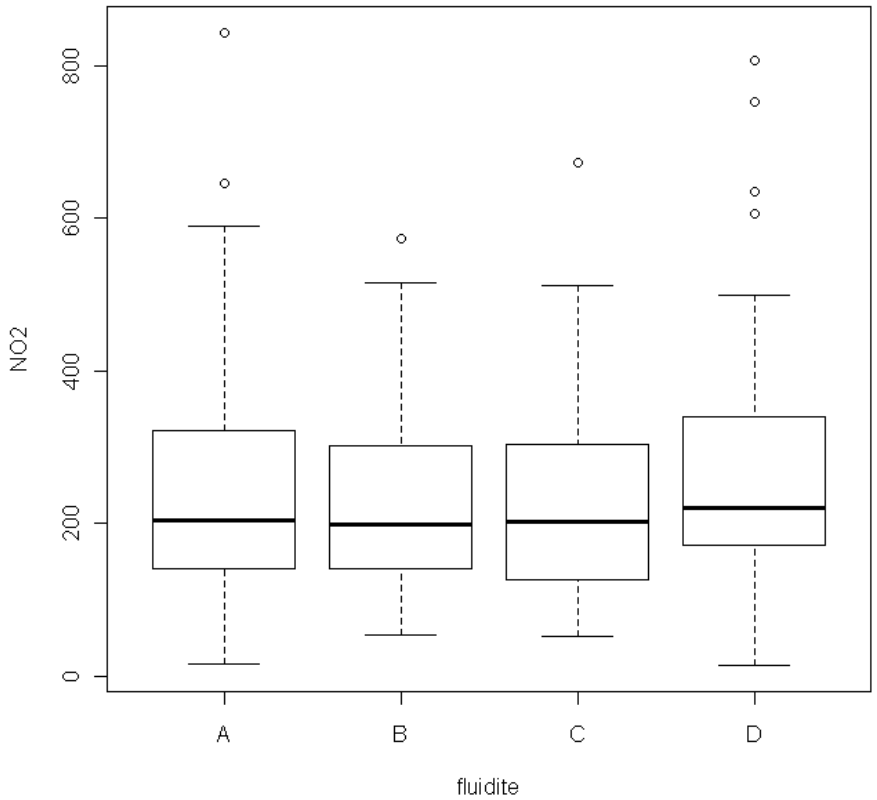
Variable quantitative/ Variable qualitative : graphes

X : variable quantitative
Y : facteur à I modalités

Graphiquement, on effectue des boxplots de X par modalité de Y.

Sous R : boxplot(X∼ Y)

Exemple : dans "NO2trafic", la concentration NO2 en fonction de "fluidite"



X : variable quantitative
Y : facteur à I modalités contenant chacune n_i individus ($\sum_{i=1}^{I} n_i = n$).
x_{ij} : valeur de X pour l'individu j se trouvant dans la modalité i de Y.
On note $\bar{x}_i$ la moyenne de X dans la modalité i et $\bar{x}$ la moyenne totale, i.e.

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{x}_i$$

Formule de décomposition de la variance : La variance totale est la somme de la variance inter-modalités et de la variance intra-modalités, ce qui s'écrit :

$$\frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2}_{S_T^2} = \frac{1}{n}\underbrace{\sum_{i=1}^{I} n_i(\bar{x}_i-\bar{x})^2}_{S_{inter}^2} + \frac{1}{n}\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}_{S_{intra}^2}$$

autrement dit $S_T^2 = S_{inter}^2 + S_{intra}^2$.

Le lien entre X et Y est parfois mesuré par le **rapport de corrélation eta**:

$$\hat{\eta}^2 = \frac{S_{inter}^2}{S_T^2} = \frac{\sum_{i=1}^{I} n_i(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2}.$$

On a ainsi $0 \leq \hat{\eta}^2 \leq 1$.

Le coefficient $\hat{\eta}^2$ estime son équivalent théorique η^2 défini par

$$\eta^2 = \frac{\mathbb{V}(\mathbb{E}(X|Y))}{\mathbb{V}(X)}.$$

Test d'analyse de la variance

En notant $\mu_i = \mathbb{E}(X|Y=i)$ pour $i = 1, \ldots, I$, on souhaite tester

$$H_0 : \mu_1 = \cdots = \mu_I \qquad (\Leftrightarrow \eta^2 = 0)$$

contre H_1 : X est différent (en espérance) dans au moins deux modalités de Y.

Si les x_{ij} sont issus d'une loi Gaussienne de même variance pour tout i, j, alors

$$F = \frac{S^2_{inter}/(I-1)}{S^2_{intra}/(n-I)} = \frac{\hat{\eta}^2/(I-1)}{(1-\hat{\eta}^2)/(n-I)} \sim F(I-1, n-I) \quad \text{sous } H_0.$$

D'où la région critique au niveau α

$$RC_\alpha = \{F > f_{I-1,n-I}(1-\alpha)\}$$

où $f_{I-1,n-I}(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ d'une loi $F(I-1, n-I)$.

Sous R : fonction aov(X∼Y) pour obtenir S^2_T, S^2_{inter} et S^2_{intra}, et summary du résultat pour effectuer le test.
Pour $I = 2$, cela correspond au test de Student d'égalité des moyennes (t.test).