Année 2008-2009

Biostatistiques

Rappels de cours et travaux dirigés

auteur : Jean-Marc Labatte
jean-marc.labatte@univ-angers.fr

SOMMAIRE

Table des matières

Accès direct aux chapitres

#I STATISTIQUES DESCRIPTIVES

#II – TESTS ELEMENTAIRES SUR VARIABLES QUALITATIVES

#III – TESTS ELEMENTAIRES SUR VARIABLES QUANTITATIVES

#IV – Régression linéaire simple

#V Régression linéaire multiple

#VI – Comparaison de moyennes ANOVA à un facteur

#VII – Comparaison de moyennes ANOVA à deux facteurs

#VIII – Analyse en composante principale

#IX Analyse factorielle des correspondances

#X Classification

Introduction

Ce cours s'adresse principalement aux étudiants de master technologie végétale et plus généralement à tout étudiant en biologie. Compte tenu de la variabilité des cursus antérieurs et du volume horaire restreint, le parti pris est celui de rappeler les principales bases succinctement et de développer certaines techniques statistiques parmi les plus utilisées: régression, analyse de variance, analyse factorielle.

Ce cours est donc loin d'être complet. Il manque volontairement les bases de probabilités et de statistiques accessibles dans tout ouvrage de statistiques ainsi qu'un certain nombre de développement. **La spécificité de ce cours tient en une présentation aussi simple et pratique que possible et à l'utilisation systématique du logiciel :** 

Tout développement théorique est donc proscrit et nous privilégierons l'étude de situations concrètes issues de la biologie par utilisation systématique du logiciel R interfacé avec R commander ou ade4. Nous renvoyons aux sites cités dans les liens pour les approfondissements possibles et souhaitables.

Logiciel  , R-commander, ade4 :

Le cours et les TD seront illustrés à l'aide du logiciel  , version libre du logiciel S-plus. Ce logiciel développé par des statisticiens pour des statisticiens est particulièrement bien adapté à l'analyse statistique descriptive, inférentielle et exploratoire. Pour faciliter son utilisation, nous utilisons un interface appelé R commander. Pour l'analyse factorielle nous recommandons l'usage de ade4 développé par l'université de Lyon.

L'apprentissage de ce logiciel  est sans doute plus délicat que les logiciels commerciaux mais permet à tous de disposer d'un outil gratuit, très performant, en perpétuelle évolution, exploitable dans toutes les circonstances et permettant à l'utilisateur une totale liberté dans ces choix d'analyse.

Quelques liens en statistiques :

http://spiral.univ-lyon1.fr/mathsv/ : mathSV de l'université de Lyon reprend l'ensemble des bases de mathématiques, probabilités et statistiques que devraient posséder les étudiants de biologie. Présentation simple, rigoureuse, adaptée aux biologistes.

http://pbil.univ-lyon1.fr/R/enseignement.html : Site développé par l'Université de Lyon1 sur l'utilisation des statistiques en Biologie, le meilleur site exploitant le logiciel R.

http://www.inra.fr/internet/Departements/MIA/T//schnu/FPstat/FP2.html : Plan de formation à la pratique statistique développé par la formation permanente de l'INRA, bien adaptée à la biologie et pouvant être utilisé à différents niveaux.

http://rfv.insa-lyon.fr/~jolion/STAT/ : Cours de statistiques appliqués assez complet

http://www.fsagx.ac.be/si/ : Ensemble de documents statistiques très intéressants appliqués principalement à l'agronomie.

http://www.cons-dev.org/elearning/stat/index.html : Présentation simple et agréable

http://www.dagnelie.be / Ouvrages en accès libre de Dagnelie (en particulier pour les protocoles expérimentaux)

Remerciements : Merci aux auteurs cités précédemment auxquels j'ai emprunté exemples et analyses.

Méthodes présentées

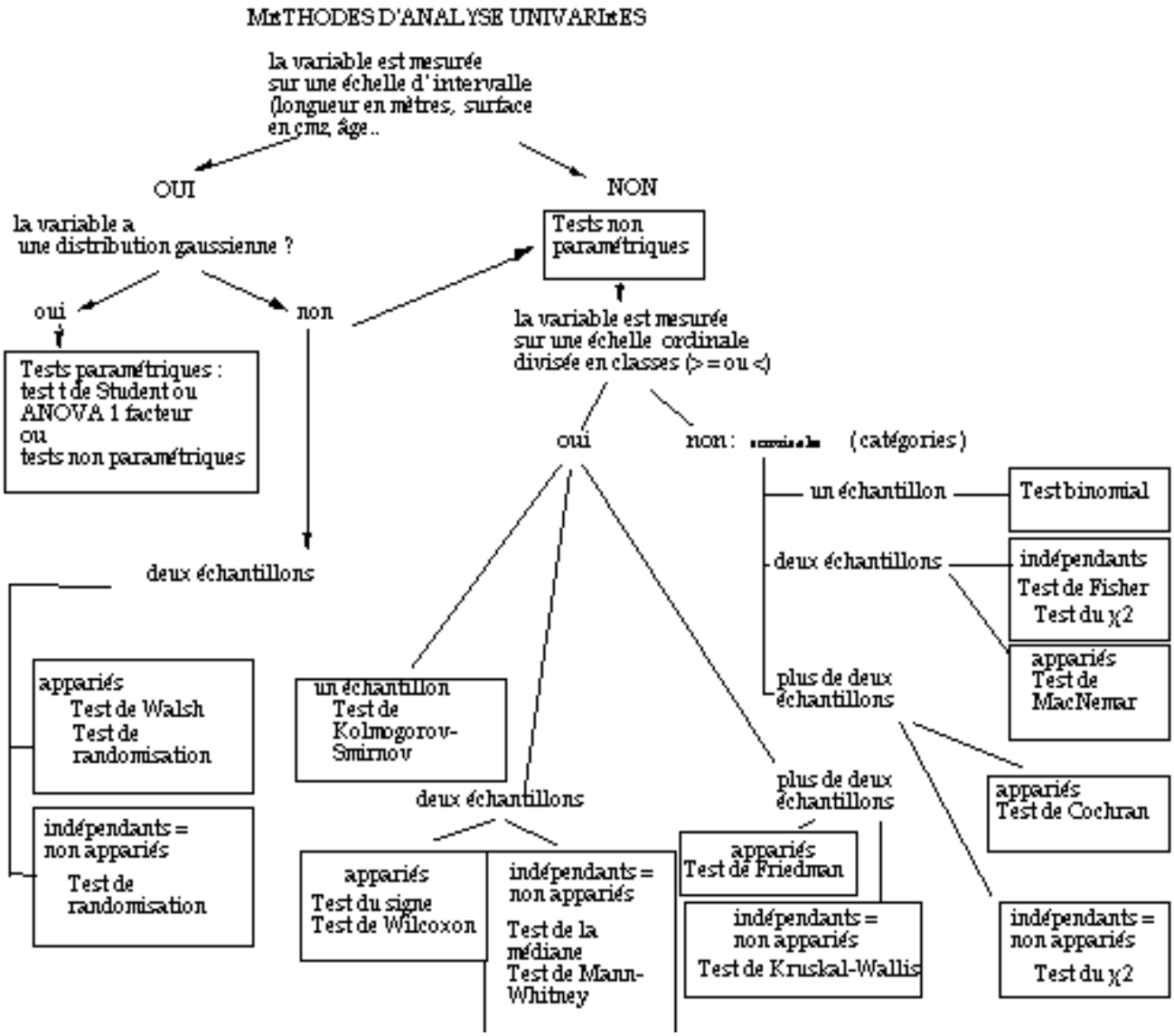
- **Une variable qualitative :**
  - test de conformité à une fréquence,
  - test de conformité du χ^2 à une distribution,
  - Test du χ^2 de comparaison de deux ou plusieurs distributions

- **Deux variables qualitatives** :
  - Test du χ^2 d'indépendance, AFC avec un grand tableau de contingence

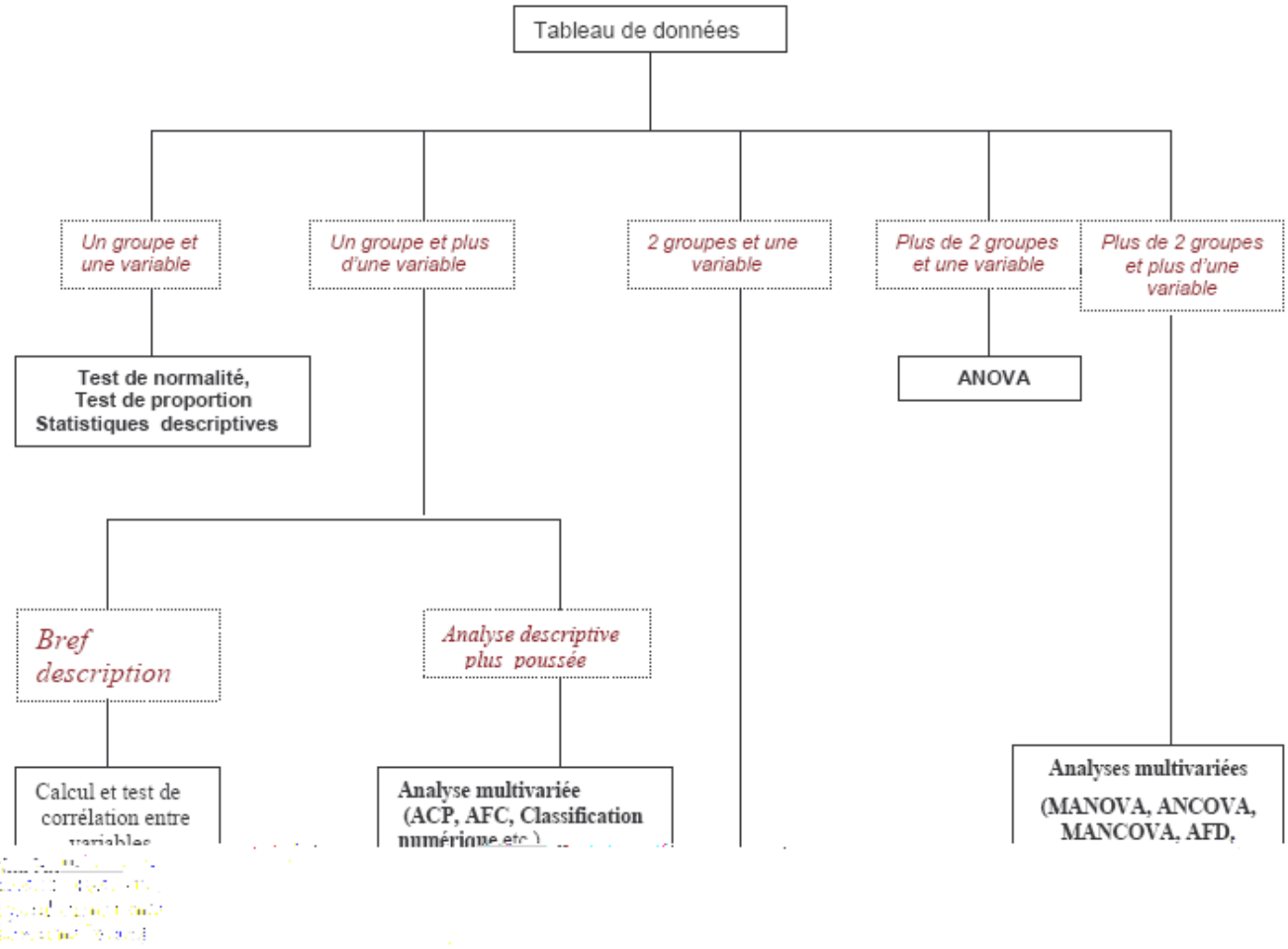
- **Une variable quantitative**
  - Test de conformité à une moyenne avec une population
  - Test t de comparaison de moyennes avec deux populations
  - ANOVA avec plusieurs populations

- **Deux variables quantitatives ou plus**
  - Nuage de points, test de la corrélation entre les deux variables
  - Régression linéaire simple avec deux variables
  - Régression linéaire multiple, ACP, classification avec plus de deux variables

- **Une variable quantitative et une variable ou plusieurs variables qualitatives**
  - ANOVA à un ou deux facteurs

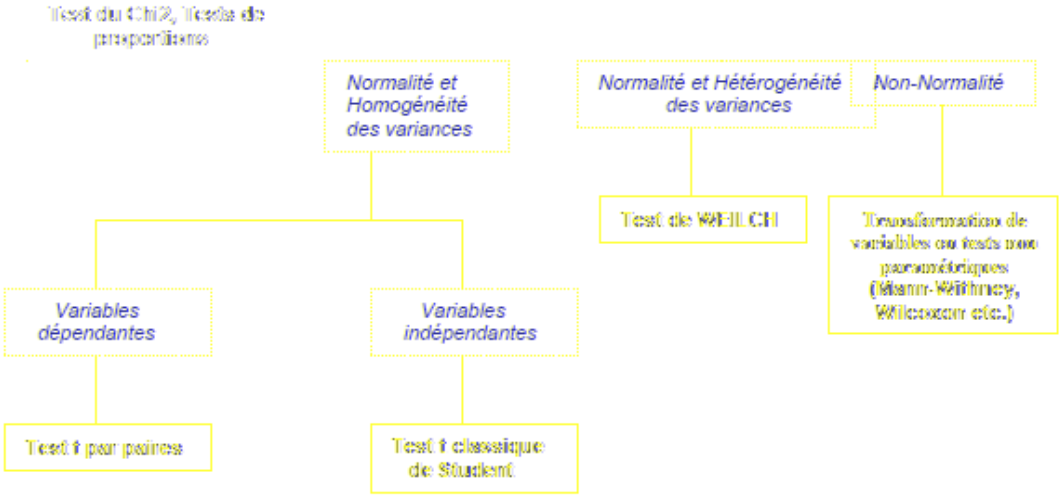


GALAXIE DES METHODES STATISTIQUES



Variable de comptage
Ex. Nombre d'individus de 2 espèces d'oiseaux

Donnée numérique
(Poids des individus de 2 espèces d'oiseaux)



5

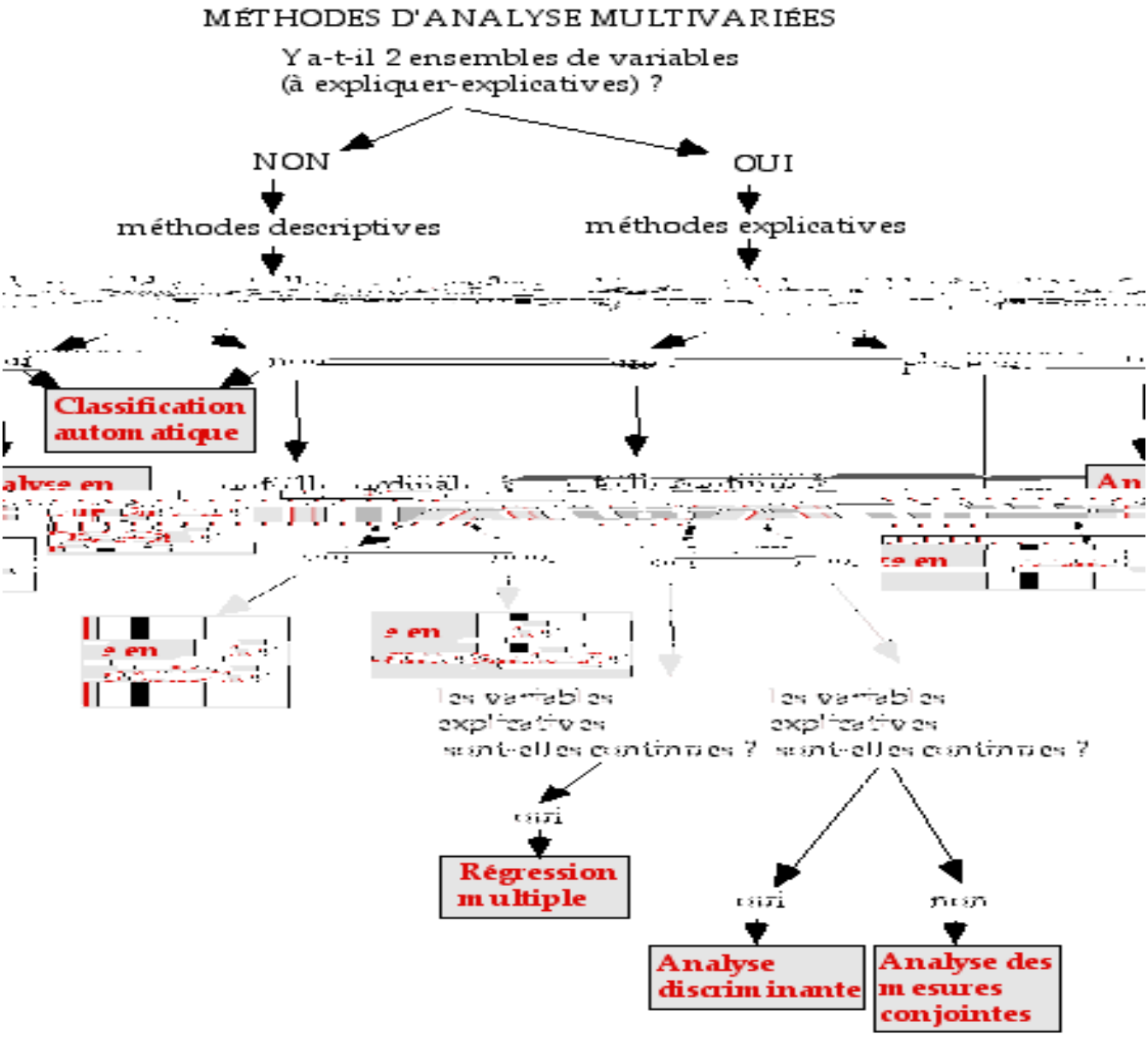
Analyse des données

L'objectif est ici d'analyser de grands tableaux de données. Trois approches sont possibles :

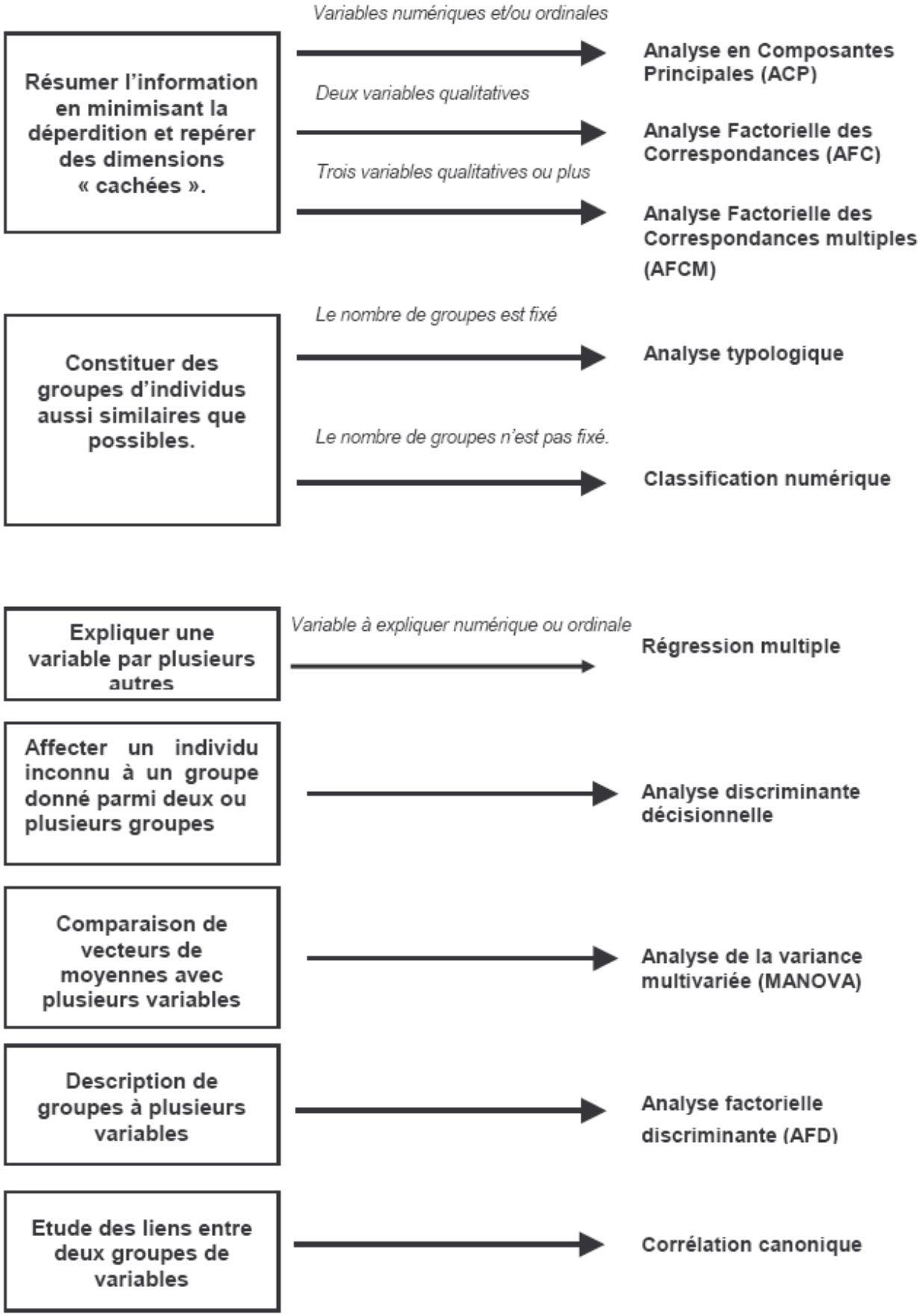
- **les méthodes factorielles** qui consistent à représenter les individus sur un plan, en perdant le moins d'information possible. On peut discerner trois méthodes principales :
  - l'**analyse en composantes principales** (plusieurs variables quantitatives),
  - l'**analyse des correspondances** (deux variables qualitatives, représentées par un tableau de contingence),
  - l'analyse des correspondances multiples (plus de deux variables qualitatives).

- **les méthodes de classification**, qui tentent de regrouper les individus en groupes homogènes. On peut discerner trois méthodes principales :
  - la **classification ascendante hiérarchique** (emboîtement de partitions sous forme d'arbre)
  - les centres mobiles, k-means …(partition en k groupes)
  - les méthodes mixtes (un mélange des deux précédentes)

- **les méthodes discriminantes** qui consistent à discriminer le mieux possible des individus décrits par des variables quantitatives en général et dont on connaît la classification.



Méthodologie de choix d'une méthode multivariée



I STATISTIQUES DESCRIPTIVES

Les statistiques descriptives recouvrent les différentes techniques de description des données, synthèse sous forme de paramètres ou de tableaux, représentations graphiques....

Pour les grands tableaux, les techniques peuvent devenir plus complexes. Elles seront abordées ultérieurement :

- Analyse en composantes principales (ACP) dans le cas de plusieurs variables quantitatives,
- Analyse des correspondances (AFC) dans le cas de grands tableaux de contingence,
- Classification (CAH)

Les statistiques descriptives sont importantes pour présenter les données, réaliser une première analyse et interprétation ...

On opppose les statistiques descriptives aux statistiques inférentielles dont l'objectif est de mettre en place des règles de décision afin de réaliser des tests statistiques. Nous aborderons ce domaine dans les chapitres suivants.

Fiche 1 – Variable qualitative

Un caractère est dit qualitatif

- s 'il est mesuré dans une échelle **nominale** (couleur, marque), les modalités ne sont pas hiérarchisées,
- s'il est mesuré dans une échelle **ordinale** (stades d'une maladie, classes d'âge ...), les modalités sont alors hiérarchisées.

Les graphiques et les tests dépendront de la nature des caractères.

Paramètre statistique Fréquence $f = \frac{effectif\ observé}{effectif\ total}$

Réprésentations graphiques :

nominale : diagramme circulaire

ordinale : diagramme en bâton ou rectangle.

Etude d'un exemple : Après inoculation d'un pathogène dans une plante, le stade de la maladie a été mesuré sur 320 plantes. Le stade est noté de I (aucun symptôme) à V (plante morte). Le résultat de l'étude est décrit dans le tableau suivant :

stade	I	II	III	IV	V	Total
Effectif	15	55	215	25	15	
fréquence						

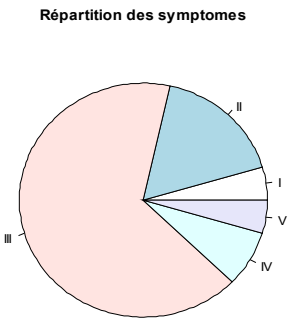
a. Sous quelle forme étaient les données brutes ?

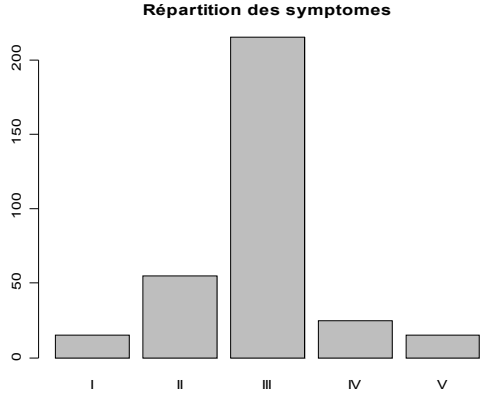
b. Créer un tableau patho1 *[donnees nouveau]*.
On notera stade et effectif les colonnes.

c. Calculer les fréquences.

d. Représenter sous R graphiquement ce tableau.

```
>attach(patho1)
>pie(effectif,labels=stade)
>barplot(effectif,names.arg=stade,main="Répartition des symptomes")
```





Fiche 2 – Couple de variables qualitatives

Etude d'un exemple : Le tableau résume la présence ou l'absence d'une infection bactérienne en fonction de l'utilisation ou non d'une antibiothérapie ou d'un placebo.

a) Compléter le tableau en ajoutant les fréquences ainsi que les effectifs marginaux.

b) Construire le tableau sous l'hypothèse d'indépendance entre les deux caractères.

Tableau observé

	Antibio		Placebo		Total	
	n	*f*	*n*	*f*	*n*	*f*
absence	75		27			
présence	10		29			
total						

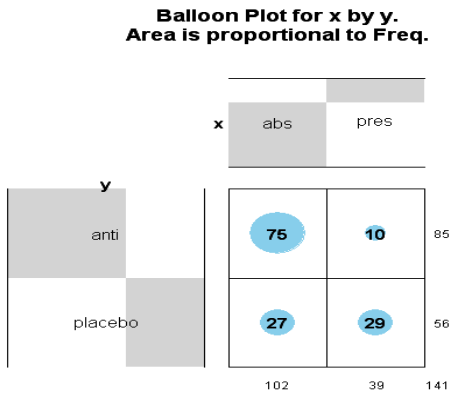
Tableau théorique sous l'hypothèse d'indépendance

	Antibio		Placebo		Total	
	n	*f*	*n*	*f*	*n*	*f*
absence						
présence						
total						

Représentations graphiques

On peut utiliser un diagramme en bâtons bidimensionnel ou la représentation suivante dit balloon :

```
library(gdata)
library(gtools)
library(gplots)
patho2=as.table(matrix(c(75,27,10,29),ncol=2,byrow=TRUE))
rownames(patho2) = c("abs","pres")
colnames(patho2) = c("anti","placebo")
balloonplot(patho2, dotsize = 10)
```



Fiche 3 – Variable quantitative

On appelle variable quantitative une variable qui se mesure dans une échelle discrète ou continue.

1. Tableau de répartition, calcul de fréquence, histogramme

En présence d'une variable continue, les observations sont souvent regroupées par classe. Le nombre de classe N à construire est défini pour que les classes aient en général une amplitude identique et un effectif minimal suffisant.
Ce nombre N peut être calculé empiriquement à partir :

- de la règle de Sturge $N=1+3{,}3\ \log_{10} n$ ou,
- de la règle de Yale $N= 2{,}5\ \sqrt[4]{n}$.

On détermine ensuite les classes $]x_i,x_{i+1}]$ avec $x_{i+1}-x_i = \dfrac{amplitude}{N}$, en arrondissant si besoin.

Exemple : La mesure de la hauteur de 15 étudiants donne les résultats suivant :
172 175 176 181 185 183 168 178 175 172 179 177 182 187 174
Présenter un tableau synthétique et le représenter graphiquement.

2. Paramètres de position et de dispersion

La définition de la moyenne μ et de la variance σ^2 est :

$$\boldsymbol{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i} \qquad \text{et} \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Dans la pratique, on utilise un échantillon de la population pour estimer ces coefficients. On note $\overline{x}$ et s les estimateurs de ces paramètres. L'estimateur de la variance σ^2 calculé en divisant par n étant biaisé, on l'estime par :

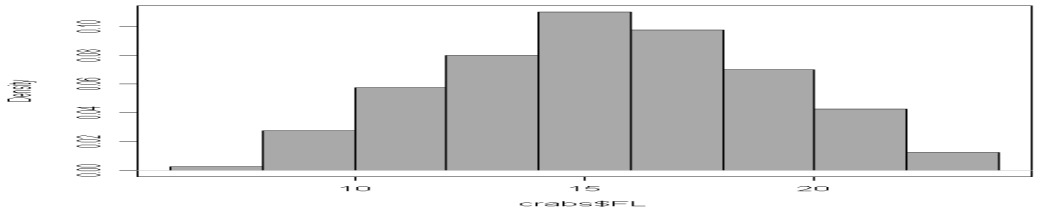
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

La médiane est le réel séparant la population en deux groupes de même effectif. On donne une définition analogue pour les quartiles, déciles ...

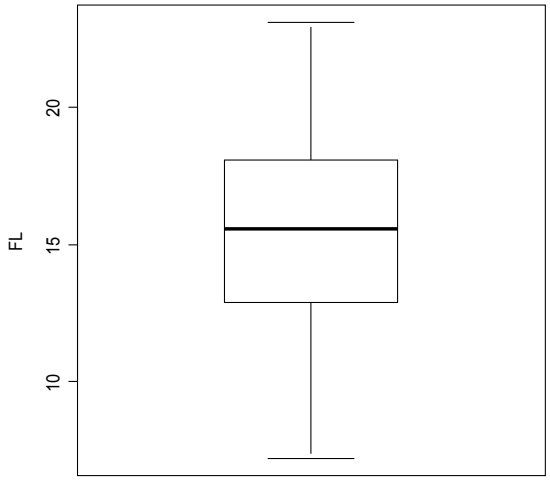
Exemple : On mesure le poids de 10 graines en mg : 1 1,1 1,2 1,5 1,2 1 0,9 0,8 1,1 1,3
Calculer la moyenne, la médiane et l'écart-type de cet échantillon.

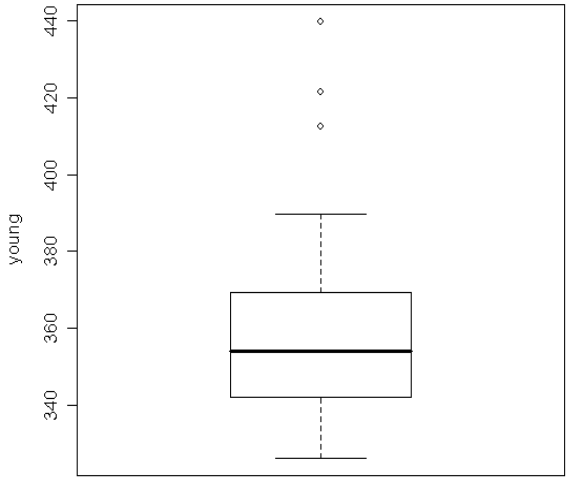
3. Représentations graphiques : Histogramme, boxplot ...

Histogramme



Boxplots

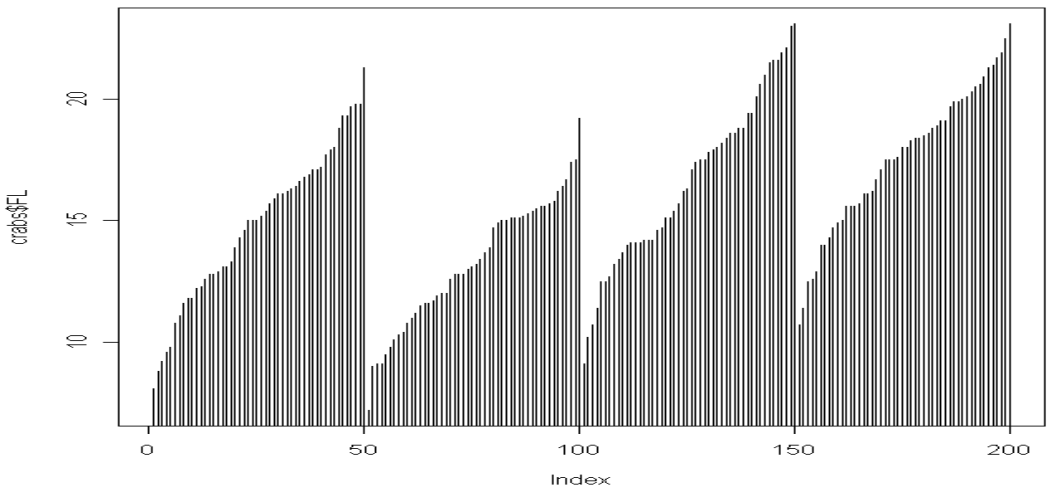




Tige et feuille (FL de crabs)

```
      n: 200
 1     7 | 2
 3     8 | 18
12     9 | 011125688
20    10 | 12347788
33    11 | 0124456667889
51    12 | 002355566678888899
64    13 | 0111223447799
82    14 | 000111222336677799
(29)  15 | 00000011111223444566666777789
89    16 | 11112222334467789
72    17 | 111124455555567899
54    18 | 000023444566688889
36    19 | 1123344778899
23    20 | 01135669
15    21 | 0334566799
 5    22 | 15
 3    23 | 011
```

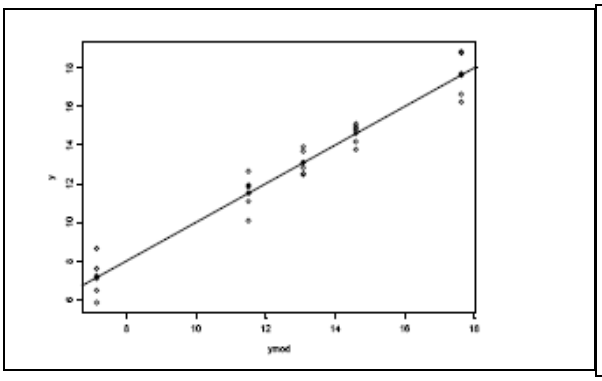
`Indexé :`



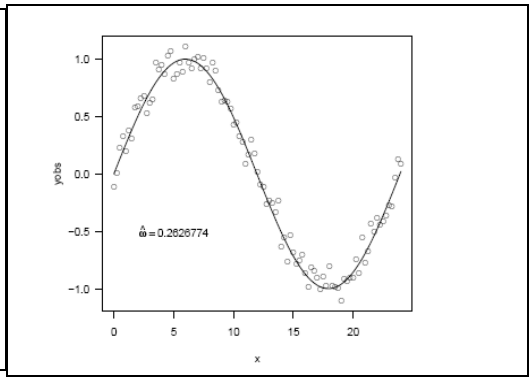
Fiche 4 – Couple de variables quantitatives - Corrélation

Sur chaque individu de l'échantillon sont mesurées maintenant deux variables quantitatives X et Y. Après un analyse de chacune de ces variables (fiche Variable quantitative), on procède à l'étude de la relation entre ces deux variables. Existe-t-il une liaison entre les deux variables ? linéaire ou non linéaire ?

La description de la liaison entre les deux variables se fait en premier lieu par un examen du nuage de points (x_i,y_i). L'observation du nuage permet de justifier la linéarité ou non de la relation (des tests existent mais nécessitent des conditions d'application rarement atteintes).



relation fonctionnelle linéaire



relation fonctionnelle non linéaire

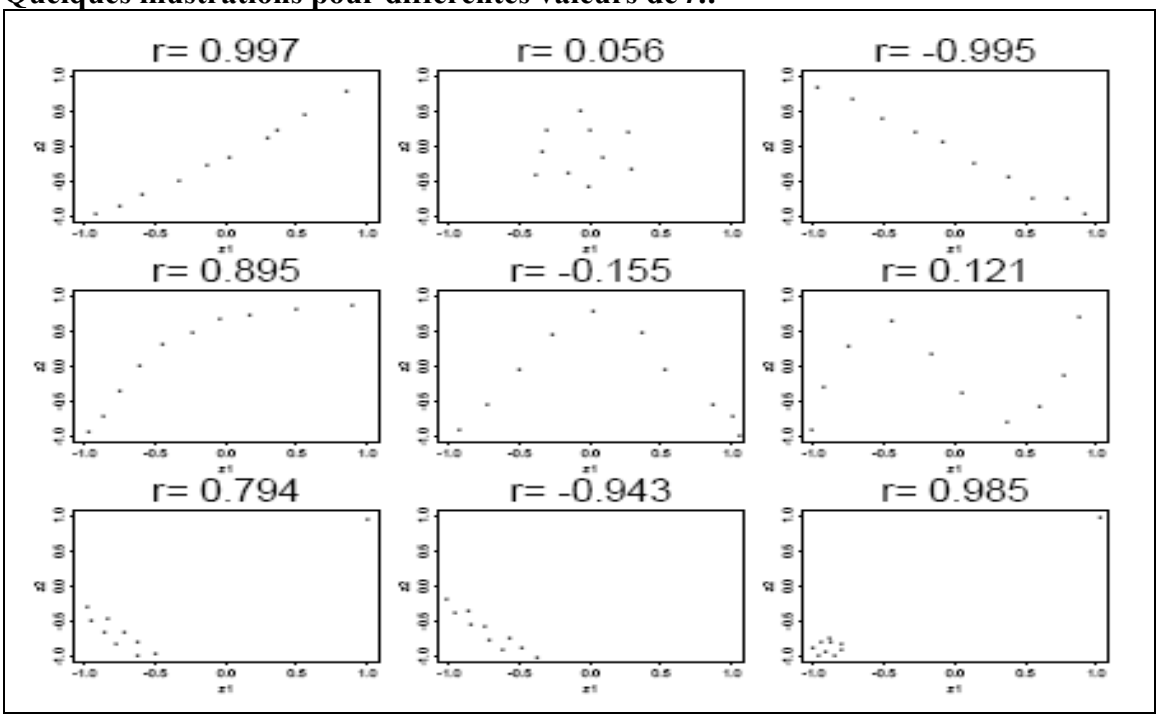
On procède dans le cas où la liaison est linéaire au calcul d'un paramètre de la liaison linéaire entre les deux variables, le coefficient de corrélation de Pearson :

$$r = \frac{\sum \left((x_i - \overline{x})(y_i - \overline{y}) \right)}{\sigma_x \sigma_y}$$

Ce coefficient est compris entre -1 et 1.

- Plus r est proche de -1 ou de 1, plus les variables sont respectivement négativement ou positivement corrélées.
- Si r est proche de 0, les variables ne sont pas corrélées mais il n'y a pas nécessairement indépendance des variables.

Quelques illustrations pour différentes valeurs de *r*.:

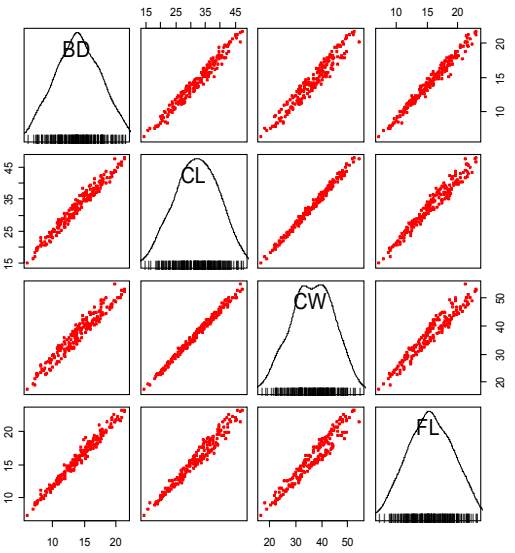


Il existe un test statistique permettant de tester l'hypothèse nulle H_0 « $r = 0$» (fiche #correlation) dans le cas d'une liaison linéaire. Dans le cas où H_0 est rejetée, on réalise en général une régression linéaire (chapitre III) de Y en X ou de X en Y.

Etude de l'exemple crabs (library MASS)**.**

a. Etudier les relations entre les variables BD, FL, CL, CW.

Matrice des nuages de points

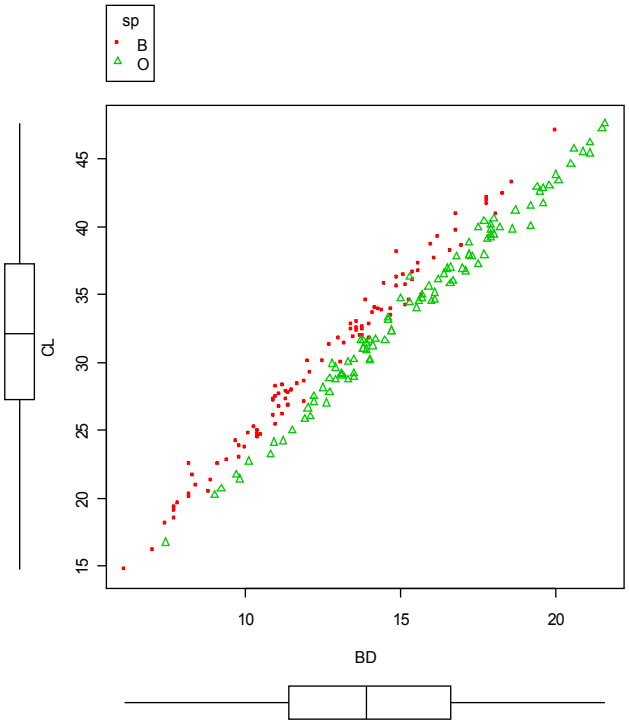


Matrice des corrélations (stat - resume)

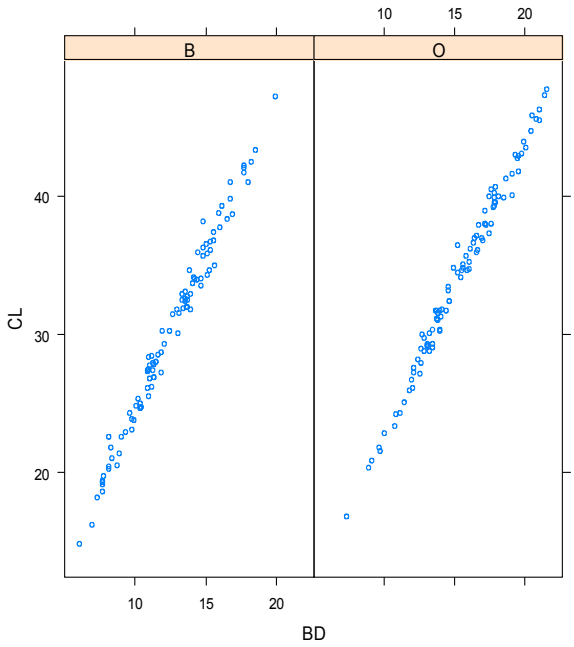
	BD	CL	CW	FL
BD	1.0000000	0.9832038	0.9678117	0.9876272
C CL	0.9832038	1.0000000	0.9950225	0.9788418
CW CW	0.9678117	0.9950225	1.0000000	0.9649558
FL FL	0.9876272	0.9788418	0.9649558	1.0000000

b. La relation est-elle similaire en fonction de l'espèce ?

Nuages de points en fonction de sp.



Graphes en ligne en fonction de sp.



Fiche 5 – Couple variable quantitative – variable qualitative

Il arrive fréquemment que l'on étudie une variable quantitative dans différentes populations. L'objectif est donc ici d'observer d'éventuelles différences entre les populations (ou modalités d'un facteur).

Exemple : On reprend l'exemple de la variable FL de crabs. On étudie les différence entre espèces.

Statitiques - summary

```
    mean       sd  0%    25%   50%    75% 100%   n
B 14.056 3.019610 7.2 11.800 14.45 16.125 21.3 100
O 17.110 3.275575 9.1 14.525 17.50 19.475 23.1 100
```

Boîte de dispersion

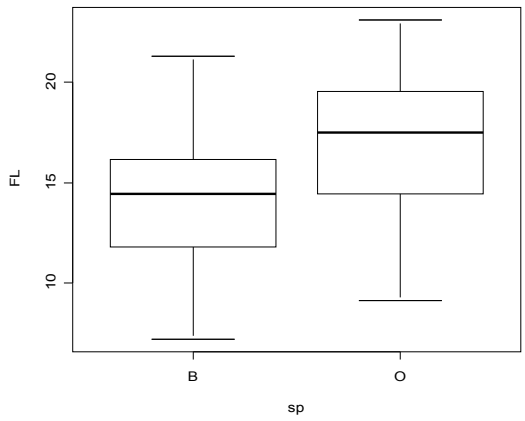
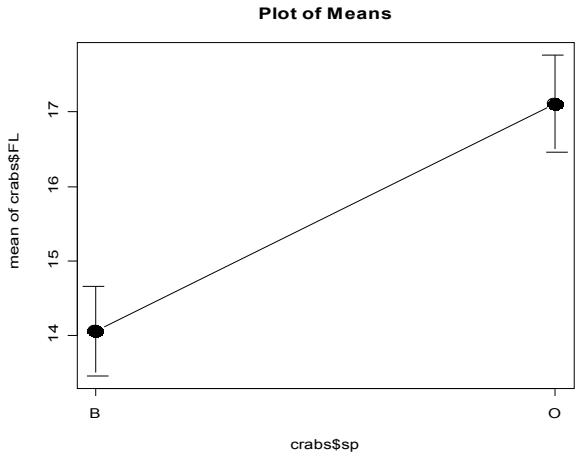
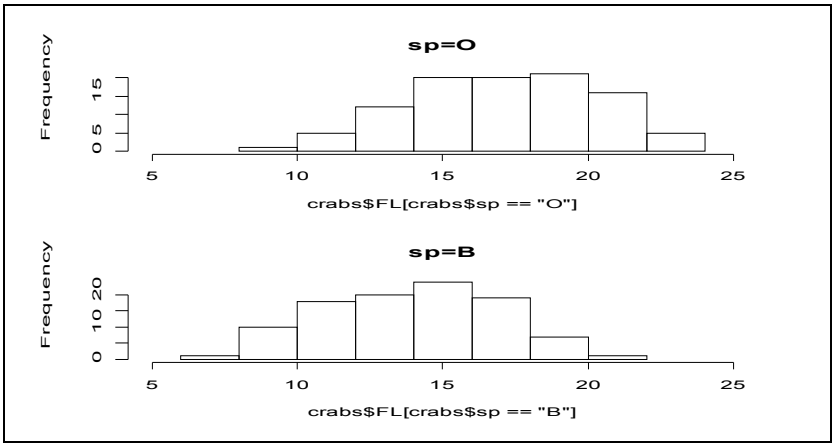


Diagramme de la moyenne avec IC 95%



Histogramme (noter que les échelles en abscisse sont les mêmes pour comparer les distributions)

```
> par(mfrow=c(2,1))
> hist(crabs$FL[crabs$sp=="O"],main="sp=O", xlim=c(5,25))
> hist(crabs$FL[crabs$sp=="B"],main="sp=B",xlim=c(5,25))
```



Fiche 6 – Estimation ponctuelle d'une moyenne et d'un écart-type, Intervalle de confiance

On dispose en général d'un échantillon prélevé dans une population pour laquelle la variable quantitative X a pour moyenne μ et variance σ^2 inconnues.

Règle pour l'estimation ponctuelle : Soit une variable quantitative X mesurée sur un échantillon de n individus,

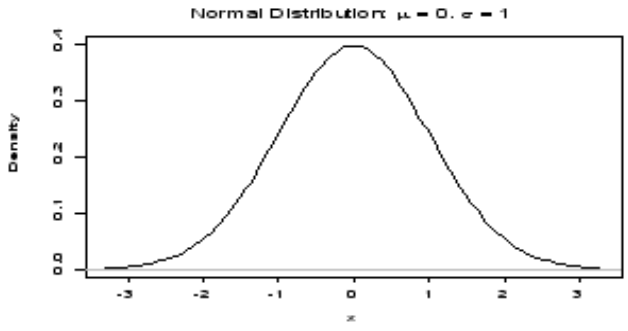
- la moyenne μ est estimée par $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- la variance σ^2 est estimée par $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$

D'un échantillon à l'autre, les estimations ponctuelles vont variées d'autant plus que sa taille n est faible. Pour affiner l'estimation de ces paramètres, on détermine alors un intervalle de confiance dans lequel les valeurs réelles μ ou σ^2 ont une probabilité déterminée à l'avance de se trouver.
Cet intervalle de confiance, noté *IC,* permet ainsi de prendre en compte la variabilité de l'estimation ponctuelle.

Exemple : Le poids des graines de radis est supposé suivre une loi normale. On a pesé 5 graines et obtenu les valeurs 1,2; 1,4; 1,5; 1,6; 1,8 en mg.
Estimer la moyenne et la variance du poids

Propriété des estimateurs $\overline{x}$ et s_X^2 :

- **cas 1** : $n < 30$ **et la variable** X **suit une loi normale** (fiche #Normalité)
  - Si σ^2 est connue, alors $u = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{(n)}}}$ suit la loi normale centrée réduite
  - Si σ^2 est inconnue, alors $t = \frac{\overline{x} - \mu}{\frac{s_X}{\sqrt{(n)}}}$ suit la loi de Student à $n-1$ degrés de liberté (ddl).
- **Cas 2 : Pour** $n > 30$ **(application du théorème central limite)**
  - $t = \frac{\overline{x} - \mu}{\frac{s_X}{\sqrt{(n)}}}$ suit la loi normale centrée réduite



Construction d'un intervalle de confiance :

On recherche toutes les valeurs possibles de μ pour lesquelles $t = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{(n)}}}$ soit compris entre $t_{\alpha/2}$ et $t_{1-\alpha/2}$

(par symétrie $t_{\alpha/2} = -t_{1-\alpha/2}$).

$t_{1-\alpha/2}$ est la valeur dans la table pour laquelle $P(t<t_{1-\alpha/2})=1-\alpha/2$ donc $P(t_{\alpha/2} < t < t_{1-\alpha/2})=1-\alpha$.

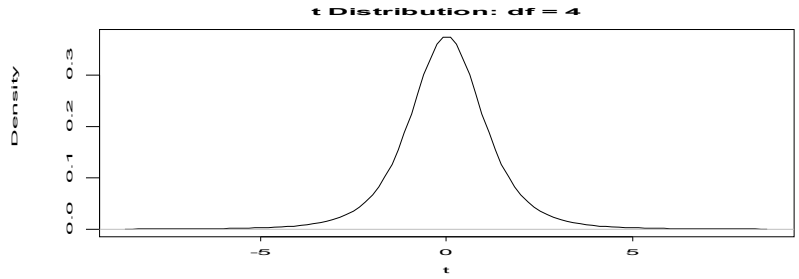
On a alors l'intervalle de confiance à 1-α pour μ : $\bar{x} - t_{1-\alpha/2} \frac{s_X}{\sqrt{(n)}} < \mu < \bar{x} + t_{1-\alpha/2} \frac{s_X}{\sqrt{(n)}}$

Pour α = 5%, ce résultat signifie que "la vraie moyenne, μ", de la population a une probabilité de 95% d'être dans cet intervalle. On notera par commodité cet intervalle de confiance IC_{95}.

Exemple : Reprendre l'exemple précédent et en déduire un intervalle à 95% de la moyenne.

```
mean        sd  0% 25% 50% 75% 100% n
  1.5 0.2236068 1.2 1.4 1.5 1.6  1.8 5
```

Distribution de t :



quantile de la loi de Student pour p=0,975 et 4 ddl (distribution – continuous – student - quantile)
[1] 2.776445

Calcul à la main de IC_95 *:*

Calcul avec R de IC_95 *:* statistiques – moyenne

One Sample t-test data: graine1$mg

```
t = 0, df = 4, p-value = 1

alternative hypothesis: true mean is not equal to 1.5

95 percent confidence interval:    1.222355 1.777645
sample estimates:                  mean of x : 1.5
```

Fiche 7 – Estimation ponctuelle d'une fréquence, Intervalle de confiance

Si une population contient une proportion f d'individus possédant un caractère donné, l'estimateur de ce paramètre est la fréquence du caractère dans l'échantillon, noté $\hat{f}$.

La fréquence $\hat{f}$ dans un échantillon de n individus (n assez grand, >100 et $\hat{f}$ entre 0,1 et 0,9) suit la loi normale $\mathcal{N}(f, \frac{f(1-f)}{n})$.

Dans les autres cas, n<100 ou $\hat{f}$ < 0,1, il faut utiliser un modèle exact (binom.test dans R).

Pour construire l'intervalle de confiance de f, on reprend un raisonnement analogue à la fiche précédente. Dans la pratique, comme on ne connaît pas f, on la remplace par $\hat{f}$.

Propriété : Pour un échantillon tel que *n* >100 et *n* $\hat{f}$ >10 et n(1-$\hat{f}$)>10,
L'intervalle de confiance à 1-α d'une proportion est :

$$\left]\ \hat{f} - u_{1-\alpha/2}\sqrt{\left(\frac{\hat{f}(1-\hat{f})}{n}\right)};\ \hat{f} + u_{1-\alpha/2}\sqrt{\left(\frac{\hat{f}(1-\hat{f})}{n}\right)}\right[$$

où $u_{1-\alpha/2}$ représente le quantile de la loi normale centrée réduite. Pour α=5%, $u_{1-\alpha/2}$=1,96.

Exemple : Un laboratoire d'agronomie a effectué une étude sur le maintien du pouvoir germinatif des graines de *Papivorus subquaticus* après une conservation de 3 ans. Sur un lot de 80 graines, 47 ont germé. Calculer la probabilité de germination des graines de *Papivorus subquaticus* après trois ans de conservation avec un coefficient de confiance de 95%.

Calcul à la main de IC_{95} :

Calcul avec R : `prop.test(47,80,p=47/80)`

```
1-sample proportions test without continuity correction

data:  47 out of 80, null probability 47/80

X-squared = 0, df = 1, p-value = 1

alternative hypothesis: true p is not equal to 0.5875

95 percent confidence interval:     0.4780404 0.6889414
sample estimates:                  p = 0.5875
```

II – TESTS ELEMENTAIRES SUR VARIABLES QUALITATIVES

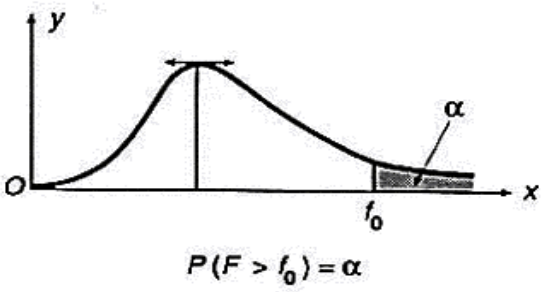
La majorité des tests élémentaires repose sur le principe suivant :

1. On définit une hypothèse nulle notée H_0 contre l'hypothèse alternative H (H_0 n'est pas vérifiée). Le test a pour objectif d'accepter ou de rejeter H_0 avec un risque connu à partir des données dont on dispose.

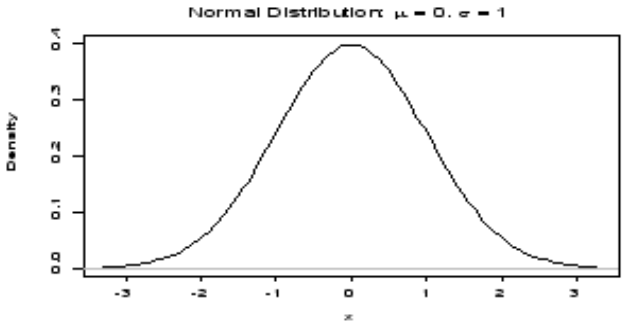
2. On détermine alors une statistique qui est une variable aléatoire dépendant des données

   - Sous H_0, cette variable suit une loi de probabilité connue.

   - On détermine alors l'intervalle dans lequel doit tomber la statistique avec une probabilité donnée, 95% le plus souvent.



Unilatéral (test F, χ^2)



Bilatéral (test t)

3. On définit alors la règle de décision suivante :

   - Si la statistique tombe dans l'intervalle, on accepte H_0.
     - Attention, cela ne veut pas dire que H_0 est vraie, mais que les données et le test ne permettent pas de voir un écart significatif à H_0.

   - Si la statistique ne tombe pas dans l'intervalle, on rejette H_0 avec un risque de première espèce estimé.
     - On peut donc rejeter H_0 alors que H_0 est vraie, mais on connait le risque que l'on prend.
     - En général, les ordinateurs calculent la valeur p de ce risque. On concluera alors que l'on rejette H_0 avec un risque de première espèce p. Plus p est faible (<0,05), plus le rejet est significatif.

4. La validité du test dépend souvent de certaines conditions sur les données, appelées conditions d'utilisation du test. Ces conditions doivent être vérifiées avant de pouvoir conclure.

5. Bilan

   Après avoir défini H_0 et le test en conséquence, on calcule la statistique et p la probabilité d'observer une valeur supérieure à cette statistique sous H_0 (en valeur absolue).
   La conclusion dépend alors de p:

   - Si $p > \alpha$ (p>0,05), on accepte H_0.
   - Si $p < \alpha$ (p<0,05), on rejette H_0 avec un risque de première espèce égal à α (ou p).

Fiche 8 – Comparaison d'une proportion à une référence

Objectif : Tester si une fréquence est conforme à une fréquence attendue.

Soit une variable *X* prenant deux modalités (absence/présence).

Le but est de savoir si un échantillon de fréquence observée $\hat{f} = \frac{cas\ favorables}{total}$, estimateur de *f*, appartient à une population de référence connue de fréquence f_0 (H_0 vraie) ou à une autre population inconnue de fréquence $f \neq f_0$ (*H* vraie).

Données : Une variable qualitative avec deux modalités.

Hypothèse nulle : La fréquence est f_0.

Principe du test :

On calcule la statistique $u = \frac{(f - f_0)}{\sqrt{\frac{f_o(1-f_0)}{n}}}$ qui suit sous H_0 la loi normale centrée réduite.

On calcule alors la probabilité p d'observer une valeur supérieure ou égale sous H_0 (en valeur absolue).

Conditions d'utilisation :

- Le test est applicable si n $\hat{f}_0 \geq 10$ et $n(1-\hat{f}_0) \geq 10$ (approximation par la loi normale). Si cette condition n'est pas vérifiée, on utilise un test exact (binom.test).
- Les individus sont indépendants.

Test : On teste H_0 $f = f_0$ contre H $f \neq f_0$.

- Si p >0,05, on accepte H_0.
- Si p <0,05, on rejette H_0 avec un risque de première espèce p.

Exemple : La proportion de plantes résistantes observée est de 57 résistantes sur 65. Ce taux est-il significativement différent de 0.9 ?

```
> prop.test(57,65,0.9) (essayer avec binom.test)
      1-sample proportions test with continuity cor.
      data:  57 out of 65, null probability 0.9

X-squared = 0.17, df = 1, p-value = 0.679

95 percent confidence interval:     0.7663630 0.9416156
sample estimates:                   p = .876923
```

Fiche 9 –Test de conformité à une distribution : test du Chi2 χ^2 (génétique)

Objectif : On considère une variable X prenant k modalités, $k > 2$ (pour $k = 2$, voir la fiche précédente).
L'objectif du test est de vérifier que les modalités se distribuent suivant des probabilités attendues. On utilise un tel test en génétique par exemple pour vérifier :

- les lois de Mendel, (répartition $\frac{1}{4}, \frac{2}{4}, \frac{1}{4}$ pour F2)
- le modèle de Hardy Weinberg. (répartition $p_1^2, 2\,p_1\,p_2, p_2^2$) .

Données : Les données sont regroupées dans un tableau de contingence de la forme :

Variable qualitative	Modalité 1	Modalité 2	....
effectif	n^1_{obs}	n^2_{obs}	....

Hypothèse nulle : Les fréquences observées sont conformes aux probabilités attendues.

Principe du test :

Le principe du test du χ^2 est d'estimer à partir d'une loi de probabilité connue (ou estimée à partir de l'échantillon), les effectifs théoriques pour les différentes modalités du caractère étudié et les comparer aux effectifs observés dans un échantillon. Deux cas peuvent se présenter :

- soit **la loi de probabilité est spécifiée *a priori*** car elle résulte par exemple d'un modèle déterministe tel que la distribution mendélienne des caractères.
- soit **la loi de probabilité théorique n'est pas connue *a priori*** et elle est déduite des caractéristiques statistiques mesurées sur l'échantillon (estimation de p_1 et p_2 dans le cas du modèle de Hardy Weinberg).

Le test du χ^2 consiste à mesurer l'écart qui existe entre la distribution théorique et la distribution observée et à tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.

- On calcule les effectifs théoriques $n^1_{\text{th } eor}$, $n^2_{\text{th } eor}$ ... attendus sous l'hypothèse où la distribution est conforme à celle attendue.
- On calcule ensuite la statistique : $\widehat{\chi^2} = \sum_{i=1}^{k} \frac{\left(n^i_{obs} - n^i_{\text{th } eor}\right)^2}{n^i_{\text{th } eor}}$ qui suit sous H_0 la loi du χ^2 à ν degrés de liberté. On rejette alors H_0 dans le cas où $\widehat{\chi^2}$ dépasse la valeur seuil $\chi^2_{1-\alpha}(\nu)$.
- Le nombre de ddl ν est $k - c$, k représente le nombre de modalités et c celui des contraintes.
  - Si la distribution théorique est entièrement connue *a priori* (lois mendeliennes), la seule contrainte est que la somme des probabilités vaut 1, donc $\nu = k - 1$.
  - Sinon, il faut estimer des probabilités sur l'échantillon et augmenter d'autant les contraintes. Par exemple avec le modèle de Hardy Weinberg, la somme des probabilités vaut 1 et il faut estimer p_1, soit c=2, donc $\nu = k - 2$.

Test : On teste l'hypothèse H_0 (conforme à la distribution attendue)

-si $\hat{\chi}^2 < \chi^2_{1-\alpha}(v)$, on accepte H_0

-sinon on rejette H_0 avec un risque de première espèce α (ou p).

Conditions d'application : Les effectifs théoriques doivent être supérieurs à 5 ($n^i_{\text{th } eor} \geq 5$). Dans le cas contraire, on peut regrouper les classes les plus faibles, utiliser un test du χ^2 corrigé, utiliser le test exact de Fisher...

Exemple 1 : Soit le locus biallélique codant pour la glucose 6 phosphate déhydrogénase (G6PDH), enzyme participant au métabolisme énergétique, l'analyse électrophorétique des génotypes chez l'anophèle, vecteur de la malaria, donne la répartition suivante : FF = 44, FS = 121, SS = 105.
La répartition des génotypes est-elle conforme au modèle de Hardy-Weinberg ?

Exemple 2 : Les résultats observés de Mendel sur le croisement de pois Jaune-Rond et Vert Ridé en F2 ont été :

Jaune Ronde : 315 Jaune Ridée : 101 Verte Ronde : 108 Verte Ridée : 32

1. Quels sont les allèles dominants, le génotype en F1 ?
2. Les lois de Mendel sont-elles vérifiées ?
3. Quels sont les effectifs attendus ?

```
> chisq.test(c(315,101,108,32),p=c(9/16,3/16,3/16,1/16))
          Chi-squared test for given probabilities
          data:  c(315, 101, 108, 32)

X-squared = 0.47, df = 3, p-value = 0.9254
```

Fiche 10 – Comparaison de deux ou plusieurs distributions : test du Chi2 χ^2

Objectif : Soit une variable qualitative X prenant deux modalités (absence/présence) ou plusieurs modalités. Le but est de savoir si deux échantillons ou plusieurs échantillons ont une même distribution.
Cet objectif revient à faire un test d'indépendance, la distribution entre les modalités est indépentante de la population dont est issue l'échantillon (fiche Chi2).

Données : Les données sont regroupées dans un tableau de contingence de la forme :

Variable qualitative	Modalité 1	Modalité 2	.....
Effectif population 1	n_{obs}^{11}	n_{obs}^{12}	
Effectif population 2			
...			n_{obs}^{ij}

Hypothèse nulle : La distribution est la même dans les différentes populations. Dans le cas où il n'y a que deux modalités, H_0 devient « la fréquence est la même dans les différentes populations ».

Principe du test :
Le principe est le même que pour le test du χ^2 (fiche Chi2).

Test : On teste H_0 , la distribution est la même dans les différentes populations contre H il existe des différence entre les populations.

- Si p >0,05, on accepte H_0.
- Si p <0,05, on rejette H_0 avec un risque de première espèce p.

Conditions d'utilisation: L'effectif théorique doit être supérieur à 5.

Exemple 1 : Deux préparations à un concours sont prévues. Les taux de réussite sont de 126 sur 180 candidats et de 129 sur 150 candidats respectivement. Les taux sont-ils semblables ?

Exemple 2 : La répartition des groupes sanguins dans trois communautés est présentée ci-dessous. Peut-on considérer que les proportions sont égales ?

	A	B	AB	O
France	54	14	6	55
Roumanie	45	14	8	32
Proche-Orient	33	34	12	33

Calculs avec R :

```
Pearson's Chi-squared test

X-squared = 11.9266, df = 1, p-value = 0.0005534
```

Fiche 11 – Test d'indépendance : test du Chi2 χ^2

Objectif : Le test du χ^2 est largement utilisé pour l'étude de l'indépendance entre deux caractères qualitatifs. La présentation des résultats se fait sous forme d'un tableau de contingence à deux entrées. Chaque entrée représente les modalités d'une des variables. On détermine alors le tableau attendu sous l'hypothèse d'indépendance (fiche #Couple de variables qualitatives)

Données : Deux variables qualitatives sont mesurées sur n individus puis présentées sous forme d'un tableau de contingence (tableau à deux entrées) :
Par exemple :

		tabac		
cancer		présence	absence	total
	présence			
	absence			
	total			

Hypothèse nulle H_0 : Les deux caractères sont indépendants

L'indépendance entre deux caractères qualitatifs signifie que les modalités d'un caractère se distribuent de la même façon pour chacune des modalités de l'autre caractère.
Par exemple, la couleur des yeux est-elle indépendante de la couleur des cheveux ? Si oui, les caractères sont indépendants et on a autant de chance d'avoir les yeux bleus sachant qu'on a les cheveux blond ou qu'on a les cheveux brun.
Ce test prend toute son importance dans la recherche des facteurs d'une maladie par exemple, indépendance entre cancer du poumon et le tabac, cancer du foie et alcool…(travail et réussite)

Principe du test : On calcule les effectifs théoriques sous l'hypothèse H_0. Les effectifs marginaux (totaux à la marge en ligne ou en colonne) et fréquences marginales du tableau restent inchangés.

$$n^{ij}_{\text{th } eor} = \frac{n^{i\cdot}_{obs} \times n^{\cdot j}_{obs}}{n}$$

avec n^{ij}_{theor} l'effectif théorique,
$n^{i\cdot}_{obs}$ et $n^{\cdot j}_{obs}$ les effectifs marginaux ligne et colonne,
n l'effectif total.

On calcule alors la statistique $\hat{\chi}^2 = \sum_{ij} \frac{\left(n^{ij}_{\text{obs}} - n^{ij}_{\text{th } eor}\right)^2}{n^{ij}_{\text{th } eor}}$

Sous H_0, cette statistique suit la loi du χ^2 à v= (l-1)(c-1) ddl avec l le nombre de lignes et c le nombre de colonnes.

Test : On teste l'hypothèse H_0 "indépendance des deux caractères" contre H "dépendance entre les deux caractères" :

-si $\hat{\chi}^2 < \chi^2_{1-\alpha}(v)$, on accepte H_0

-sinon on rejette H_0 avec un risque de première espèce α (ou p).

Conditions d'utilisation: L' effectif théorique calculé sous l'hypothèse H_0 doit être supérieur à 5.

Remarque : Pour des effectifs faibles, il existe des corrections à ce test ou d'autres tests (test exact de Fisher).

Exemple : Dans une formation végétale homogène, on choisit 38 points au hasard. En chaque point, on écoute pendant un temps fixé les oiseaux présents. On repère la pie Pica pica 17 fois et la corneille Corvus corone 19 fois. Sachant que les deux espèces ont été enregistrées simultanément 11 fois, peut-on affirmer que la présence d'une espèce influence la présence de l'autre ?

Tableau de contingence observé :

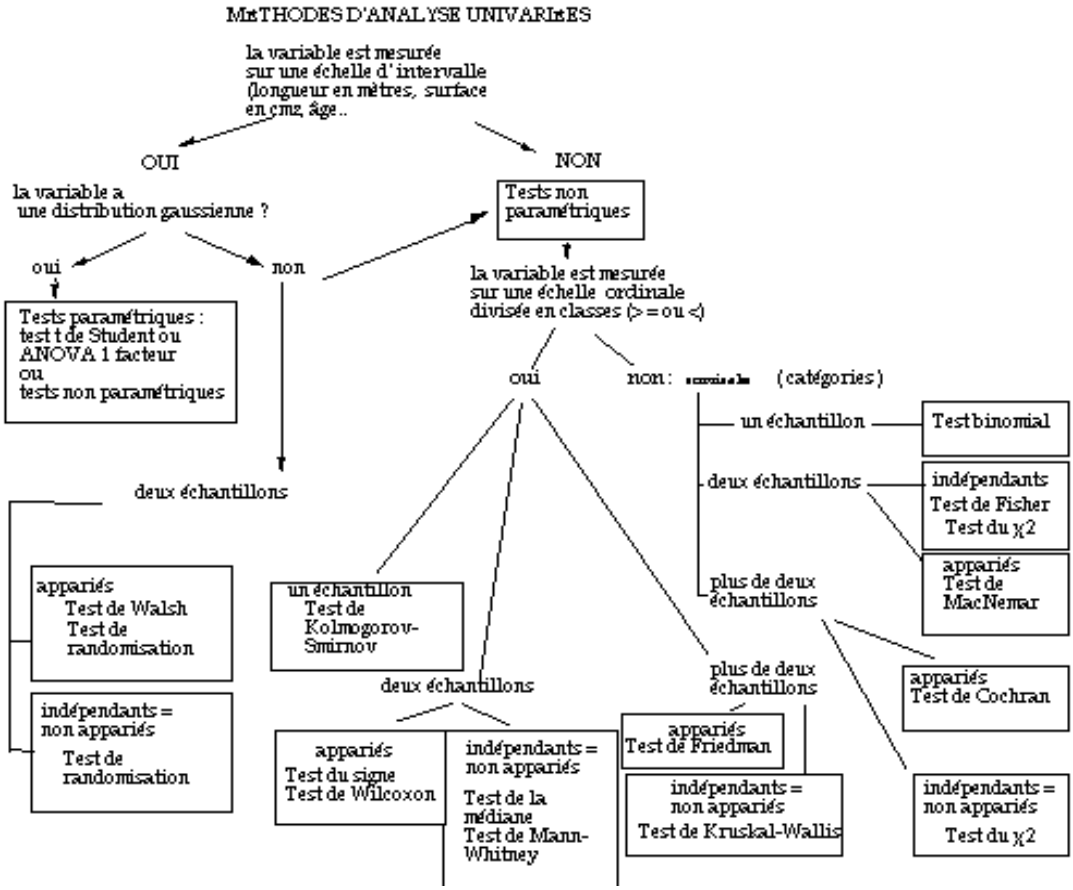
		corneille		
		présence	absence	total
Pie	présence	11		
	absence			
	total			

Sous l'hypothèse d'indépendance:

		corneille		
		présence	absence	total
Pie	présence			
	absence			
	total			

```
Pearson's Chi-squared test

X-squared = 2.6611, df = 1, p-value = 0.1028


Calculs à la main :
```

III – TESTS ELEMENTAIRES SUR VARIABLES QUANTITATIVES



Fiche 12 – Comparaison d'une moyenne à une valeur référence

Objectif : L'objectif est de comparer une moyenne à une valeur de référence. On qualifie un tel test de test de conformité.

Données : On dispose d'une variable quantitative X mesurée sur n individus.

Hypothèse nulle H_0 : « $\mu = \mu_0$ »

Conditions d'utilisation:

- Un échantillon de *n* **individus indépendants**
- La variable suit une **loi normale ou *n* >30**.

Principe du test :

Pour une population de moyenne et variance inconnue, nous avons déjà vu que si les conditions sont respectées :

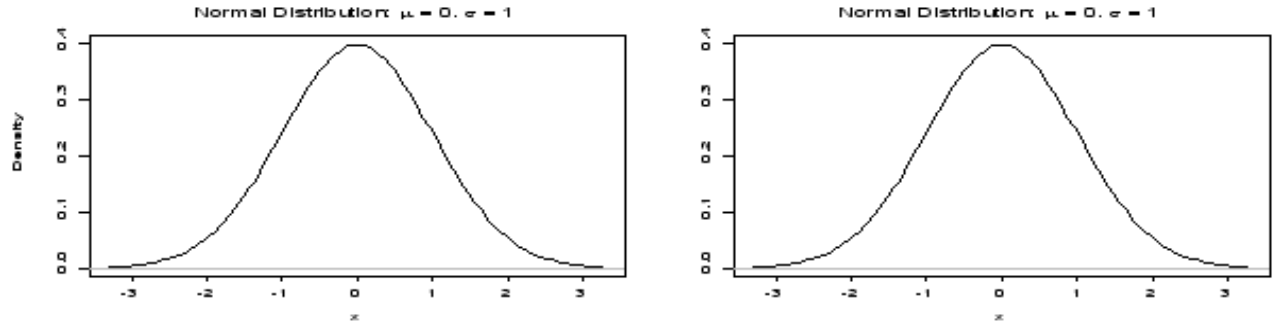
$$t = \frac{\overline{x} - \mu_0}{\frac{s_X}{\sqrt{(n)}}}$$ suit sous H_0 une loi de Student à *n-1* ddl.

Test bilatéral: On teste H_0 **:** « $\mu = \mu_0$ » contre $\boldsymbol{H}$ **:** « $\mu \neq \mu_0$»

- si , $\mu_0 - t_{1-\alpha/2} \frac{\overline{x} - \mu_0}{\frac{s_X}{\sqrt{(n)}}} < \overline{x} < \mu_0 + t_{1-\alpha/2} \frac{\overline{x} - \mu_0}{\frac{s_X}{\sqrt{(n)}}}$ on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce α (ou *p*).

Test unilatéral: H_0 « $\mu > \mu_0$ » contre H « $\mu \leqslant \mu_0$»

- si $\mu_0 - t_{1-\alpha} \frac{\overline{x} - \mu_0}{\frac{s_X}{\sqrt{(n)}}} > \mu$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce α (ou *p*).



bilatéral **unilatéral**

Exemple : Le poids des graines de radis est supposé suivre une loi normale. On a pesé 5 graines et obtenu les valeurs 1,2; 1,4; 1,5; 1,6; 1,8 en mg. La moyenne est-elle plus grande que 1,3 ? différente de 1,3 ?

Fiche 13 – Comparaison de deux moyennes : t - test

Objectif : Comparer les moyennes obtenues dans deux populations.

Données : On dispose d'une variable quantitative mesurée sur n_1 individus d'une population 1 et sur n_2 individus d'une population 2.

	X	Population
ind 1	x_1	1
ind 2	x_2	2
...		
ind i	x_i	1

X : variable quantitative
Population : variable qualitative (facteur sous R)

Hypothèse nulle H_0 : « $\mu_1 = \mu_2$ »

Principe du test :

La variable $d = \overline{x_1} - \overline{x_2}$ a pour variance estimée $s_d^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$.

Si les conditions sont respectées, la statistique $t = \frac{\overline{x_1} - \overline{x_2}}{s_d}$ suit sous H_0 une loi de Student à $n_1 + n_2 - 2$ ddl.

On construit alors le test suivant:

Test bilatéral: On teste H_0 : « $\mu_1 = \mu_2$ » contre $\boldsymbol{H}$: « $\mu_1 \neq \mu_2$ »

- si $-t_{1-\frac{\alpha}{2}} \; s_d < d < t_{1-\frac{\alpha}{2}} \; s_d$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce α (ou p).

Test unilatéral: On teste H_0 : « $\mu_1 > \mu_2$ » contre $\boldsymbol{H}$: « $\mu_1 \leqslant \mu_2$ »

- si $d > -t_{1-\alpha} \; s_d$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce α (ou p).

Conditions d'utilisation:

- Deux échantillons de n_1 *et* n_2 individus **indépendants**.
- La variable suit une **loi normale** dans chaque population ou $\boldsymbol{n_1}$ *et* $\boldsymbol{n_2}$ >30 : fiche Normalité
- La variable a la **même variance** dans les deux populations : fiche Test F

Compléments:

- Pour **comparer plusieurs populations** dans les mêmes conditions : **analyse de variance**

- Si les **hypothèses** de normalités ou d'égalité des variances **ne sont pas vérifiées**, on utilise
  - soit un **test non-paramétrique** fiche #Mann Whitney ,
  - soit un **changement de variable** fiche #Remédiation

Exemple 1 : On a mesuré les dimensions d'une tumeur chez des souris traitées ou non avec une substance antitumorale. On a obtenu les résultats suivants :

Souris témoins : $n_1 = 20$ $\overline{x_1} = 7{,}075$ cm² $s_1 = 0{,}576$ cm²

Souris traitées : $n_2 = 18$ $\overline{x_2} = 5{,}850$ cm² $s_2 = 0{,}614$ cm²

La différence observée est-elle significative ? Quuelles sont les hypothèses à vérifier ?

Exemple 2 : Reprendre l'exemple de la variable FL, fichier crabs. L'espèce a-t-elle une influence sur cette variable ? L'analyse est-elle biaisée ?

```
>    t.test(FL~sp,    alternative='two.sided',    conf.level=.95,    var.equal=TRUE,
data=crabs)

t = -6.8551, df = 198, p-value = 8.842e-11

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:     -3.932543 -2.175457
sample estimates:                  mean in group B mean in group O
                                         14.056          17.110
```

Fiche 14 – Comparaison de deux moyennes : t-test apparié

Objectif : Comparer les moyennes obtenues dans le cas où les observations sont appariées (avant-après sur un même individu, mesure par deux méthodes).
Chaque individu est décrit par un couple de variables (X_1, X_2).

Données : On dispose de deux variables quantitatives X_1 et X_2 mesurées sur n individus d'une population.

	X_1	X_2
ind 1	x_{11}	x_{21}
ind 2	x_{12}	x_{22}

Hypothèse nulle H_0 : « $\mu_1 = \mu_2$ »

Principe du test :
On construit une nouvelle variable $Z = X_2 - X_1$.

Si les conditions sont respectées, la variable $t = \dfrac{\bar{z}}{\dfrac{s_z}{\sqrt{(n)}}}$ suit sous H_0 une loi de Student à *n-1 ddl*.

Test bilatéral: On teste H_0 : « $\mu_1 = \mu_2$ » contre $\boldsymbol{H}$: « $\mu_1 \neq \mu_2$ »

- si $-t_{1-\frac{\alpha}{2}} < t < t_{1-\frac{\alpha}{2}}$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce égal à α (ou p).

Conditions d'utilisation:
- Les individus sont **indépendants**.
- Les variables X_1 et X_2 suivent une **loi** normale ou ***n*** >30 : fiche #Normalité
- Les variables ont la **même variance** : fiche #Test F

Compléments:
- Si les hypothèses de normalités ou d'égalité des variances ne sont pas vérifiées, on utilise un **test non-paramétrique** (Wilcoxon, fiche #Wilcoxon) ou un **changement de variable** ($\log(X)$..., fiche #Remédiation) .

Exemple : Une étude sur l'efficacité d'une crème amincissante conduite sur 10 hommes est présentée ci-dessous. On suppose que le poids suit une loi normale. Le régime est-il efficace ?

	1	2	3	4	5	6	7	8	9	10
avant	65	75	72	69	71	82	63	69	70	69
après	64	72	72	65	72	78	60	68	64	68

```
>   t.test(poids$avant,   poids$apres,   alternative='two.sided',   conf.level=.95,
paired=TRUE)

t = 3.2359, df = 9, p-value = 0.01023

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:     0.6620289 3.7379711
sample estimates:                   mean of the differences
                                                        2.2
```

Fiche 15 – Comparaison de moyennes : tests non paramètriques de Mann Whitney - Wilcoxon

L'utilisation d'un test non paramétrique ne repose pas sur l'hypothèse d'une distribution particulière de la variable étudiée. Sa seule condition d'application est **l'indépendance des individus**.

Ces tests sont en général **moins puissants** que les tests paramétriques lorsqu'ils existent (#t - test). La valeur de p est en effet plus grande avec ces tests. De plus les tests paramétriques permettent de tester des hypothèses en générale plus complexes (anova, comparaisons multiples ...). On réalise donc préférentiellement un test paramétrique.

On utilise donc de tels tests lorsque la distribution ne suit pas une loi normale, lorsque l'égalité des variances n'est pas vérifiée, en cas de faibles effectifs... La facilité d'exécution de ces tests avec un logiciel permet de confirmer en cas de doute sur les conditions d'application la signification du test paramétrique. Il ne faut pas hésiter à doubler le test paramétrique par un test non-paramétrique.

Test de comparaison des moyennes de deux populations de Mann-Whitney

Principe du test :

Les données sont rangées dans l'ordre croissant. Dans l'hypothèse d'égalité des moyennes, les deux populations s'ordonnent aléatoirement.

La statistique étudiée est la somme des rangs d'une des 2 populations, W. Cette somme suit une loi normale sur laquelle repose la règle de décision.

On note m_1 l'effectif de la population 1 et m_2 celui de la population 2, m_1 et m_2 >10, W la somme des rangs de la population 1.

On a alors $\mu_W = \frac{\left(m_1\left(m_1+m_2+1\right)\right)}{2}$ et $\sigma^2{}_W = \frac{m_1\, m_2\left(m_1+m_2+1\right)}{12}$

Exemple : Reprendre l'exemple FL en fonction de sp

```
> wilcox.test(FL ~ sp, alternative="two.sided", data=crabs)
      Wilcoxon rank sum test with continuity correction
data:  FL by sp

W = 2540.5, p-value = 1.868e-09
alternative hypothesis: true location shift is not equal to 0
```

Test de comparaison des moyennes de données appariées de Wilcoxon

Principe du test :

Il repose sur le signe de la différence

Exemple : Reprendre l'exemple des crèmes amincissantes

```
> wilcox.test(poids$var1, poids$var2, alternative='two.sided', paired=TRUE)

      Wilcoxon signed rank test with continuity correction data:

V = 42.5, p-value = 0.01955
alternative hypothesis: true location shift is not equal to 0
```

Fiche 16 – Comparaison de deux variances : Test F

Objectif : L'hypothèse d'égalité des variances est indispensable pour tester l'égalité de deux moyennes avec le test *t* (#t - test).

Données : une variable quantitative X mesurée dans deux populations. Sous R, le tableau se présente sous la forme :

X	Facteur
x_1	1
x_2	1
x_3	2

avec X la variable quantitative et Facteur la variable qualitative indiquant la population (facteur).

Hypothèse nulle H_0 : Les variances sont égales « $\sigma_1^2 = \sigma_2^2$ »

Conditions d'utilisation:

- Deux populations de moyennes et variances inconnues.
- Deux échantillons de **n_1 et n_2 individus indépendants,**
- Les variables suivent des **lois normales ou chacun des effectifs est supérieur à 30**

Principe du test : On souhaite tester l'égalité des variances de deux populations, H_0 est "$\sigma_1^2=\sigma_2^2$"

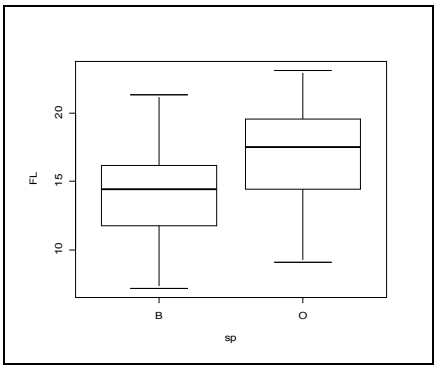
Le quotient ($\frac{s_1^2}{s_2^2}$ >1) ou ($\frac{s_2^2}{s_1^2}$ >1) suit sous H_0 la loi de Fisher-Snedecor à *n_1-1 et n_2-1 ddl*

Test : On teste l'hypothèse H_0 ($\sigma_1^2 = \sigma_2^2$) contre H ($\sigma_1^2 \neq \sigma_2^2$)

- si $\frac{s_1^2}{s_2^2} < F_{1-\frac{\alpha}{2}}$ (n_1-1,n_2-1), on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce égal à α (ou p).

Compléments :

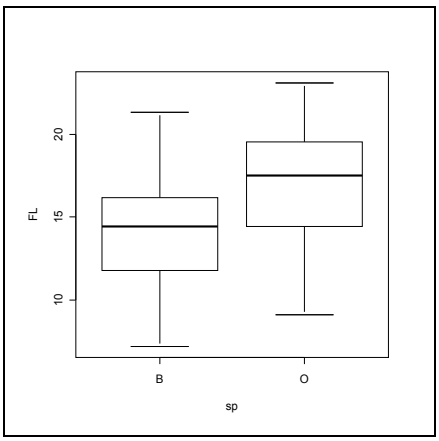
- Pour comparer plusieurs variances, on utilise le test de Bartlett ou de Lévène (chap V ANOVA).
- On peut appréhender les différences de variabilité à l'aide d'histogramme ou de boxplot.



Exemple 1 : Un biologiste effectue des dosages par une méthode de mesure de radioactivité et ne dispose donc que d'un nombre très limité de valeurs. Les concentrations C_1 et C_2 mesurées sur deux prélèvements ont donné les valeurs suivantes : C_1 : 3,9 – 3,8 – 4,1 – 3,6 C_2 : 3,9 – 2,8 – 3,1 – 3,7 – 4,1

La variabilité des valeurs obtenues pour les deux prélèvements est-elle similaire ?

Exemple 2 : Reprendre l'exemple crabs et la variable FL. La variabilité des valeurs obtenues pour les deux espèces est-elle similaire ?




```
F test to compare two variances :
                                FL by sp
F = 0.8498, num df = 99,
denom df = 99, p-value = 0.4196

alternative hypothesis: true ratio of
variances is not equal to 1

95 percent confidence interval:
                  0.5717937 1.2630299
sample estimates:  ratio of variances
                        0.8498192
```

Fiche 17 –Test de conformité à une distribution : test du χ^2

Objectif : On considère une variable X quantitative.
Si X est continue, les observations sont réparties par classes, assimilées à des modalités
Si X est discrète, chaque valeur est assimilée à une modalité. Dans le cas d'un nombre infini de valeurs, les valeurs peuvent également être réparties par classe.

L'objectif est de tester si la répartition des individus dans les différentes classes est conforme à celle attendue pour une distribution théorique :

- cette distribution théorique peut être définie *a priori*,
- ou être estimée par rapport à l'échantillon. Dans le cas de la loi normale, on calcule la moyenne et l'écart-type de l'échantillon et on vérifie que l'échantillon se distribue suivant la loi normale de mêmes moyenne et écart-type.

Données : Les données sont regroupées dans un tableau de la forme :

Variable quantitative X	classe 1	classe 2	....
effectif	n^1_{obs}	n^2_{obs}	....

Hypothèse nulle : Les fréquences observées sont conformes aux probabilités attendues.

Principe du test :

Le principe du test du χ^2 est d'estimer à partir d'une loi de probabilité connue (ou estimée à partir de l'échantillon), les effectifs théoriques pour les différentes classes du caractère étudié et les comparer aux effectifs observés dans un échantillon. Deux cas peuvent se présenter :

- soit **la loi de probabilité est spécifiée *a priori*.** Le nombre de ddl est alors le nombre de classes k moins 1, $\nu = k$-1,

- soit **la loi de probabilité théorique n'est pas connue *a priori*** et elle est déduite des caractéristiques statistiques mesurées sur l'échantillon. Le nombre de ddl est alors le nombre de modalités moins un et moins le nombre de paramètres estimés. Pour l'adéquation à une loi normale inconnue, $\nu = k - 1 - 2$.

Test : On teste l'hypothèse H_0 (conforme à la distribution attendue)
-si $\hat{\chi}^2 < \chi^2_{1-\alpha}(\nu)$, on accepte H_0
-sinon on rejette H_0 avec un risque de première espèce α (ou p).

Conditions d'application : Il est impératif d'avoir $n \geq 50$ et $n^i_{\text{th } eor} \geq 5$

Compléments :

- Pour tester la normalité d'une distribution, voir la fiche correspondante.
- Il existe aussi le **test de** Kolmogorov-Smirnov `(>ks.test(X))`. Ce test compare la fonction de répartition théorique supposée continue et entièrement déterminée avec la fonction de répartition de l'échantillon empirique.

Exemple : On lance 60 fois un dé et on obtient 12 pour la face 1, 8 pour 2, 15 pour 3, 7 pour 4, 8 pour 5 et 10 pour 6. Le dé est-il pipé ?

Fiche 18 – Normalité d'une distribution

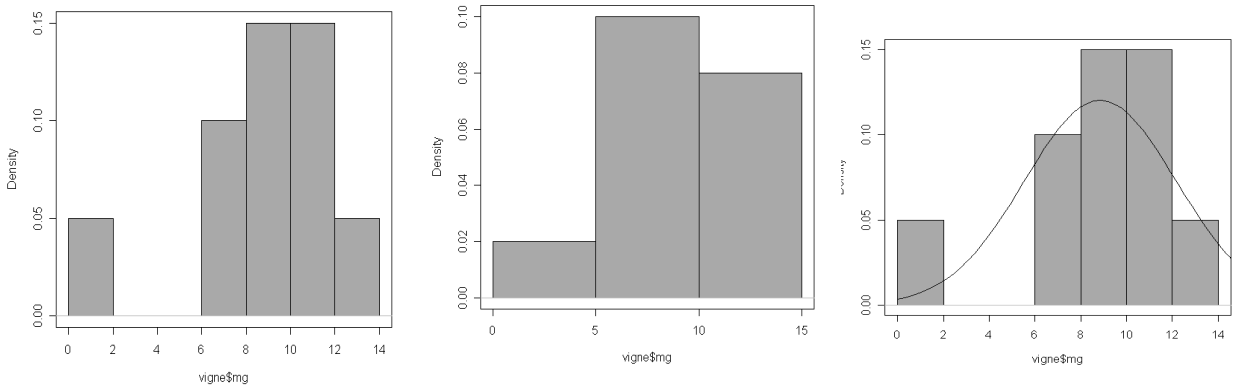
Objectif : La distribution suivant une loi normale d'une variable X est un prérequis nécessaire à la majorité des tests paramétriques (ANOVA, régression ...). Les méthodes de vérification ne sont pas entièrement satisfaisantes (faible puissance) notamment du fait des faibles effectifs souvent étudiés.
On est donc conduit à croiser plusieurs approches, graphiques et tests, pour évaluer cette hypothèse.

Exemple 1 : On a prélevé un échantillon de 10 feuilles de vigne puis on a fait la minéralisation (données ci-dessous). La quantité de minéraux dans les feuilles de vigne suit-elle une loi normale et ce, avec un risque d'erreur de 5% (créer le tableau vigne)?

1,08 7,68 8,28 8,23 7,63 11,74 10,30 11,72 12,87 9,02

Représentations graphiques :

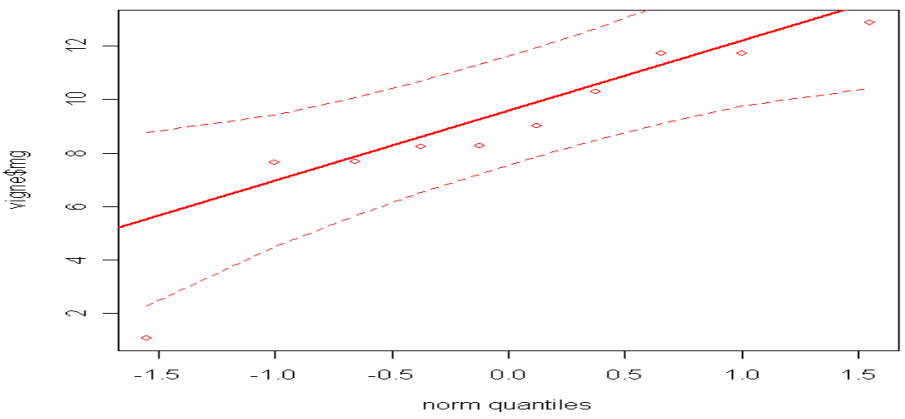
1. **Symétrie et unimodalité de la distribution**
   On réalise ici un histogramme. L'existence de deux « pics » ou une forte dissymétrie est un bon indice d'une non normalité. On peut également représenter sur l'histogramme la fonction de densité de la loi normale correspondante et évaluer les écarts.




```
> x=seq(0,15,le=50)
> lines(x,dnorm(x,mean(vigne),sd(vigne)))
```

2. **Droite de Henry**

La droite de Henry représente les quantiles (x_i) de la loi empirique en fonction des quantiles de la loi normale centrée réduite (t_i).
Si la loi empirique suit une loi normale, les points sont alignés ($x_i = \sigma\ t_i + \mu$).



Tests statistiques :

Il existe différents tests pour étudier la normalité : Test de Jarque Bera , Test d'adéquation du χ^2, test de Lilliefor (`> library(nortest) > lillie.test(X))`, test de shapiro Wilks. La multitude des tests indique qu'aucun n'est entièrement satisfaisant. Nous nous limiterons au dernier parmi les plus utilisés.

- **Test de Shapiro & Wilks :**

On retiendra que le test de Shapiro et Wilks porte sur la corrélation au carré qu'on voit sur un qqplot [*qqnorm(X) sous R et graphique quantile-quantile sous R commander*]. La corrélation est toujours très forte, la question est toujours "l'est-elle assez ?" La probabilité critique est la probabilité pour que la statistique soit inférieure ou égale à l'observation.

```
> shapiro.test(vigne$mg)
        Shapiro-Wilk normality test

W = 0.8688, p-value = 0.09689
```

Exemple 2 : Reprendre l'exemple crabs et la variable FL.

```
> shapiro.test(crabs$FL)
        Shapiro-Wilk normality test

W = 0.9904, p-value = 0.2023
```

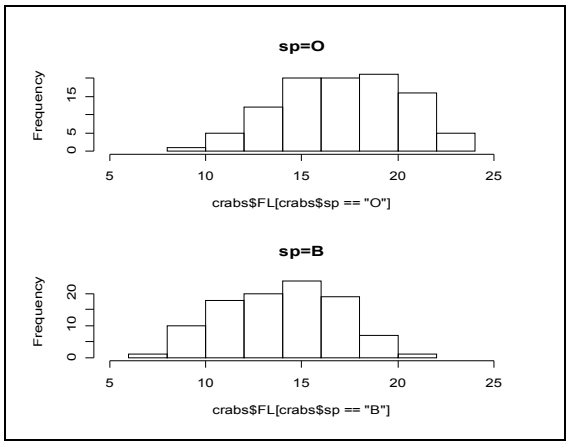
Le test n'est pas correct ici car cette variable dépend de l'espèce. Il faut donc tester la normalité au sein de chaque espèce sinon on réalise le test sur un mélange de deux distributions.

On teste la normalité pour chacune des populations X1=crabs$FL[crabs$sp=="O"] et X2=crabs$FL[crabs$sp=="B"].

```
> shapiro.test(crabs$FL[crabs$sp=="O"])
        Shapiro-Wilk normality test

W = 0.981, p-value = 0.1592
```

Examen des deux distributions

```
> par(mfrow=c(2,1))
>hist(crabs$FL[crabs$sp=="O"],main="sp
=O", xlim=c(5,25))
> hist(crabs$FL[crabs$sp=="B"],
main="sp=B",xlim=c(5,25))
```



Fiche 19 – Test du coefficient de corrélation

Soient (X,Y) un couple de variables quantitatives. La description de la liaison entre les deux variables se fait préalablement par un examen du nuage de points (x_i,y_i). Voir la fiche #correlation

Si le nuage de points décrit une relation linéaire entre les deux variables, on peut calculer comme indicateur de la liaison linéaire entre les deux variables, le coefficient de corrélation de Pearson :

$$r = \frac{\sum \left((x_i - \overline{x})(y_i - \overline{y}) \right)}{\sigma_x \sigma_y}$$

Si la relation entre les variables n'est pas linéaire, il est possible d'utiliser un autre coefficient de corrélation (par exemple le coefficient de corrélation de Spearman basé sur les rangs des observations).

En présence de plusieurs variables, on construit la matrice des corrélations ainsi que la matrice des nuages de points (voir régression multiple).

Objectif : La liaison linéaire entre les variables est significative si le coefficient de corrélation peut être considéré comme significativement non nul.

Données : Un couple de variables quantitatives :

X	Y
x_1	y_1
x_2	y_2

Hypothèse nulle : H_0 "le coefficient de corrélation de Pearson est nul" ou "Les variables X et Y ne sont pas corrélées".

Principe du test : Sous H_0, la statistique *r* suit une loi tabulée à *n-2 ddl*. On construit alors une zone d'acceptation centrée sur 0.

Test : On teste H_0 « r= 0 » contre H « r $\neq$ 0 ».

- Si p >0,05, on accepte H_0.
- Si p <0,05, on rejette H_0 avec un risque de première espèce p.

Conditions d'application : Elles reposent sur la distribution multinormale du couple *(X,Y)*.

Exemple : Reprendre l'exemple crabs et tester la signification de la corrélation entre FL et BD.

```
> cor.test(crabs$BD, crabs$FL, alternative="two.sided", method="pearson")

t = 88.6184, df = 198, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:     0.9836734 0.9906280
sample estimates:                        cor : 0.9876272
```

Remarque : Il faut vérifier la linéarité de la relation (nuage de points) et la normailité des variables (Shapiro Wilks).

IV – REGRESSION LINEAIRE SIMPLE

Objectif : La régression linéaire simple s'applique à des tableaux décrivant pour chaque individu deux variables quantitatives. L'analyse de ces tableaux peut se limiter à l'analyse des liaisons entre variables (corrélation, ACP) mais on recherche souvent à expliquer une des variables en fonction d'autres variables.

On distingue alors la variable à expliquer Y et les variables explicatives X_i. Les variables explicatives peuvent être fixées par l'expérimentateur ou aléatoires. Dans tous les cas, la mesure de X_i est considérée comme exacte (à l'errreur de mesure près). Par contre la variable Y est considérée comme une variable aléatoire avec une variabilité naturelle qui est décrite par un modèle dit modèle d'erreur (loi normale le plus souvent). Le rôle des variables n'est donc pas symétrique et le choix de Y est le plus souvent naturel.

Exemple : Choisir X_i et Y

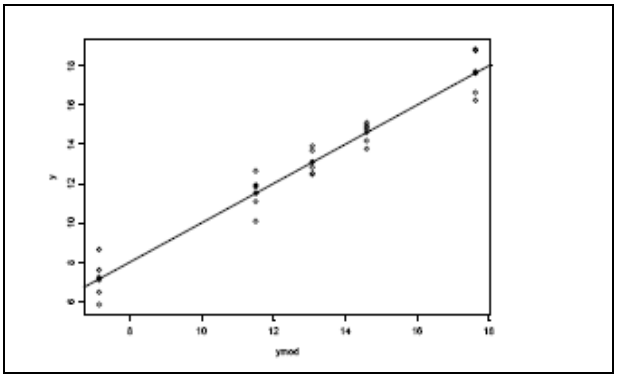
1. poids taille **2.** poids frais poids sec 3. CO_2 nb de voitures nb habitants

L'objectif de la régression est de déterminer, si elle existe, une relation fonctionnelle entre la variable à expliquer Y et une ou plusieurs variables explicatives X_1, X_2 …

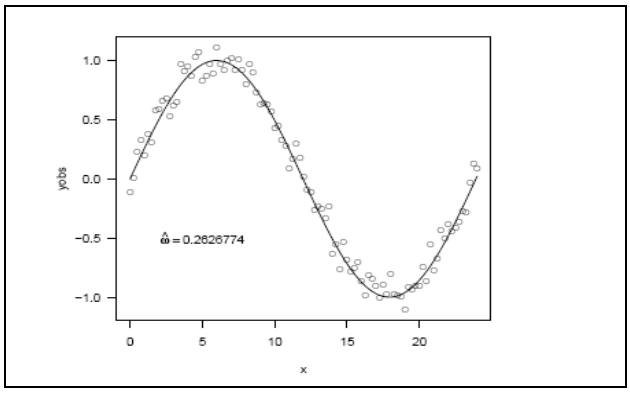
Données :

	Y	X_1
individu 1	y_1	x_{11}
individu 2	y_2	x_{12}

Représentation graphique : La première étape est d'observer le nuage de point pour déceler une éventuelle relation fonctionnelle.



relation fonctionnelle linéaire

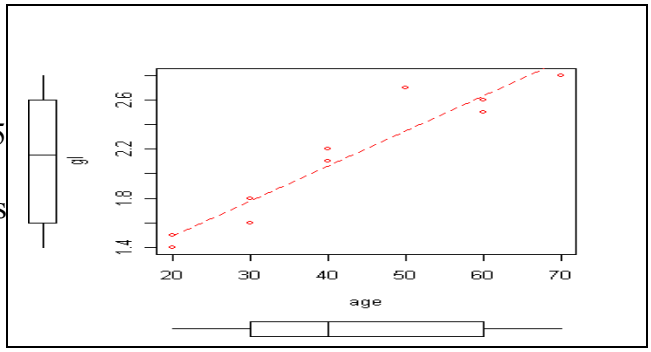


relation fonctionnelle non linéaire

Exemple : Sur un échantillon de 10 sujets d'âges différents, on a recueilli l'âge et la concentration sanguine du cholestérol (en g/L) de 10 individus :

age (xi)	30	60	40	20	50	30	40	20	70	60
gl (yi)	1.6	2.5	2.2	1.4	2.7	1.8	2.1	1.5	2.8	2.6

Le taux de cholestérol est-il fonction de l'âge?
L'inverse a-t-il un sens ?
La relation fonctionnelle est-elle linéaire ?
Peut-on prévoir le taux de cholestérol attendu à 35 ans, 75 ans?
Quels taux de cholestérol peuvent être considérés comme normaux à un âge donné?



Fiche 20 – Modèle de régression linéaire simple - Ajustement

On utilisera une régression linéaire simple dans le cas ou :

- la relation fonctionnelle peut être considérée comme **linéaire entre *Y* et *X*** (observation du nuage de points),
- la **corrélation est significativement différente de 0** (fiche #correlation).
  Dans le cas contraire, il n'existe pas de relation (linéaire) significative entre Y et X et l'utilisation d'un modèle de régression linéaire n'a aucun intérêt.

On réalisera donc toujours ces deux vérifcations au préalable et dans l'ordre avant de se lancer dans une régression linéaire.

Dans de nombreux cas, la relation fonctionnelle entre Y et X ne peut pas être considérée comme linéaire :

- on peut soit revenir à un modèle linéaire par changement de variables (fiche #Remédiation),
- soit utiliser une régression non linéaire (non abordé).

Le modèle

On définit ainsi un modèle stochastique :

$y_i = f(x_i) + \varepsilon_i$ *avec* f la relation fonctionnelle entre X *et* Y
ε_i une variable aléatoire décrivant la partie aléatoire de la variable Y

Pour le modèle linéaire, on obtient :

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \textbf{ou} \quad y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

avec ε_i une variable aléatoire suivant une loi normale centrée $\mathcal{N}(0,\sigma^2)$

L'intérêt du modèle linéaire est sa simplicité, son ubiquité et les différents outils statistiques qui s'y rattachent : diagnostic, intervalle de prédiction, test sur les coefficients …

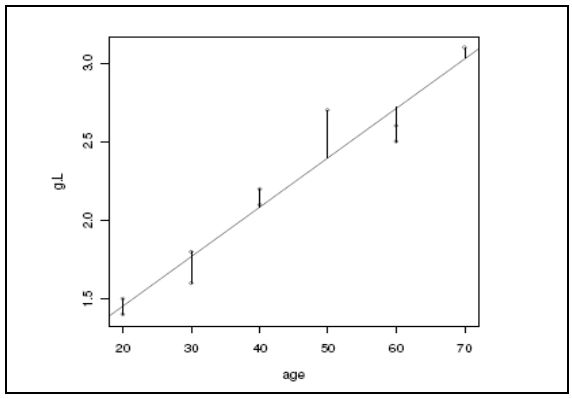
Ajustement du modèle:

On recherche dans une famille fixée de fonctions, la fonction f pour laquelle les y_i sont les plus proches des $f(x_i)$. La proximité se mesure en général comme une *erreur quadratique moyenne* :

$$\textbf{Critère des moindres carrés} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2$$

On parle alors de régression *au sens des moindres carrés.* Les différences entre les valeurs observées y_i et les valeurs prédites par le modèle $f\left(x_i\right)$, notée e_i, s'appellent les *résidus, notés* e_i.

Ecart résiduel : $e_i = y_i - f(x_i)$ *ou* $e_i = y_i - \hat{y}_i$ *avec* $\hat{y}_i = f\left(x_i\right)$



L'ajustement du modèle consiste donc à déterminer dans une famille de fonctions données, quelle est celle qui minimise l'erreur quadratique.

Dans le cadre du modèle linéaire, on notera *a, b, s²* les estimations des paramètres α, β et σ^2.
L'étude théorique conduit à :

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\rho_{XY}}{s_X s_Y}$$ avec ρ_{XY} le coefficient de covariance de X et Y.

$$a = \overline{y} - b\,\overline{x}$$

$$s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum e_i^2$$

Exemple : reprendre l'exemple cholesterol et estimer les paramètres à la main et avec R

A la main avec les fonction var, cov, mean:

$\overline{x}$= $\overline{y}$= ρ_{XY}= s_X= s_Y=

b= a= s^2=

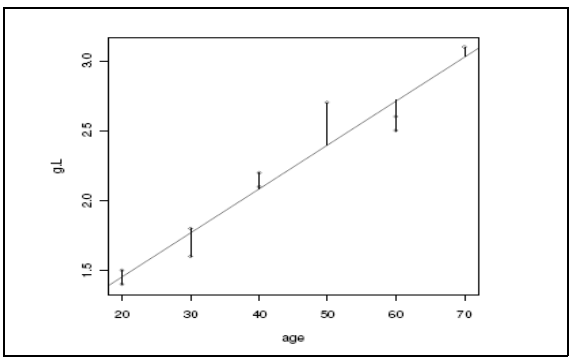
Avec R :
menu : Statistique - Ajustement de modèle – régression simple

Attention Y est choisi en premier (Y~X) au contraire du nuage de point

```
lm(formula = mg ~ age, data = cholesterol)
Residuals:
     Min       1Q   Median       3Q      Max
-0.17826 -0.11141 -0.01304  0.03315  0.35217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.923913   0.141793   6.516 0.000185 ***
age         0.028478   0.003139   9.071 1.75e-05 ***

Residual standard error: 0.1649 on 8 degrees of freedom
Multiple R-Squared: 0.9114,     Adjusted R-squared: 0.9003
F-statistic: 82.29 on 1 and 8 DF,  p-value: 1.748e-05
```

Illustration du principe d'ajustement

```
attach(cholesterol)
lm.chol <- lm(gl~age)
attach(cholesterol)
plot(age,gl)
abline(lm.chol)
segments(age,g.L,age,fitted(lm.chol))
```



Fiche 21 – Validation du modèle de régression linéaire simple

On se place dans le cadre d'une relation linéaire entre deux variables (examen du nuage de points) et d'une liaison linéaire significative entre ces deux variables (coefficient de corrélation significativement non nul).

Les **hypothèses du modèle** de régression linéaire simple nécessaire à la construction des principaux tests statistiques (inférence) sont :

- l'indépendance des observations,
- la distribution normale centrée de l'écart résiduel,
- l'homoscédasticité, à savoir que l'écart résiduel suit la même loi indépendamment des valeurs de x_i ou $\widehat{y}_i$.

Dans le cas où ces hypothèses sont vérifiées, il est possible de construire des intervalles de confiances pour les paramètres estimés, des intervalles de confiance pour la prédiction, comparer les modèles, ...

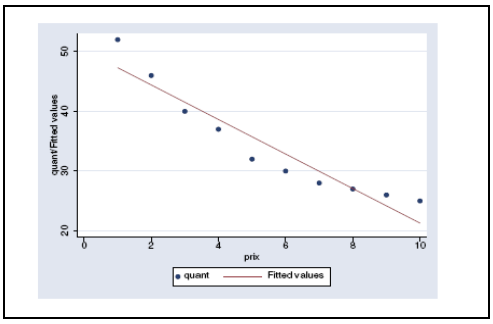
La vérification de ces hypohèses n'est pas toujours évidente. Il est préférable de croiser différentes méthodes, graphiques et tests, pour évaluer l'existence d'écarts aux hypothèses. Aucune méthode n'est entièrement satisfaisante.

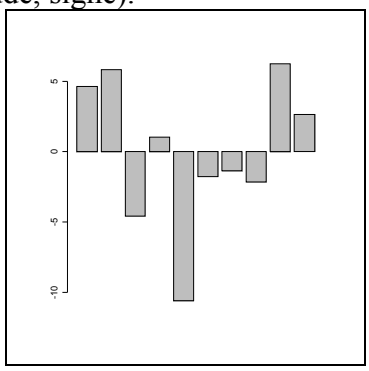
Exemple : Les analyses sont illustrées sur l'exemple cholesterol.

1. Indépendance des écarts résiduels

Le problème d'indépendance est important notamment dans le cas de séries chronologiques pour lesquels il existe des modèles spécifiques (ARIMA…).

Observations graphiques : On observe sur l'ajustement du nuage par une droite ou sur le diagramme des écarts l'éventuelle apparition de tendances cycliques (saisons, cycles économiques,...), une relation non linéaire, une répartition non aléatoire des résidus (amplitude, signe).





Test de Durbin-Watson : On teste l'existence d'une relation autorégressive de type

$$\varepsilon_{i+1} = \rho\,\varepsilon_i + \tau \qquad \text{avec } \tau \sim \mathcal{N}(0,\sigma^2).$$

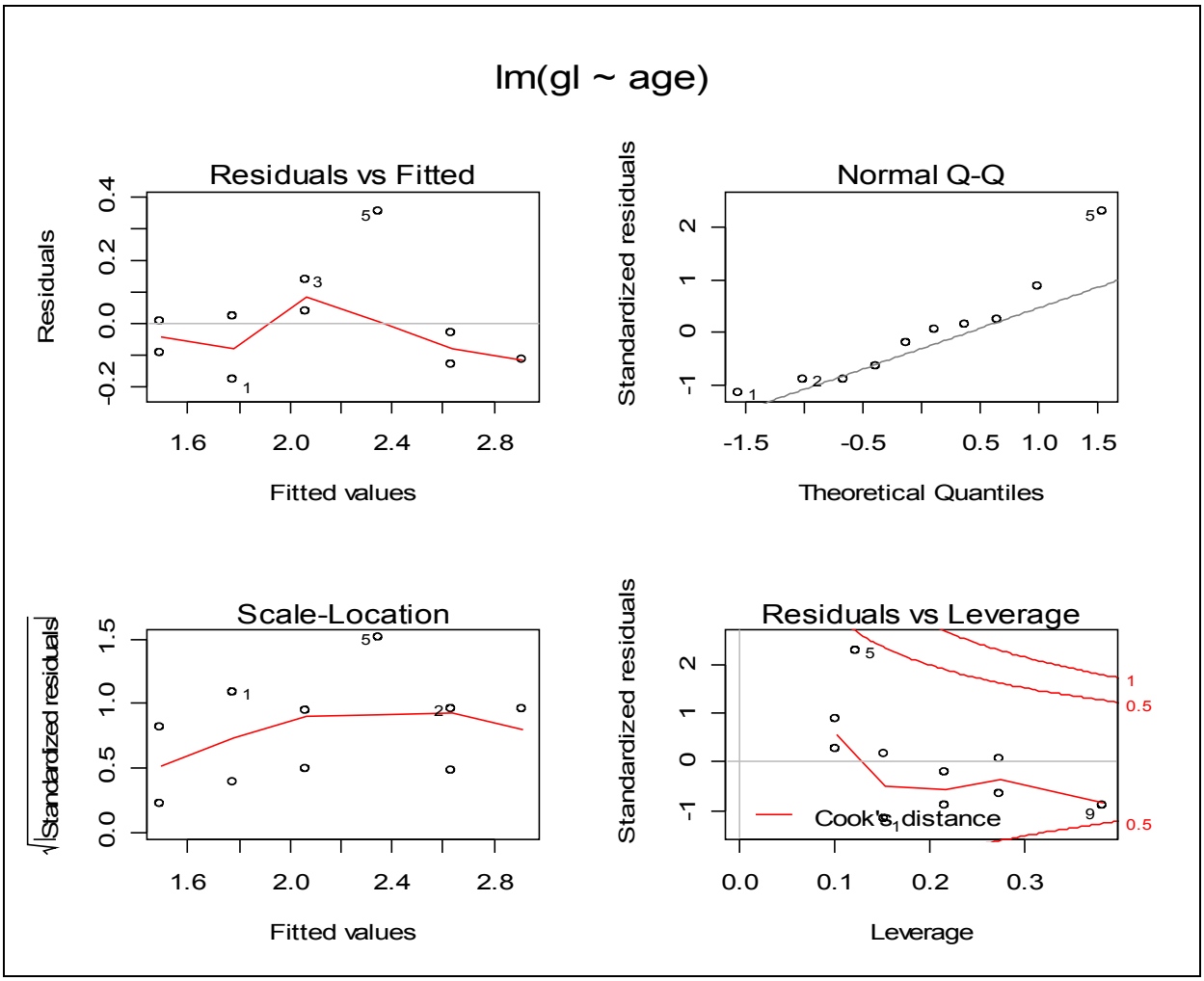
On utilise la statistique $\dfrac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}$ estimateur de $2(1-\rho)$

```
data:  gl ~ age
DW = 2.1109, p-value = 0.8063
alternative hypothesis: true autocorelation is not 0
```

2. Homoscédasticité

Un des problèmes récurrents est l'existence d'une relation entre l'écart résiduel et la valeur de $\widehat{y}_i$ ou celle de x_i. L'écart e_i a parfois tendance à croître avec $\widehat{y}_i$ ou x_i.
Il existe différents tests complémentaires concernant l'hypothèse H_0 d'homoscédasticité..

Observations graphiques :
On représente les écarts e_i en fonction de $\widehat{y}_i$ou x_i. Les écarts ne doivent pas croître en fonction de $\widehat{y}_i$ou x_i mais toujours rester du même ordre de grandeur.



Test de Breusch-Pagan : Ce test utilise la statistique nR^2 qui suit sous l'hypothèse H_0 d'homoscédasticité un χ^2.

Breusch-Pagan test
data: gl ~ age
`BP = 0.4573, df = 1, p-value = 0.4989`

Remarque : Il existe d'autres tests, test de White par exemple.

3. Normalité des écarts résiduels

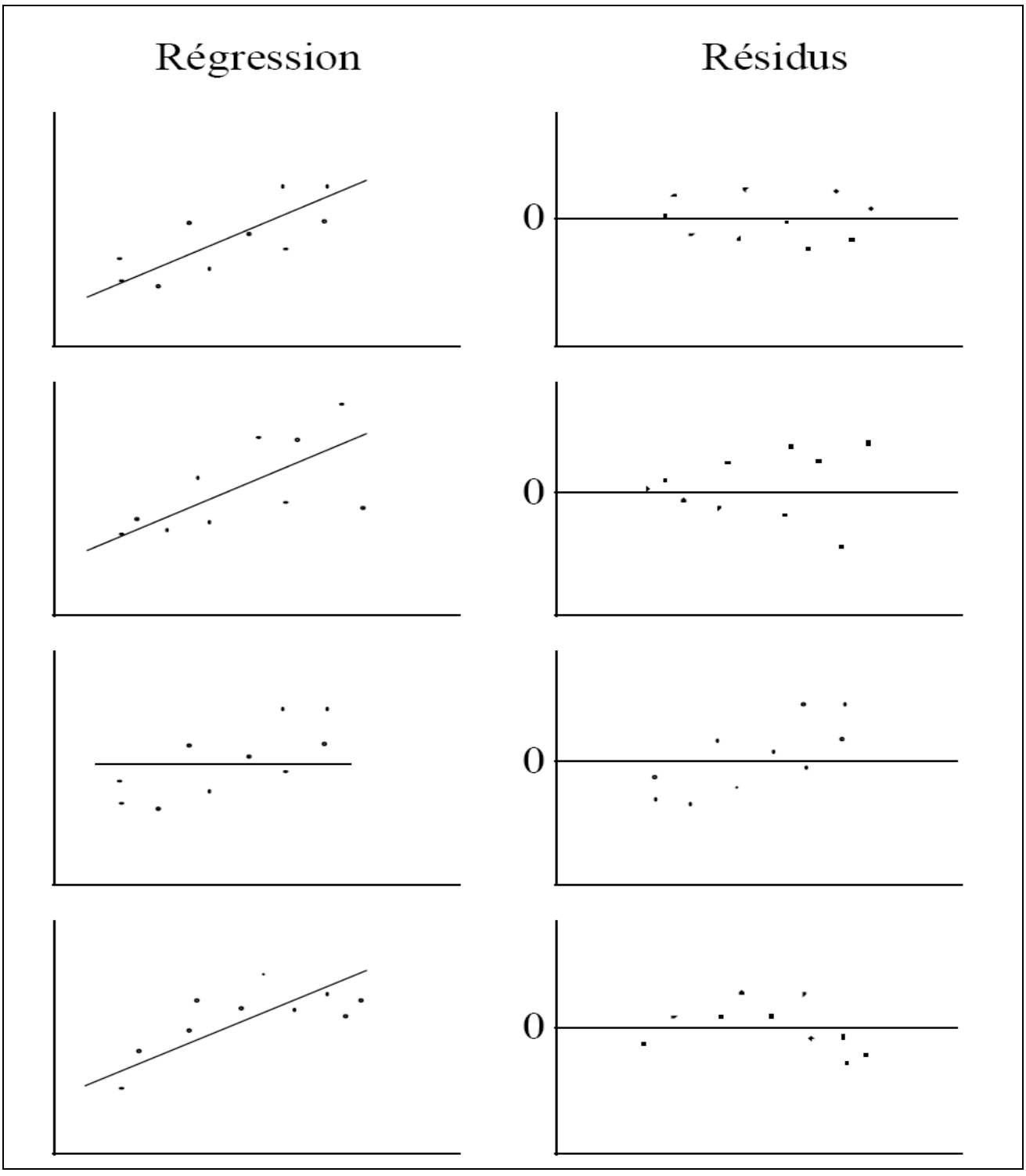
voire la fiche Normalité

Observations graphiques :
On peut utiliser les qqplot, histogramme, histogramme des résidus réduits ou studentisés (95% des valeurs entre -2 et 2)...

Test de shapiro : (chap III)

```
shapiro.test(cholesterol$res)
        Shapiro-Wilk normality test

data:  cholest$res
W = 0.9389, p-value = 0.5413
```

4. Quelques exemples



Fiche 22 – Remédiation par changement de variables

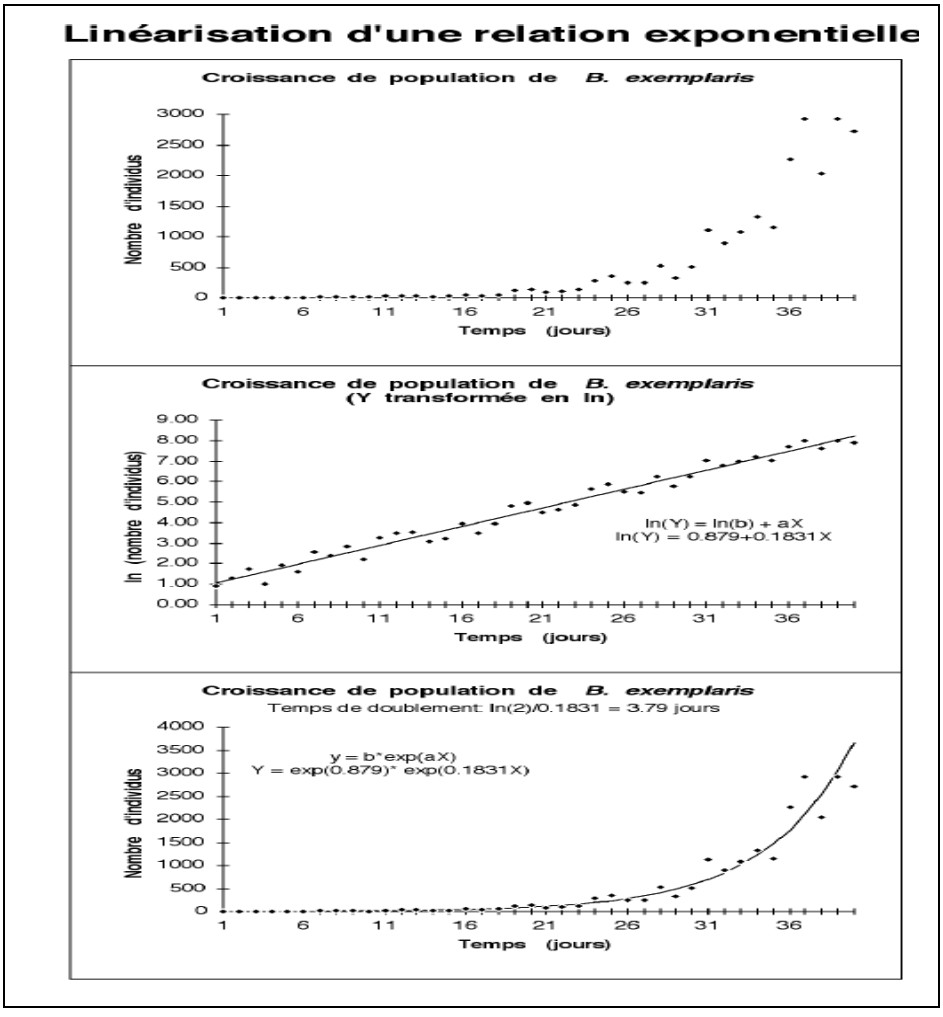
Il existe différentes techniques pour remédier aux violations de certaines hypothèses. Nous insisterons ici essentiellement sur le changement de variables qui permet par exemple de rendre linéaire une relation entre deux variables, de rendre normale la distribution des écarts résiduels ou de stabiliser la variance (homoscédasticité).

Si un tel changement ne permet pas de corriger l'écart, on peut :

- pour une relation non-linéaire utiliser un modèle non linéaire,
- pour l'hétéroscédasticité, pondérer les observations,
- pour la non normalité, changer le modèle d'erreur...

Pour certaines familles de fonctions, on transforme le problème de manière à se ramener à une régression linéaire. Voici quelques cas fréquents.

Famille	Fonctions	Transformation	Forme affine
exponentielle	$y = ae^{bx}$	$y' = \log(y)$	$y' = \log(a) + bx$
puissance	$y = ax^b$	$y' = \log(y)\ x' = \log(x)$	$y' = \log(a) + bx'$
inverse	$y = a + b/x$	$x' = 1/x$	$y = a + bx'$
logistique	$y = 1/(1 + e^{-(ax+b)})$	$y' = \log(y/(1-y))$	$y' = ax + b$
proportion	y est une proportion	$y' = \arcsin(y)$	$y' = ax + b$



Fiche 23 – Tests sur les paramètres - ANOVA

Après vérification des hypothèses, il est possible d'envisager la construction de tests.

Intervalle de confiance des paramètres

Les estimateurs *a, b, s* des coefficients suivent des lois connues permettant de calculer des intervalles de confiance, de construire des tests…

```
Call:
lm(formula = gl ~ age, data = cholest)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17826 -0.11141 -0.01304  0.03315  0.35217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.923913   0.141793   6.516 0.000185 ***
age         0.028478   0.003139   9.071 1.75e-05 ***

Residual standard error: 0.1649 on 8 degrees of freedom
Multiple R-Squared: 0.9114,  Adjusted R-squared: 0.9003
F-statistic: 82.29 on 1 and 8 DF,  p-value: 1.748e-05
```

Déterminer un intervalle de confiance à 99% des coefficients :

```
                0.5 %     99.5 %
(Intercept) 0.44814383 1.39968225
age         0.01794455 0.03901197
```

Inférence sur les coefficients :

Analyse de variance

Décomposition de la variance :

L'écart total $(y_i - \overline{y})$ se décompose en deux parties : l'écart expliqué par le modèle et l'écart résiduel.

On a la relation : écart total = écart expliqué + écart résiduel

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$$

On montre que les sommes des carrés des écarts (SCE) vérifient la même relation :
La somme des carrés des écarts totale (SC_{tot}) est égale à la somme des SC due au modèle (SC_{reg}) et des SC des résidus (SC_{res}) :

$$SCE_{tot} = SCE_{reg} + SCE_{res}$$

$$\sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n}(y_i - \widehat{y}_i)$$

Exemple : faire les calculs sur cholesterol

Coefficient de détermination R^2:

Pour quantifier l'intérêt du modèle, on calcule le coefficient de détermination, noté R^2, égal au rapport de la somme des carrés des écarts expliqués par le modèle par la somme des carrés des écarts totaux :

$$R^2 = \frac{SCE_{reg}}{SCE_{tot}}$$

On montre que ce coefficient est égal au coefficient de corrélation au carré : $R^2=r^2$. Plus R^2 est proche de 1, plus le modèle permet d'expliquer une grande partie de la variabilité de Y.

Exemple : faire les calculs sur cholesterol

Analyse de variance :

Sous l'hypothèse H_0 "b=0" (absence de liaison entre X et Y), on montre que $\frac{SCE_{tot}}{n-1}$, $\frac{SCE_{reg}}{1}$ et $\frac{SCE_{res}}{n-2}$ sont trois estimateurs sans biais de σ^2 qui suivent des lois du χ^2 à respectivement à *n-1, 1 et n-2 ddl.*

On montre alors que la statistique $F = \boxed{\frac{SC_{reg}/1}{SCres/n-2}}$ suit la loi de Fisher à (1,n-2) ddl. Il est alors possible de construire un test pour tester H_0 à l'aide de la distribution de F.

Exemple : faire les calculs sur cholesterol

Tableau d'analyse de variance :

```
Response: gl
           Sum Sq Df F value    Pr(>F)
age       2.23839  1   82.29 1.748e-05 ***
Residuals 0.21761  8
```

Fiche 24 – Prédiction

Le modèle de régression linéaire permet de construire des "enveloppes" de confiance pour $\widehat{y}_i$:

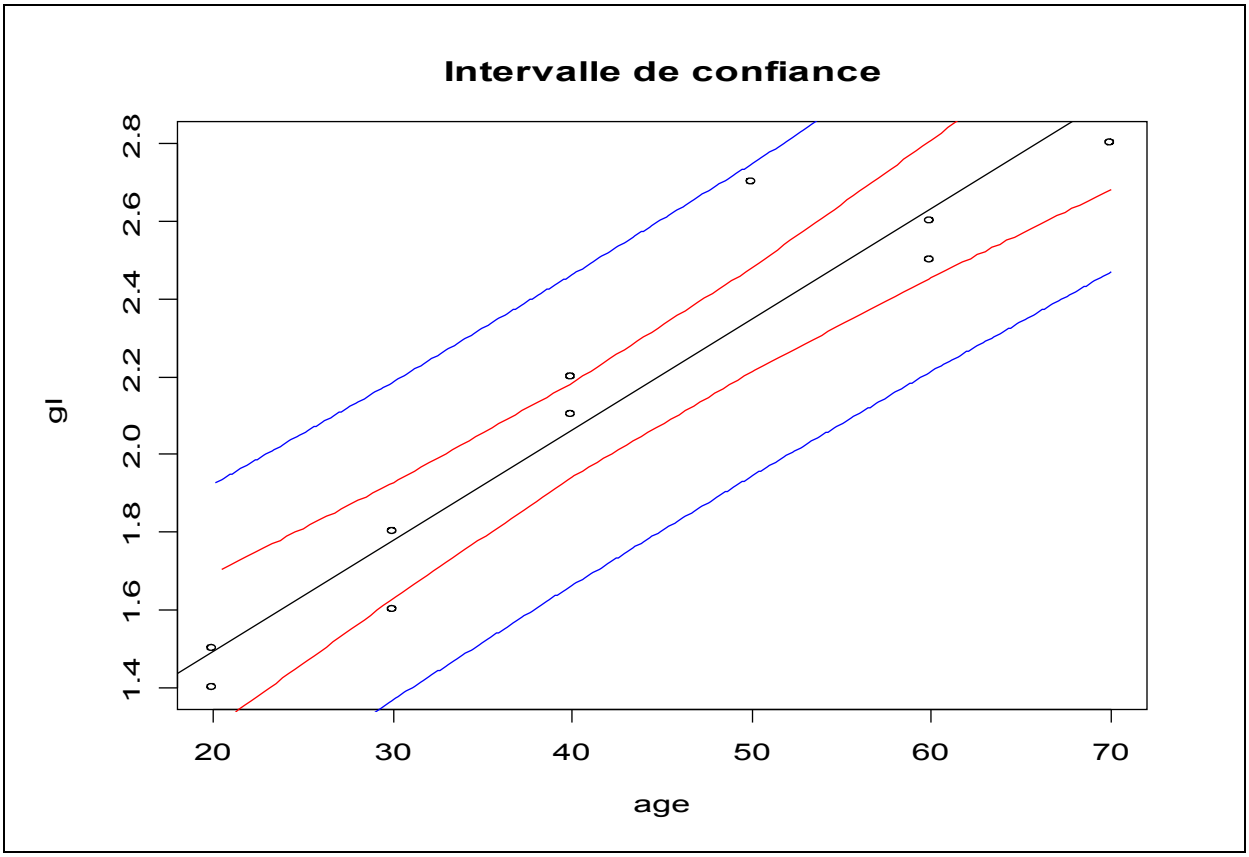
$$\widehat{y}_i \pm s \times \sqrt{\left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_X^2}\right)} \times t$$

ou de prédiction pour pour y_i :

$$\widehat{y}_i \pm s \times \sqrt{\left(1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_X^2}\right)} \times t$$

avec t la valeur du *t* de student correspondant à l'IC souhaité (prendre 2 pour 95%).

```
predc.col <- predict(lm.chol,int="confidence")


predp.col <- predict(lm.chol,int="prediction")
plot(chol,main="Intervalle de confiance");abline(lm.chol)
matlines(sort(age),predc.col[order(age),2:3],lty=c(2,2),col="red")
matlines(sort(age),predp.col[order(age),2:3],lty=c(3,3),col="blue")
```

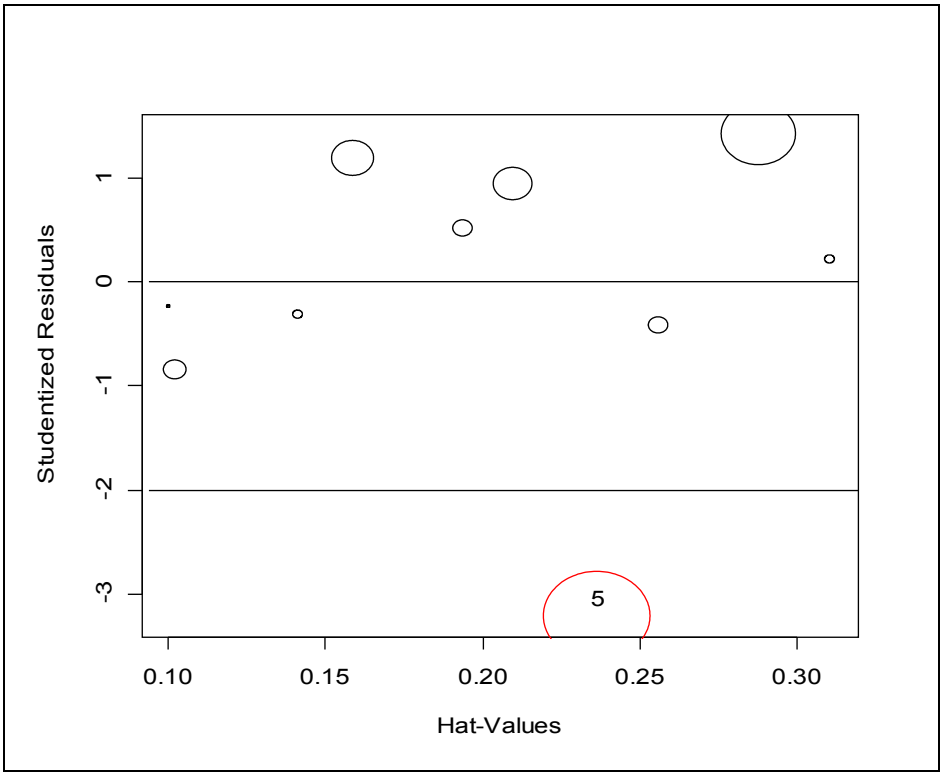


Fiche 25 – Diagnostics des points

Certains points apparaissent parfois très éloignés des valeurs attendues. Un examen graphique permet de les identifier et de vérifier alors la possibilité d'une erreur de saisie ou d'une valeur aberrante.
Il existe également des techniques permettant d'identifier les points ayant une forte influence sur les coefficients de la droite. On appelle ces points, des points leviers. Une façon intuitive d'évaluer leur influence est de construire la droite de régression avec ou sans ce point et d'observer la différence entre les droites de régression obtenues.

Distance de Cook : Cette distance calculée pour chaque point est d'autant plus grande que le point a une influence importante et/ou présente un écart important.

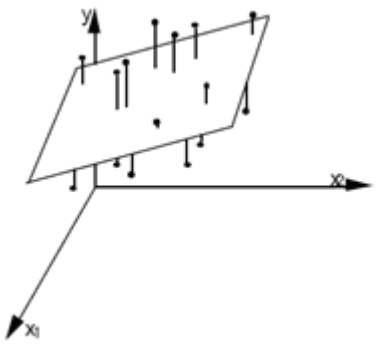
hat values h_{ii}: Les hat values sont également des indicateurs des observations influentes.



V REGRESSION LINEAIRE MULTIPLE

Principe et objectifs : Il arrive qu'on veuille expliquer les variations d'une variable dépendante par l'action de plusieurs variables explicatives. Lorsqu'on a des raisons de penser que la relation entre ces variables est linéaire (faire des nuages de points), on peut étendre la méthode de régression linéaire simple à plusieurs variables explicatives.
S'il y a deux variables explicatives, le résultat peut être visualisé sous la forme d'un plan de régression dont l'équation est : $y_i = a_1 x_{1i} + a_2 x_{2i} + e_i$

Le plan est ajusté selon le principe des moindres carrés où les sommes des carrés des erreurs d'estimation de la variable dépendante sont minimisées (construire un nuage 3D).



La régression multiple peut être utilisée à plusieurs fins:

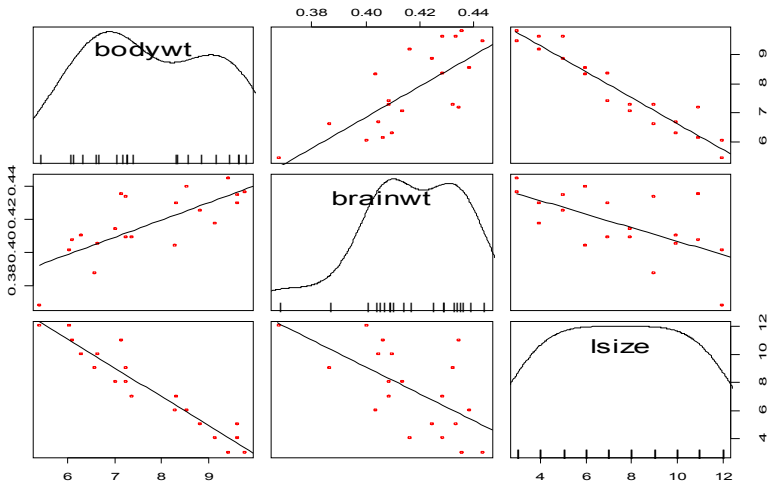
• Trouver la meilleure équation linéaire de prévision (modèle) et en évaluer la précision et la signification.

• Estimer la contribution **relative** de deux ou plusieurs variables explicatives sur la variation d'une variable à expliquer; déceler l'effet complémentaire ou, au contraire, antagoniste entre diverses variables explicatives.

• Juger de l'importance relative de plusieurs variables explicatives sur une variable dépendante en lien avec une **théorie causale** sous-jacente à la recherche (attention aux abus: une corrélation n'implique pas toujours une causalité; cette dernière doit être postulée *a priori*).

Exemple : Fichier litters de la library DAAG décrivant la taille de la portée (litter size), le poids de l'animal (bodywt) et le poids du cerveau (brainwt). L'objectif est d'étudier le poids du cerveau en fonction des deux autres variables.

Matrice des nuages de points (scatterplot matrix)



Matrice des corrélation :

	bodywt	brainwt	lsizebodywt
bodywt	1.0000000	0.7461485	-0.9548494
brainwt	0.7461485	1.0000000	-0.6214719
lsize	-0.9548494	-0.6214719	1.0000000

Modèle de régression :

Il est possible de proposer trois modèles de régression linéaire pour expliquer brain wt.
Utiliser le menu pour les construire : **statistiques – ajustement du modèle – régression linéaire**

Préciser les 3 modèles possibles et la valeur de R^2 correspondant:

Quel est le modèle le plus intéressant ?

Validation du modèle : La validation du modèle reste similaire : indépendance, homoscédasticité, normalité des résidus. Il est important de vérifier également la linéarité des relations à l'aide de nouveaux graphiques de contrôle (fiche validation).

Compléments

- **Régression sur variables centrées-réduites**

Une pratique courante en régression consiste à **interpréter les coefficients de régression centrés-réduits**, c'est-à-dire ceux qu'on obtient en centrant-réduisant toutes les variables (y compris la variable dépendante). En exprimant toutes les variables en unités d'écart-type, on rend les coefficients de régression insensibles à l'étendue de variation des variables explicatives, leur permettant ainsi d'être interprétés directement en termes de "poids" relatif des variables explicatives. Notez aussi que la plupart des logiciels courants fournissent de toute manière les "coefficients de régression centrés réduits"(*standardized regression coefficients*) en plus des coefficients calculés pour les variables brutes.

- **Régression sur les coordonnées principales.**

On réalise au préalable une analyse en composantes principales sur les variables explicatives. On construit ainsi de nouvelles variables non corrélées.

Bibliographie

Analyse de régression appliquée., Y Dodge V Rousson, Paris, Dunod 2nde ed., 2004.
Applied regression analysis., Draper N. R. Smith H. NY, J Wiley Sons, 3eme ed., 1998

Fiche 26 – Test d'un modèle

Comme en régression linéaire simple, on mesure la **variance expliquée** par la régression à l'aide du **coefficient de détermination multiple R^2**:

$$R^2 = \frac{SCE_{reg}}{SCE_{tot}}$$

Test de signification globale du modèle de régression multiple

La signification du modèle de régression multiple peut être testée par une statistique F qui, sous H_0, est distribuée suivant la loi de Fisher à p et $(n–p-1)$ degrés de liberté, p désigne le nombre de variables explicatives.

Les hypothèses du test sont:

H_0 : la variable Y est linéairement indépendante des variables X_k

H : la variable Y est expliquée linéairement par au moins une des variables X_k

L'expression de F est :

$$F = \frac{\frac{SCE_{reg}}{p}}{\frac{SCE_{res}}{n-p-1}}$$

On vérifie ainsi que le modèle est globalement significatif en comparant la valeur de F à celle de $F_{1-\alpha}(p,n-p-1)$.

Test de signification des coefficients

La question se pose ensuite de vérifier que chaque variable a une contribution significative. On teste la signification des coefficients relatifs à chaque variable à l'aide d'un test t de Student. Si le coefficient n'est pas significativement différent de 0, alors on supprime la variable du modèle.

Exemple : On reprend l'exemple litters

```
lm(formula = brainwt ~ bodywt + lsize, data = litters)
Residuals:
       Min         1Q     Median         3Q        Max
-0.0230005 -0.0098821  0.0004512  0.0092036  0.0180760

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.178247   0.075323   2.366  0.03010 *
bodywt      0.024306   0.006779   3.586  0.00228 **
lsize       0.006690   0.003132   2.136  0.04751 *

Residual standard error: 0.01195 on 17 degrees of freedom
Multiple R-Squared: 0.6505,   Adjusted R-squared: 0.6094
F-statistic: 15.82 on 2 and 17 DF,  p-value: 0.0001315

> step(RegModel.1)
Start:  AIC=-174.32     brainwt ~ bodywt + lsize

         Df Sum of Sq      RSS      AIC
<none>                   0.002 -174.323
- lsize   1     0.001    0.003 -171.568
- bodywt  1     0.002    0.004 -165.058

lm(formula = brainwt ~ bodywt + lsize, data = litters)
Coefficients:
(Intercept)       bodywt        lsize
    0.17825      0.02431      0.00669
```

Fiche 27 – Recherche du meilleur modèle : stepwise regression

R^2 ajusté

Une des propriétés de la régression multiple est que l'ajout de chaque variable explicative au modèle permet d'"expliquer" plus de variations, et cela même si la nouvelle variable explicative est complètement aléatoire. Par conséquent, le R^2 calculé comme ci-dessus comprend une composante déterministe, et une composante aléatoire d'autant plus élevée que le nombre de variables explicatives est élevé dans le modèle de régression. Plus on utilise de variables plus on explique mécaniquement la variabilité sans que la variable n'ait forcément un rôle explicatif.

Pour contrer cet effet, et donc éviter de surestimer le R^2, plusieurs auteurs ont proposé un ***R*² ajusté**, qui tient compte du nombre de variables explicatives du modèle de régression. La formule la plus couramment utilisée est la suivante:

$$\widehat{R^2} = \frac{(n-1)\,R^2 - p}{n - p + 1}$$

où n = nombre d'observations et p = nombre de variables explicatives

Ce coefficient de détermination ajusté $\widehat{R^2}$ est pertinent pour comparer la qualité de deux modèles, indépendamment du nombre de variables pris en compte. On prendra donc en compte son augmentation pour comparer deux modèles.

Sélection pas à pas *(stepwise regression)*

On rencontre parfois des situations dans lesquelles on dispose de *trop* de variables explicatives, soit parce que le plan de recherche était trop vague au départ (on a mesuré beaucoup de variables "au cas où elles auraient un effet"), soit parce que le nombre d'observations (et donc de degrés de liberté) est trop faible par rapport au nombre de variables explicatives intéressantes.

Une technique est parfois employée pour sélectionner un nombre réduit de variables qui explique pourtant une quantité raisonnable de variation.

Cette procédure, la plus complète, consiste à faire entrer les variables l'une après l'autre dans le modèle par sélection progressive et à chaque étape :

- on retient la variable qui permet la plus forte augmentation du $\widehat{R^2}$,
- on vérifie que les coefficients relatifs à chaque variable sont significativement non nuls.

Trois cas sont possibles :

- La dernière variable introduite a un coefficient non significativement non nul, on s'arrête au modèle précédent.
- Toutes les variables introduites ont un coefficient significativement non nul, on continue en ajoutant une nouvelle variable.
- Le coefficient d'une variable précédemment introduite a un coefficient non significativement non nul, on élimine cette variable et on reprend le stepwise.

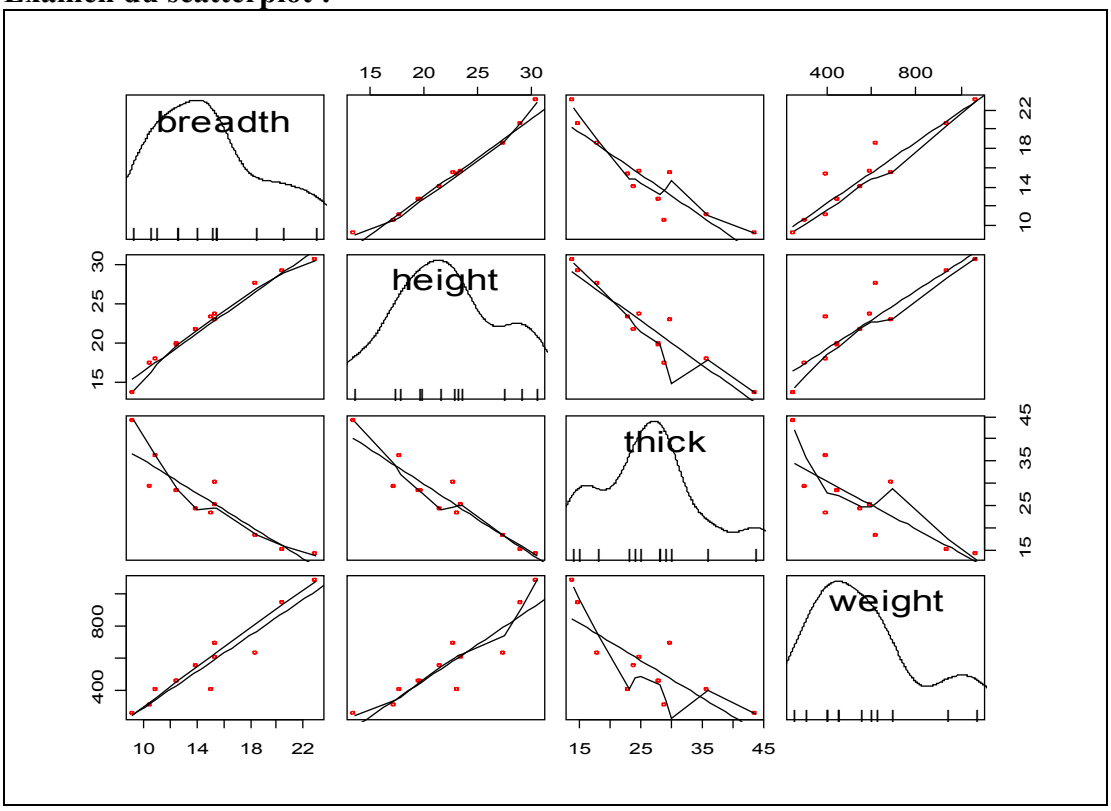
Etude d'un exemple : Utiliser les données oddbooks de la library DAAG. Nous allons étudier la variable weight en fonction des trois autres.

```
   thick height breadth weight
1     14   30.5      23   1075
```

Premier pas : Recherche de la meilleure variable explicative

```
          breadth     height      thick     weight
breadth  1.0000000  0.9859209 -0.8980836  0.9430565
height   0.9859209  1.0000000 -0.9392100  0.9080642
thick   -0.8980836 -0.9392100  1.0000000 -0.7897682
weight   0.9430565  0.9080642 -0.7897682  1.0000000
```

Examen du scatterplot :



Modèle 1 :

```
lm(formula = weight ~ breadth, data = oddbooks)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -272.95      96.27  -2.835   0.0177 *
breadth        56.21       6.27   8.965 4.28e-06 ***

Residual standard error: 86.2 on 10 degrees of freedom
Multiple R-Squared: 0.8894,   Adjusted R-squared: 0.8783
F-statistic: 80.38 on 1 and 10 DF,  p-value: 4.284e-06
```

Second pas : Recherche d'une seconde variable explicative

```
lm(formula = weight ~ breadth + height, data = oddbooks)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -104.47     162.07  -0.645   0.5353
breadth       101.85      36.39   2.799   0.0207 *
height        -38.10      29.95  -1.272   0.2352

Residual standard error: 83.65 on 9 degrees of freedom
Multiple R-Squared: 0.9062,   Adjusted R-squared: 0.8854
F-statistic: 43.48 on 2 and 9 DF,  p-value: 2.369e-05

lm(formula = weight ~ breadth + thick, data = oddbooks)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -731.748    372.119  -1.966 0.080804 .
breadth       72.031     13.831   5.208 0.000558 ***
thick          8.565      6.724   1.274 0.234671

Residual standard error: 83.64 on 9 degrees of freedom
Multiple R-Squared: 0.9063,   Adjusted R-squared: 0.8854
F-statistic:  43.5 on 2 and 9 DF,  p-value: 2.365e-05
```

Conclusion et vérification avec la fonction step de R:

Fiche 28 – Validation du modèle

La régression linéaire multiple est soumise aux mêmes contraintes que la régression linéaire simple:

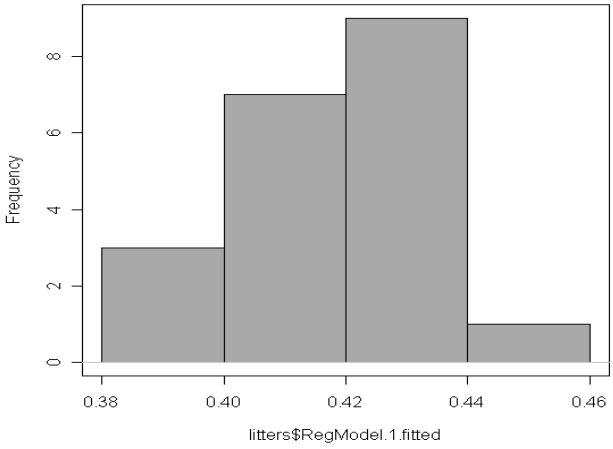
- Distribution normale de la variable dépendante (normalité des résidus)
- Homosécdasticité
- Indépendance des résidus
- Linéarité des relations entre la variable dépendante Y et chacune des variables explicatives X_k.

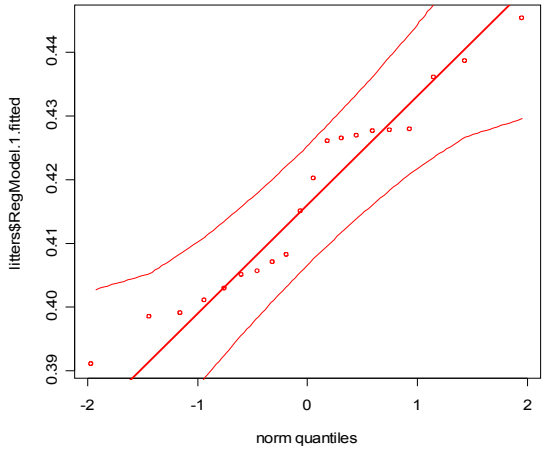
On utilise le « ajouter statistiques» du modèle pour calculer les résidus RegModel.1.residuals e_i et les valeurs ajustées RegModel.1.fitted $\widehat{y_i}$:

Exemple : Modèle litters

1. Normalité

```
Shapiro-Wilk normality test
data:  litters$RegModel.1.residuals
W = 0.9743, p-value = 0.8416
```

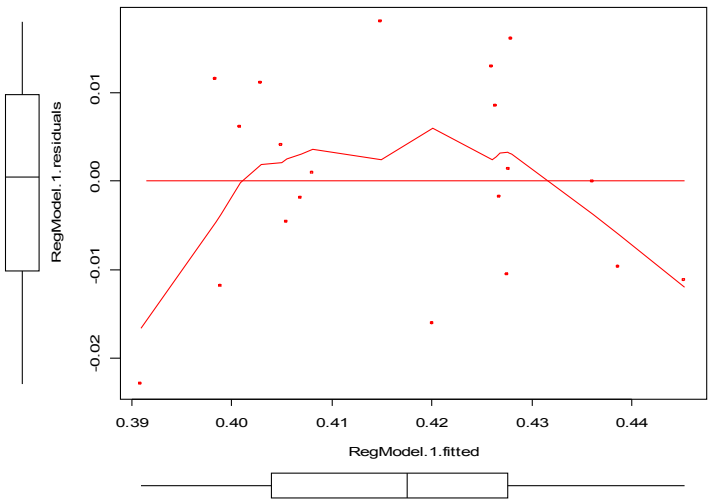




On peut également vérifier que les résidus studentisés se retrouvent en majorité entre -2 et 2 (4.).

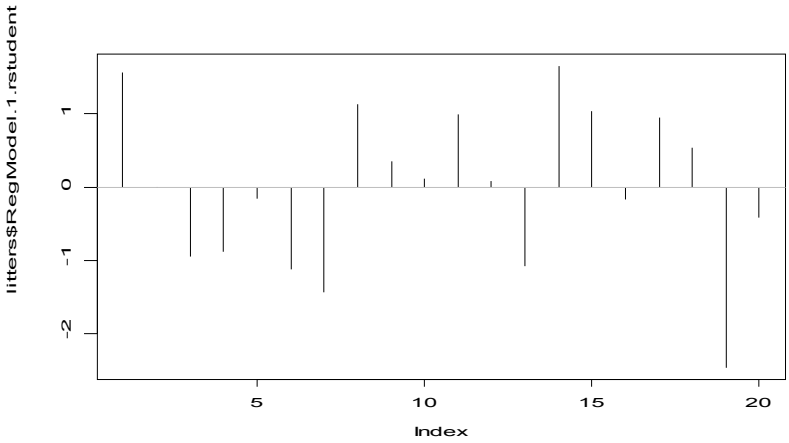
2. Homoscédasticité

```
Breusch-Pagan test
data:  brainwt ~ bodywt + lsize
BP = 0.6121, df = 1, p-value = 0.434
```

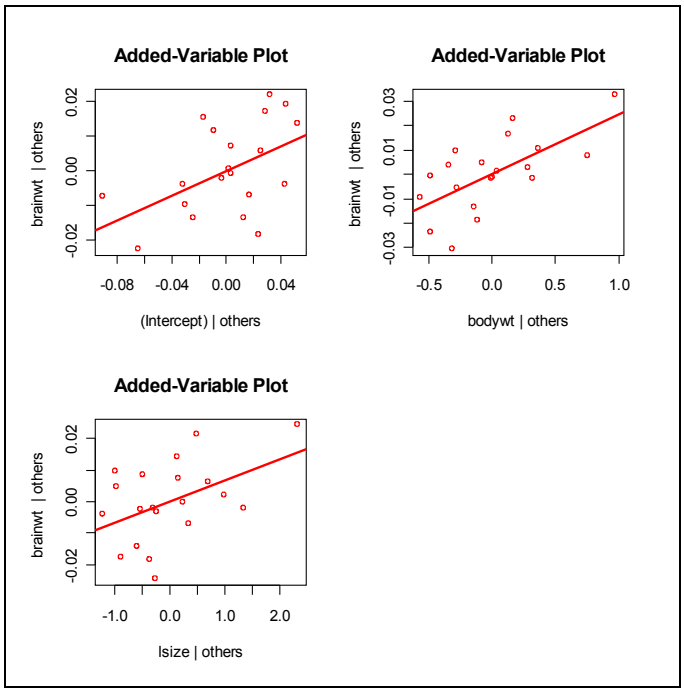


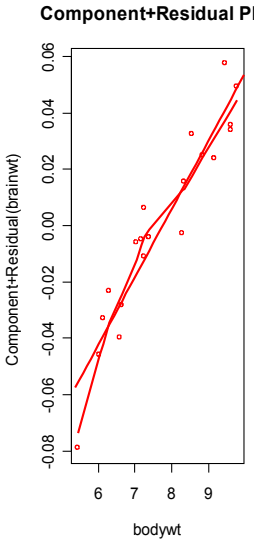
3. Indépendance

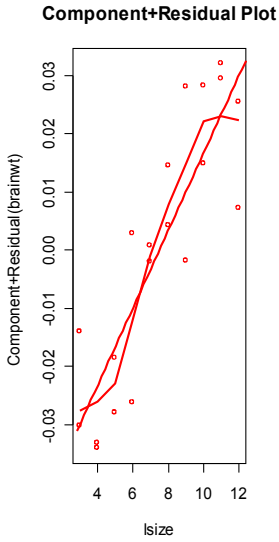
```
Durbin-Watson test
data:  brainwt ~ bodywt + lsize
DW = 1.7547, p-value = 0.426
alter hyp: true autocorelation is not 0
```



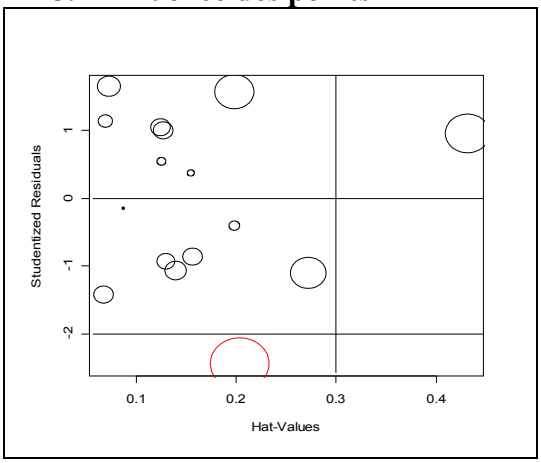
4. Linéarité







5. Influence des points



VI – COMPARAISON DE MOYENNES : ANOVA A UN FACTEUR

Les données et objectifs :

Cette technique s'applique à des tableaux décrivant pour chaque individu une variable quantitative *Y* en fonction d'un facteur. On appelle facteur une variable qualitative prenant plusieurs modalités dont on étudie l'influence sur la variable *Y*. Par exemple, le facteur peut être la variété, le dosage d'un apport nutritif, le type d'engrais, un traitement …

Le tableau type sous R est :

Y	Facteur	Facteur est une colonne déclarée en facteur
y_{11}	1	Dans Rcmdr, construire facteur en numérique (1, 2 , 3 ...)
y_{12}	1	puis convertir en facteur (donnees, gérer les variables)
$y_{2\,k}$	2	k représente la répétion de la mesure pour la modalité 2

L'objectif est d'évaluer si le facteur conditionne significativement la variable *Y*.

Pour tester l'hypothèse nulle H_0 "toutes les moyennes sont égales", on a le plus souvent recours à l'analyse de variance (ANOVA) développée par Fischer.

Le modèle linéaire pour un facteur

En présence d'un seul facteur, on considère que la variable *Y* suit pour chaque modalité *i* une loi normale $\mathcal{N}(\mu_i,\sigma^2)$.

On écrit alors : $y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$ **ou** $y_{ik} = \mu_i + \varepsilon_{ik}$

avec μ la moyenne générale de *Y*

α_i l'effet du la modalité i sur la moyenne $(\mu + \alpha_i = \mu_i)$

ε_i une variable aléatoire suivant une loi normale centrée $\mathcal{N}(0,\sigma^2)$

Ajustement du modèle:

Les coefficients sont estimés en minimisant l'*erreur quadratique moyenne* :

$$\text{Critère des moindres carrés} = \frac{1}{n}\sum_{ik}\left(y_{ik}-\overline{y_i}\right)^2$$

Les différences entre les valeurs observées y_{ik} et les valeurs prédites par le modèle notée $\widehat{y_{ik}} = \overline{y_i}$, s'appellent les *résidus, notés* $e_{ik} = y_{ik} - \overline{y_i}$

Les estimations des coefficients sont :

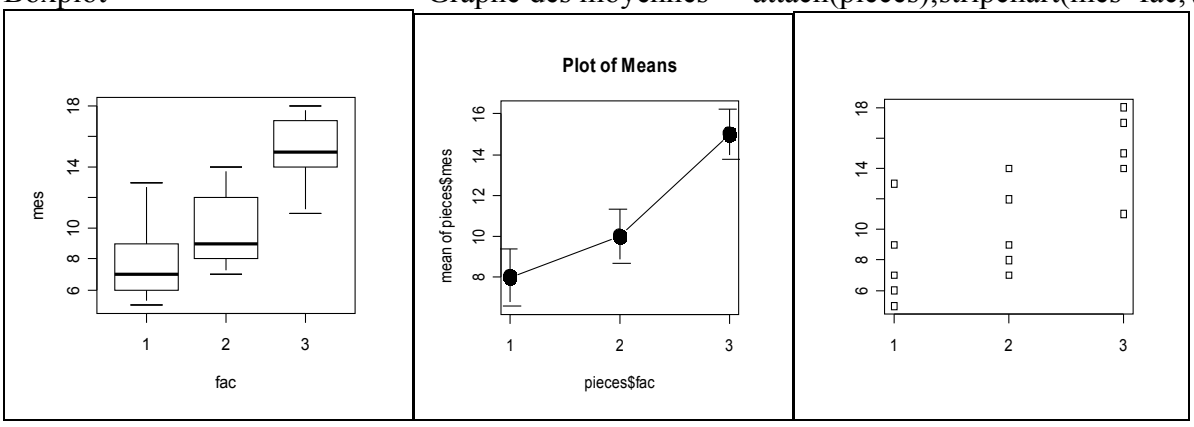
- $\overline{y} = \frac{1}{n}\sum_{ik} y_{ik}$ pour μ
- $\overline{y_i} = \frac{1}{n_i}\sum_{k} y_{ik}$ pour μ_i **soit** $a_i = \overline{y_i} - \overline{y}$ **pour** α_i
- $s^2 = \frac{1}{n-q}\sum_{ik}\left(y_{ik}-\overline{y_i}\right)^2$ pour σ^2 avec q le nombre de modalités

Exemple : Cinq pièces sont prélevées au hasard dans la production de trois machines, A, B et C. Chacune des pièces est ensuite mesurée par un seul opérateur. Les mesures sont présentées dans le tableau ci-dessous:

facteur	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C
mesure	5	7	6	9	13	8	14	7	12	9	14	15	17	18	11

Construire le fichier pieces et convertir le facteur.

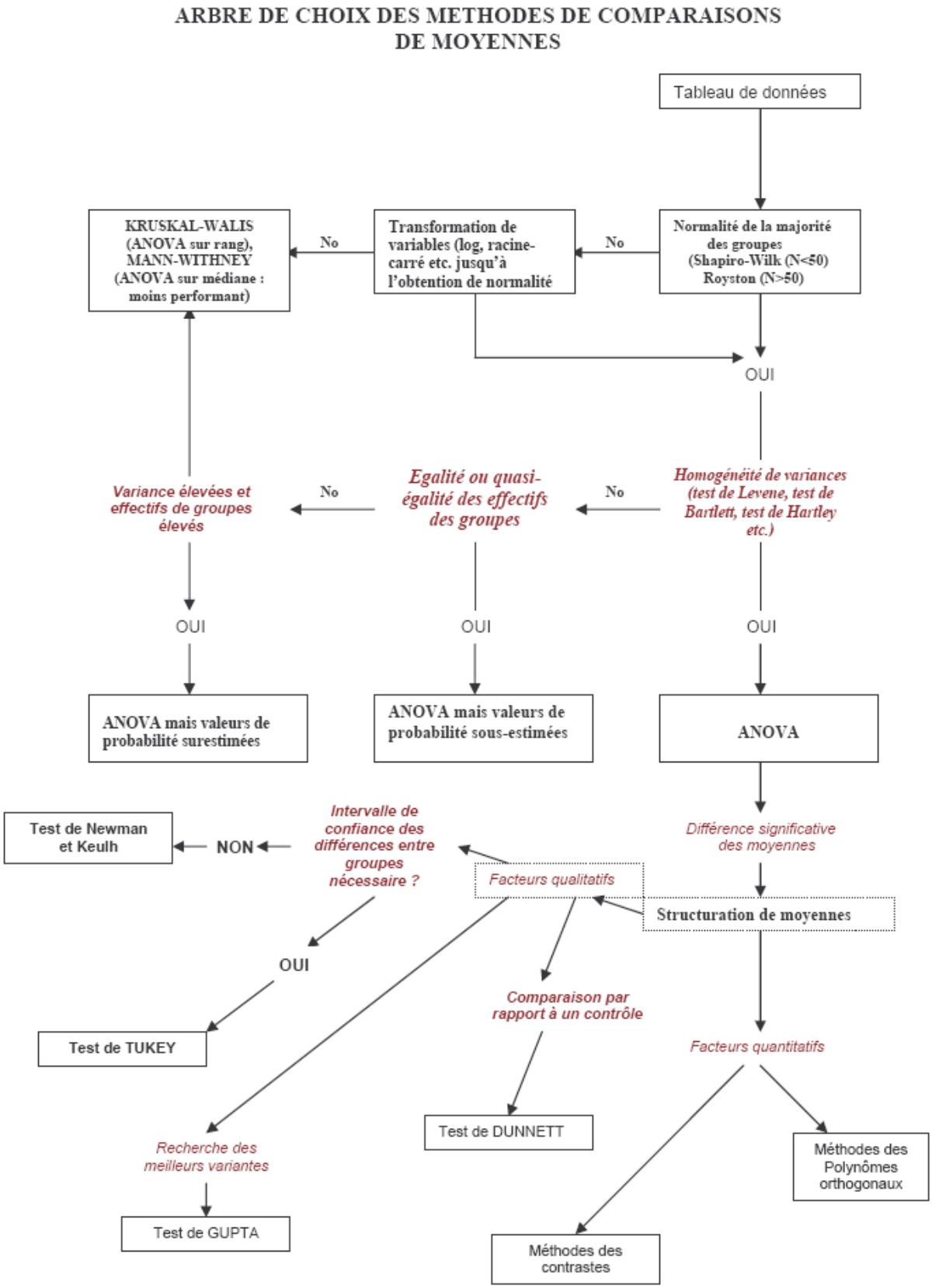
Boxplot Graphe des moyennes attach(pieces);stripchart(mes~fac,vert=T)



Calculer les moyennes et écart-type de chaque modalité

Calculer les estimations α_i et s^2 :

Bilan ANOVA


ARBRE DE CHOIX DES METHODES DE COMPARAISONS DE MOYENNES
Tableau de données
Normalité de la majorité des groupes (Shapiro-Wilk (N<50) Royston (N>50)
No
Transformation de variables (log, racine-carré etc. jusqu'à l'obtention de normalité
No
KRUSKAL-WALIS (ANOVA sur rang), MANN-WITHNEY (ANOVA sur médiane : moins performant)
OUI
Homogénéité de variances (test de Levene, test de Bartlett, test de Hartley etc.)
No
Egalité ou quasi-égalité des effectifs des groupes
No
Variance élevées et effectifs de groupes élevés
OUI
OUI
OUI
ANOVA
ANOVA mais valeurs de probabilité sous-estimées
ANOVA mais valeurs de probabilité surestimées
Différence significative des moyennes
Structuration de moyennes
Facteurs qualitatifs
Intervalle de confiance des différences entre groupes nécessaire ?
NON
Test de Newman et Keulh
OUI
Test de TUKEY
Comparaison par rapport à un contrôle
Test de DUNNETT
Recherche des meilleurs variantes
Test de GUPTA
Facteurs quantitatifs
Méthodes des Polynômes orthogonaux
Méthodes des contrastes

Fiche 29 – Test ANOVA

Objectif : L'objectif du test est de montrer l'existence de différences significatives entre les moyennes.

Hypothèse nulle : H_0 « les moyennes sont toutes égales » contre H « les moyennes ne sont pas toutes égales ».

Il est important de comprendre que l'ANOVA n'est pas un test permettant de « classer » des moyennes. On étudiera certains tests dits « tests de comparaison multiples » permettant de répondre à ce problème au paragraphe III.

Principe du test :

L'écart total e_T se décompose en un écart expliqué par le modèle, e_B et un écart résiduel e_W, soit :

$$\begin{array}{ccccc} e_T & = & e_B & + & e_W \\ y_{ik} - \overline{y} & = & (\overline{y_i} - \overline{y}) & + & (y_{ik} - \overline{y_i}) \end{array}$$

On utilise l'écriture anglosaxonne avec :

B pour between groups (entre groupes)

W pour within group (intra groupe)

On montre que l'on a alors l'égalité :

$$\begin{array}{ccccc} SCE_T & = & SCE_B & + & SCE_W \\ \sum_{ik} (y_{ik} - \overline{y})^2 & = & \sum_{ik} (\overline{y_i} - \overline{y})^2 & + & \sum_{ik} (y_{ik} - \overline{y_i})^2 \end{array}$$

En notant SCE_T la somme des carrés des écarts totaux, SCE_B la somme des carrés des écarts inter-groupes et SCE_W la somme des carrés des écarts intra-groupe.

On obtient les différentes variances, ou carrés moyens, en divisant les sommes de carrés d'écart par leurs degrés de liberté. Total: $n-1$ Inter : $q-1$ Intra : $n-q$

On vérifie que $q-1+n-q=n-1$. Les degrés de liberté se décomposent de manière additive comme les sommes de carrés d'écarts. On obtient alors les carrés moyens (mean squares) ou variances par les formules suivantes :

$$CM_T = \frac{SCE_T}{n-1} \qquad CM_B = \frac{SCE_B}{q-1} \qquad CM_W = \frac{SCE_W}{n-q}$$

avec n l'effectif total et q le nombre de modalités.

On montre alors que la statistique $F = \frac{CM_B}{CM_W}$ suit la loi de Fisher à (q-*1;n-q*) ddl sous H_0.

Test : On teste H_0 « les moyennes sont toutes égales » contre H « les moyennes ne sont pas toutes égales

- si $F < F_{1-\alpha}(q\text{-}1, n\text{-}q)$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce égal à α (ou p).

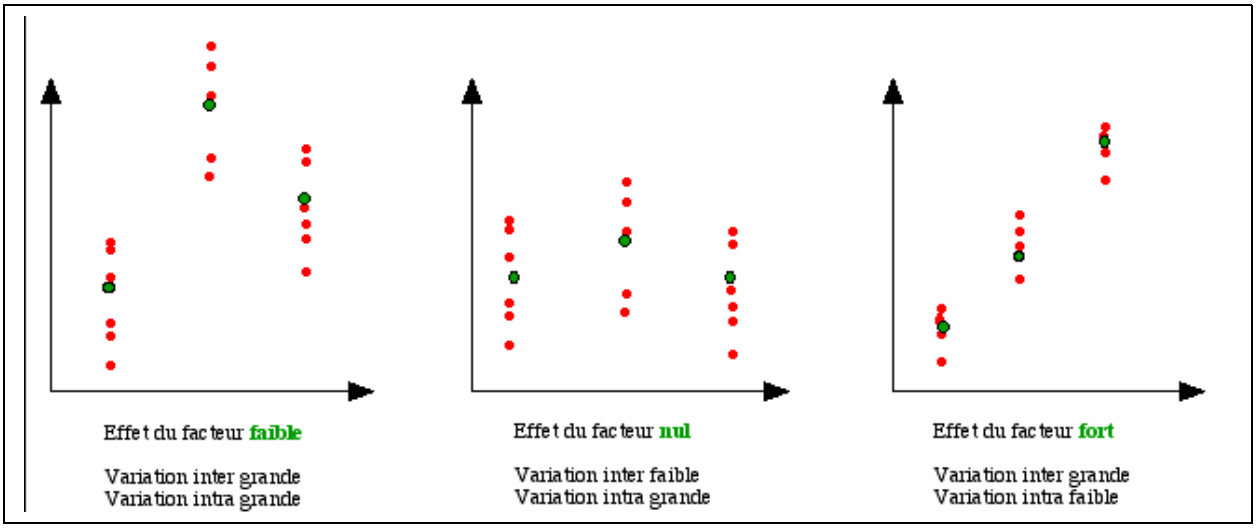
Tableau d'analyse de variance : les résultats du test se présente sous forme d'un tableau :

```
Response: Y
          Sum Sq Df F value    Pr(>F)
facteur      130  2     7.5 0.007707 **
Residuals    104 12
```

Conditions d'utilisation : On retrouve les trois conditions :

- les observations sont indépendantes,
- la variable Y suit la loi normale au sein de chaque modalité,
- la variance de Y est la même pour toutes les modalités.

Interprétation graphique :



Exemple : Construire le tableau d'analyse de variance pour l'exemple pièces

A la main :

Fiche 30 – Validation du modèle

L'hypothèse principale de cette méthode est **l'indépendance des données**. Cette propriété conduit à construire des protocoles expérimentaux permettant justement de contrôler d'éventuels biais : gradient (terrain, lumière), rôle de l'expérimentateur, du terrain ....
Ne pas respecter cette propriété conduit à mesurer et tester autre chose que l'effet étudié, autant dire les données deviennent **inexploitables**.

La décomposition de la variance est toujours valable, quelle que soit la distribution des variables étudiées. Cependant, lorsqu'on réalise le test final (test F), on admet **la normalité des distributions** (puisque le F est le rapport de deux khi-deux, qui sont des sommes de carrés de lois normales). L'ANOVA fait donc l'hypothèse de normalité. Elle est cependant assez robuste à la non normalité, ce qui permet de l'utiliser dans une grande variété de conditions.

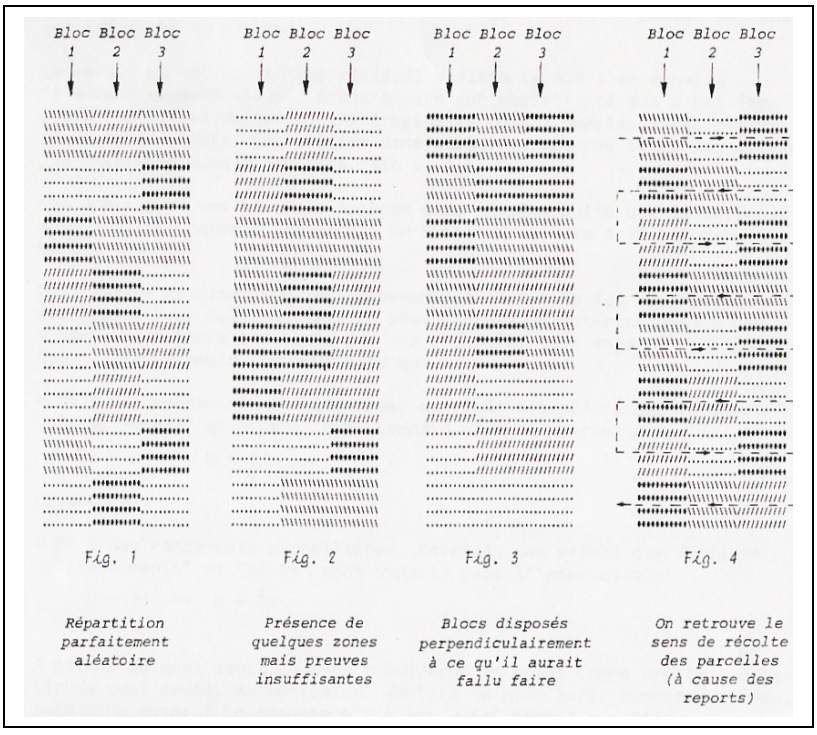
A l'opposé, l'ANOVA fait une autre hypothèse très forte et moins évidente. Il est en effet nécessaire que la variance dans les différents groupes soit la même. C'est l'hypothèse d'**homoscedasticité**. L'ANOVA y est sensible. Il est donc nécessaire de la tester avant toute utilisation.

L'analyse des résidus e_{ij} est particulièrement utile pour répondre aux hypothèses de normalité et d'homoscédasticité.

1. Indépendance

L'indépendance entre les différentes valeurs de la variable mesurée y_{ij} est une condition essentielle à la réalisation de l'analyse de variance. Les q échantillons comparés sont indépendants. L'ensemble des n individus est réparti au hasard (randomisation) entre les q modalités du facteur contrôlé.

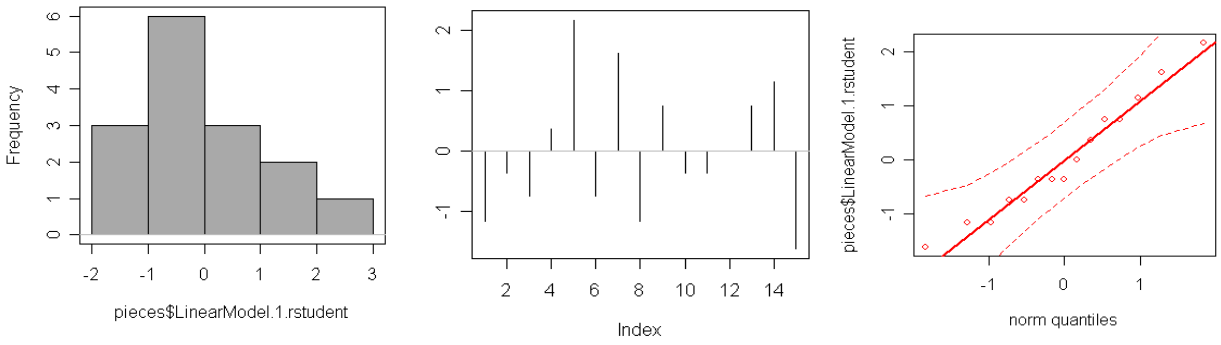
On est conduit à construire des protocoles expérimentaux basés par exemple sur la construction de bloc, de carré latin ...
La cartographie des résidus est un moyen d'évaluer d'éventuels problèmes:



2. Normalité

La variable quantitative étudiée doit suivre une loi normale dans les *q* populations comparées.
La normalité de la variable pourra être testée à l'aide du test de Shapiro-Wilk si les effectifs sont suffisamment importants. Sinon le test non paramétrique de Lilliefors permet de tester l'ajustement à loi normale lorsque les effectifs sont faibles.

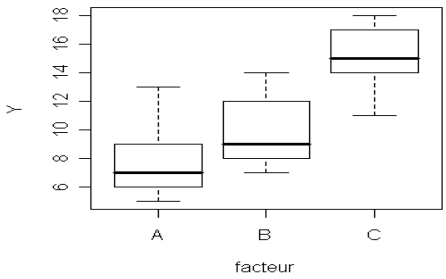
Observations graphiques :




```
shapiro.test(pieces$LinearModel.1.residuals)
        Shapiro-Wilk normality test
data:  pieces$LinearModel.1.residuals
W = 0.9535, p-value = 0.5816
```

3. Homoscédasticité de la variance

Observation graphique:



On dispose de différents tests pour vérifier l'égalité de la variance dans les différentes populations :

- Le **test de Lévène** est le test le plus satisfaisant pour effectuer la comparaison multiple de variances mais sa réalisation est assez longue car il correspond à une ANOVA sur les résidus e_{ij}.
- Le **test de Bartlett** est dédié à la comparaison multiple de variances avec un nombre de répétitions n_i différent selon les modalités i du facteur. Mais ce test est très sensible à l'hypothèse de normalité des *p* populations (peu robuste).
- Le **test de Hartley** est dédié à la comparaison multiple de variances avec un nombre de répétitions n_i identiques selon les modalités i du facteur. Mais ce test est très sensible à l'hypothèse de normalité des *p* populations (peu robuste).

```
bartlett.test(mes ~ fac, data=pieces)
Bartlett test of homogeneity of variances
Bartlett's K-squared = 0.0752, df = 2, p-value = 0.963

levene.test(pieces$mes, pieces$fac)
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  2  0.0175 0.9826
      12
```

Fiche 31 – Robustesse de l'ANOVA

De nombreux travaux ont étudié la robustesse de l'ANOVA1 vis à vis des écarts aux hypothèses faites.

Hypothèses	**Test**	**Robustesse de l'ANOVA**
Normalité de *Y*	Test de Shapiro Wilk	Très robuste si indépendance et égalité des variances
Homoscédasticité des *p* distributions	Test de Levène ou de Bartlett	Très robuste à l'inégalité des variances
Indépendance des *p* distributions	Plan expérimental	Pas robuste

Remarque : L'analyse de variance à un facteur contrôlé est relativement peu sensible à l'inégalité des variances ainsi qu'à la non normalité lorsque les échantillons comparés sont de grandes tailles.

Dans le cas de forts écarts aux hypothèses de normalités ou d'homoscédasticité, on peut :

- **Utiliser un changement de variables** pour vérifier les hypothèses (fiche #Remédiation)
- **Utiliser un test non-paramétrique** comme le test de Kruskal Wallis pour comparer les moyennes. le test de Kruskal Wallis, qui est lui-même une extension du test non paramétrique de Wilcoxon Mann Whitney (du nom de ses trois auteurs). Ce test utilise des données du même type, mais se limite au cas "un facteur". Pour le cas à deux facteurs sans répétition, on peut envisager d'utiliser le test de Friedman.

```
kruskal.test(mes ~ fac, data=pieces)
        Kruskal-Wallis rank sum test
data:  mes by fac
Kruskal-Wallis chi-squared = 7.7566, df = 2, p-value = 0.02069
```

Fiche 32 – Planification expérimentale

Afin de tester l'effet d'un ou plusieurs facteurs, il est indispensable au préalable de planifier l'expérience. La planification expérimentale correspond aux choix à effectuer concernant :

- le **nombre de facteurs** à étudier et le **nombre de niveaux** par facteur,
- le **nombre d'unités expérimentales**,
- la **façon de regrouper les unités**.

Le plan expérimental va dépendre naturellement du nombres d'unités possibles (coût, disponibilité en place) et donc limitera le choix des facteurs et des niveaux en conséquence. Le nombre d'unités conditionnent directement la puissance du test, à savoir la capacité du test statistiques à mettre en évidence une différence donnée.

Il va de soit que **la planification est l'étape la plus importante d'une expérimentation** (après il est trop tard). Il est donc nécessaire de bien définir au préalable l'objectif de l'expérimentation, les dispositifs possibles, les contraintes, les biais possibles ...

La planification expérimentale repose sur trois principes :

- **La randomisation** : les niveaux des facteurs sont répartis au hasard dans les unités expérimentales afin de limiter des biais, par exemple :
  - l'existence de gradient dans un champ,
  - l'existence de gradient dans une serre (lumière, arrosage),
  - l'influence du notateur s'il y a plusieurs expérimentateurs.

  Ces différents biais conduisent à mesurer un effet autre que le facteur étudié.
- **Les répétitions** : le nombre de répétitions augmentent sensiblement la puissance du test statistique. Ainsi une même différence pourra être non significative avec 5 répétitions mais significative avec 10 répétitions. Il existe des table permettant d'évaluer le nombre de répétitions nécessaires pour mettre en évidence une différence donnée.
- **Le contrôle de l'erreur** : L'erreur résiduelle doit être la plus faible possible en limitant ainsi l'effet de facteurs non contrôlés afin d'augmenter la puissance du test (le *F*). Par exemple, on choisira des unités expérimentales de petite taille regroupées au sein de blocs.

Nous renvoyons à l'ouvrage de P. Dagnelie en libre accès internet pour une information complète et nous nous limiterons à présenter succintement deux plans expérimentaux :

- **La méthode des blocs**

On désigne par blocs des ensembles dans lesquels sont regroupées les unités expérimentales de telle sorte qu'elles soient aussi semblables que possible à l'intérieur de chaque bloc. On peut s'attendre ainsi à ce que l'erreur expérimentale soit moindre que pour un même nombre d'unités aléatoirement situées à l'intérieur de la totalité de l'espace expérimental.

Les blocs sont généralement déterminés pour tenir compte, outre les causes contrôlables définies par les facteurs étudiés, d'autres causes qu'il peut être difficile, voire impossible, de maintenir constantes sur la totalité des unités expérimentales de l'expérience.

Les variations entre les blocs sont alors éliminés lorsque l'on compare les effets des facteurs. Cette méthode peut être comparée à une analyse de variance à deux facteurs croisés. Le premier facteur étant le facteur étudié, le second se rapportant aux blocs.

Si toutes les situations sont représentées dans l'expérience réalisée, on dit qu'on utilise un plan à blocs complets; si ce n'est pas le cas, c'est un plan à blocs incomplets.

Exemple : si on compare le rendement de quatre variétés de maïs en les semant sur un lot de parcelle (six par exemple); les différences de fertilité de ces dernières vont introduire une variabilité parasite, nuisible pour la comparaison. L'idéal serait de découper chaque parcelle en quatre, de répartir

aléatoirement chaque variété dans chaque quart pour comparer la productivité de chaque espèce de maïs au sein de chaque parcelle, et finalement résumer ces six comparaisons en une seule conclusion.

Parcelle 1 (bloc 1)	Rendement Maïs **2**	Rendement Maïs **1**	Rendement Maïs **4**	Rendement Maïs **3**
Parcelle 2 (bloc 2)	Rendement Maïs **1**	Rendement Maïs **3**	Rendement Maïs **2**	Rendement Maïs **4**
Parcelle 3 (bloc 3)	Rendement Maïs **2**	Rendement Maïs **3**	Rendement Maïs **1**	Rendement Maïs **4**
Parcelle 4 (bloc 4)	Rendement Maïs **4**	Rendement Maïs **2**	Rendement Maïs **3**	Rendement Maïs **1**
Parcelle 5 (bloc 5)	Rendement Maïs **3**	Rendement Maïs **4**	Rendement Maïs **1**	Rendement Maïs **2**
Parcelle 6 (bloc 6)	Rendement Maïs **1**	Rendement Maïs **4**	Rendement Maïs **2**	Rendement Maïs **3**

Une analyse de variance à deux facteurs (le premier facteur correspond au rendement; le second à l'effet bloc) pourra nous dire si, après élimination des effets de bloc, il existe une différence significative entre les variétés de maïs.

- **La méthode des carrés latins**

Le carré latin est un dispositif qui permet de contrôler l'hétérogénéité du matériel expérimental dans deux directions. Dans certaines expériences, il arrive qu'une série de k traitements soit donnée à des sujets à des moments différents (ou à des endroits différents du corps s'il s'agit de crèmes), et que l'ordre (ou le lieu d'application) dans lequel est donnée la séquence soit potentiellement important. Il est alors indispensable de tenir compte dans l'analyse d'un effet "ordre (ou lieu) d'administration" et faire attention à ce que chaque traitement soit donné de façon équilibrée en $1^{ère}$, $2^{ème}$, ..., $k^{ème}$ position. L'utilisation des carrés latins répond à cet impératif.

Prenons l'exemple de 4 traitements donnés à 4 moments différents de la journée. Les sources d'erreur sont :

- les moments de la journée

- l'ordre d'administration

Dans la figure suivante sont représentés par des lettres les 4 traitements. Les lignes du tableau représente les moments; les colonnes, l'ordre.

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Chaque traitement doit apparaître une fois dans chaque ligne et dans chaque colonne. Dans un carré latin, le nombre de lignes doit être égal au nombre de colonnes. Ainsi le carré latin sera toujours de type 3 x 3 ou 4 x 4 …

Pour un carré latin 3 x 3, il y a donc 12 configurations possibles; pour un carré latin 4 x 4, 576; pour un carré latin 5 x 5, 161.280 combinaisons différentes …

La méthode des carrés latins est assimilée à une analyse de variance à trois facteurs. En effet, le premier facteur est le facteur traitement; les deux autres correspondent aux sources d'erreur (facteur ligne et facteur colonne).

Fiche 33 – Comparaisons multiples

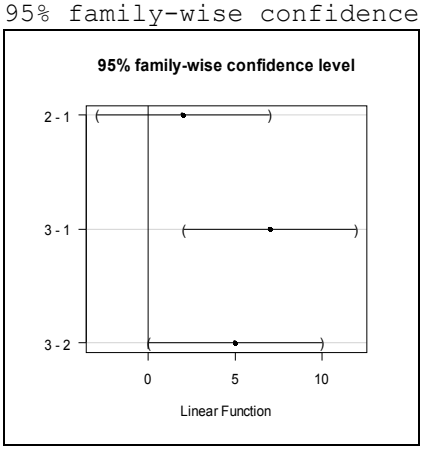
Lorsqu'une ANOVA a rejeté l'hypothèse nulle à un certain niveau de risque α. Ce rejet se fait "en bloc", sans aucun détail sur les raisons qui l'ont provoqué. Lorsqu'on planifie plusieurs comparaisons, il est nécessaire de prendre des mesures pour limiter le taux d'erreur de l'ensemble. Ainsi par exemple, il serait incorrect de réaliser des tests t classiques pour réaliser des comparaisons multiples car cela provoquerait une inflation très importante du risque de commettre au moins une erreur de première espèce. La solution la plus simple pour éviter ce problème consiste simplement à fixer un taux d'erreur α par comparaison beaucoup plus sévère que 0,05 de telle sorte que le taux d'erreur de l'ensemble reste raisonnable. C'est cette technique simple qui est à la base du test t de Dunn-Bonferroni (parfois dénommé simplement **le test t de Bonferroni**).

D'autres tests (dits "*post hoc*", ou "*a posteriori*") ont été développés dans le but d'analyser plus finement la situation, et de comprendre ce qui a provoqué ce rejet. Par exemple :

- Le test de Dunnett s'intéresse à la situation où un des groupes est un groupe "témoin", et où l'on cherche à mettre en évidence le groupe dont la moyenne est significativement différente de celle du groupe témoin (typiquement, un groupe "placebo").

- Le test de Newman-Keuls qui réduit le nombre de comparaisons deux-à-deux en garantissant comme non significatives les différences entre moyennes prises "en sandwich" entre deux moyennes dont les différences ne sont elles-mêmes pas significatives. Le test de Newman-Keuls est un test de comparaison de moyennes par paire, pratiqué à l'issue d'une ANOVA.

```
Simultaneous Confidence Intervals for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = mes ~ fac, data = pieces)
Estimated Quantile = 2.6669
Linear Hypotheses:
           Estimate lwr      upr
2 - 1 == 0  2.00000 -2.96544  6.96544
3 - 1 == 0  7.00000  2.03456 11.96544
3 - 2 == 0  5.00000  0.03456  9.96544

95% family-wise confidence level
```




```
> pairwise.t.test(mes,fac,P.adj="bonf")
        Pairwise comparisons using t tests with pooled SD
data:  mes and fac
  1      2
2 0.3039 -
3 0.0082 0.0397

P value adjustment method: holm
```

VII – COMPARAISONS DE MOYENNES : ANOVA à 2 FACTEURS

Objectif et données : On étudie maintenant une variable quantitative Y en fonction de deux facteurs, le rendement en fonction de la variété et de l'engrais utilisé par exemple. L'objectif est alors de tester la signification de l'effet moyen de chaque facteur et de leur intéraction.

Le tableau de donnée se présente sous R sous la forme :

Y	Facteur I	Facteur II	
y_{111}	1	1	
y_{235}	2	3	5ème répétition de la combinaison 2-3
y_{ijk}	i	j	k ème répétition de la combinaison $i-j$

Sous Rcmdr, il faut créer les facteurs en numérique (1, 2, 3...) puis les convertir en facteur (données – gérer - convert)

La construction du test est similaire à celle de l'ANOVA à un facteur. Les modèles utilisés ici sont :

Modèle sans répétition ou dispositif en bloc complets :

On écrit : $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

avec
- μ la moyenne générale de Y
- α_i l'effet du à la modalité i du facteur A sur la moyenne
- β_j l'effet du à la modalité j du facteur B sur la moyenne
- ε_{ijk} une variable aléatoire suivant une loi normale centrée $\mathcal{N}(0,\sigma^2)$

Modèle avec répétition et interaction

On écrit : $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

avec
- μ la moyenne générale de Y
- α_i l'effet du à modalité i du facteur A sur la moyenne
- β_j l'effet du à modalité j du facteur B sur la moyenne
- γ_{ij} l'effet du à l'interaction de i et j sur la moyenne
- ε_{ijk} une variable aléatoire suivant une loi normale centrée $\mathcal{N}(0,\sigma^2)$

Principe du test :

On décompose la somme des carrés des écarts et à l'aide du test F on peut tester la signification de :

- l'effet moyen de chaque facteur,
- l'intraction entre les deux facteurs.

Conditions d'utilisation : (fiche #Validation du modèle):

- Indépendance des observations
  Conditionnée par le protocole expérimental, test avec la cartographie des résidus
- Normalité des écarts résiduels
  test global sur les résidus et si effectifs suffisants test dans chaque groupe,
- Homoscédasticité
  test de Bartlett pour chaque facteur, test sur les combinaison de facteurs si effectifs suffisants

Compléments : Parmi les méthodes qui découlent de l'ANOVA, citons :

- les plans d'expériences avec facteurs emboîtés : l'influence d'un facteur dépend d'un autres facteurs (par exemple l'impact de différents traitements sur les individus d'une portée),
- Analyse de la covariance : la variable étudiée est influencée par des facteurs et d'autres variables quantitatives,
- Dispositifs factoriels : combinaisons de plusieurs facteurs ...

Fiche 34 – Test ANOVA

Test d'ANOVA

Le principe du test est similaire. On décompose la somme des carrés des écarts en fonction du facteur étudié et on construit une statistique qui suit la loi de Fisher.

Il est ainsi possible de tester si une interaction est significative, si l'effet moyen d'un facteur est significatif …

Tableau d'analyse de variance

Source	ddl	SCE	CM	*F*	*p*
facteur 1	$q_1 - 1$	SCE_{F1}	$\frac{SCE_{F1}}{q_1 - 1}$	$\frac{CM_{F1}}{CM_{res}}$	
facteur 2	$q_2 - 1$	SCE_{F2}	$\frac{SCE_{F2}}{q_2 - 1}$	$\frac{CM_{F2}}{CM_{res}}$	
Interaction	$(q_1 - 1)(q_2 - 1)$	$SCE_{F1\,F2}$	$\frac{SCE_{F1\,F2}}{(q_1 - 1)(q_2 - 1)}$	$\frac{CM_{F1\,F2}}{CM_{res}}$	
résiduelles	$n - q_1 - q_2 - 1$	SCE_{res}	$\frac{SCE_{res}}{n - q_1 - q_2 - 1}$		

- La signification de l'effet d'un facteur indique qu'en moyenne ce facteur à un effet significatif (il existe en moyenne une différence entre au moins deux modalités)
- La signification de l'interaction indique que les deux facteurs interagissent.
- L'analyse de ces résultats se fait grace à un diagramme des interactions (voir la fiche)

Exemple : On souhaite étudier le rendement d'une céréale en fonction de l'engrais et de la nature du terrain. On cultive p=2 types de terrain T1 et T2 et q=3 types d'engrais E1, E2 et E3. Chacune des combinaisons T*E est répétée 4 fois. (fichier **rendement.txt**)

Terrain	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2
Engrais	E1	E1	E1	E1	E2	E2	E2	E2	E3	E3	E3	E3	E1	E1	E1	E1	E2	E2	E2	E2	E3	E3	E3	E3
Rendt	61	76	47	77	34	30	67	62	85	104	74	75	77	47	54	77	46	50	29	53	69	70	75	85

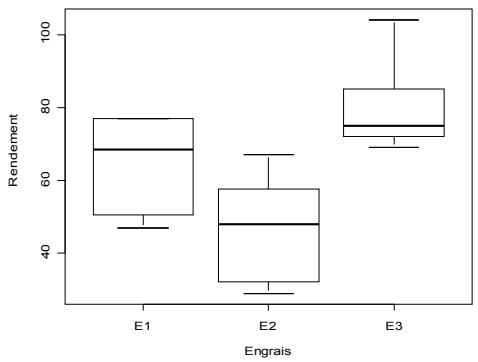
Examen des données

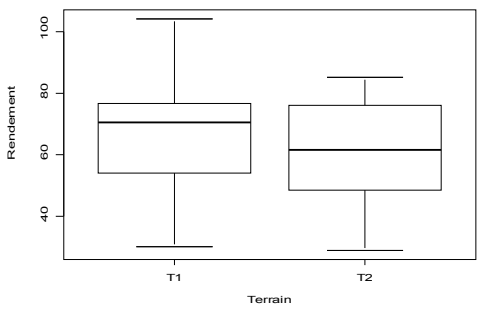
Facteur Engrais

```
     mean       sd n
E1 64.500 13.81511 8
E2 46.375 14.39184 8
E3 79.625 11.53798 8
```

Facteur Terrain

```
     mean       sd  n
T1     66 21.09287 12
T2     61 16.80909 12
```





Moyennes croisées (stat – résumé – tableau)

```
      Terrain
Engrais    T1    T2
     E1 65.25 63.75
     E2 48.25 44.50
     E3 84.50 74.75
```

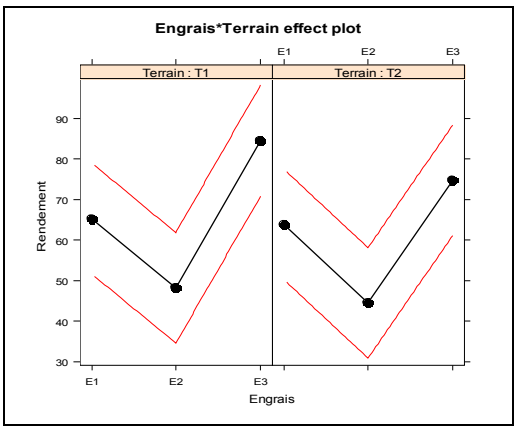
écarts type croisés

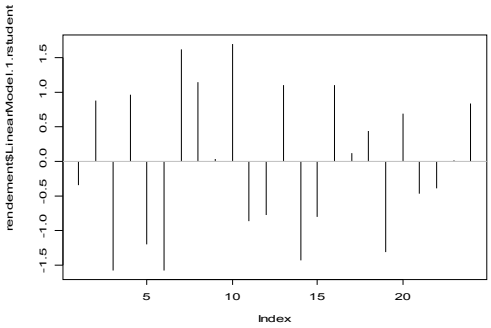
```
      Terrain
Engrais       T1        T2
     E1 14.19800 15.564382
     E2 18.94510 10.723805
     E3 13.91642  7.320064
```

Modèle 1 deux facteurs avec interaction :

```
lm(formula = Rendement ~ Engrais * Terrain, data = rendement)
               Sum Sq Df F value   Pr(>F)
Engrais        4434.2  2 11.4187 0.000628 ***
Terrain         150.0  1  0.7725 0.391018
Engrais:Terrain  72.8  2  0.1873 0.830759
Residuals      3495.0 18
```

```
Coefficients:                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         65.250      6.967   9.365 2.43e-08 ***
Engrais[T.E2]                      -17.000      9.853  -1.725   0.1016
Engrais[T.E3]                       19.250      9.853   1.954   0.0665 .
Terrain[T.T2]                       -1.500      9.853  -0.152   0.8807
Engrais[T.E2]:Terrain[T.T2]         -2.250     13.934  -0.161   0.8735
Engrais[T.E3]:Terrain[T.T2]         -8.250     13.934  -0.592   0.5612
```

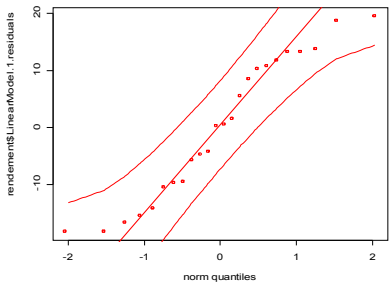




Validation

Normalité

```
> shapiro.test(rendement$LinearModel.1.residuals)
data:  rendement$LinearModel.1.residuals
W = 0.9327, p-value = 0.1119
```



Homoscédasticité

```
> bartlett.test(Rendement ~ Engrais, data=rendement)
data:  Rendement by Engrais
Bartlett's K-squared = 0.3483, df = 2, p-value = 0.8402
```

```
> bartlett.test(Rendement ~ Terrain, data=rendement)
data:  Rendement by Terrain
Bartlett's K-squared = 0.5377, df = 1, p-value = 0.4634
```

Modèle 2 simplifié à un facteur :

```
Anova Table (Type II tests)
          Sum Sq Df F value    Pr(>F)
Engrais   4434.3  2  12.524 0.0002628 ***
Residuals 3717.7 21
```

```
Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.500      4.704  13.711    6e-12 ***
Engrais[T.E2]  -18.125      6.653  -2.724   0.0127 *
Engrais[T.E3]   15.125      6.653   2.273   0.0336 *
```

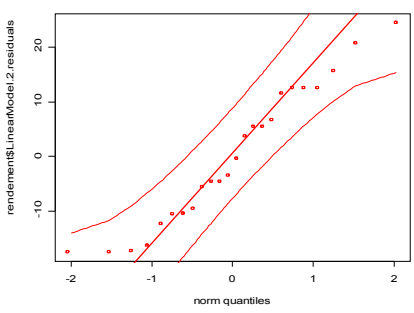
```
Residual standard error: 13.31 on 21 degrees of freedom
Multiple R-Squared: 0.5439,    Adjusted R-squared: 0.5005
```

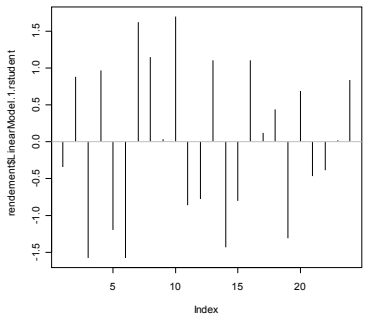
F-statistic: 12.52 on 2 and 21 DF, p-value: 0.0002628

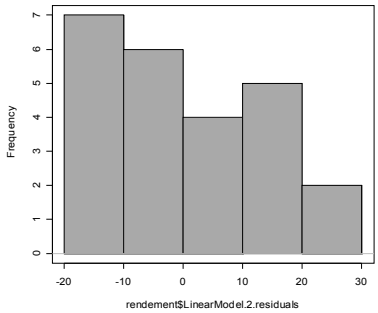
Validation :

Normalité

```
> shapiro.test(rendement$LinearModel.2.residuals)
data:  rendement$LinearModel.2.residuals
W = 0.9459, p-value = 0.2206
```



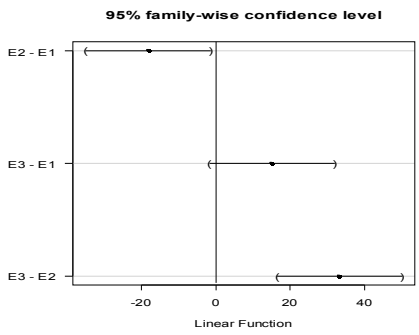




Homoscédasticité

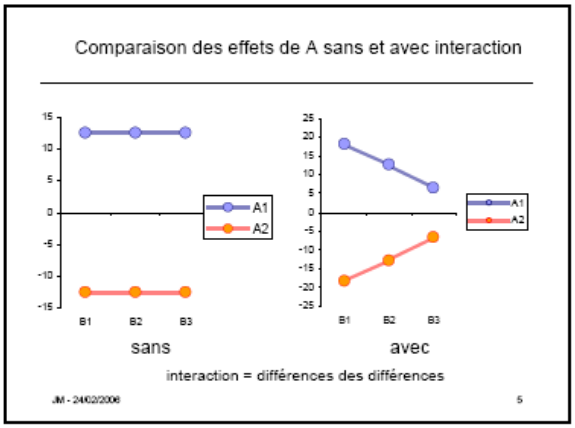
```
> bartlett.test(Rendement ~ Engrais, data=rendement)
data:  Rendement by Engrais
Bartlett's K-squared = 0.3483, df = 2, p-value = 0.8402
```

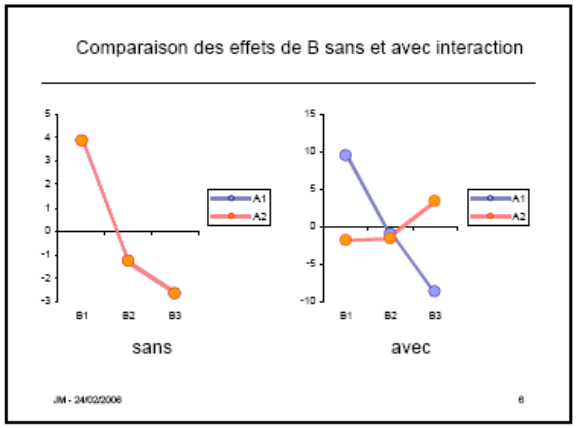
Comparaisons multiples

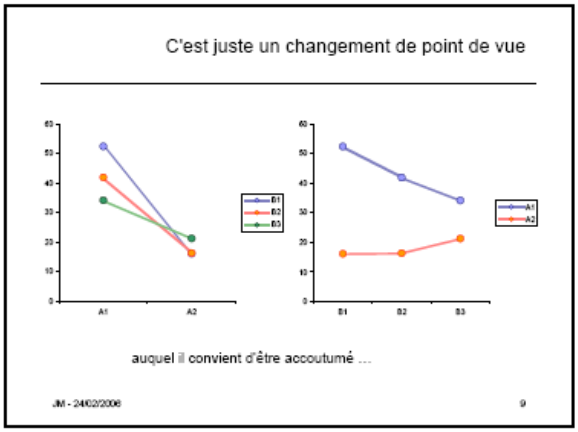


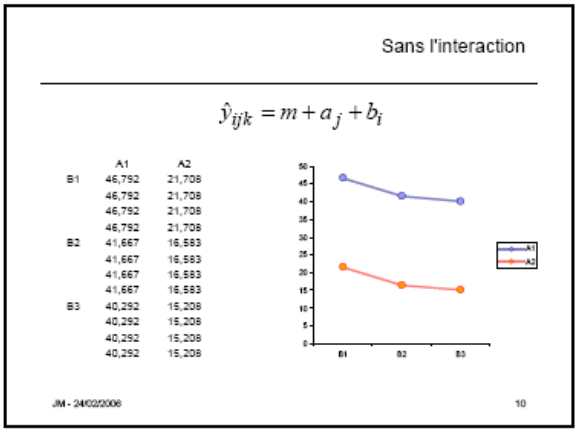
Fiche 35 – Interaction entre facteurs

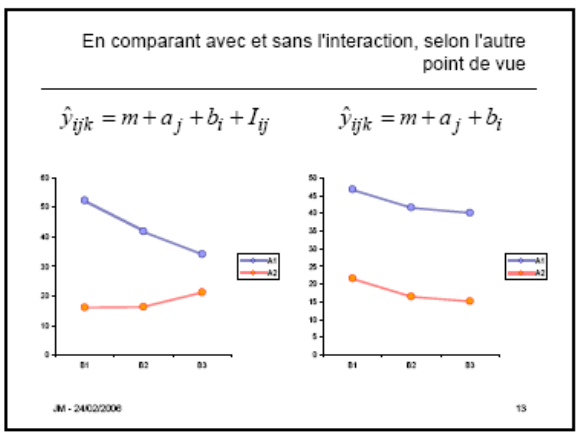
Interaction entre facteurs

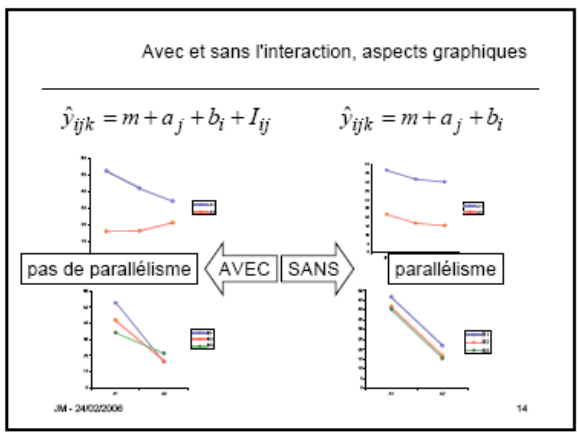












VIII – ANALYSE EN COMPOSANTES PRINCIPALES

Objectifs et données pour l'ACP

Cette technique s'applique à des tableaux décrivant chaque individu par p variables quantitatives X_k.
Les techniques classiques ne permettent que l'étude de la liaison entre deux variables : corrélation, régression et nuage de points par exemple.

L'**objectif** est ici de faire une synthèse de l'ensemble du tableau afin de :

- **synthétiser les liaisons entre variables** (cercle des corrélations), définir les variables qui vont dans le même sens, dans un sens opposé, indépendantes ...
- représenter dans un plan les individus afin de déterminer les individus proches ou éloignés, les regrouper en classe homogène, ... On parle de **topologie des individus**.
- construire de **nouvelles variables**, appelées composantes principales, non corrélées et qui permettent de synthétiser l'information

Ainsi, au lieu d'analyser le tableau à travers p variables, on se limitera à l'étude de quelques variables synthétiques, les composantes principales. La difficulté sera de donner un sens à ces variables et de proposer une analyse des résultats.

Le **tableau** se présente sous la forme :

	X_1	...	X_j	...	X_p
individu 1	x_{11}		$x_{1\,j}$		$x_{1\,p}$
...					
individu i	x_{i1}		x_{ij}		x_{ip}
...					
individu n	x_{n1}		x_{nj}		x_{np}

Exemple : Nous étudions dans cette partie les masses de différentes parties d'un groupe de 23 bovins constitué de 12 charolais (1 à 12) et 11 zebus (13 à 23).
Les variables représentent: poids vif. poids de la carcasse. poids de la viande de première qualité. poids de la viande totale. poids du gras. poids des os.

Retrouver à l'aide du logiciel R et son interface R-commander les différents résultats ci-dessous et expliquer leur apport respectif dans l'analyse des données.

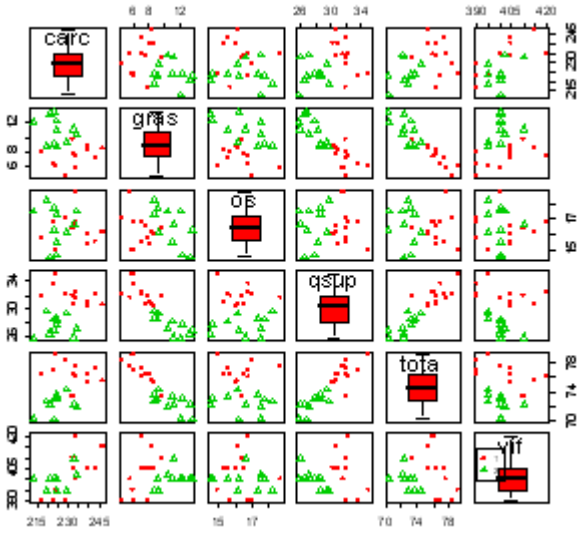
1. Paramètres statistiques: Moyenne et écart-type par race

```
Variable: carc
      mean       sd  n
1 233.0000 8.790491 12
2 224.2727 6.018154 11

Variable: os
      mean       sd  n
1 16.30833 0.9949494 12
2 16.51818 1.2584261 11

Variable: tota
      mean       sd  n
1 76.60000 1.502120 12
2 72.56364 1.297130 11

Variable: gras
       mean       sd  n
1  7.258333 1.439986 12
2 10.845455 1.758615 11

Variable: qsup
      mean       sd  n
1 31.99167 1.344658 12
2 27.66364 1.343334 11

Variable: vif
      mean       sd  n
1 402.5000 9.885711 12
2 399.7273 4.221159 11
```

Matrice des corrélations

```
       vif  carc  qsup  tota  gras    os
vif   1.00  0.64 -0.09 -0.13  0.16 -0.06
carc  0.64  1.00  0.28  0.39 -0.33 -0.09
qsup -0.09  0.28  1.00  0.89 -0.86 -0.06
tota -0.13  0.39  0.89  1.00 -0.91 -0.12
gras  0.16 -0.33 -0.86 -0.91  1.00 -0.27
os   -0.06 -0.09 -0.06 -0.12 -0.27  1.00
```

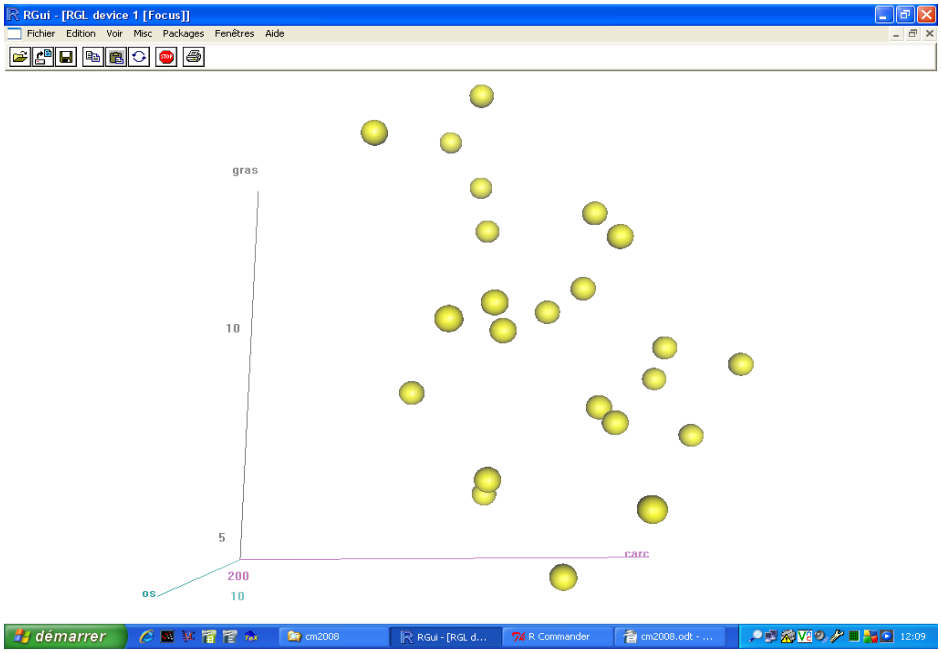
2. Représentation graphique



3. Nuage de points

Dans R, on peut utiliser les commandes suivantes pour construire des nuages 3d. On pourra changer les 3 variables. A défaut, utiliser R commander.

```
>library(rgl)
>attach(zebu)
>col <- ifelse(race==1,"blue","red")
>plot3d(gras,tota,qsup,type="s",col=col)            les échelles sont différentes
suivant les axes
>plot3d(gras,tota,qsup,type="s",col=col,xlim=c(0,450),ylim=c(0,450),zlim=c(0,450))
les échelles sont les mêmes
```



Quelques liens :

http://pbil.univ-lyon1.fr/R/enseignement.html
http://www.unilim.fr/pages_perso/vincent.jalby/m1sm/documents/m1sm_S_03.pdf
http://infolettres.univ-brest.fr/~carpenti/2006-2007/Ana-mult-1-2007.doc
http://www.lirmm.fr/~guindon/dess/acp.df

Fiche 36 – Principe de la méthode ACP

Chaque individu est décrit ici par p=6 variables quantitatives. Un individu est représenté par un point dont les coordonnées sont les valeurs prises par les 6 variables (espace à p=6 dimensions). On peut ainsi mesurer la distance entre deux individus à l'aide d'une distance classique entre deux points.

Le principe de l'ACP répond simultanément aux deux objectifs suivant :

- **Pour les individus**

L'objectif de la méthode ACP est de projeter les individus sur des axes appelés axes factoriels en conservant le mieux possible les distances entre individus. Cela revient à **déformer le moins possible le nuage de points initial lorsqu'on le projette sur un axe ou un plan**.
Dans la pratique, la projection sur l'axe F_1 permet d'obtenir le maximum de dispersion (=inertie = variance en une dimension) des points projetés sur l'axe.

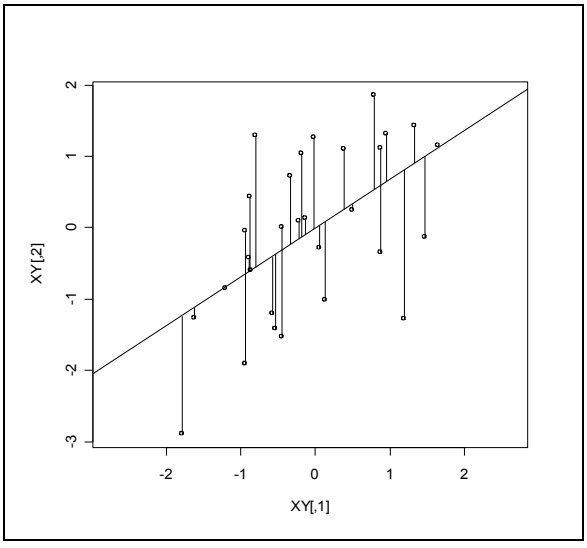
- **Pour les variables**

Cela revient à **construire des variables, appelées composantes principales, par combinaison linéaire des variables initiales et telles que ces nouvelles variables aient la plus grande variance possible.** Les composantes principales sont de plus non corrélées.

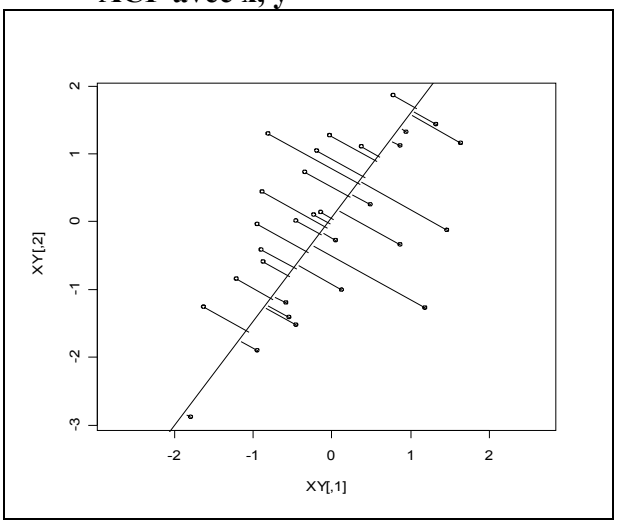
On ne s'intéresse alors qu'aux composantes principales qui ont la plus forte **variance (=valeur propre de l'axe**). On construit ensuite des nuages de points des individus en fonction de ces composantes principales dans les plans factoriels F_1 F_2, ou F_1 F_3 ...

Interprétation graphique de l'ACP

Régression linéaire de y en x



ACP avec x, y

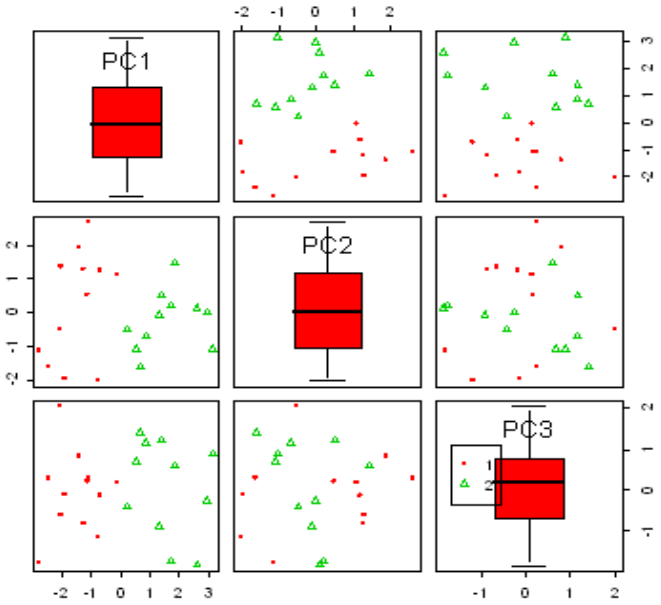


Remarque importante : En général, du fait de l'**hétérogénéité des variables initiales** et de leurs unités, **on réduit ces variables**. On parle alors d'**ACP normée**.
Une variable est dite réduite quand sa variance vaut 1. De la sorte, chaque variable initiale aura une même importance dans l'analyse car sa contribution est proportionnelle à sa variance.
Dans la pratique, on normalise presque toujours et surtout lorsque les variables sont exprimées dans des unités différentes.

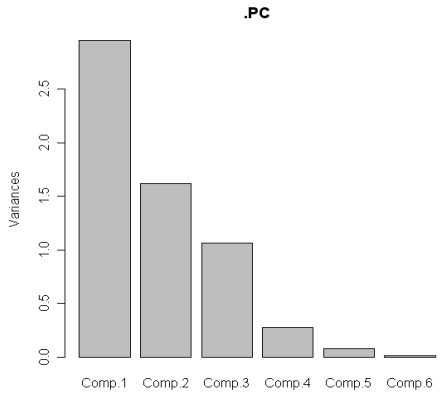
L'objectif est ainsi de construire des variables qui synthétisent la dispersion du nuage. Si plusieurs variables initiales sont ainsi fortement corrélées entre elles, celles-ci sont alors représentées par une

composante principale qui les résume. Au final, au lieu de travailler avec p=6 variables, on peut espérer travailler sur 2 ou 3 variables synthétiques qui résument l'essentiel de l'information.
On retrouve une partie des résultats de l'acp dans Rcmdr, statistiques, ajustement multivarié.

Projection des individus dans les plans définis par les axes F_1 F_2 F_3



Projection de l'inertie sur les différents axes factoriels



Mais il est fortement conseillé sous R d'utiliser la librairie ade4.

>library(ade4)
>zebu.acp <- dudi.pca(zebu[,1:6])
>zebu.acp$eig (valeurs propres)
>zebu.acp$li (composantes principales)
>zebu.acp$co (coordonnées des variables) …. voir l'aide en ligne

Guide pratique de l'analyse ACP

- **Etape 1 :** Sélection des axes et des plans retenus principalement par rapport aux valeurs propres.

- **Etape 2 :** Projection des variables et individus dans un plan donné (F_1 F_2 en premier)
  - Examen des *qlt* dans le plan pour éliminer les individus mal représentés
  - Bilan des *ctr* pour un axe afin de donner un sens à cet axe (opposition, tendance ...)
  - Topographie des variables et individus afin d'identifier des groupes, des oppositions, des tendances notamment à l'aide de la fonction s.class
  - Utiliser ses connaissances sur le sujet pour proposer des explications sur les résultats de l'analyse
  - Utiliser des individus ou variables supplémentaires ou des profils type (moyenne des H et des F par exemple)

Fiche 37 – Aides à l'interprétation

1. Valeurs propres λ et choix des axes

Pour définir le nombre d'axes étudiés, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projetée sur un axe donné.

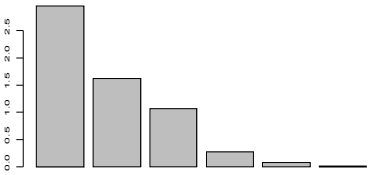
Remarque importante: La somme des valeurs propres est égale à l'inertie totale du nuage (= nombre de variables en ACP normé). On caractérise ainsi chaque axe par le pourcentage d'inertie qu'il permet d'expliquer.

On ne retient donc que les axes avec les plus fortes valeurs propres. Le choix des axes retenus est un peu délicat. On peut donner quelques règles :

- **Règle de Kaiser en ACP normée**: on ne s'intéresse qu 'aux axes avec une valeur propre supérieure à 1 (= inertie d'une variable initiale).
- **Règle de l'inertie minimale** : On sélectionne les premiers axes afin d'atteindre un % donné d'inertie expliquée (70% par exemple).
- **Règle du coude** : On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme. On retient les axes avant le décrochage.
- **Règle de bon sens** : On analyse les plans et axes et on ne retient que ceux interprétables.

Exemple zebu

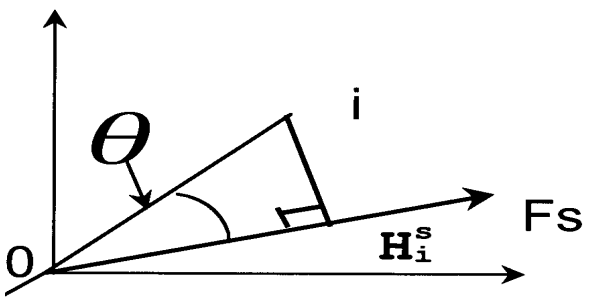
```
> round(acp$eig,2)
[1] 2.95 1.62 1.07 0.27 0.08 0.01
> round(acp$eig/sum(acp$eig)*100)
[1] 49 27 18  5  1  0
```



Avec ade4: réaliser l'acp >acp=dudi.pca(tableau)
Extraire les valeurs propres >acp$eig
Calcul des % >acp$eig/sum(acp$eig)*100

2. Qualité de représentation *qlt*

Les individus représentés dans un plan factoriel ne sont pas forcément correctement représentés.



Qualité de représentation *qlt*:
Si l'angle θ est grand, le point initial est éloigné de sa projection. On utilise le paramètre $\cos^2 \theta$ pour caractériser la qualité de représentation (*qlt*) sur un axe.

- Plus qlt_i est proche de 1 plus il est bien représenté.
- Plus qlt_i est proche de 0 plus il est mal représenté.
- Dans un plan, on calcule la somme des deux *qlt*, par exemple $qlt_{F1} + qlt_{F2}$ pour le plan F_1 F_2.
- *qlt* correspond en fait au rapport de l'inertie du projeté sur l'inertie du point initial.

Qualité globale : Dans un plan donné, on définit également la qualité globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la *qlt* d'un individu ou d'une variable.

Remarque : La *qlt* des variables peut s'analyser de la même façon mais l'utilisation du cercle des corrélation est plus intuitive.

Bilan : On commencera donc toujours l'analyse d'un plan factoriel en précisant l'existence (ou non) d'individus ou variables mal représentés et en justifiant par les *qlt*.

3. Contribution *ctr*

Lors de la construction d'un axe factoriel, certaines variables et certains individus ont des rôles plus importants. On calcule un paramètre appelé contribution, *ctr*, qui permet de calculer cette influence.

Définition: La contribution *ctr* est définie comme la proportion de l'inertie de l'axe expliquée par la variable ou l'individu.

Règles d'interprétation :

- L'analyse se fait axe par axe, en parallèle sur les variables et les individus.
- Plus *ctr* est grande, plus l'influence de l'individu est grande. On ne retient donc que les plus fortes valeurs (il y a souvent un décrochage après quelques valeurs).
- *ctr* est considéré comme positif si l'individu est dans la partie positive de l'axe.
- *ctr* est considéré comme négatif si l'individu est dans la partie négative de l'axe.
- Le bilan des *ctr* peut être présenter pour un axe donné sous forme d'un tableau avec les principales *ctr* + et – des individus et des variables, en précisant la valeur de *ctr* :

ctr axe F_1 :	-	+
Variables :		
Individus :		

On réalise ensuite une interprétation.

Sous R, ces paramètres sont obtenus avec les commandes suivantes :

Pour les lignes (individus)	`>inertieL<-inertia.dudi(acp, row.inertia=TRUE)`
ctr des lignes en %	`>inertieL$row.abs/100`
qlt des lignes en %	`>inertieL$row.rel/100`
Pour les colonnes (variables)	`>inertieC<-inertia.dudi(acp, col.inertia=TRUE)`
ctr des colonnes en %	`>inertieC$col.abs/100`
qlt des colonnes en %	`>inertieC$col.rel/100`

Fiche 38 – Représentation graphique des variables

Les composantes principales sont construites comme des combinaisons linéaires des variables initiales. Pour visualiser les liaisons entre la composante principale et les variables initiales, on représente en ACP normée les variables dans les plans factoriels. Les coordonnées des variables sont les coefficients de corrélation de ces variables avec les composantes principales.

Les règles de lecture du cercle des corrélations sont :

- On ne prend en compte **que les variables proches du cercle des corrélations**. Dans le cas contraire, la variable est non corrélée à la composante principale et est donc mal représentée.

- **La liaison entre variables bien représentées s'analyse à travers la direction et le sens de leur vecteur** :
  - si les vecteurs ont même direction et même sens, les variables sont corrélées positivement,
  - si les vecteurs ont même direction mais de sens contraire, les variables sont corrélées négativement,
  - si les vecteurs sont perpendiculaires, les variables sont non corréles.

- On synthétise chaque axe en précisant les variables qui contribuent le plus en positif ou en négatif (étude des ctr)

Exemple :

```
inertieV=inertia.dudi(acp,col.inertia=TRUE)
```

Composante principale

```
> round(acp$co,2)
   Comp1 Comp2
FL -0.99 -0.05
RW -0.94  0.35
CL -0.99 -0.10
CW -0.99 -0.07
BD -0.99 -0.10
```

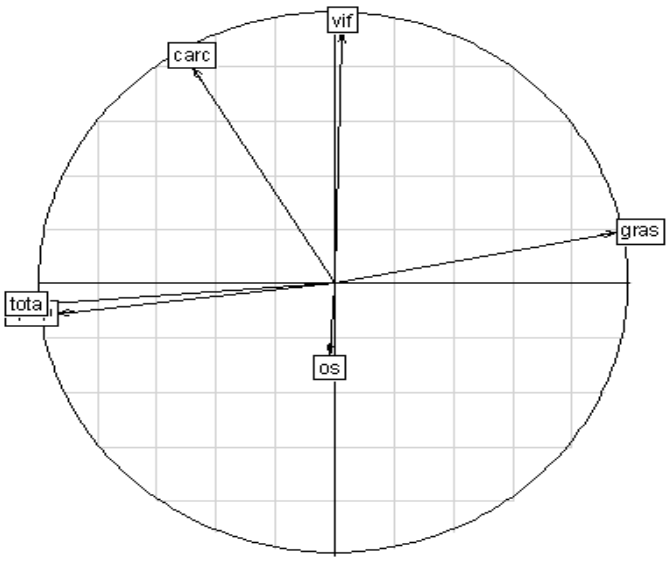
ctr / 10 000

```
 inertieV$col.abs
   Comp1 Comp2
FL  2043   189
RW  1833  8066
CL  2054   720
CW  2035   326
BD  2035   699
```

qlt / 10 000

```
inertieV$col.rel
   Comp1 Comp2 con.tra
FL -9786   -29    2000
RW -8776  1224    2000
CL -9835  -109    2000
CW -9745   -49    2000
BD -9746  -106    2000
```

```
> s.corcircle(zebu.zebu.acp$co,xax=1,yax=2)
```



Fiche 39 – Représentation graphique des individus

Les individus sont associés à des points de l'espace dont les coordonnées sont les variables. On peut mesurer la **distance entre ces individus** en utilisant simplement la distance euclidienne classique entre ces deux points (comme au collège...).
La construction des composantes principales conduit à **rendre minimale la déformation des distances entre individus lorsque l'on projette les individus dans le plan factoriel** $F_1 F_2$. Ainsi les distances que l'on observe entre les individus dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces individus.
L'analyse des plans factoriels permet ainsi d'observer les individus proches entre eux ou au contraire éloignés. Il est ainsi possible de construire des groupes, d'observer des tendances ...

Les règles de lecture des plans factoriels sont :

- **Seuls les individus bien représentés sont pris en compte** dans l'interprétation.
  - On calcule la somme des *qlt* dans le plan et on vérifie que cette somme n'est pas trop faible par rapport à la qualité moyenne du plan.

- On réalise le **bilan en positif et en négatif des individus qui ont la plus forte contribution** pour un axe donné.
  - On donne ainsi en parallèle avec l'analyse des variables une signification concrête à ces axes en terme d'opposition entre individus et variables ou tendance particulière.

- **On réalise des groupes**, à l'aide éventuelle de la fonction `s.class`, dans le cas de groupes préexistants (homme-femme par exemple) ou on construit arbitrairement ces groupes en raison des proximités entre individus.

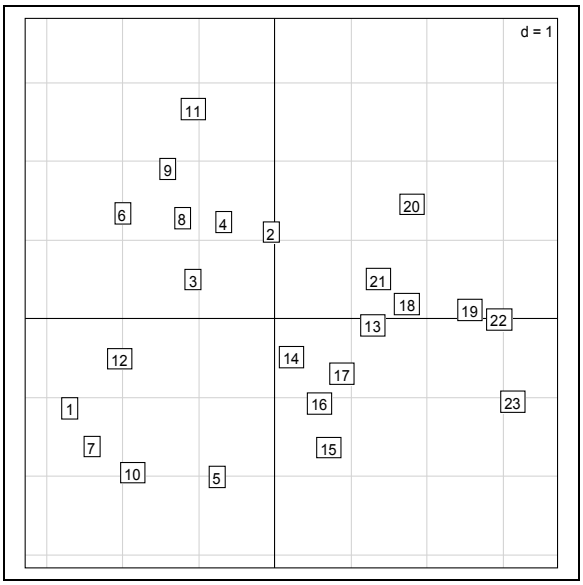
- En présence de trop nombreux individus, on peut utiliser des **individus type** et réaliser une analyse sur ce individus.

- L'utilisation d'**individus supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

```
inertie <-inertia.dudi(acp, row.inertia=TRUE)
Composantes principales
>  round(acp$li,3)
   Axis1  Axis2  Axis3
1  -2.691 -1.137 -1.786
2  -0.050  1.102  0.180
3  -1.072  0.499  0.202
4  -0.671  1.228 -0.149
5  -0.756 -2.009 -1.152
6  -1.999  1.337 -0.616
7  -2.402 -1.625  0.276
8  -1.213  1.278 -0.819
9  -1.401  1.906  0.823
10 -1.869 -1.954 -0.105
11 -1.065  2.663  0.276
12 -2.032 -0.507  2.028
13  1.287 -0.082 -0.898
14  0.214 -0.485 -0.409
15  0.713 -1.635  1.418
16  0.586 -1.076  0.701
17  0.880 -0.699  1.164
18  1.735  0.187 -1.741
19  2.561  0.113 -1.835
20  1.801  1.457  0.593
21  1.365  0.502  1.204
22  2.949 -0.008 -0.253
23  3.130 -1.055  0.897

[CTR en 10000ième]
> inertie$row.abs
   Axis1 Axis2 Axis3
1   1067   347  1302
2      0   326    13
3    169    67    17
4     66   405     9
5     84  1083   541
6    589   480   155
7    850   708    31
8    217   438   274
9    289   975   276
10   515  1024     4
11   167  1904    31
12   609    69  1677
13   244     2   329
14     7    63    68
15    75   718   820
16    51   311   200
17   114   131   553
18   443     9  1236
19   966     3  1374
20   478   570   143
21   275    68   591
22  1281     0    26
23  1444   298   329

[Qlt en 10000ième]
> inertie$row.re
   Axis1 Axis2 Axis3 con.tra
1  -5975 -1068 -2634     878
2    -10  4836   129     182
3  -7332  1591   260     114
4  -1994  6672   -98     164
5   -948 -6695 -2202     437
6  -6224  2785  -591     465
7  -6356 -2907    84     658
8  -3728  4140 -1701     286
9  -2274  4212   785     625
10 -4653 -5085   -15     544
11 -1359  8497    91     605
12 -4712  -294  4690     635
13  6643   -27 -3231     181
14   556 -2848 -2028      60
15   945 -4975  3742     389
16  1612 -5446  2310     154
17  2909 -1836  5094     193
18  4939    57 -4972     442
19  6551    13 -3365     725
20  5421  3547   588     434
21  4490   606  3494     301
22  9644     0   -71     654
23  8109  -920   667     876
```

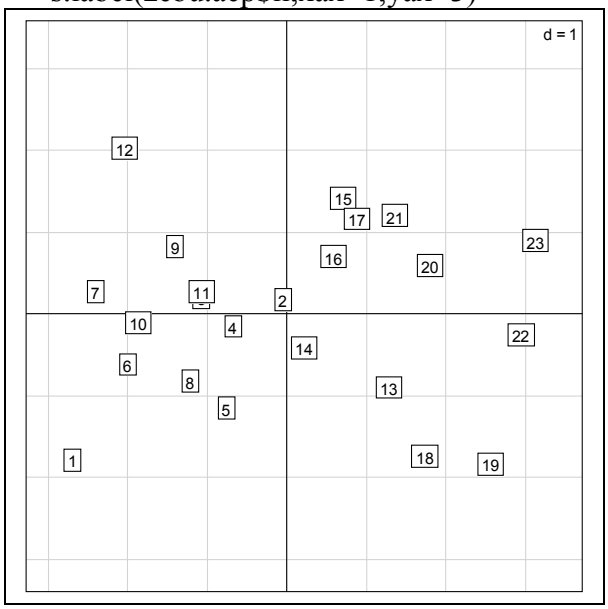
Plan F_1 F_2

> s.label(zebu.acp$li,xax=1,yax=2)

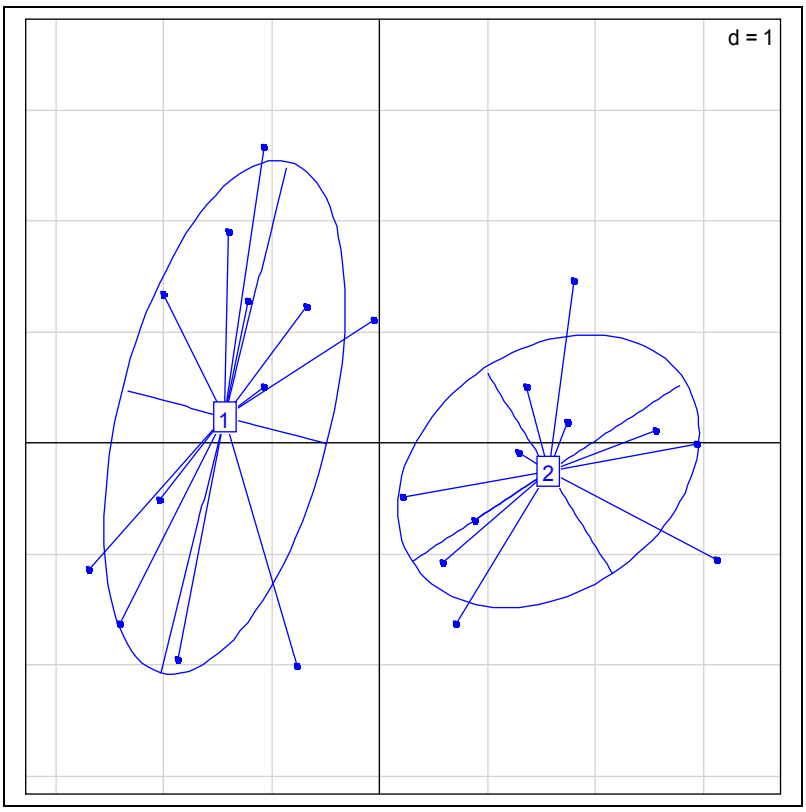


Plan F_1 F_3

> s.label(zebu.acp$li,xax=1,yax=3)



>s.class(dfxy=zebu.acp$li,fac=race,col=col,xax=1,yax=2)



IX ANALYSE FACTORIELLE DES CORRESPONDANCES

Objectifs et données pour l'AFC

Cette technique s'applique à des tableaux de contingence croisant deux variables qualitatives avec de nombreuses modalités chacune. Les données sont donc les effectifs des individus croisant deux modalités données. Pour de tels tableaux nous disposons du test d'indépendance du χ^2.

L'**objectif** est ici de faire une synthèse de l'ensemble du tableau afin de répondre aux questions :

- **Pour une variable donnée**, certaines modalités sont-elles proches ou éloignées.
  - La proximité de deux modalités se mesure en comparant leur distribution par rapport à l'autre variable.
    Par exemple, yeux bleus et verts sont proches si les deux groupes ont les mêmes distribution de couleurs de cheveux.
- **Entre les deux variables**, certaines modalités « s'atirent-elles » davantage ou au contraire « se repoussent ».
  - On compare la fréquence observée par rapport à la fréquence attendue sous l'hypothèse d'indépendance, si la fréquence observée est plus forte il y a une plus forte association entre les deux et inversement.
    Par exemple, les yeux bleus et les cheveux blond « s'attirent », au contraire des yeux noirs et des cheveux blond.

Remarque : L'AFC n'a d'intérêt que si il y a dépendance entre les deux variables, en cas contraire elle n'apporte pas d'information.

Le **tableau** se présente sous la forme :

		variable qualitative 2		
		modalité 1	... modalité j	modalité J
variable qualitative 1	modalité 1	n_{11}	n_{1j}	n_{1J}
	...			
	modalité i	n_{i1}	n_{ij}	n_{iJ}
	...			
	modalité I	n_{I1}	n_{Ij}	n_{IJ}

On constate la symétrie du tableau contrairement au tableau utilisé en ACP.

Exemple 1 (trivial)

On examine la répartition des couleurs de cheveux et d'yeux.

Tableau de contingence

\cheveux yeux	blond	roux	brun	total
bleu	20	5	5	
vert	0	15	5	
marron	5	5	40	
total				

Tableau théorique sous H_0 (Indépendance)

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Test du χ^2 d'indépendance

Fiche 40 – Principe de l'AFC

Pour mesurer les distances entre modalités, il est nécessaire de calculer au préalable la distribution de chaque modalité d'une variable en fonction de l'autre variable.

On définit ainsi **les profils colonnes et les profils ligne qui sont les distributions respectives des modalités des deux variables.**

Un profil ligne (colonne) est déterminé en divisant la ligne (colonne) par le total de la ligne (colonne).

Profils lignes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Profils colonnes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Chaque profil est alors assimilé à un point de coordonnées les proportions par rapport aux modalités de l'autre variable. Le nuage des profils ligne est alors projeté sur des axes factoriels en conservant le maximum d'inertie et il en est de même pour le nuage des profils colonne.

Les principales différences avec l'ACP sont :

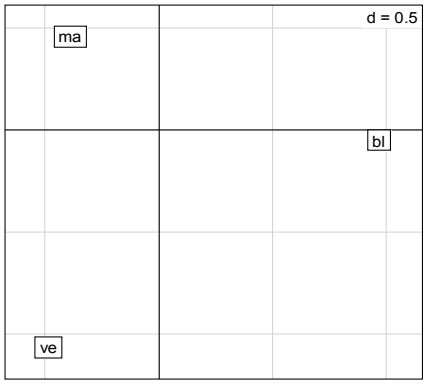
- **Lignes et colonnes sont transformées au préalable en profil et jouent un rôle symétrique**.
- La distance entre deux profils est calculée à l'aide de la **distance du** χ^2 (distance entre deux distributions) :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

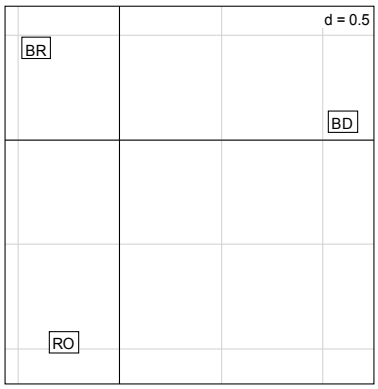
- Les profils sont tous dans un hyperplan car la somme de leurs coordonnées est 1. On a donc un axe de moins qu'en ACP et une valeur propre en moins.
- **Les valeurs propres (inertie projetée sur l'axe) sont inférieures à 1**.
- Les deux nuages représentent des profils et il est d'**usage de représenter les deux nuages dans un même plan** (le profil d'une modalité est « quasi » le barycentre des profils des modalités de l'autre variable).
- **La proximité entre modalités des deux variables indique une attirance entre ces modalités** (l'effectif observé est supérieur à celui attendu sous H_0).
- **La proximité entre modalités d'une même variable indique que les distributions sont voisines** pour ces deux modalités au regard de l'autre variable.

Exemple 1 : Les résultats de l'ajustement donnent :

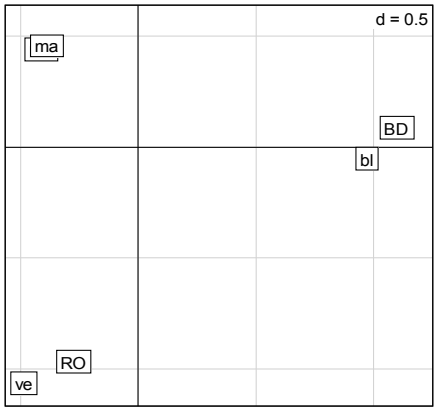
```
> afc$eig
[1] 0.4050000 0.3333333
> (afc$eig/sum(afc$eig))
[1] 0.5485327 0.4514673
```



Profils ligne dans le plan F_1 F_2



Profils colonne



Projection simultanée

Exemple 2 : CSP

Le tableau décrit la consommation annuelle en francs d'un ménage pour différentes denrées alimentaires en 1972. MA, EM, CA indiquent la catégorie socio-professionnelle et 2,3,4,5 la taille du foyer.

```
> csp=read.table("csp.txt")
    pain legu frui vian vola lait vin
MA2  332  428  354 1437  526  247 427
EM2  293  559  388 1527  567  239 258
      ...
```

Statistiques élémentaires

```
chisq.test(csp)
        Pearson's Chi-squared test
X-squared = 1290.386, df = 66, p-value < 2.2e-16
```

Calcul des profils :

Profils ligne

```
> round(csp/apply(csp,1,sum),2)
    pain legu frui vian vola lait  vin
MA2
EM2 0.08 0.15 0.10 0.40 0.15 0.06 0.07
CA2 0.07 0.15 0.11 0.37 0.18 0.04 0.08
MA3 0.10 0.14 0.08 0.37 0.13 0.08 0.10
EM3 0.09 0.15 0.10 0.36 0.14 0.08 0.09
CA3 0.07 0.14 0.11 0.39 0.19 0.04 0.06
MA4 0.12 0.14 0.08 0.35 0.14 0.09 0.09
EM4 0.09 0.14 0.10 0.37 0.15 0.08 0.08
CA4 0.07 0.13 0.11 0.40 0.19 0.05 0.05
MA5 0.12 0.14 0.08 0.34 0.14 0.09 0.09
EM5 0.10 0.17 0.09 0.35 0.15 0.09 0.05
CA5 0.07 0.15 0.12 0.37 0.16 0.08 0.04
```

Profils colonne

```
> round(t(t(csp)/apply(t(csp),1,sum)),2)
    pain legu frui vian vola lait  vin
MA2 0.06 0.05 0.06 0.06 0.05 0.06
EM2 0.05 0.06 0.06 0.07 0.06 0.06
CA2 0.07 0.09 0.09 0.09 0.10 0.05
MA3 0.08 0.06 0.06 0.07 0.06 0.08
EM3 0.07 0.07 0.07 0.07 0.06 0.07
CA3 0.08 0.10 0.11 0.10 0.12 0.06
MA4 0.10 0.08 0.06 0.07 0.07 0.10 0.09
EM4 0.09 0.08 0.08 0.08 0.08 0.09 0.09
CA4 0.07 0.09 0.10 0.10 0.12 0.07 0.06
MA5 0.12 0.09 0.07 0.08 0.08 0.12 0.11
EM5 0.11 0.11 0.09 0.09 0.09 0.12 0.07
CA5 0.10 0.12 0.15 0.12 0.12 0.13 0.06
```

Poids des lignes et colonnes

```
> round(apply(csp,1,sum)/sum(csp),2)
 MA2  EM2  CA2  MA3  EM3  CA3  MA4  EM4  CA4  MA5  EM5  CA5
0. 0.06 0.09 0.07 0.07 0.10 0.08 0.08 0.10 0.09 0.10 0.12
```

```
> round(apply(csp,2,sum)/sum(csp),2)
   pain legu frui vian vola lait  vin
   0.   0.14 0.10 0.37 0.16 0.07 0.07
```

Comment décrire ce tableau ? Quelles sont les relations entre variables et modalités de chaque variables ? Comment présenter ce tableau à un public non averti ?

Fiche 41 – Aides à l'interprétation

1. Valeurs propres λ et choix des axes

Pour définir le nombre d'axes étudiés, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projeté sur un axe donné.

Remarques importantes:

- La somme des valeurs propres est toujours égale à l'inertie totale du nuage. On caractérise ainsi chaque axe par le % d'inertie qu'il permet d'expliquer.
- En AFC, les **valeurs propres sont toutes inférieures à 1**.

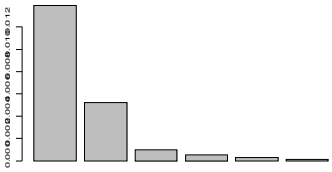
On ne retient donc que les axes avec les plus fortes valeurs propres. Le choix des axes retenus est un peu délicat. On peut donner quelques règles :

- **Règle du coude :** On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme. On retient les axes avant le décrochage.
- **Règle de l'inertie minimale** : On sélectionne les premiers axes afin d'atteindre un % donné d'inertie expliquée (70% par exemple).
- **Règle du bon sens :** On analyse les plans et axes et on ne retient que ceux interprétables.

Exemple 2 CSP

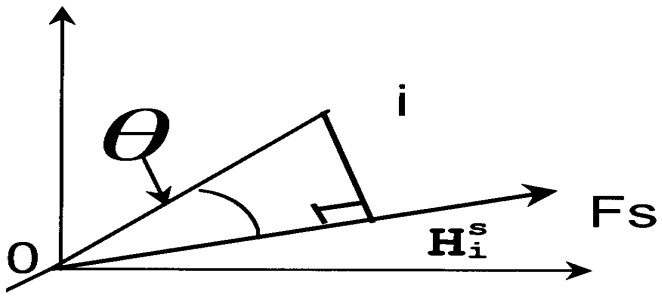
Avec ade4: réaliser l'afc >afc=dudi.coa(tableau)
Extraire les valeurs propres >afc$eig
Calcul des % >afc$eig/sum(afc$eig)

```
> round(afc$eig,3)
[1] 0.014 0.005 0.001 0.001 0.000 0.000
> round(afc$eig/sum(afc$eig)*100)
[1] 66 25  5  2  1  1
```



2. Qualité de représentation *qlt*

Les profils projetés dans un plan factoriel ne sont pas forcément correctement représentés. Dans ce cas, les interprétations sont erronées. Il est indispensable de vérifier la bonne représentation au préalable.



Qualité de représentation *qlt*:
Si l'angle θ est grand, le point initial est éloigné de sa projection. On utilise le paramètre $\cos^2 \theta$ pour caractériser la qualité de représentation (*qlt*) sur un axe.

- Plus qlt_i est proche de 1 plus il est bien représenté.
- Plus qlt_i est proche de 0 plus il est mal représenté.
- Dans un plan, on calcule la somme des deux *qlt*, par exemple $qlt_{F1} + qlt_{F2}$ pour le plan $F_1 F_2$.
- *qlt* correspond en fait au rapport de l'inertie du projeté sur l'inertie du point initial.

Qualité globale : Dans un plan donné, on définit également la qualité globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la *qlt* d'un profil.

Bilan : On commencera donc toujours l'analyse d'un plan factoriel en précisant l'existence (ou non) de profils mal représentés et en justifiant par les *qlt*.

3. Contribution *ctr*

Lors de la construction d'un axe factoriel, certains profils ont des rôles plus importants. On calcule un paramètre appelé contribution, *ctr*, qui permet de calculer cette influence.

Définition: La contribution *ctr* est définie comme la proportion de l'inertie de l'axe expliquée par le profil pour un axe donné.

Règles d'interprétation :

- L'analyse se fait axe par axe, en parallèle sur les profils ligne et colonne.
- Plus *ctr* est grande, plus l'influence du profil est grande. On ne retient donc que les plus fortes valeurs (il y a souvent un décrochage après quelques valeurs).
- *ctr* est considéré comme positif si le profil est dans la partie positive de l'axe.
- *ctr* est considéré comme négatif si le profil est dans la partie négative de l'axe.
- Le bilan des *ctr* peut être présenter pour un axe donné sous forme d'un tableau avec les principles *ctr* + et – des profils, en précisant la valeur de *ctr* :

ctr axe F_1 :	-	+
Profils colonnes :		
Profils lignes :		

On réalise ensuite une interprétation.

Sous R, ces paramètres sont obtenus avec les commandes suivantes :

Pour les lignes `>inertieL<-inertia.dudi(acp, row.inertia=TRUE)`
ctr des lignes en % `>inertieL$row.abs/100`
qlt des lignes en % `>inertieL$row.rel/100`

Pour les colonnes `>inertieC<-inertia.dudi(acp, col.inertia=TRUE)`
ctr des colonnes en % `>inertieC$col.abs/100`
qlt des colonnes en % `>inertieC$col.rel/100`

Fiche 42 – Représentation graphique des profils

Les profils ligne (respectivement colonne) sont associés à des points de l'espace dont les coordonnées sont les distributions conditionnelles. On peut mesurer la **distance entre ces profils** en utilisant la distance du χ^2 entre ces deux points.

La construction des composantes principales conduit à **rendre minimale la déformation des distances entre profils lorsque l'on projette les profils dans le plan factoriel** $F_1 F_2$. Ainsi les distances que l'on observe entre les profils dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces profils.

L'analyse des plans factoriels permet ainsi d'observer les profils proches entre eux (distribution similaire) ou au contraire éloignés. Il est ainsi possible de construire des groupes, d'observer des tendances ...

Les règles de lecture des plans factoriels sont :

- **Les profils ligne et colonne sont projetés simultanément** afin notamment d'observer les modalités des deux variables présentant de fortes (attraction) ou faibles (répulsion) associations.

- **Seuls les profils bien représentés sont pris en compte** dans l'interprétation.
  - On calcule la somme des qtl dans le plan et on vérifie que cette somme n'est pas trop faible par rapport à la qualité moyenne du plan.

- On réalise le **bilan en positif et en négatif des profils qui ont la plus forte contribution** pour un axe donné.
  - On donne ainsi en parallèle sur les lignes et colonnes une signification concrête à ces axes en terme d'attraction, de similarité, entre modalités ou en tendance particulière.

- **On réalise des groupes**, à l'aide éventuelle de la fonction `s.class`, dans le cas de groupes préexistants (homme-femme par exemple) ou on construit arbitrairement ces groupes en raison des proximités entre profils.

- L'utilisation de **profils supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

Guide pratique de l'analyse AFC

- **Etape 1 :** Sélection des axes et des plans retenus principalement par rapport aux valeurs propres.

- **Etape 2 :** Projection des profils ligne et colonne dans un plan donné (F_1 F_2 en premier)
  - Examen des *qlt* dans le plan pour éliminer les profils mal représentés
  - Bilan des *ctr* pour un axe afin de donner un sens à cet axe (opposition, tendance ...)
  - Topographie des profils afin d'dentifier des groupes, des oppositions, des tendances notamment à l'aide de la fonction s.class
  - Utiliser ses connaissances sur le sujet pour proposer des explications sur les résultats de l'analyse
  - Utiliser des profils supplémentaires ou des profils type (moyenne des H et des F par exemple)

Exemple 2 csp

Profils lignes et colonnes

```
> inertie <-inertia.dudi(afc, row.inertia=TRUE)
```

Coordonnées profils ligne

```
> round(afc$li,2)
     Axis1 Axis2
MA2 -0.10 -0.14
EM2  0.04 -0.02
CA2  0.08 -0.09
MA3 -0.13 -0.05
EM3 -0.08 -0.01
CA3  0.16 -0.05
MA4 -0.15  0.02
EM4 -0.05 -0.01
CA4  0.17 -0.02
MA5 -0.17  0.03
EM5 -0.03  0.12
CA5  0.11  0.10
```

colonne

```
> round(afc$co,2)
     Comp1 Comp2
pain -0.19  0.04
legu  0.01  0.07
frui  0.13  0.01
vian  0.04 -0.03
vola  0.12 -0.02
lait -0.19  0.15
vin  -0.23 -0.19
```

[CTR en %]

```
> round(inertie$row.abs/100)
     Axis1 Axis2
MA2      4    24
EM2      1     0
CA2      4    14
MA3      8     3
EM3      3     0
CA3     18     4
MA4     13     1
EM4      2     0
CA4     20     1
MA5     18     2
EM5      1    27
CA5     10    24
```

[CTR en %]

```
> round(inertie$col.abs/100)
     Comp1 Comp2
pain    22     3
legu     0    13
frui    11     0
vian     4     5
vola    17     1
lait    18    31
vin     29    48
```

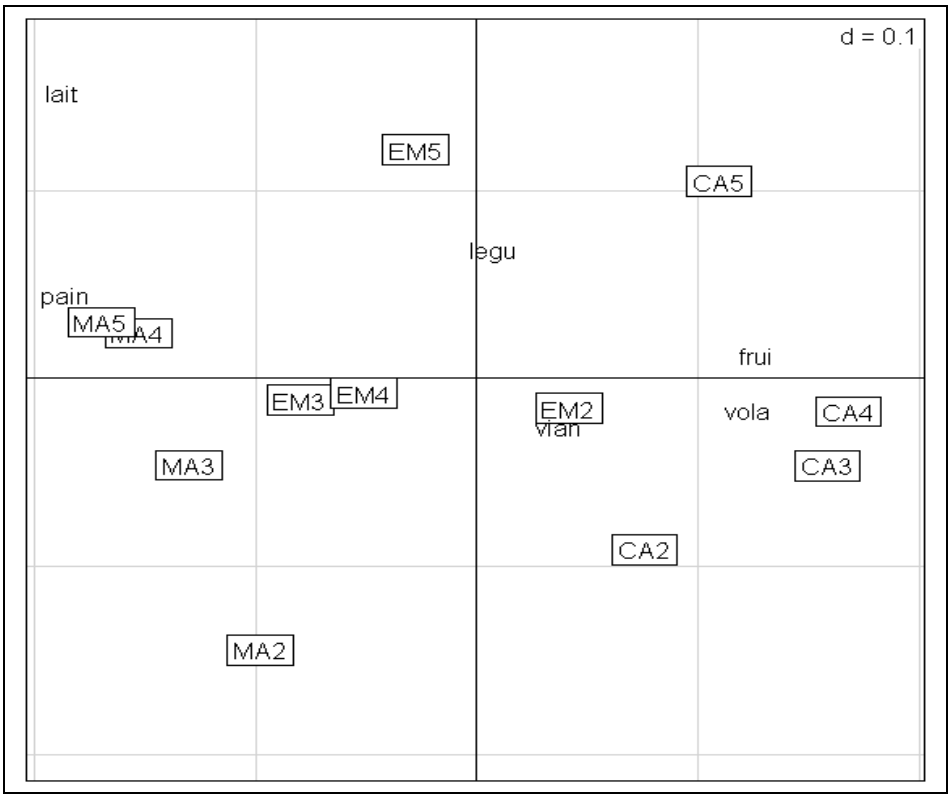
[QLT en %]

```
> round(inertie$row.re/100)
     Axis1 Axis2 con.tra
MA2    -29   -64      10
EM2     31    -5       2
CA2     34   -49       7
MA3    -86   -11       6
EM3    -80    -2       3
CA3     87    -7      14
MA4    -94     2       9
EM4    -81    -2       1
CA4     90    -1      14
MA5    -91     3      13
EM5     -5    89       8
CA5     46    43      14
```

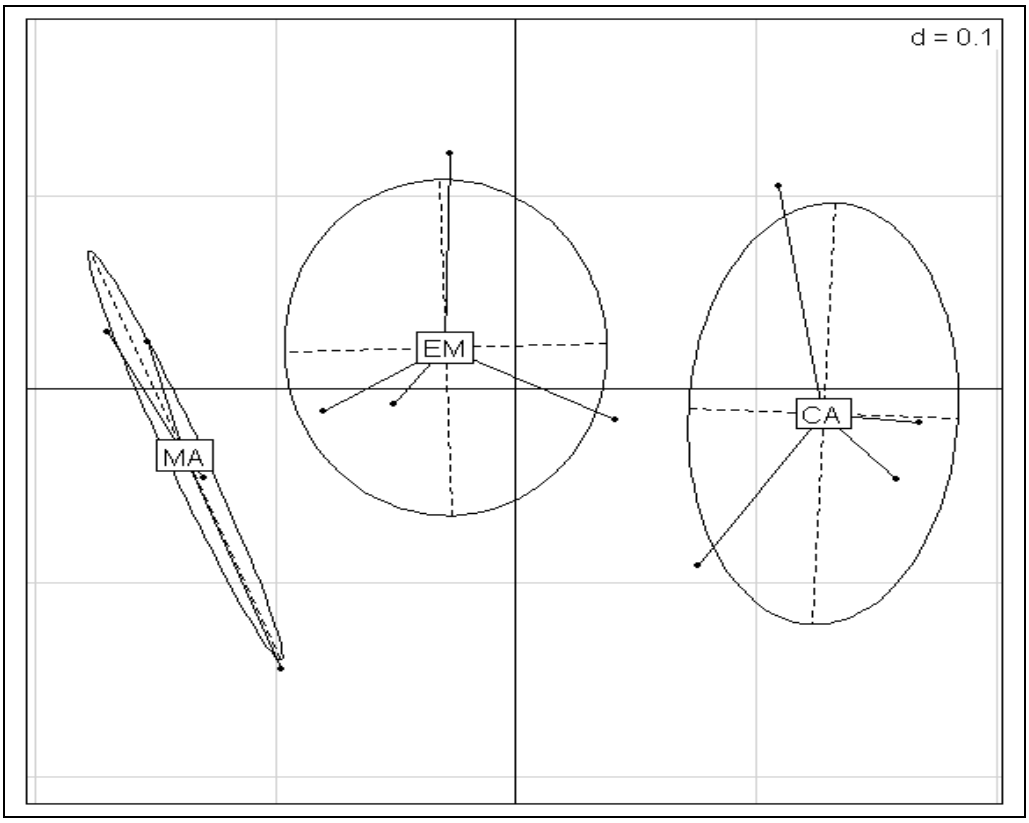
[QLT en %]

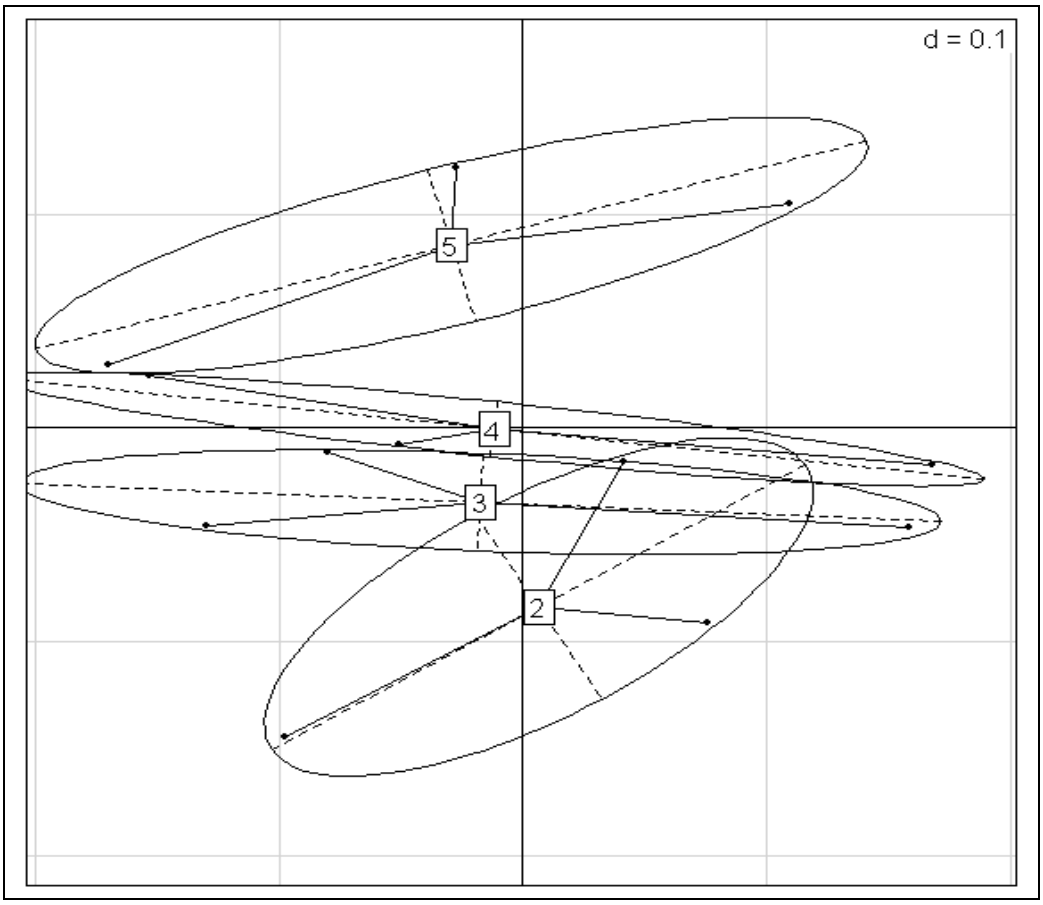
```
> round(inertie$col.re/100)
     Comp1 Comp2 con.tra
pain   -87     5      17
legu     1    66       5
frui    79     1      10
vian    51   -25       5
vola    84    -2      13
lait   -58    38      20
vin    -61   -38      31
```

```
> s.label(afc$li,xax=1,yax=2)
> s.label(afc$co,xax=1,yax=2,add.plot=T,boxes=F)
```



```
> pro <- as.factor(c("MA","EM","CA","MA","EM","CA","MA","EM","CA","MA","EM","CA"))
> nb <- as.factor(c(2,2,2,3,3,3,4,4,4,5,5,5))
> s.class(afc$li,xax=1,yax=2,fac=pro)
> s.class(afc$li,xax=1,yax=2,fac=nb)
```





X CLASSIFICATION

Objectifs et données :

On dispose de n objets caractérisés par des variables ou par les distances entre ces objets (tableau de distance). L'objectif de la classification est de construire des groupes d'objets homogènes avec comme principe :

- au sein d'un groupe, les objets sont les plus similaires possibles,
- entre les groupes, les différences sont le plus grandes possibles.

Lien avec l'AFC et l'ACP : L'analyse factorielle et la classification sont souvent conduites en parallèle. On peut ainsi effectuer une classification sur les profils, les variables, les individus. Les résultats de ces classifications peuvent ainsi être mis en parallèle avec l'analyse factorielle et améliorer l'interprétation.

Les méthodes de classification repose sur la notion de similarité ou de dissimilarité (distance) entre les objets que l'on souhaite regrouper en classes homogènes. On est donc souvent conduit à manipuler des tableaux de distance.

Soit E un ensemble de n objets.

- Une application d de $E \times E$ dans $\mathbb{R}^+$ est appelée **dissimilarité** si elle vérifie pour tout couple (i,j) de E^2 :
  - $d(i,j) = d(j,i)$
  - $d(i,j) \geq 0$ et $d(i,i) = 0$
  - si de plus $d(i,j) \leq d(i,k) + d(k,i)$, d est appelée **distance**

- Une application s de $E \times E$ dans $\mathbb{R}^+$ est appelée **similarité** si elle vérifie pour tout couple (i,j) de E^2 :
  - $s(i,j) = s(j,i)$
  - $s(i,j) \geq 0$
  - $s(i,i) \geq s(i,j)$

Exemples de distance et similarité

- **Données numériques (individus et variables quantitatives)**

Pour des objets décrits par des variables quantitatives, on peut utiliser la distance euclidienne classique (ACP)

- **Tableaux de contingence (profils)**

Pour des profils de modalités dans un tableau de contingence, on utilisera la distance du χ^2 (AFC).

- **Pour des données binaires** (présence/absence d'une caractéristique)

Plusieurs indices de similarité existent, comme l'indice de Jaccard : $J = \dfrac{a}{a+b+c}$

avec a le nombre de caractéristiques communes,
b celui possédé par i et pas par j
c celui possédé par j et pas par i.

Exemple : Construire le tableau des indices de Jaccard pour les individus A à E

	C1	C2	C3	C4
A	0	1	1	0
B	1	1	1	1
C	1	0	0	1
D	1	1	1	0

- **Distance génétique** (Chessel et al., http://pbil.univ-lyon1.fr/R/enseignement.html)

Soit **A** un tableau de fréquences alléliques avec t lignes (populations) et m colonnes (formes allèliques). Soit v le nombre de loci. Le locus j a $m(j)$ formes allèliques.

$$m = \sum_{j=1}^{v} m(j)$$

1 — Distance de Nei [22] (Voir [23]) :

$$D_1(a,b) = -Ln\left(\frac{\sum_{k=1}^{v}\sum_{j=1}^{m(k)} p_{aj}^{k} p_{bj}^{k}}{\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{aj}^{k}\right)^2}\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{bj}^{k}\right)^2}}\right)$$

Principe de la Classification Ascendante Hiérarchique (CAH)

La CAH est une méthode itérative.

1. Au départ chaque objet représente un groupe.
2. A chaque étape est construite une nouvelle partition de moins en moins fine, contenant une classe de moins à chaque étape. A la première étape, on regroupe les deux objets les plus proches et on ne dispose plus alors que de $n-1$ classes.

Cette méthode aboutit à un emboîtement de partitions visualisé graphiquement par un arbre hiérarchique indicé (dendrogramme).

On regroupe à chaque étape les 2 objets ou les 2 groupes d'objets dont **la ressemblance** est **la plus forte**. Cette méthode nécessite ensuite de **recalculer la distance des objets restant au groupe ainsi formé**. Ces méthodes de calculs s'appellent le **critère d'agrégation** et définissent la méthode :

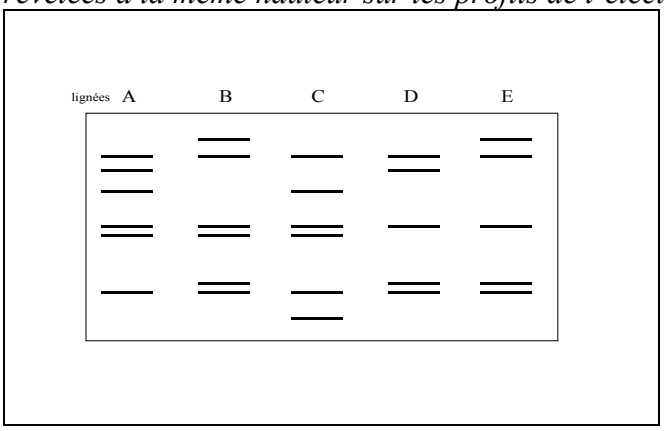
- **Critère du saut minimal :** la distance entre h et h' est celle définie par les deux éléments les plus proches.
- **Critère du diamètre ou distance maximale:** la distance entre h et h' est celle définie par les deux éléments les plus éloignées
- **Critère de la moyenne (UPGMA):** la distance entre h et h' est celle définie par la distance moyenne entre les éléments de h et de h'.
- **Critère de Ward** : Le critère est calculé pour h et h' par $\Delta_{hh'} = \frac{p_h \times p_{h'}}{p_h + p_{h'}} d^2(g_h, g_{h'})$. Le critère de Ward correspond à l'augmentation de l'inertie intra résultant du regroupement de h et h'.

Dans la pratique, le critère de Ward est le plus largement utilisé en CAH.

Liens sur internet :
http://biom3.univ-lyon1.fr/R/stage/stage7.pdf
http://zoonek2.free.fr/UNIX/48_R_2004/06.html
http://www.lsp.ups-tlse.fr/Carlier/.Hyper/plaq2html/node130.html
http://www.univ-orleans.fr/SCIENCES/LIFO/Manifestations/FD2004/FDD_Vrain.pdf

Fiche 43 – Exemple manuel de CAH

Un laboratoire veut étudier la ressemblance génétique de cinq lignées de sorgho (A, B, C, D, E). Pour cela, il réalise une analyse par RFLP avec l'enzyme de restriction *Eco* RI et une sonde d'origine inconnue.
Les fragments d'ADN ainsi amplifiés sont ensuite séparés par electrophorèse. Les résultats de l'electrophorèse donnent les profils suivants:
NB : Deux lignées de sorgho présentent autant de fragments d'ADN identiques que de bandes révélées à la même hauteur sur les profils de l'electrophorèse.



- Question 1 : Construisez la matrice de similarité entre les cinq lignées de sorgho à partir du coefficient de similarité de Dice (Sxy) défini ci-dessous.

$$S_{XY} = \frac{N_{XY}}{N_X + N_Y}$$

Nx = nombre de bandes de la lignée X

Ny = nombre de bandes de la lignée Y
Nxy = nombre de bandes identiques entre les lignées X et Y

Similarité	A	B	C	D	E
A					
B					
C					
D					
E					

- Question 2 :

Déterminez la matrice de dissimilarité en utilisant la fonction de similitude linéaire : Dxy = 1 – Sxy

Dissimilarité	A	B	C	D	E
A					
B					
C					
D					
E					

- Question 3 : Effectuez sur la matrice de dissimilarité la CAH en utilisant le critère du saut maximal.

Dij					Dij				Dij		

Dendrogramme

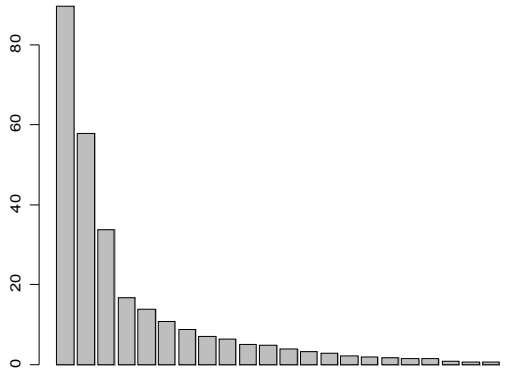
Fiche 44 – Exemple de CAH avec ade4

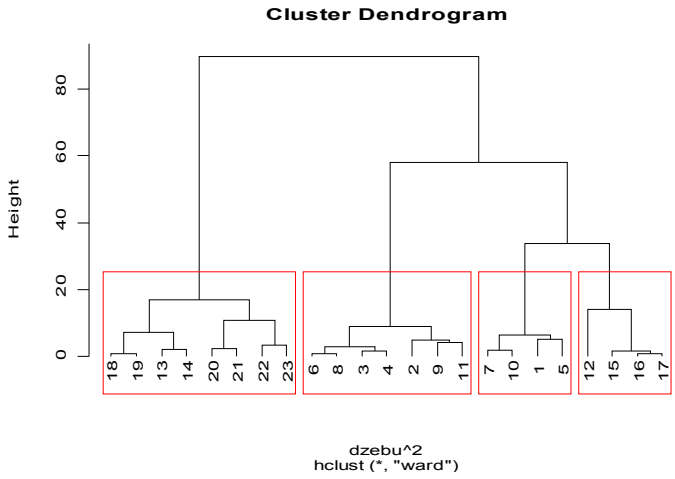
Exemple zebu

```
> library(ade4)
> acp = dudi.pca(zebu[,1:6])
```

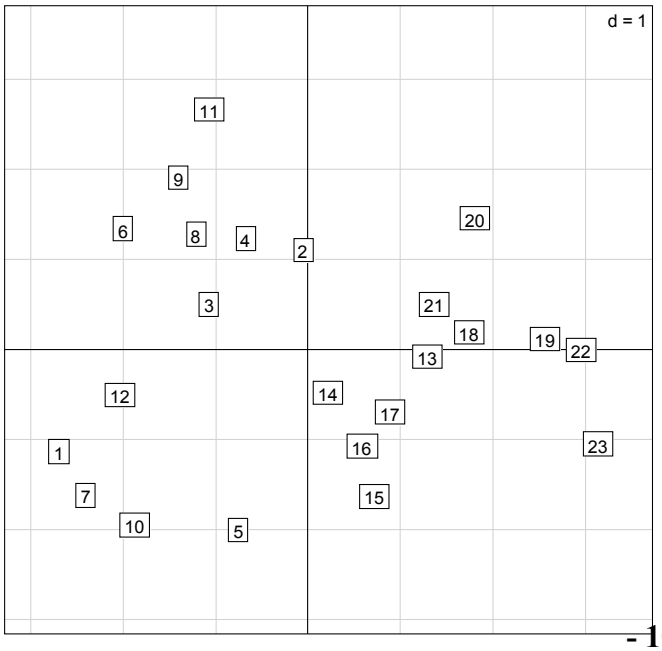
CAH sur les individus

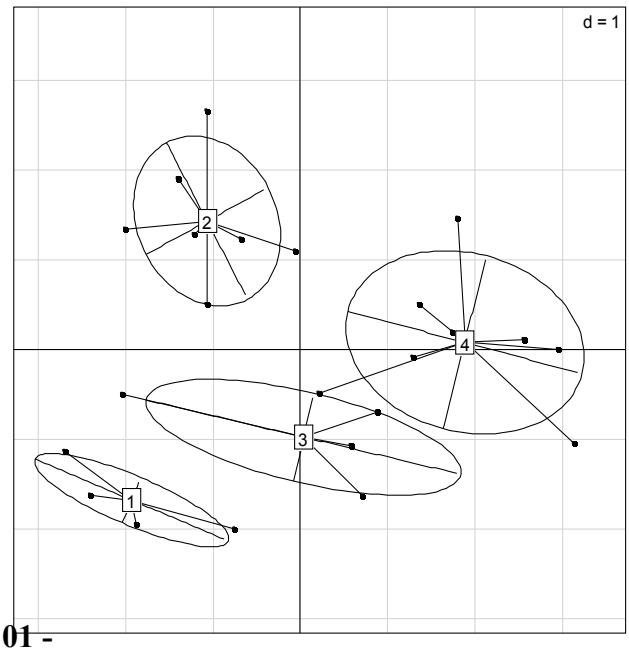
```
> dzebu = dist.dudi(acp, amongrow=TRUE)
> round(dzebu,1)     (extrait pour les 10 premiers individus)
     1   2   3   4   5   6   7   8   9  10
2  4.1
3  3.1 1.8
4  3.7 1.6 1.2
5  2.2 3.6 2.9 3.5
6  3.0 2.6 1.7 1.4 3.7
7  2.5 4.1 2.6 3.4 2.4 3.2
8  3.2 2.1 1.4 0.9 3.4 0.8 3.4
9  4.3 2.1 2.1 2.5 4.6 2.6 4.3 2.6
10 2.1 3.7 2.7 3.5 1.6 3.4 1.3 3.4 4.3

> cah=hclust(dzebu^2,method="ward")

> barplot(cah$height[order(cah$height,decreasing=TRUE)])

> plot(cah,hang=-1)
> rect.hclust(cah,h=20)
```





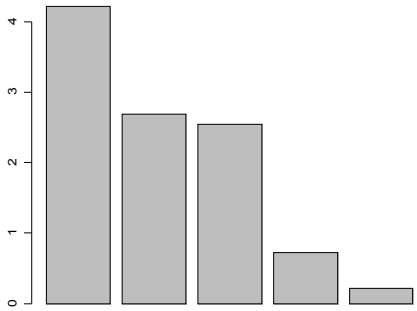

```
> cutree(cah,h=20)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 1  2  2  2  1  2  1  2  2  1  2  3  4  4  3  3  3  4  4  4  4  4  4
> s.label(acp$li)
> s.class(dfxy=acp$li,fac=as.factor(cutree(cah,h=20)))
```

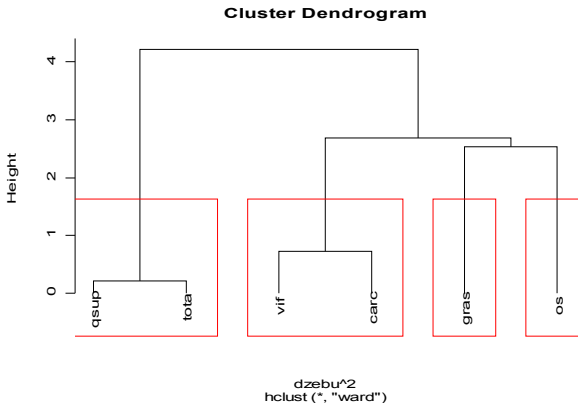




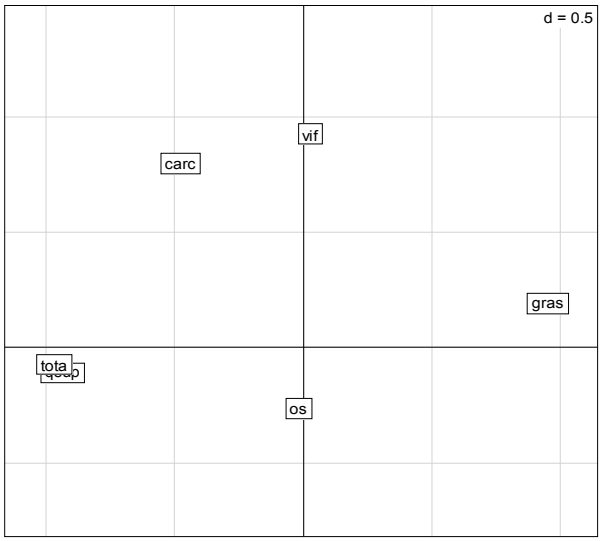
CAH sur les variables

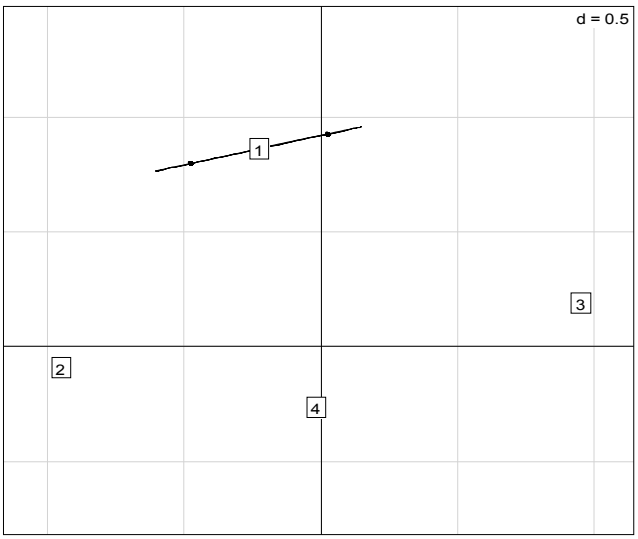
```
> dzebu = dist.dudi(acp, amongrow=FALSE)
> round(dzebu,1)     (extrait pour les 10 premiers individus)
      1   2   3   4   5   6   7   8   9  10
2   4.1
3   3.1 1.8
4   3.7 1.6 1.2
5   2.2 3.6 2.9 3.5
6   3.0 2.6 1.7 1.4 3.7
7   2.5 4.1 2.6 3.4 2.4 3.2
8   3.2 2.1 1.4 0.9 3.4 0.8 3.4
9   4.3 2.1 2.1 2.5 4.6 2.6 4.3 2.6
10 2.1 3.7 2.7 3.5 1.6 3.4 1.3 3.4 4.3

> cah=hclust(dzebu^2,method="ward")

> barplot(cah$height[order(cah$height,decreasing=TRUE)])

> plot(cah,hang=-1)
> rect.hclust(cah,h=1)
```






```
> cutree(cah,h=1)
 vif carc qsup tota gras   os
   1    1    2    2    3    4
> s.label(acp$co)
> s.class(dfxy=acp$co,fac=as.factor(cutree(cah,h=1)))
```





ANNEXES

Annexe A : Lois de probabilité usuelles
Annexe B : construction d'un test
Annexe C : Installation du logiciel R

Pour être à peu près complet, il serait souhaitable d'ajouter en annexe quelques bases de probabilité, des notions sur les protocoles expérimentaux, les bases mathématiques de l'analyse des données ...

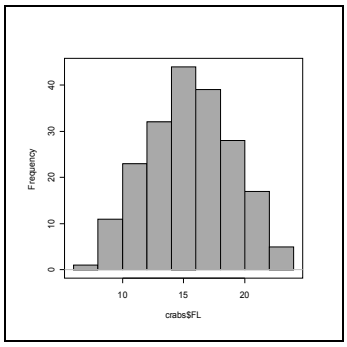
Toutes ces notions sont développées dans les ouvrages classiques ou sur internet.

Annexe A : Lois de probabilités usuelles

Dans la majorité des processus étudiés, le résultat d'une étude peut se traduire par une grandeur mathématique (variable quantitative discrète ou continue, variable binaire). Cette grandeur est assimilée à une variable aléatoire réelle discrète ou continue (fonction d'un espace probabilisé dans $\mathbb{R}$).
A partir de ces grandeurs, le statisticien construit des statistiques (t du test de Student, χ^2 du test du Khi Deux, F du test de Fisher, …) qui suivent des lois connues à partir desquelles sont construites les règles de décision.

1. Loi normale

Etude d'un exemple : télécharger le fichier crabs
[donnees dans le package lire package :MASS fichier : crabs]
Dans R, taper help(crabs) pour une description du fichier.
Nous nous intéressons pour l'instant à la variable FL.
Représenter l'histogramme de cette variable. Que peut-on dire de sa distribution ?



Dans de nombreux cas, les variables quantitatives se distribuent suivant une loi normale (courbe en cloche). Deux arguments expliquent ce phénomène :

- Une variable quantitative qui dépend d'un grand nombre de causes indépendantes dont les effets s'additionnent et dont aucune n'est prépondérante suit une loi normale.
- Le théorème "central limit" conclut que la moyenne S_n de n variables quantitatives suivant une même loi de moyenne μ et d'écart-type σ converge vers la loi normale de mêmes paramètres.

La loi normale est définie sur $\mathbb{R}$ et a pour fonction de densité $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$.

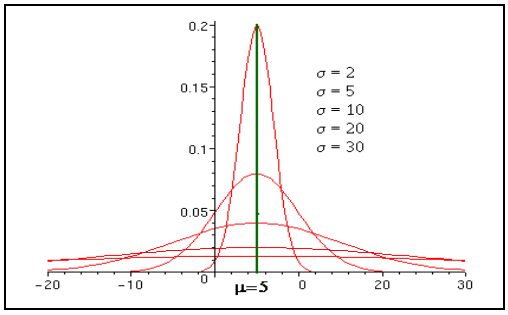
En particuliers, la loi normale centrée (moyenne nulle) réduite (variance unitaire) a pour fonction de densité $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x^2}{2}\right]$.

Dans la pratique on revient toujours à cette loi en changeant la variable X en la variable $X_{cr} = \frac{X-\mu}{\sigma}$.

X_{cr} est alors dite centrée réduite.

Vous pouvez retrouver l'allure des fonctions de densités en utilisant le menu *[distribution continue loi normale]* :

Retrouver le quantile pour lequel la loi normale centrée réduite vérifie P(X > q) =0.975.



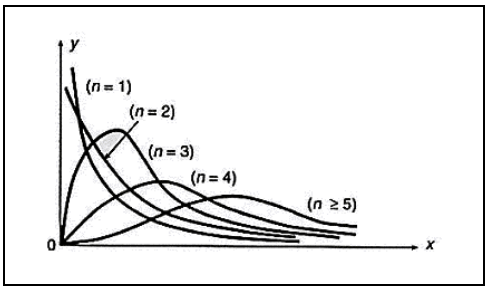
2. Loi de Pearson

La loi de Pearson ou loi du χ^2 (Khi deux) trouve de nombreuses applications dans le cadre de la comparaison de proportions, des tests de conformité d'une distribution observée à une distribution théorique et le test d'indépendance de deux caractères qualitatifs.

Définition : Soit X_{cr1}, ..., X_{crn}, n variables normales centrées réduites indépendantes, on appelle χ^2 la variable aléatoire définie par : $\chi^2 = \boxed{X_{cr1}^2 + ... + X_{crn}^2}$.
Cette variable aléatoire suit la loi de Pearson (ou du χ^2) à n degré de liberté (*ddl*).

S'il existe k relations entre ces variables alors le nombre de *ddl* devient n-k. Par exemple, si les variables sont centrées (calcul de la moyenne), le nombre de ddl devient n-1.

En utilisant R commander, étudier l'allure de la fonction de densité de la loi du χ^2 en fonction du degré de liberté. Quel est le quantile, q, vérifiant P($X > q$) =0.05 pour 1 *ddl*, 2 *ddl*, 20 *ddl* ?

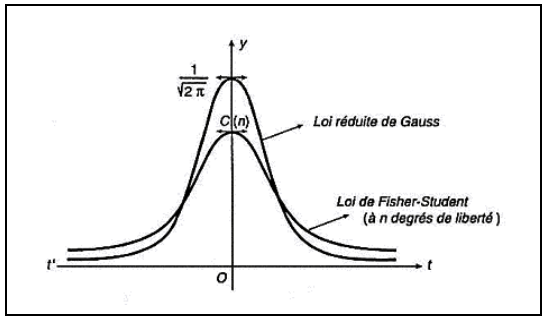


3. Loi de Student

La loi de Student est utilisée lors des tests de comparaison de paramètres comme la moyenne et dans l'estimation de paramètres de la population à partir de données sur un échantillon (Test de Student).

Définition : Soit U une variable aléatoire suivant une loi normale centrée réduite et V une loi du χ^2 à n *ddl*. U et V étant indépendante on dit alors que $t = \boxed{\frac{U}{\sqrt{V/n}}}$ suit une loi de Student à n *ddl*.

En utilisant R commander, étudier l'allure de la fonction de densité de la loi de Student en fonction du degré de liberté.
Quel est le quantile, q, vérifiant P($X > q$) =0.05 pour 1 *ddl*, 2 *ddl*, 20 *ddl*.



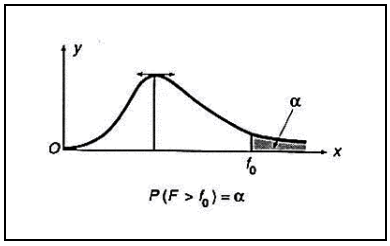
4. Loi de Fisher-Snedecor

La loi de Fisher-Snedecor est utilisée pour comparer deux variances observées et sert surtout dans les tests d'analyse de variance et de covariance.

Définition : Soit U et V deux variables aléatoires suivant des lois de Pearson respectivement à n et m *ddl*.

La statistique $F = \dfrac{U/n}{V/m}$ suit une loi de Fisher-Snedecor à *(n,m) ddl*.

En utilisant R commander, étudier l'allure de la fonction de densité de la loi de Fisher-Snedecor en fonction de n et m.



Annexe B : Construction d'un test statistique

Différentes étapes doivent être suivies pour tester une hypothèse :

(1) définir l'hypothèse nulle (notée H_0) à contrôler,

(2) choisir un test statistique ou une statistique pour contrôler H_0,

(3) définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée »,

(4) définir le niveau de signification du test ou région critique notée α,

(5) calculer, à partir des données fournies par l'échantillon, la valeur de la statistique

(6) prendre une décision concernant l'hypothèse posée et faire une interprétation biologique

L'hypothèse nulle est par exemple : l'absence de différence entre deux moyennes, l'adéquation d'une distribution à une distribution donnée, l'indépendance de deux caractères …
L'hypothèse alternative est alors la négation de cette hypothèse.

Les tests statistiques sont généralement construits de la façon suivante :

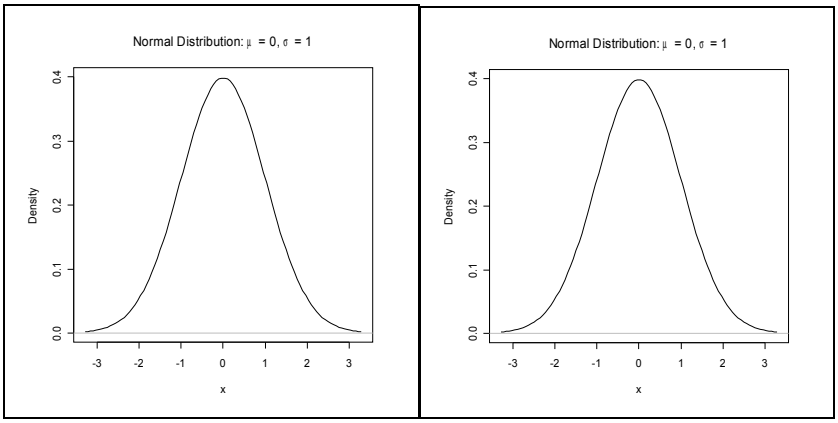
a. On suppose l'hypothèse H_0 vraie, par exemple $\mu=\mu_0$.

b. On construit alors une statistique *T* à partir des observations, par exemple $T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$,

qui va suivre une loi connue pour H_0 vraie, ici *T* suit la loi de Student à *n-1 ddl.*

c. D'après les tables de cette loi, on calcule la probabilité *P* d'obtenir une valeur égale ou supérieure à celle observée.

d. On construit une règle de décision : - Si $P < 0.05$ on rejette H_0
- Si $P \geq 0.05$ on accepte H_0



Un tel test fait intervenir deux types d'erreurs :

Le risque de première espèce α qui consiste à rejeter H_0 alors qu'elle est vrai, ici $\alpha=0.05$. C'est le risque que l'on prend en rejetant H_0.

Le risque de seconde espèce β qui consiste à accepter H_0, alors qu'elle est fausse. Ce risque est souvent inaccessible et dépend du protocole expérimental ainsi que de l'hypothèse alternative. La valeur 1- β est appelée puissance d'un test (à montrer une différence donnée).

Annexe C : Installation de R

Créer un répertoire logiciel R dans lequel vous allez installer le logiciel.

Méthode d'installation : Vous pouvez installer ce logiciel gratuit sur votre ordinateur personnel en le téléchargeant à l'adresse suivante :

http://cran.miroir-francais.fr/

Lancer le fichier .exe, vous devez ensuite télécharger les packages Rcmdr, ade4 …

Pour installer Rcmdr, taper dans R avec une liaison internet :

```
install.packages("Rcmdr", dependencies=TRUE)
install.packages("ade4", dependencies=TRUE)
```

Rcmdr comme ade4 nécessitent de nombreux autres packages qui sont chargés automatiquement grace à `dependencies=TRUE`

Une aide sur le logiciel R est disponible à l'adresse suivante :

http://www.r-project.org/

Une aide sur l'interface R commander est disponible sur le site :

http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/

Pour activer l'interface, entrer dans R puis taper library("Rcmdr").

Une aide sur ade4 est disponible sur le site :

http://pbil.univ-lyon1.fr/ADE-4/

Création d'un fichier dans R :

Il est possible pour un petit tableau de le créer directement dans R avec R cmdr par exemple.

Pour un plus grand fichier, mieux vaut le créer dans un tableur de type excel, puis le sauvegarder en format txt dans le répertoire de travail. Il faut faire en sorte que la virgule soit un point dans le fichier txt.

On récupère alors le fichier dans R en tapant :

>tableau=read.table(« fichier.txt », header=TRUE) en absence de légende des individus

piece.txt

facteur	mesure
A	5
A	7
A	6

```
> piece=read.table("piece.txt",header=TRUE)
  facteur mesure
1       A      5
2       A      7
3       A      6
```

>tableau=read.table(« fichier.txt ») avec une légende pour chaque individu (la colonne des individus n'a pas de nom dans .txt)

soclib.txt

	SALA	EXPO	PRES	PREL	TXEP
RFA6	90.1	30.6	20.2	37.8	12.4
FRA6	84.2	22.5	27.5	45.6	12.1
ITA6	74.1	24	24	34.7	21.9
GBR6	89.2	29.7	15.3	38.1	5.4
USA6	91.5	9.9	12.7	29.2	3.3
JAP6	82.8	15.9	13.9	28	16.8

```
>soclib=read.table("soclib.txt")
     SALA EXPO PRES PREL TXEP
RFA6 90.1 30.6 20.2 37.8 12.4
FRA6 84.2 22.5 27.5 45.6 12.1
ITA6 74.1 24.0 24.0 34.7 21.9
GBR6 89.2 29.7 15.3 38.1  5.4
USA6 91.5  9.9 12.7 29.2  3.3
JAP6 82.8 15.9 13.9 28.0 16.8
```

Installation de R, Rcmdr ade4, FactoMineR : 21/10/08

Installation du logiciel de base (windows) :

Site officiel de R : http://www.r-project.org/
puis dowload CRAN et choisir un miroir français : Lyon par exemple
Choisir :
- windows
- base
- cliquer sur le fichier .exe et c'est parti

Ou encore plus simple, taper :
http://cran.univ-lyon1.fr/bin/windows/base/R-2.7.2-win32.exe

Dire oui à toutes les étapes

Installation des packages (windows) :
Lancer le logiciel R puis taper les commandes suivantes :

```
install.packages("ade4", dependencies=TRUE)
install.packages("Rcmdr", dependencies=TRUE)
install.packages("ade4TkGUI", dependencies=TRUE)
install.packages("FactoMineR", dependencies=TRUE)
```

Installation directe de factominer et Rcmdr sous R :

>source("http://factominer.free.fr/install-facto.r")

Compléments pour l'analyse de données sous R

ade4TkGUI

Cette librairie permet d'accéder aux fonctions de ade4 interactivement.

Sous R : >library(ade4TkGUI) puis >ade4TkGUI

Compléments pour l'analyse de données sous R

FactoMineR

Sous R, taper : > source("http://factominer.free.fr/install-facto.r")

(sous réserve d'accés à internet)

Utilisation de FactoMineR sous Rcmdr

Sélectionner le tableau sous Rcmdr puis dans le menu FactoMineR choisir la méthode.

- Vous pouvez choisir les variables et les individus en supplémentaires.
- Les résultats sont disponibles sous R dans le fichier de sortie noté par défaut res.
- Vous pouvez également sauver un fichier excel avec outputs.
- Pour afficher des plans factoriels, sélectionner dans FactoMineR les axes à représenter

Aides à l'interprétations à récupérer dans le fichier R res ou le fichier excel:

ACP

- valeurs propres : res$eig
- qlt et ctr : Sous R, les qlt sont sous la forme resvarcos2 ou resindcos2, les ctr resvarcontrib ou resindcontrib.
- Tous sur les variables et individus : res$var ou res$ind
- variables supplémentaire : res$quanti.sup et res$ind.sup

AFC

- valeurs propres : res$eig
- qlt et ctr : Sous R, les qlt sont sous la forme rescolcos2 ou resROWcos2, les ctr rescolcontrib ou resrowcontrib.
- Tous sur les profils ligne et colonne : res$col ou res$var
- `profil supplémentaire : res$col.sup res$row.sup`