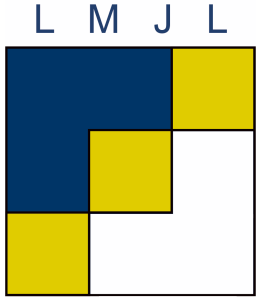
UNIVERSITÉ DE NANTES



Laboratoire Central
de Surveillance de le Qualité de l'Air

Statistiques pour données de pollution atmosphérique

Frédéric LAVANCIER

Université de Nantes,
Laboratoire de Mathématiques Jean Leray
email : frederic.lavancier@univ-nantes.fr

6 décembre 2014

Seconde partie : Liens significatifs, modélisation

5. La significativité en statistique
6. Les tests statistiques
7. Test de comparaison de 2 moyennes
   - t-test indépendant
   - t-test apparié
8. Généralisation du t-test indépendant : l'ANOVA
9. La régression linéaire
   - Principe général
   - Mise en oeuvre
   - Diagnostics
10. Aspects temporels
    - Séries temporelles et dépendance temporelle
    - Tirer profit de la dépendance
    - Analyser les cycles et la tendance
    - Précautions en présence de dépendance temporelle

Outils informatiques : Excel vs R

Excel :

- Manipulation des données assez aisée (importation ; exportation ; tri, création, suppression ou modification de variables)
- Au niveau statistique :
  - quelques graphiques descriptifs (nuage de points, courbe de séries), mais le choix est très limité et manque de souplesse,
  - quelques rares méthodes stat. sont disponibles dans Macros Complémentaies\Utilitaire d'analyse (régression linéaire, ANOVA, tests), mais sont peu pratiques à utiliser.

$\Rightarrow$ Excel est globalement mal adapté aux études statistiques.

Logiciels spécialisés payants : SPSS, SAS, Splus, Spad...

Logiciel spécialisé gratuit : R

Le logiciel R

Le logiciel R est disponible pour Windows, MacOS ou Linux sur le site

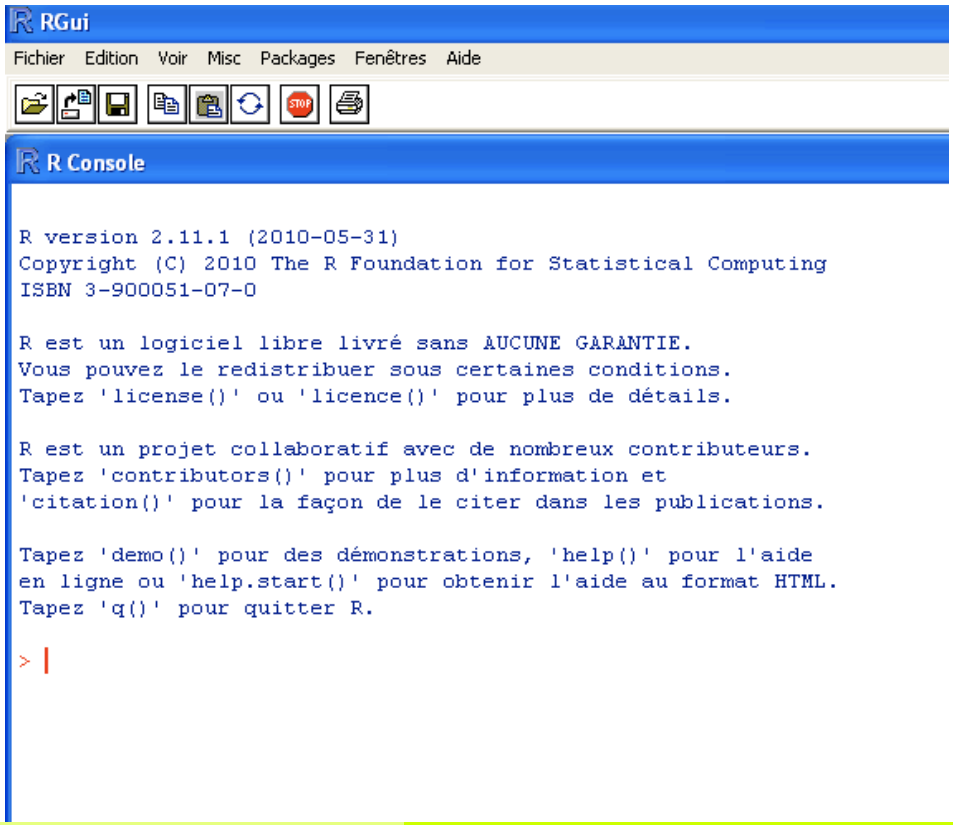
http://cran.r-project.org/

L'utilisation de R se fait principalement à l'aide de commandes que l'on entre dans la console R.

Au démarrage :

> apparaît automatiquement au début de chaque ligne de commande.

\+ apparaît si la ligne précédente est incomplète.



Le logiciel R

Les commandes font appel à des fonctions :

- soit disponibles par défaut dans R,
- soit présentes dans des packages complémentaires à télécharger :
  - dans le menu Packages\Installer le(s) package(s)...(téléchargement et installation en ligne)
  - ou en .zip depuis le site du `cran` (taper le nom du package + `cran` sur google) puis Packages\Installer le(s) package(s) depuis des fichiers zip...

ex : la commande mean(x) calcule et renvoie la moyenne de la série x.

Il existe cependant une interface graphique clique-boutons :

R Commander.

AIDE pour l'utilisation de R en lignes de commandes

Quelle fonction permet de réaliser telle ou telle procédure ?

- help.start() ou Aide\Aide HTML lance l'aide générale qui permet notamment de rechercher des fonctions.
- ??mot-clé ou Aide\Rechercher dans l'aide... présente les fonctions disponibles en lien avec le mot-clé. ex : ??regression

Une fois la fonction ciblée, comment l'utiliser ? Que fait-elle exactement ?

- ?"fonction" ou Aide\Fonctions R (texte)... lance l'aide spécifique sur la fonction "fonction" (comment l'utiliser, les différentes options, des exemples). ex : ?mean
- example("fonction") lance automatiquement les exemples présentés dans l'aide de la fonction "fonction" . Il faut en général appuyer sur "Entrée" pour passer à l'exemple suivant. ex : example(barplot)

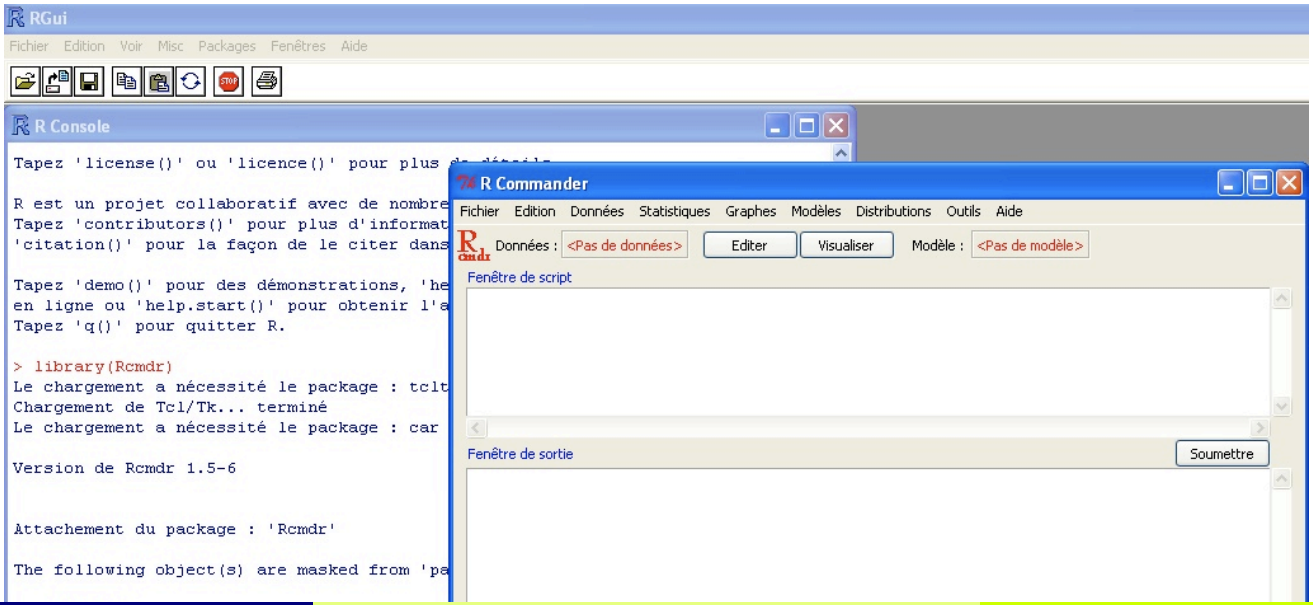
R Commander

Avantages : utilisation clique-boutons, prise en main rapide.
Inconvénients : fonctionnalités limitées, parfois instable.

Il faut d'abord télécharger le package Rcmdr (depuis le menu Packages\Installer le(s) package(s)...ou à partir d'un fichier .zip, voir p9) puis lancer R Commander par la commande :

```
library(Rcmdr)
```

(Rem : on peut le réouvrir après fermeture par la commande `Commander()`)



Importer des données

Avec R Commander : Utiliser le menu Données\Importer des données
On peut ensuite visualiser le tableau en cliquant sur Visualiser.

En ligne de commandes dans la console R :
Actualiser Fichier\Changer le répertoire courant...

- Pour des données .csv extraites de Xair :

```
mydata=read.csv("nomdefichier.csv",skip=1,sep="\t")
```

- Pour des données .xls ou .xlsx :
  library(xlsx) (pour charger le package `xlsx`)
  mydata=read.xlsx("nomdefichier.xlsx",1) (pour ouvrir la feuille 1)

Les données ainsi chargées sont aussi visibles avec R Commander (sélectionner les données à visualiser en cliquant à droite de Données).

Rem : ATMO PC a écrit une fonction R qui permet d'extraire directement des données de Xair.

Si problème : formater les données avec un tableur (préférer le .csv)

Manipuler/Modifier les données

Remarque : Il est toujours possible de modifier le jeu de données dans Excel (ou autre) avant de l'importer sous R.

Avec R Commander : En cliquant sur Editer, on peut modifier chaque cellule du tableau actif. Pour les opérations plus sophistiquées, utiliser les menus Données\Jeu de données actif et Données\Gérer les variables dans le jeu de données actif. On peut notamment :

- éliminer des lignes ou des variables
- découper une variable en classes (ex : la direction du vent)
- extraire un sous-tableau (ex : uniquement les jours ouvrés)
- recoder des facteurs (ex : regrouper des modalités)

En ligne de commandes dans la console R :

names(mydata) donne le nom des variables présentes dans la table mydata.
On accède à la variable nommée var de la table mydata par mydata$var.

Exemple : mean(mydata$var) renvoie la moyenne de var .

On peut manipuler la table à l'aide des fonctions subset, recode, bin.var, cut ...

Première partie

Analyses graphiques

2. Analyse graphique univariée
   - Coup d'oeil global
   - Variable qualitative
   - Variable quantitative
3. Analyse graphique bivariée
   - Variable quantitative/ Facteur
     - Répartition par modalités du facteur
     - Rose des vents et rose de pollution
   - Variable quantitative/ Variable quantitative
4. Lien graphique entre plusieurs variables quantitatives
   - L'ACP
   - Compléments de l'ACP : la classification des individus
   - Classification de variables

2 Analyse graphique univariée

- Coup d'oeil global
- Variable qualitative
- Variable quantitative

Vue d'ensemble

Dans R Commander, Statistiques\Résumés\Jeu de données actif renvoie un résumé de toutes les variables dans la Fenêtre de sortie.

Exemple : les données tab contiennent, pour différents trajets en voiture, la moyenne de NO2, le type de voie empruntée et la fluidité du trafic.

On trouve en sortie :

- pour les variables quantitatives (comme NO2) : différentes valeurs de position.
- pour les facteurs (comme fluidite) : les effectifs de chaque modalité.

Fenêtre de sortie

```
> showData(tab, placement='-20+200',
+   maxheight=30)

> summary(tab)
      NO2          type    fluidite
 Min.   : 13.4   A:69    A:79
 1st Qu.:147.1   P:78    B:68
 Median :206.6   T:33    C:65
 Mean   :245.3   U:75    D:74
 3rd Qu.:319.5   V:31
 Max.   :844.4
```

tab

	NO2	type	fluidite
1	378.94400	P	A
2	806.66940	T	D
3	634.57840	A	D
4	673.35140	T	C
5	589.75070	P	A
6	445.69850	V	D
7	382.80220	A	D
8	325.86510	U	D
9	194.99190	U	D
10	203.64420	P	A
11	216.00190	A	D
12	345.65270	P	D
13	275.53660	P	D

Variable qualitative (ou facteur)

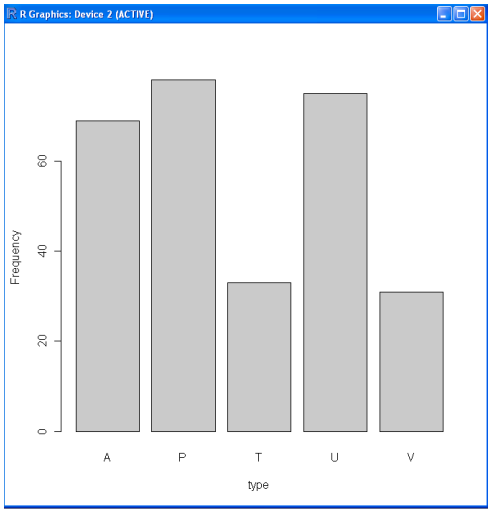
Dans R Commander :

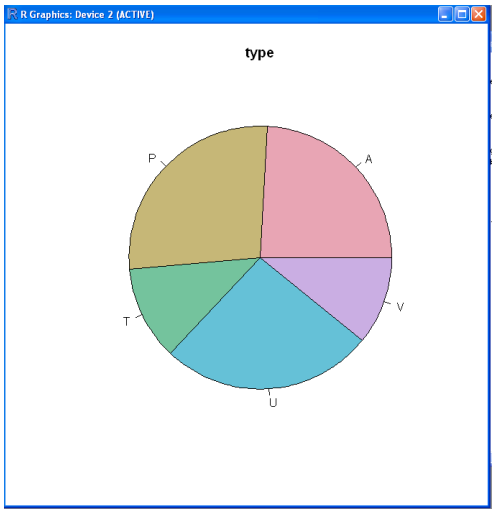
- Statistiques\Résumés\Distributions de fréquences... renvoie l'effectif et la fréquence de chaque modalité.
- Graphes\Graphe en barres... et Graphes\Graphe en camembert... résument la même information graphiquement.

Exemple : pour la variable type de la table précédente :

Fenêtre de sortie

```
> .Table <- table(tab$type)

> .Table  # counts for type

 A  P  T  U  V
69 78 33 75 31

> 100*.Table/sum(.Table)  # percentages for type

       A        P        T        U        V
24.12587 27.27273 11.53846 26.22378 10.83916

> remove(.Table)
```





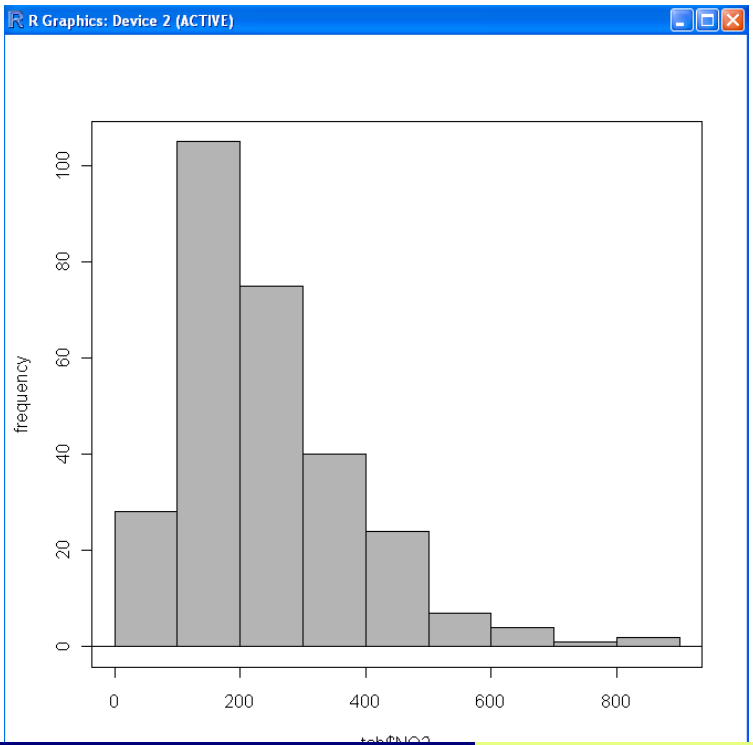
En ligne de commandes : fonctions table, barplot, pie.

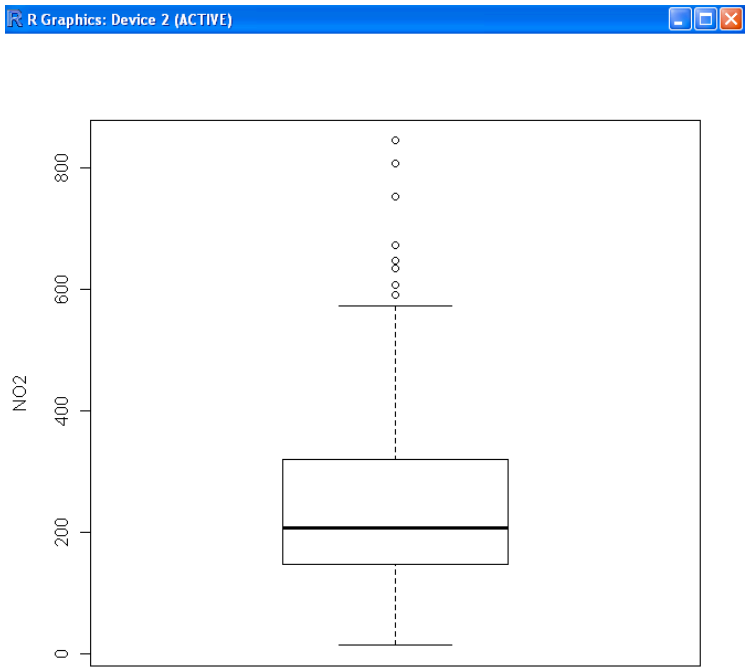
Variable quantitative

Dans R Commander :

- Statistiques\Résumés\Statistiques descriptives... renvoie la moyenne, l'écart-type, des quantiles, etc, selon le choix dans l'onglet Statistiques.
- Graphes\Histogramme... et Graphes\Boite de dispersion... résument graphiquement la distribution de la variable.

Exemple : pour la variable NO2 de la table précédente :





Répartition d'une série de mesures

Les quantités suivantes reflétent les valeurs prises par une série :

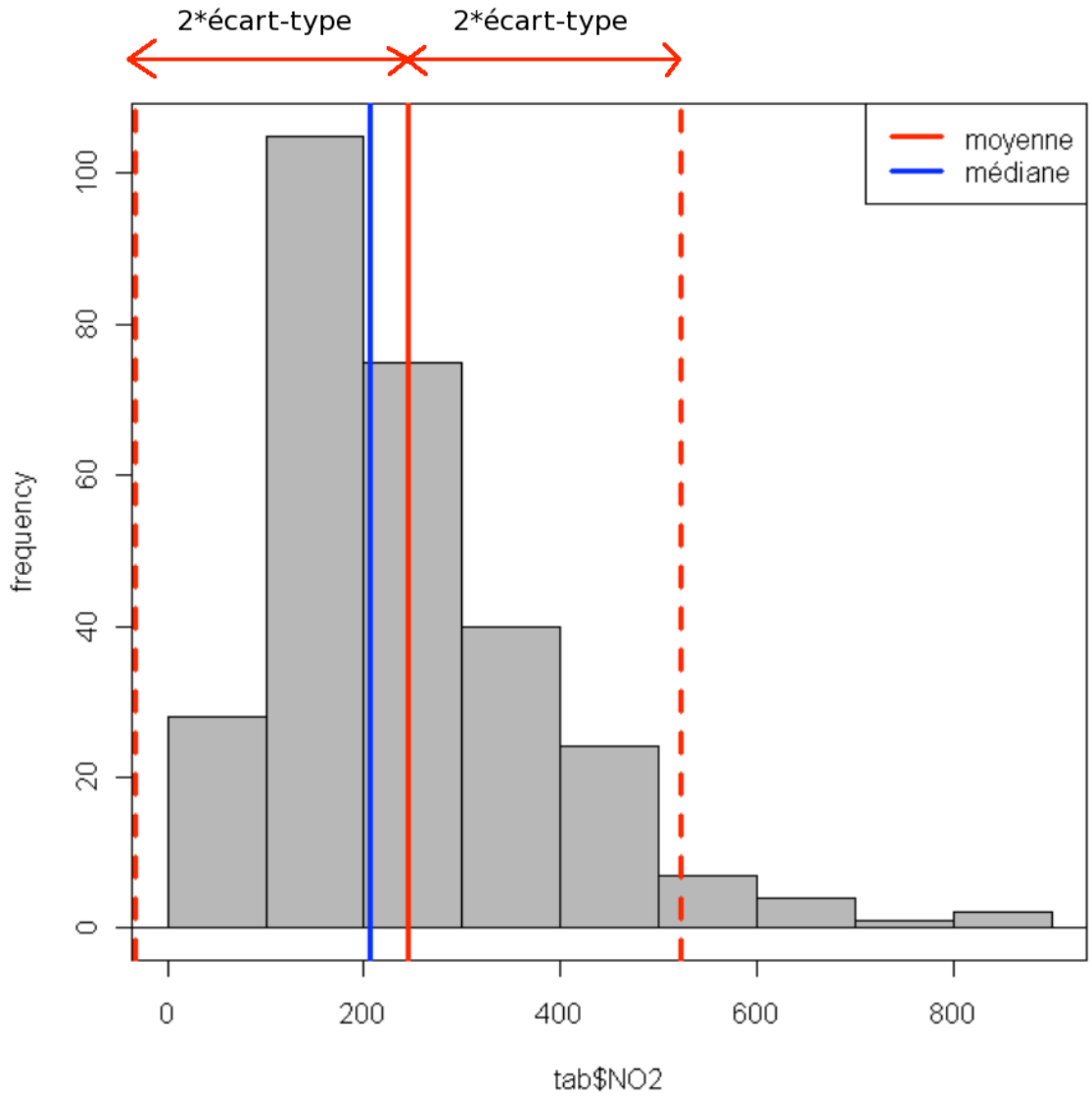
- la moyenne
- la variance : c'est la moyenne des $(x_i - m)^2$ où x_i parcourt les valeurs de l'échantillon et m représente la moyenne de l'échantillon.
- l'écart-type : c'est la racine-carrée de la variance.
- la médiane : elle sépare les valeurs de l'échantillon en 2 parties égales (50% des valeurs sont supérieures à la médiane, 50% sont inférieures)
- les percentiles (ou quantiles) : le percentile à 95% est la valeur telle que 95% des valeurs de l'échantillon lui sont inférieures (et donc 5% lui sont supérieures).

La moyenne seule résume mal les valeurs prises par l'échantillon.

L'écart-type apporte une information essentielle : il quantifie à quel point les valeurs peuvent être différentes de la moyenne.

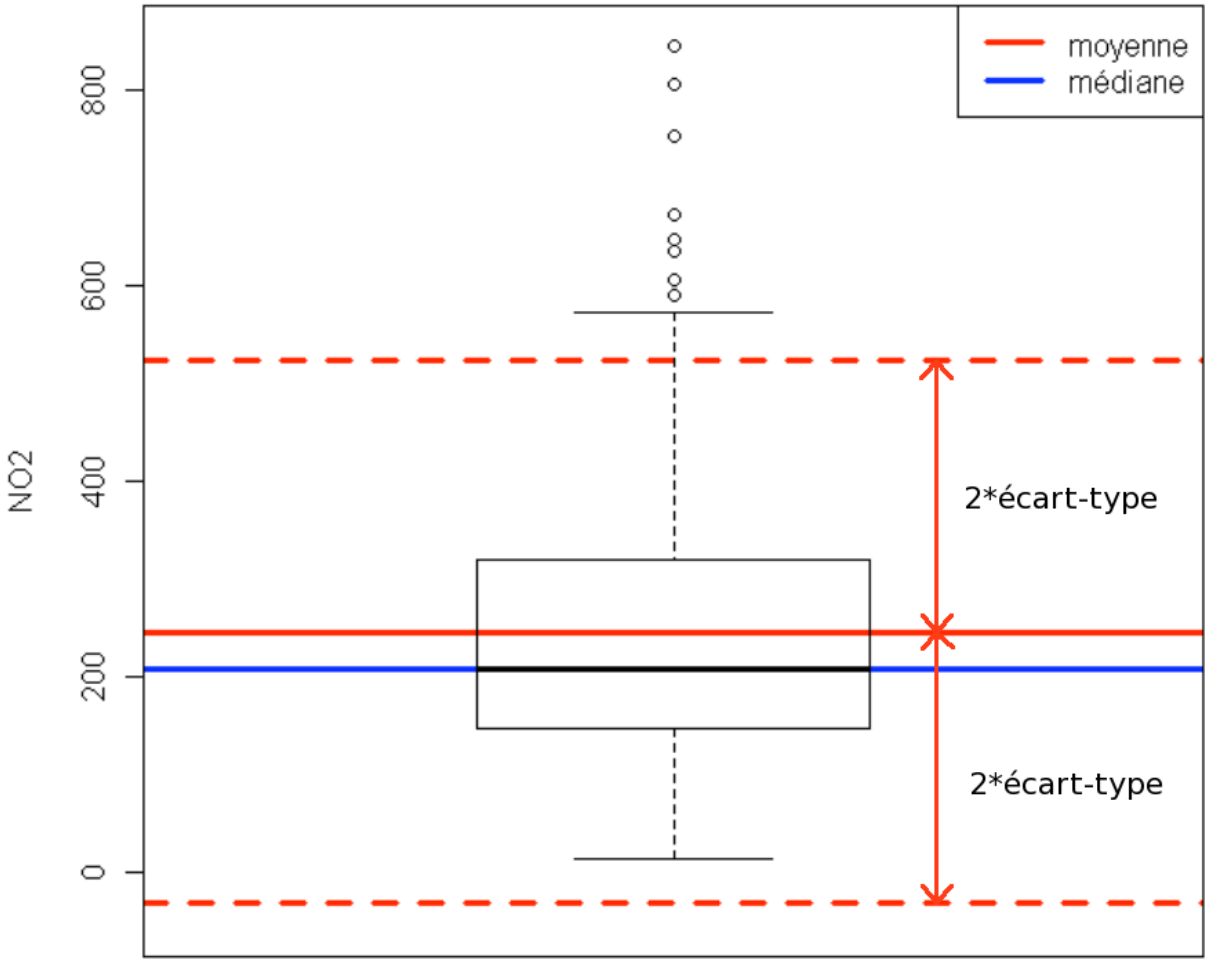
On peut comprendre l'écart-type de la façon suivante : la grande majorité des valeurs d'un échantillon (plus des trois quarts) sont à moins de 2∗ecart-type de la moyenne.

Exemple : Sur l'histogramme



On peut comprendre l'écart-type de la façon suivante : la grande majorité des valeurs d'un échantillon (plus des trois quarts) sont à moins de 2∗ecart-type de la moyenne.

Exemple : Sur la boite de dispersion (ou "boite à moustaches", ou encore "boxplot")



Les quantiles et les extrêmes

Un percentile d'ordre 90% (ou 95%) précise le seuil au delà duquel se trouvent les 10% (ou 5%) des valeurs les plus élevées.
En particulier, la médiane est le percentile à 50%.

Les valeurs extrêmes (comme le maximum) sont souvent intéressantes à étudier en soi :

- Sont-elles plausibles ou aberrantes ?
- Peut-on trouver une explication à ce comportement extrême ?

On peut repérer les valeurs extrêmes lors du tracé d'un boxplot en choisissant l'option Identifier les extrêmes à la souris.

3 Analyse graphique bivariée

- Variable quantitative/ Facteur
  - Répartition par modalités du facteur
  - Rose des vents et rose de pollution
- Variable quantitative/ Variable quantitative

Variable quantitative/ Facteur : stat. descriptives

L'option Résumer par groupe...de Statistiques\Résumés\Statistiques descriptives... permet de résumer différentes statistiques d'une série selon les modalités d'un facteur.

Exemple : La série PQVNH (ammoniac à Rouen) donne par année (remarque : on a demandé également le percentile à 90%)

```
> numSummary(a[,"pqvnh"], groups=a$annee, statistics=c("mean", "sd",
+   "quantiles"), quantiles=c(0,.25,.5,.75,0.9,1))
         mean       sd 0% 25% 50% 75% 90% 100%   n  NA
2005 1.909091 2.960661  0   0   1   2 5.0   20 253 112
2006 1.172107 1.544978  0   0   1   2 3.0    9 337  28
2007 2.582781 3.153562  0   0   2   4 6.9   23 302  63
2008 2.035813 2.334307  0   1   1   2 5.0   20 363   3
2009 1.602410 2.122757  0   0   1   2 4.0   26 332  33
```

En commande : by(a$pqvnh,a$annee,summary) ou by(a$pqvnh,a$annee,quantile).

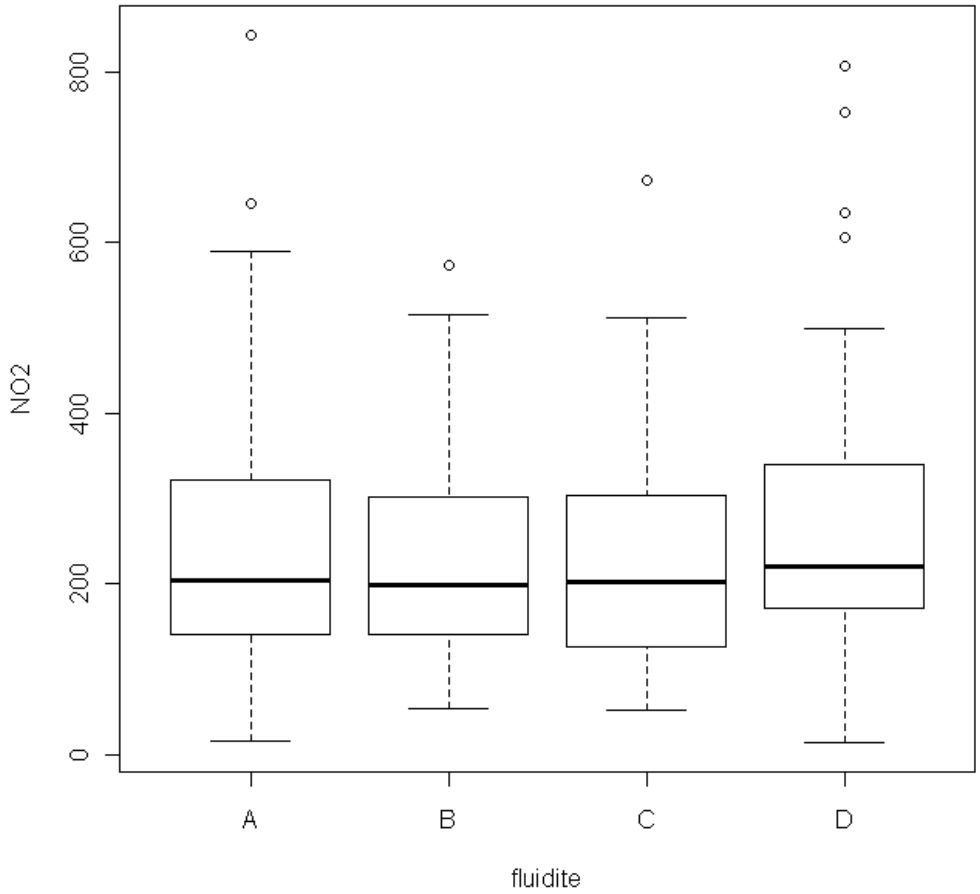
Variable quantitative/ Facteur : graphes

L'option Graphe par groupe... de Graphes\Boite de dispersion... permet une comparaison rapide de la répartition d'une série selon les modalités d'un facteur.

Exemple : Pour la série NO2 précédente, classée selon la fluidité du trafic, de fluide ("A") à congestionné ("D").

Remarque : les différences semblent peu convaincantes malgré des moyennes distinctes.

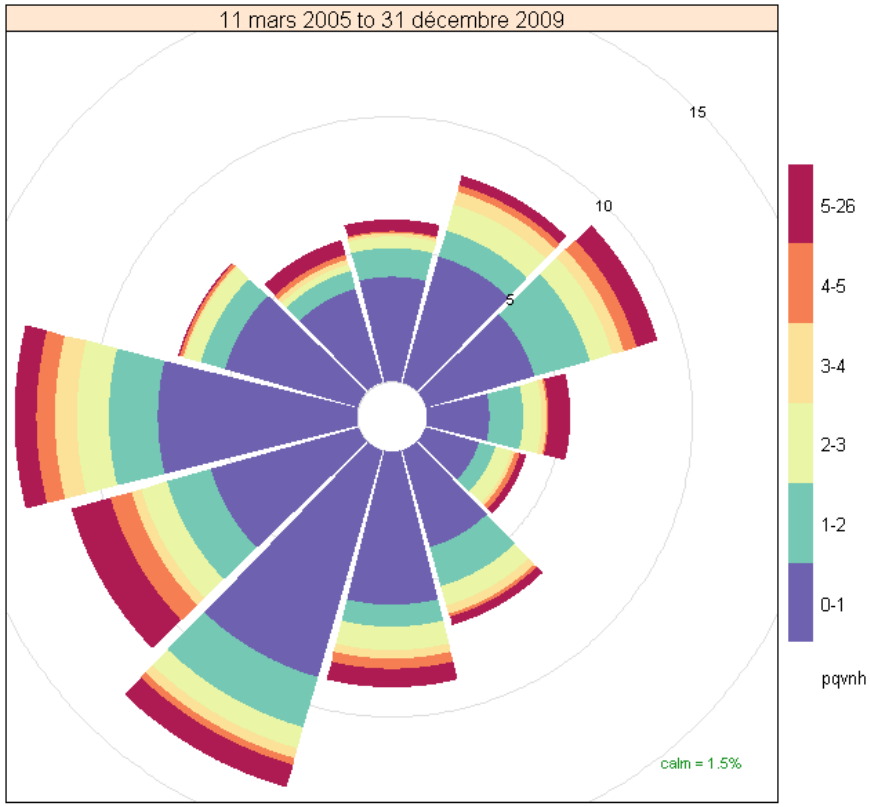
	mean	n
A	244.1691	79
B	232.5729	68
C	235.1250	65
D	266.9988	74

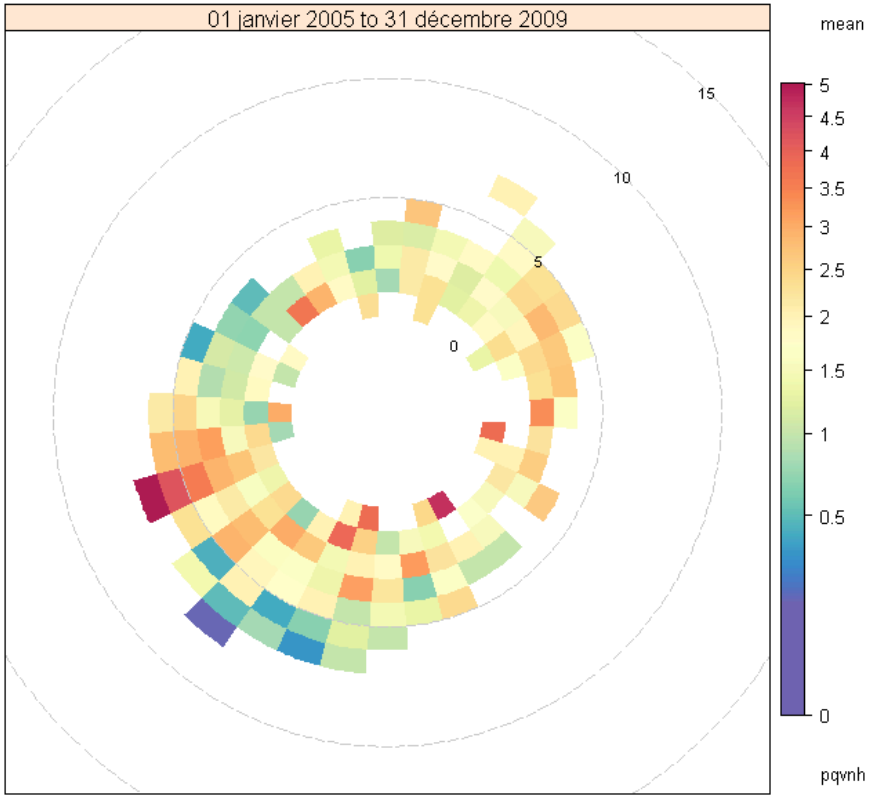


En commande : boxplot(NO2~fluidite, data=tab)

Rose des vents

Uniquement en ligne de commandes : avec le package openair , on peut faire les roses suivantes (exemple pour l'ammoniac "PQVNH" à Rouen) :





- Taille des pétales = fréquence de la direction
- Couleurs = valeurs du polluant dans cette direction.

- Une cellule par direction et vitesse du vent (qui est représentée par les cercles)
- Couleur de la cellule = moyenne du polluant pour cette direction et cette vitesse du vent.

Rose des vents : utilisation de openair

La fonction pollutionRose

Elle permet de tracer une rose des vents colorée selon la vitesse du vent (par défaut) ou selon les valeurs d'un polluant (option pollutant)

Exemple : pour tracer la rose précédente (figure de gauche)
pollutionRose(tab,ws="speed",wd="dir",pollutant="pqvnh", paddle=F, angle=30)

- tab : nom du tableau de données ;
- "speed" : nom de la variable vitesse du vent ;
- "dir" : nom de la variable direction du vent (entre 0 et 360) ;
- "pqvnh" : nom du polluant qui colore les pétales ;
- angle : définit l'angle de chaque pétale ;
- paddle : précise la forme des pétales (T : "rames" ou F : "pétales").

La fonction windRose

Elle permet de faire une rose des vents classiques, colorée selon la vitesse du vent.

Exemple : avec les même options que ci-dessus
windRose(tab,ws="speed",wd="dir",paddle=F, angle=30)

Rose des vents : utilisation de openair

La fonction polarFreq

Elle permet de faire la représentation "en cellules" précédente (figure de droite).
Avec la version actuelle de openair (12/2014), la variable "vitesse du vent" doit obligatoirement se nommer ws (wind speed) et la "direction du vent" se nommer wd (wind direction) et être entre 0 et 360 (ces restrictions disparaîtront vraisemblablement dans une prochaine version et ces noms pourront être précisés dans la fonction).

Exemple de mise au bon format :
names(tab) renvoie le nom des variables de la table tab.
names(tab)[2]='ws' permet de nommer la 2ème variable ws.

Exemple d'utilisation :
polarFreq(tab,pollutant="pqvnh",statistic="mean",min.bin=5)
L'option statistic précise la statistique calculée dans chaque cellule, min.bin définit le nombre minimal de valeurs par cellule exigé pour valider le calcul de cette statistique.

Rose des vents : alternative à openair

Les fonctions du package openair souffrent de quelques défauts :

- elles ne permettent pas la construction de roses dont la taille des pétales représente la moyenne d'un polluant selon les directions (mais uniquement la fréquence d'apparition de ces directions). Néanmoins, la fonction polarFreq permet de bien visualiser la moyenne d'un polluant selon les directions et la vitesse du vent.
- la fréquence est proportionnelle à la longueur des pétales et non à leur aire (ce qui serait visuellement plus correct).

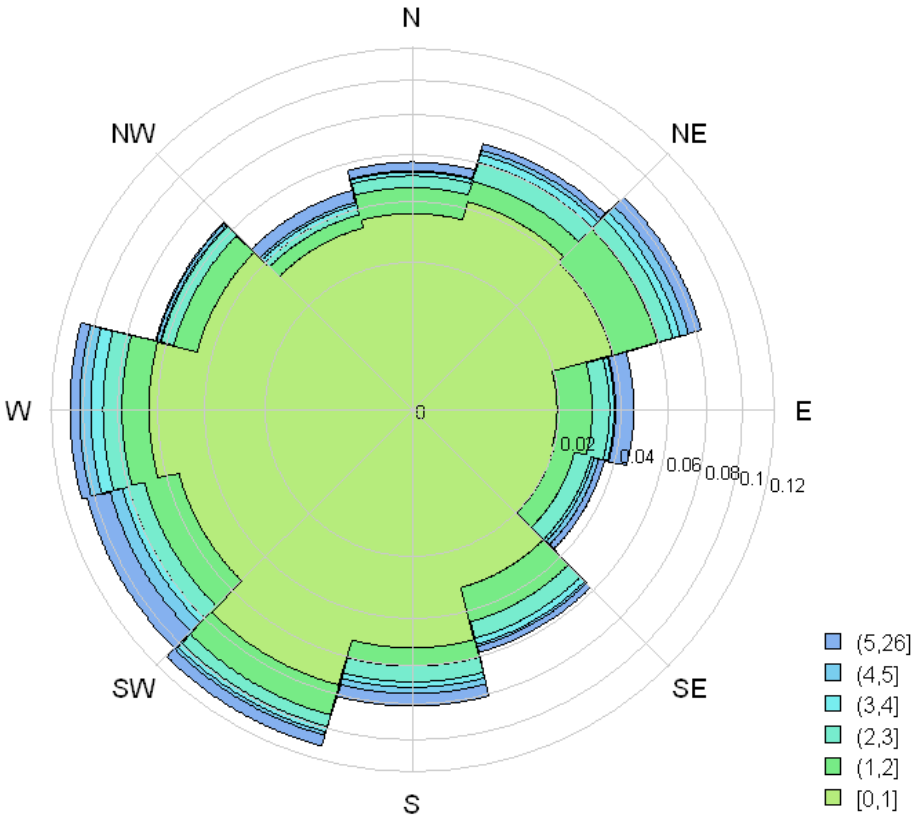
La fonction rose du package IDPmisc est une alternative plus souple (mais moins jolie).

- on peut choisir quelle quantité représentera la taille des pétales
- par défaut, les valeurs représentées sont proportionnelles à l'aire des pétales et non à leur longueur (choix tout de même possible)
- on peut choisir quelle variable sera représentée en couleur dans chaque pétale

Rose des vents : exemples avec l'ammoniac à Rouen

Même rose qu'avec openair

- Taille des pétales (**aire**) = fréquence de la direction
- Couleurs = valeurs du polluant.



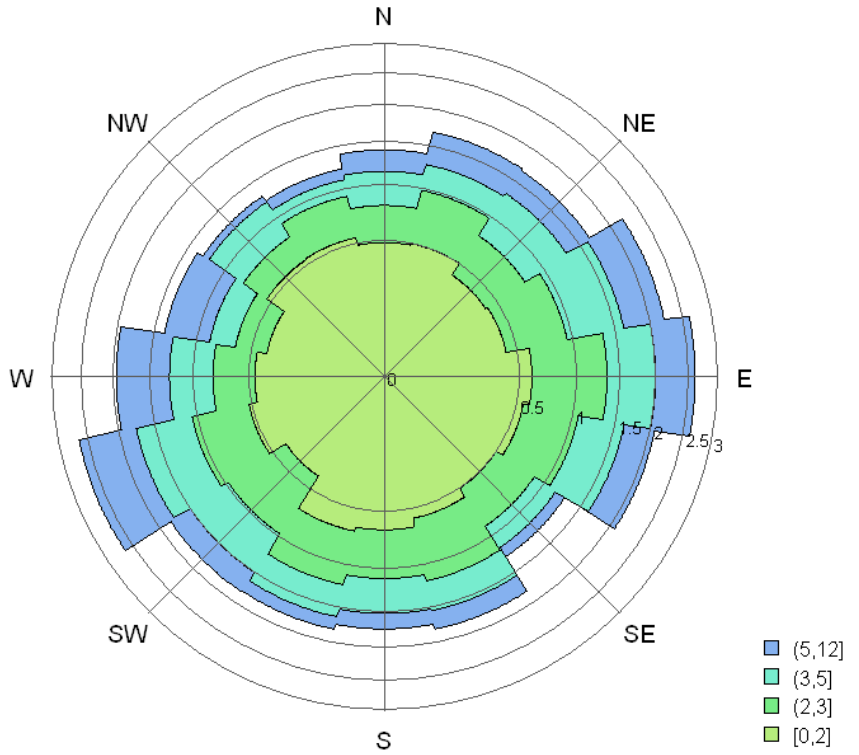
Pour cela, on calcule la rose d'un vecteur taux constant :

```
taux=rep(1,length(a$wd))/length(a$wd)
rosevent=rose(taux,FUN=sum,cut=a$pqvnh,...)
plot(rosevent,general=general.control(stacked=T))
```

Voir ?rose pour les autres options.

Rose de pollution sur 16 directions

- Taille des pétales (**aire**) = moyenne d'ammoniac dans la direction
- Couleurs = vitesse du vent.

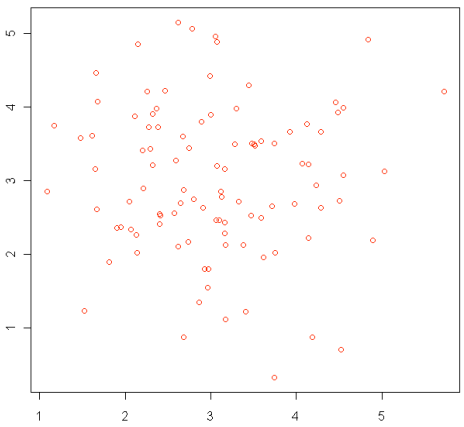


On calcule la rose de "pqvnh". On divise par quatre pour compenser l'empilement des 4 niveaux de vitesses :

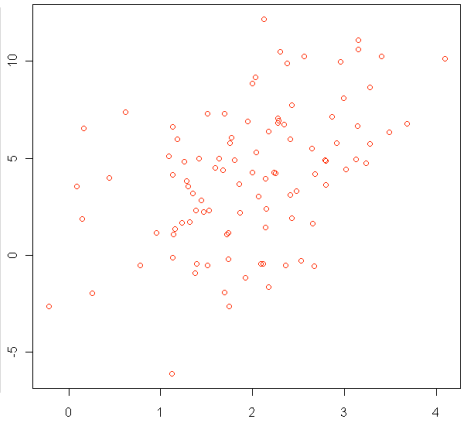
```
rosepoll=rose(a["pqvnh"]/4, FUN=mean, cut=a$ws,...)
plot(rosepoll,general=general.control(stacked=T))
```

Lien entre deux variables quantitatives

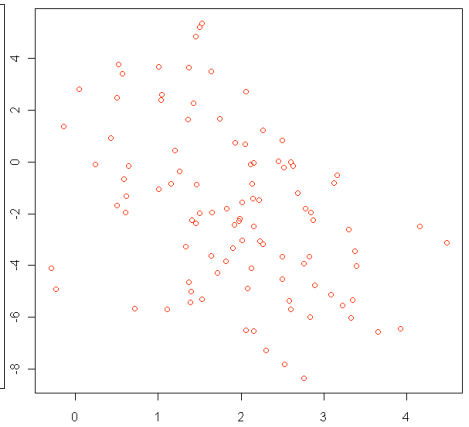
On le visualise grâce à Graphe\Nuage de points...



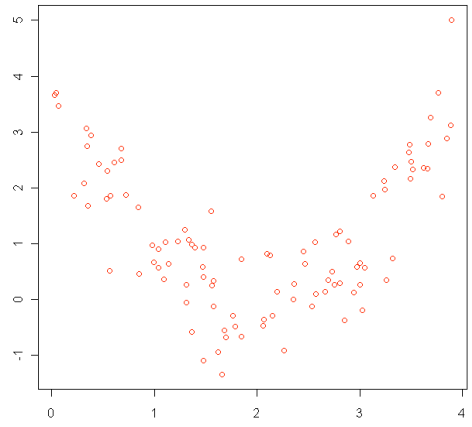
Aucun lien ($r \approx 0$)



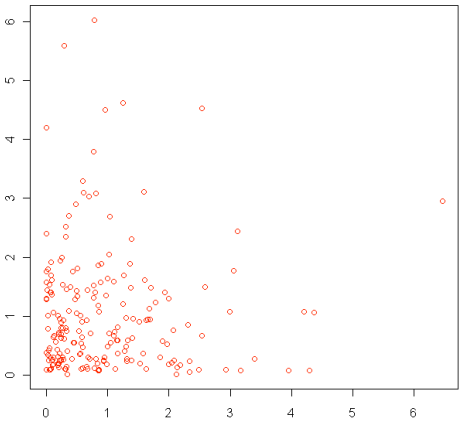
Lien linéaire ($r \approx 0.4$)



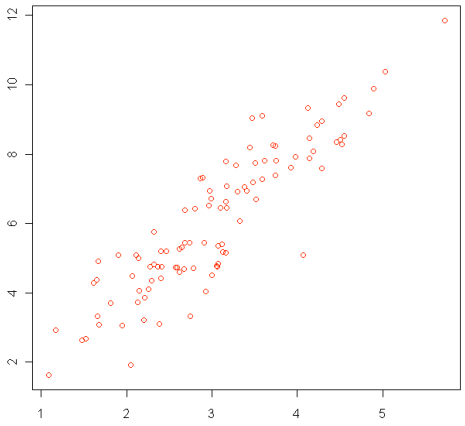
Lien linéaire ($r \approx -0.4$)



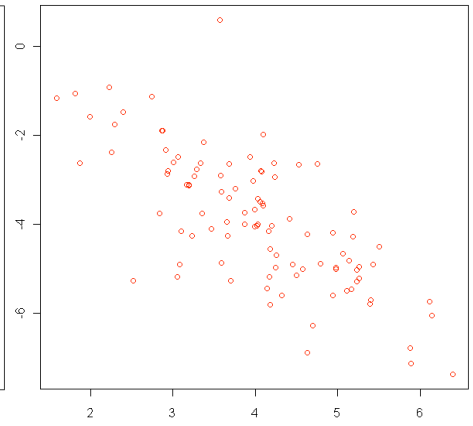
Lien non-linéaire ($r \approx 0$)



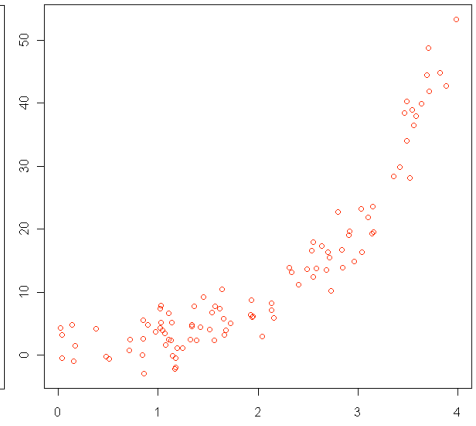
Aucun lien ($r \approx 0$)



Lien linéaire ($r \approx 0.9$)



Lien linéaire ($r \approx -0.8$)



Lien non-linéaire ($r \approx 0.8$)

Lien entre deux variables quantitatives : corrélation

Le lien **linéaire** entre deux variables quantitatives se mesure par la corrélation (ou coefficient de Pearson).

Sous R Commander : utiliser le menu Statistiques\Résumés\Matrice de corrélations...

En ligne de commandes : la corrélation entre deux séries x et y est donnée par cor(x,y), en présence de valeurs manquantes : cor(x,y, use='complete.obs')

La corrélation r est toujours comprise entre -1 et 1 :

- si $r = 1$, il y a un lien linéaire "parfait" positif
- si $r = -1$, il y a un lien linéaire "parfait" négatif
- si $r = 0$, il n'y a aucun lien linéaire.

Important : (cf les nuages de points de la colonne de droite précédente)

Il faut toujours représenter le nuage de points pour répérer d'éventuels liens non-linéaires. r peut-être négligeable alors que les variables ont un fort lien non-linéaire entre elles. Même lorsque r est élevé, il peut y avoir un lien non-linéaire plus pertinent.

Lien entre variables quantitatives : un exemple

Les données contiennent 4 séries de mesures horaires de NO2 sur 5 mois : deux en doublon sur le même site urbain "JUS" au centre-ville de Rouen, une sur le site urbain "CHS" à Rouen, et une sur le site urbain "EVT" à Evreux.

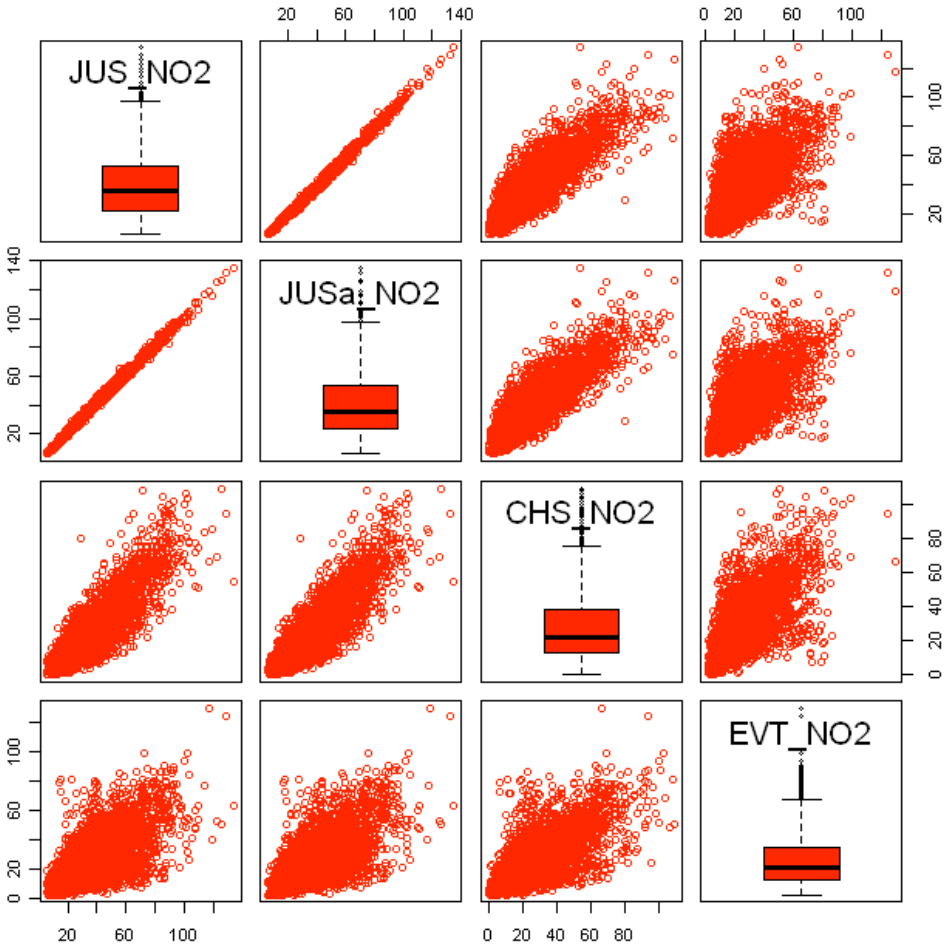
La matrice de corrélation vaut :

Fenêtre de sortie

```
> cor(tab[,c("CHS_NO2","EVT_NO2","JUS_NO2","JUSa_N
           CHS_NO2   EVT_NO2   JUS_NO2  JUSa_NO2
CHS_NO2  1.0000000 0.6989207 0.8606922 0.8618510
EVT_NO2  0.6989207 1.0000000 0.6496888 0.6519591
JUS_NO2  0.8606922 0.6496888 1.0000000 0.9977919
JUSa_NO2 0.8618510 0.6519591 0.9977919 1.0000000
```

Pour tracer plusieurs nuages de points avec R Commander : Graphes\Matrice de nuages de points...

En lignes de commandes : fonction scatterplot.matrix



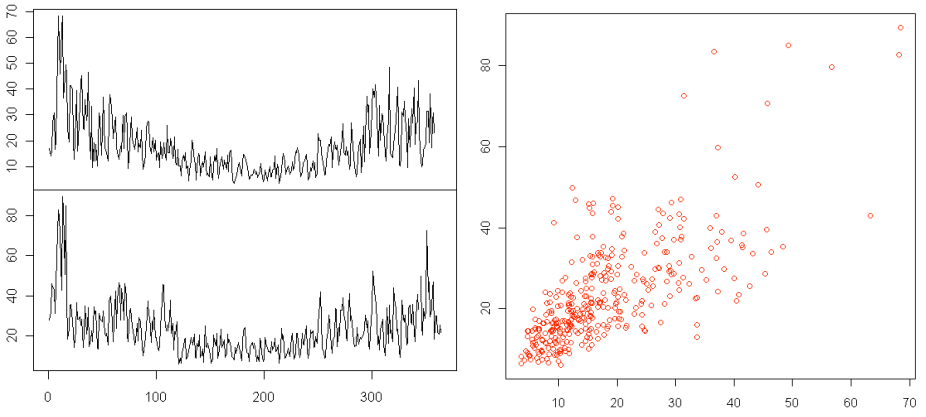
Corrélation fortuite

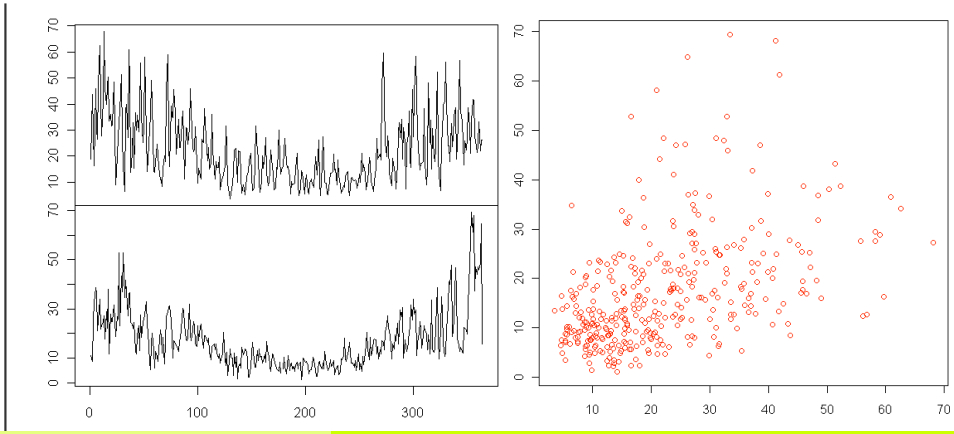
Attention : une corrélation entre deux variables ne signifie pas qu'il y a un lien de cause à effet.

- Lorsque $r > 0$, cela témoigne juste du fait que lorsqu'une des deux variables prend des grandes valeurs, l'autre a plutôt tendance à prendre aussi des grandes valeurs.
- Si $r < 0$, c'est le contaire, des grandes valeurs pour l'une correspondent à des petites valeurs pour l'autre.

Cette corrélation peut-être fortuite : elle peut-être due à une tendance globale commune (voir l'ex ci-dessous), à une cause commune cachée (comme la météo), etc.

Exemples : A gauche : séries de NO2 journalier au Puy en Velay et à Bassens (banlieue de Bordeaux) en 2009. La corrélation entre les deux séries est de $r = 0.71$.
A droite : NO2 journalier au Puy en 2007 et à Toulouse en 2009 ($r = 0.45$).
(cf partie 10 pour la représentation de séries).





4 Lien graphique entre plusieurs variables quantitatives
- L'ACP
- Compléments de l'ACP : la classification des individus
- Classification de variables

4 Lien graphique entre plusieurs variables quantitatives
- L'ACP
- Compléments de l'ACP : la classification des individus
- Classification de variables

Analyse en Composantes Principales (ACP)

Les valeurs prises par 2 variables forment un nuage de points en dimension 2 (cf ci-dessus).
Les valeurs prises par p variables forment un nuage de points en dimension p.

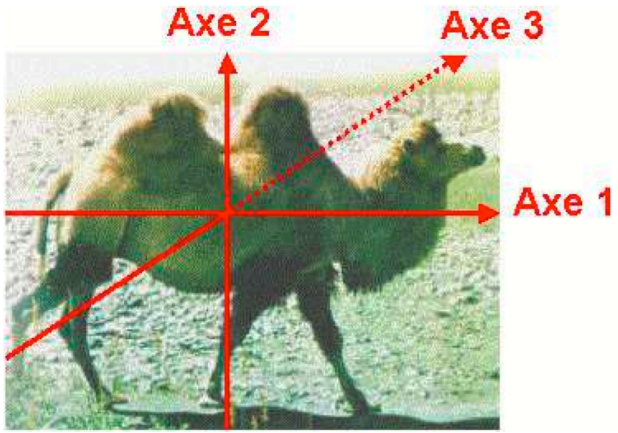
$\Rightarrow$ On ne peut donc pas visualiser ce nuage de points dès que $p > 3$.

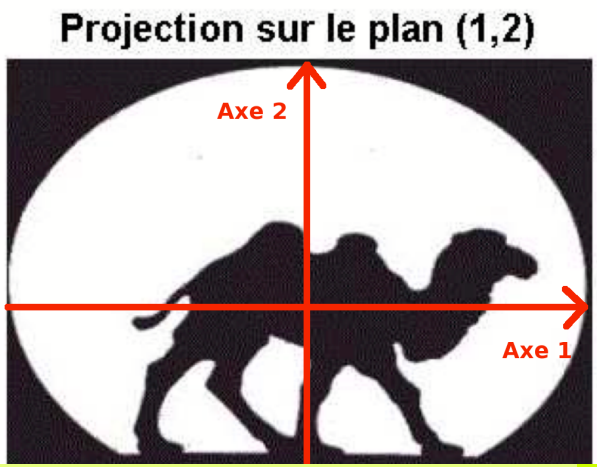
(pour $p = 3$, Graphes\Graphe 3D représente le nuage de points mais son analyse est délicate.)

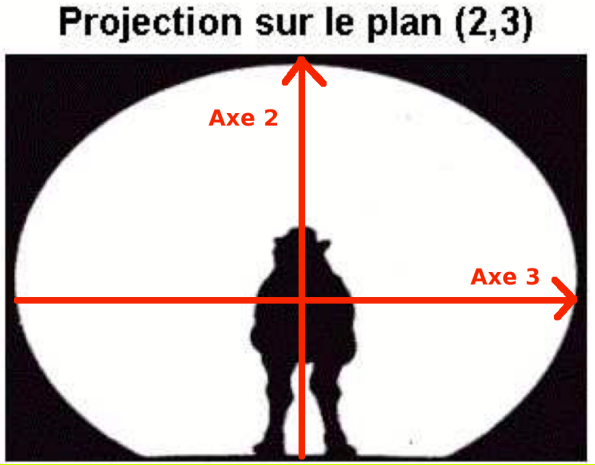
L'ACP permet de trouver et de classer (de 1 à p) **les directions** (orthogonales) **les plus informatives** dans un nuage de points en dimension p.

- On peut alors projeter et visualiser le nuage de points sur le plan formé des 2 axes les plus pertinents : cela fournit une représentation selon l'angle de vue le plus informatif possible.
- L'analyse du nuage de points peut se poursuivre en projetant sur le plan suivant, etc.

Ex : Un chameau en 3D ($p = 3$) se reconnait mieux dans le plan (1,2) que dans le plan (2,3).







Analyse en Composantes Principales (ACP)

Que représentent les axes principaux retenus par l'ACP ?

Chaque axe est une combinaison des variables initiales, chacune plus ou moins bien représentée par cet axe. Cette représentativité se mesure par la corrélation de la variable avec l'axe.
Plus généralement, les variables bien représentées par un plan sont repérables grâce au cercle des corrélations : plus une variable est proche du cercle, mieux elle est représentée dans ce plan.
Grâce à la position des variables dans le cercle des corrélations d'un plan donné, on observe :

- que certaines variables sont bien représentées par tel ou tel axe,
- que certaines variables sont en opposition sur tel ou tel axe,
- des regroupements entre variables (ce qui peut conduire à une classification).

Comment interpréter le nuage de points (les individus ou les dates) projetés sur un plan ?

On compare la position des points à celle des variables dans le cercle des corrélations de ce plan.

- S'ils sont dans la même zone, cela signifie que l'individu (ou la date) associé au point prend des valeurs relativement élevées pour ces variables.
- S'ils sont dans des zones opposées, c'est le contraire.

On peut ainsi mettre en évidence des groupes de points ayant des caractéristiques similaires ou opposés relativement aux variables bien représentées dans le cercle des corrélations.

Comment connaitre la proportion d'information contenue dans un plan ?

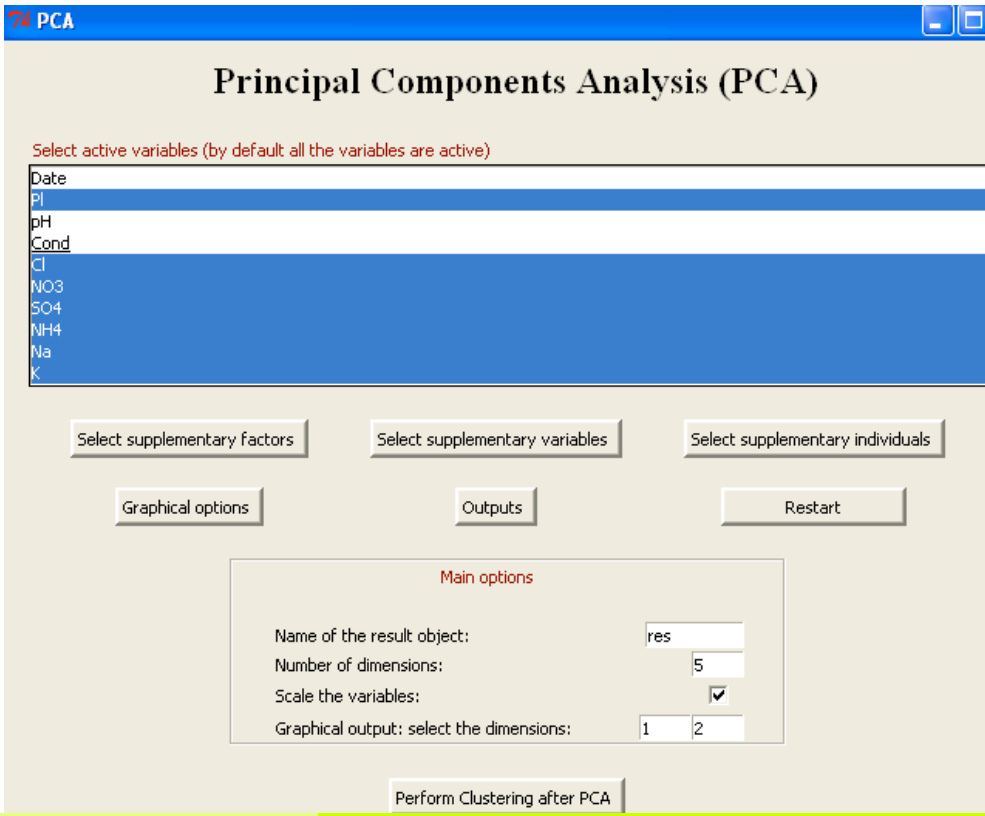
Il suffit d'additionner le pourcentage d'inertie de ses deux axes. Ce pourcentage représente la part de l'information initiale (la variabilité) conservée après projection sur cet axe. Les axes sont toujours classés du plus informatif au moins informatif.

Ex : ACP sur des mesures d'ions dans les eaux de pluie

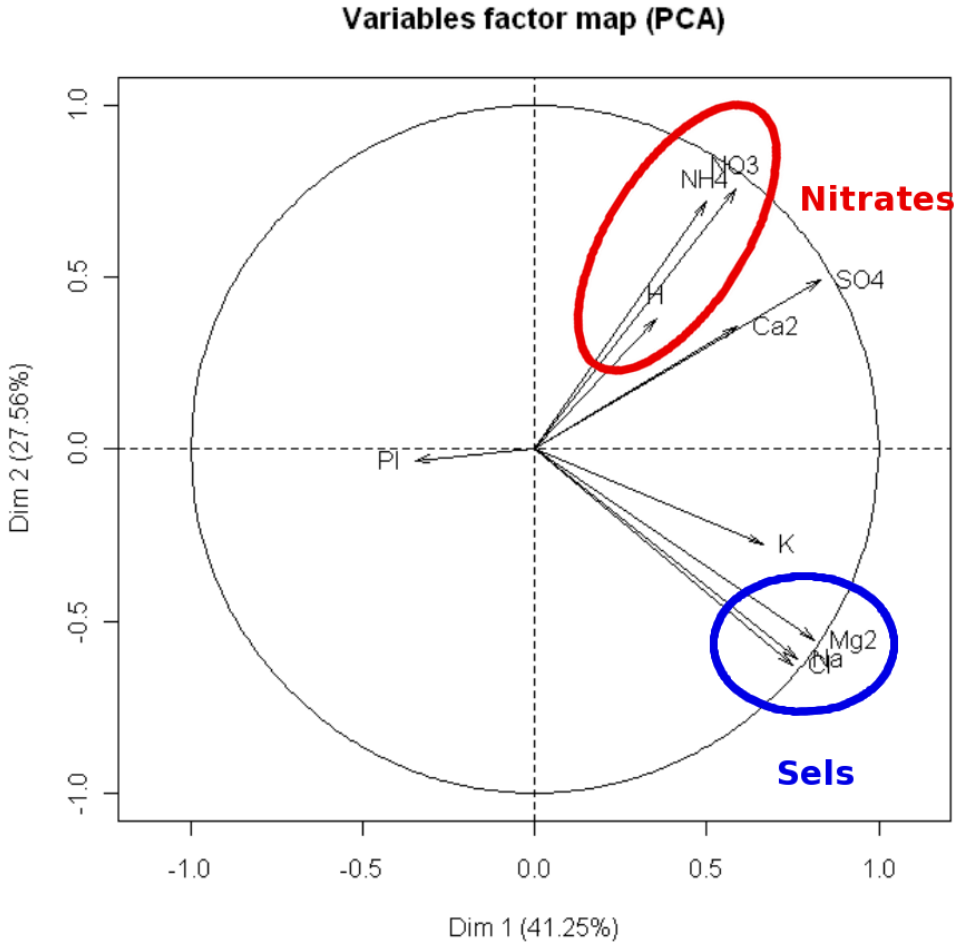
Avec R Commander, il faut ajouter le plugin FactoMineR grâce à Outils\Charger des plug-ins Rcmdr.... Cela ajoute le menu FactoMineR.
On lance une ACP par FactoMineR\Principal Component Analysis (PCA).

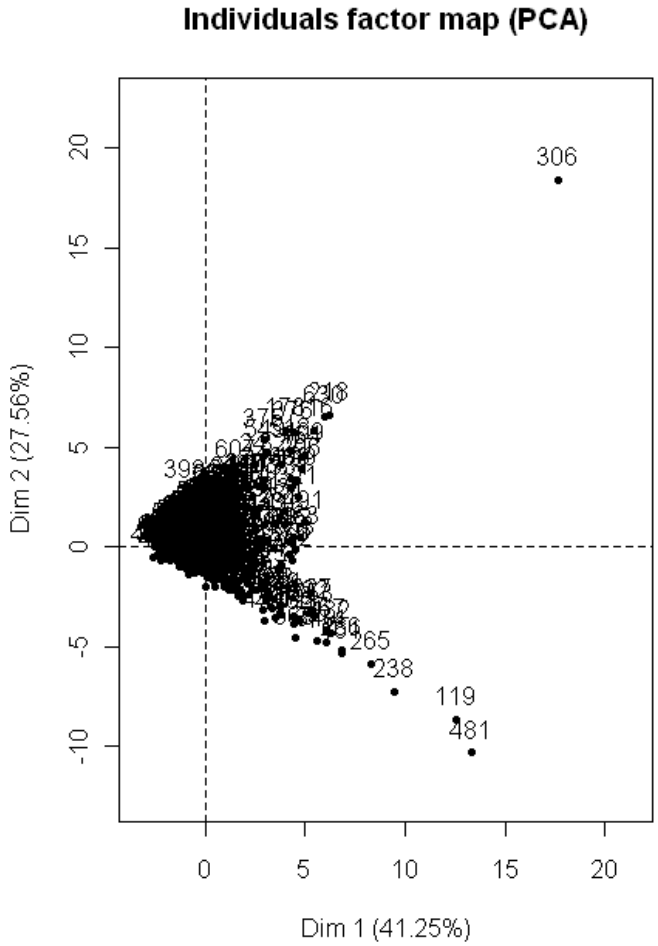
Ex : Les données contiennent la quantité de pluie tombée par jour et sa concentration en différents ions. Seuls les jours de pluie sont conservés pour l'étude.

- On choisit les variables d'intéret pour l'ACP : elles doivent toutes être **quantitatives**.
- Les individus sont ici les jours.
- Seront représentés par défaut : le cercle des corrélations et le nuage de points des individus dans le premier plan factoriel (associé à l'axe 1 et 2).



On obtient le résultat suivant :





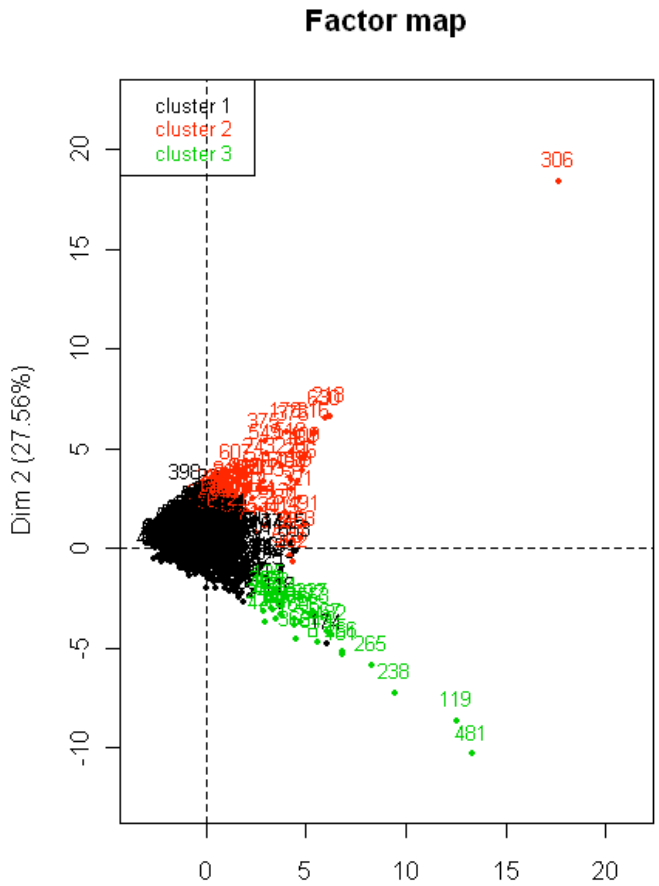
- Ce plan factoriel contient $41.25 + 27.56 = 68.81\%$ de l'information totale.
- Le premier axe met en évidence un effet dilution : les variables "concentration en ions" sont opposées à la quantité de pluie tombée.
  Le deuxième axe oppose 2 groupes d'ions : ceux liés aux sels et ceux liés aux nitrates (et aux roches).
- Le nuage de points des individus se répartit selon l'opposition précédente. Par ex, le jour 306 est un jour à forte concentrations d'ions (axe 1) avec une sur-représentation des ions "nitrates" (axe 2). Le jour 481 est un jour à forte concentrations d'ions (axe 1) avec une sur-représentation des ions "sels" (axe 2).

Compléments dans le menu PCA de FactoMineR

L'ACP ne peut se faire qu'à partir de variables quantitatives.
On peut néanmoins colorier les points selon les modalités d'un facteur en cliquant sur Select supplementary factors puis en choisissant le facteur voulu dans Coloring for individuals après avoir cliqué sur Graphical Options.
Cela peut permettre d'illustrer les groupements de points par ce facteur.

Il est par ailleurs possible de classer les dates dans des groupes homogènes en cliquant sur Perform clustering after PCA.
Pour l'ex précédent on obtient les 3 groupes ci-contre.
Le détail de chaque classe, dont notamment leurs individus représentatifs, peut être visualisé en cochant Print results for clusters.

```
$para
cluster: 1
      661       690       407       323       705
0.5833162 0.6064230 0.6259927 0.6867380 0.7263013
------------------------------------------------------------
cluster: 2
      216       444        24       515        76
1.068403 1.151242 1.302203 1.582687 1.606808
------------------------------------------------------------
cluster: 3
      349       267       365       160       482
0.6437549 0.9038389 1.0282965 1.0357527 1.0380285

$dist
cluster: 1
      398        82       653       336       453
10.871424  9.063105  8.404185  8.147447  7.800155
```



4 Lien graphique entre plusieurs variables quantitatives
- L'ACP
- Compléments de l'ACP : la classification des individus
- Classification de variables

Classification de variables

L'ACP permet de classer naturellement les individus (d'autres méthodes existent).

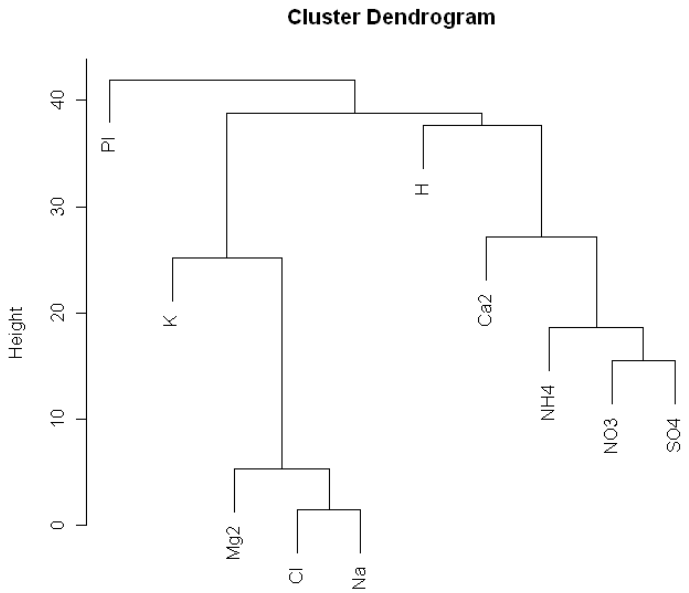
Pour classer des variables (quantitatives), ce n'est possible qu'en ligne de commandes :

- Il faut d'abord centrer et réduire chaque variable avec la fonction scale.
  ex : Pour remplacer les variables du tableau tab par leur version centrée réduite :
  tab=scale(tab)
- On calcule ensuite la distance entre les variables centrées réduites. Pour cela on utilise la fonction dist appliquée au tableau transposé : dist(t(tab))
- Le résultat peut alors être exploité par la fonction hclust qui effectue une classification hiérarchique des variables. Cela donne un arbre que l'on peut représenter avec plot.

En résumé : si tab contient toutes les variables quantitatives que l'on veut classer :

plot(hclust(dist(t(scale(tab)))))

fournira une représentation comme ci-contre (pour l'ex précédent).



Deuxième partie

Liens significatifs, Modélisation

5. La significativité en statistique
6. Les tests statistiques
7. Test de comparaison de 2 moyennes
   - t-test indépendant
   - t-test apparié
8. Généralisation du t-test indépendant : l'ANOVA
9. La régression linéaire
   - Principe général
   - Mise en oeuvre
   - Diagnostics
10. Aspects temporels
    - Séries temporelles et dépendance temporelle
    - Tirer profit de la dépendance
    - Analyser les cycles et la tendance
    - Précautions en présence de dépendance temporelle

5 La significativité en statistique

La significativité statistique

On a rarement accès aux données complètes concernant un phénomène. En général, on dispose d'un échantillon, c'est à dire d'observations partielles, par exemple :

- uniquement certaines périodes de mesures dans l'année ;
- des relevés sur une sous-population (ex : dans quelques batiments pour la pollution intérieure, dans quelques véhicules concernant l'exposition des trajets domicile-travail, etc) ;
- même pour un analyseur fixe : le polluant n'est connu que sur un historique de quelques années.

Cela engendre une **incertitude** sur les analyses conduites et les résultats obtenus.

La significativité statistique

Lorsque l'on observe une propriété sur des données, comment savoir si celle-ci est réellement présente et si elle n'est pas le fruit du hasard dû à l'incertitude de l'étude ? Autrement dit, est-elle significative ?

Par exemple, la propriété observée peut être

- des concentrations similaires d'un polluant sur deux sites
- des moyennes de pollution différentes sur deux sites
- une tendance décroissante de la concentration d'un polluant
- un lien entre la concentration et des causes extérieures (ex : densité du trafic en voiture)

Se poser cette question revient à se demander si en présence d'une nouvelle étude similaire (de nouvelles données), on retrouverait la propriété observée.

En général, on répond à cette question en effectuant un test statistique.

6 Les tests statistiques

Principe d'un test statistique

Un test statistique consiste à confronter deux hypothèses aux données.
On note généralement ces deux hypothèses :

$$H_0 \text{ (hypothèse nulle) et } H_1 \text{ (hypothèse alternative).}$$

ex : H_0 : la concentration moyenne d'un polluant est similaire sur deux sites,
H_1 : elle est significativement différente.

Le logiciel fournit en réponse une **p-value** (ou niveau de significativité).
Cette p-value estime le risque que l'on commet en décidant H_1.

Ainsi, **si la p-value est faible, on conclut H_1, sinon on conclut H_0.**
Le seuil pour la décision est souvent fixé à 5%.

Remarque :
La décision d'un test n'est jamais certaine, elle est toujours entachée d'un risque inhérent à l'incertitude des données initiales.

Ce risque vaut la p-value si on décide H_1. On ne le connait pas en général si on décide H_0 (il devient toutefois négligeable si le nombre de données est grand).

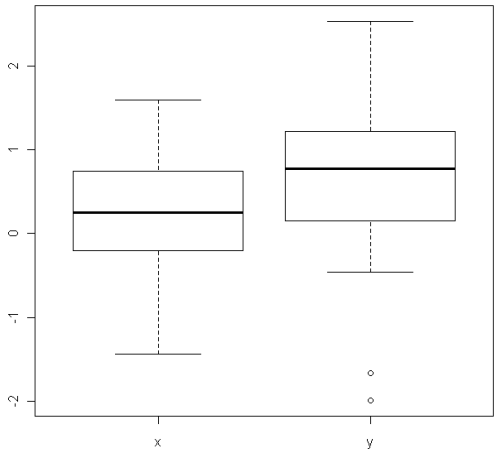
Exemple d'un test d'égalité des moyennes

La significativité ou non de la différence entre deux moyennes dépend essentiellement :

- de la variabilité des données,
- du nombre d'observations.

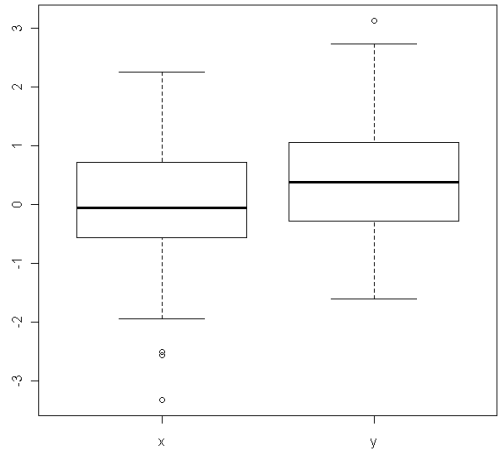
Ex : On teste l'égalité des moyennes entre deux séries indépendantes dans 2 situations. La mise en oeuvre d'un tel test est présentée dans la partie suivante. Les hypothèses testées sont : H_0 : les moyennes sont égales ; H_1 : elles sont différentes.

```
        mean  n
x 0.1687974 15
y 0.5493601 15
```




```
        Welch Two Sample t-test

data:  valeur by type
t = -0.9994, df = 25.162, p-value = 0.3271
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.164541  0.403416
sample estimates:
mean in group x mean in group y
      0.1687974       0.5493601
```

```
        mean   n
x 0.008790623 100
y 0.464863428 100
```

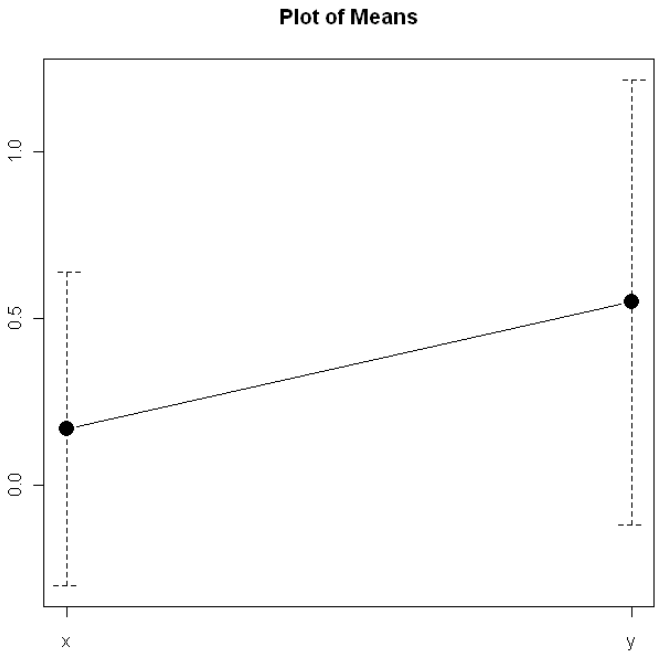


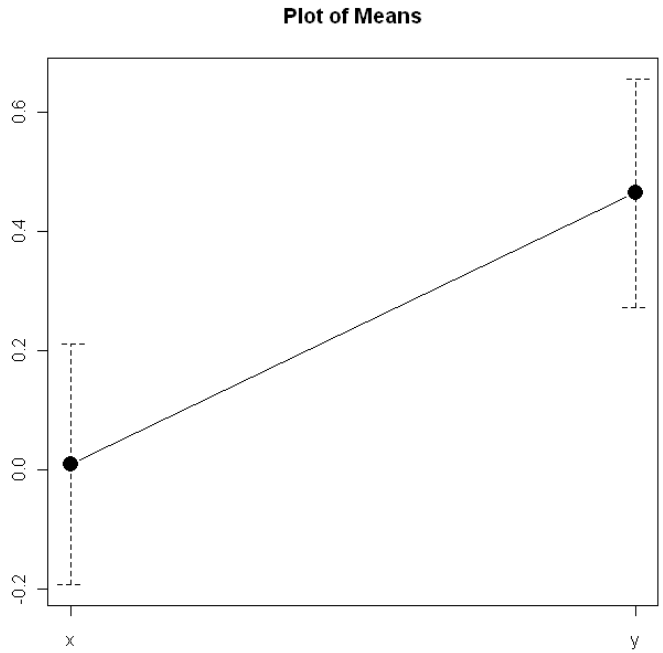

```
        Welch Two Sample t-test

data:  valeur by type
t = -3.2392, df = 197.464, p-value = 0.001407
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7337363 -0.1784093
sample estimates:
mean in group x mean in group y
    0.008790623     0.464863428
```

Dans l'exemple précédent, les boxplots semblent similaires : la variabilité et la différence entre les moyennes sont à peu près les mêmes. Pourtant, au vu de la p-value du test :

- dans le premier cas, on ne conclut pas à une différence significative des moyennes,
- dans le second cas, on conclut à une différence significative des moyennes.

La différence entre les deux situations provient du nombre d'observations.
Dans le second cas, les moyennes sont mieux estimées donc la différence est plus significative.
Une façon de confirmer visuellement ce résultat est de représenter les moyennes estimées avec leur intervalle de confiance dans Graphes\Graphe des moyennes...





Dans le premier cas, l'incertitude de l'estimation ne permet pas de conclure à une différence significative des moyennes.

7 Test de comparaison de 2 moyennes

- t-test indépendant
- t-test apparié

Les différents tests

Il existe deux types de comparaison de moyennes.

Comparaison à partir de 2 échantillons indépendants :

- On dispose de 2 séries de valeurs (pas forcément de taille égale) calculées sur des individus différents
- On désire tester si les moyennes des 2 séries coincident (sans que les valeurs prises une à une soient comparables).

  Ex : la concentration d'un polluant sur deux périodes ; l'exemple précédent.
- Le test s'appuiera pour cela sur la différence des deux moyennes.

Comparaison à partir de 2 échantillons appariés :

- On dispose de 2 séries de valeurs calculées sur les mêmes individus (elles sont donc forcément de même taille).
- On désire tester si les valeurs prises par chaque individu sont significativement différentes sur les deux séries.

  Ex : la concentration d'un polluant mesurée aux mêmes instants sur deux sites.
- Le test s'appuiera pour cela sur la moyenne de ces différences.

7 Test de comparaison de 2 moyennes
- t-test indépendant
- t-test apparié

t-test indépendant

Sous R Commander : Statistiques\Moyennes\t-test indépendant...
On doit disposer d'une variable de type facteur qui prend 2 modalités.
On désire tester si la moyenne d'une variable reste similaire selon ces 2 modalités.
En notant m_1 la moyenne selon la 1ère modalité et m_2 la moyenne selon la seconde, on peut tester :

$H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$ (cas Bilatéral)
$H_0 : m_1 \geq m_2$ contre $H_1 : m_1 < m_2$ (cas Différence < 0)
$H_0 : m_1 \leq m_2$ contre $H_1 : m_1 > m_2$ (cas Différence > 0)

Le menu demande également si les variances sont égales : par défaut, on coche Non ; si on a des bonnes raisons de le faire, Oui rend les résultats plus précis.
(Des tests d'égalité des variances sont disponibles dans Statistiques\Variances)

Remarque : pour créer une variable contenant les valeurs empilées de deux séries et un facteur qui différencie leur type, on peut utiliser Données\Jeu de données actif\Empiler les variables...

En ligne de commandes : si x et y sont les deux séries à comparer, on utilise

```
t.test(x,y,paired=FALSE)
```

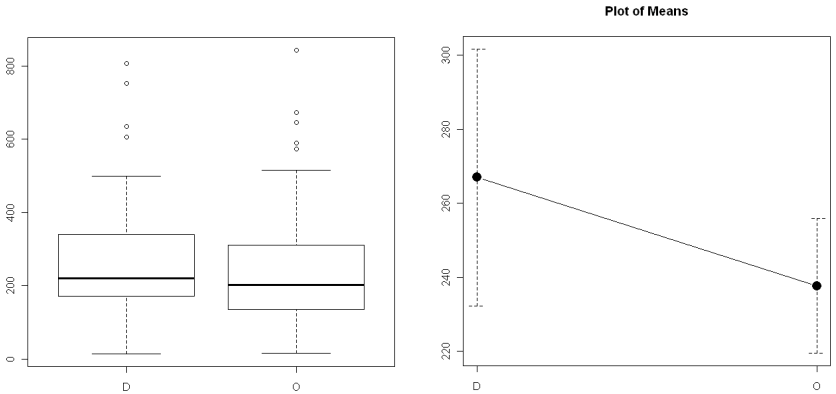
On peut préciser le type de test comme ci-dessus (bilatéral ou non), voir ?t.test

t-test indépendant : un exemple

Pour la série NO2 dans les véhicules vue p17, on teste si pour la modalité "D" (congestionné), le taux moyen de NO2 est différent que pour les autres modalités (toutes regroupées ici en "O").

Notons m_1 la moyenne associée à la modalité "D" et m_2 la moyenne associée à "O".

Le test bilatéral $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$ (à variances non égales) donne :




```
        Welch Two Sample t-test

data:  NO2 by fluiditebis
t = 1.4864, df = 116.702, p-value = 0.1399
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
 -9.748285 68.392661
sample estimates:
mean in group D mean in group O
       266.9988        237.6766
```

Le risque de se tromper en concluant H_1 est de 14% (= la p-value), donc on préfère conclure H_0 : il n'y a pas de différence significative des moyennes.
Le risque associée à cette décision est inconnu : les données étant relativement nombreuses, on peut supposer qu'il est faible.

Remarque : dans le cas du test bilatéral, il y a équivalence entre les deux propriétés suivantes :

- les intervalles de confiance (à 95%) des moyennes ne se recouvrent pas ;
- les moyennes sont significativement différentes (p-value$<$ 5%).

Le graphique des moyennes avec IC (à droite ci-dessus) confirme donc la décision du test.

t-test indépendant : un exemple (suite)

Pour le problème précédent, on peut par ailleurs effectuer les tests suivants :

- $H_0 : m_1 \geq m_2$ contre $H_1 : m_1 < m_2$. On obtient une p-value de 0.93.
  $\Rightarrow$ On conclut donc à $m_1 \geq m_2$ avec un risque inconnu.
- $H_0 : m_1 \leq m_2$ contre $H_1 : m_1 > m_2$. On obtient une p-value de 0.069.
  $\Rightarrow$ On peut conclure à $m_1 > m_2$ avec un risque de 7%.

Les deux conclusions précédentes contredisent la décision du test bilatéral.
La décision la mieux maitrisée est la dernière, car on la prend avec un risque connu : 7%.

Finalement, si ce risque semble acceptable, on peut raisonnablement conclure à $m_1 > m_2$: la moyenne de NO2 en trafic "congestionné" est significativement supérieure aux autres conditions de trafic.

Remarque 1 : Des données supplémentaires confirmeraient probablement la significativité de $m_1 \neq m_2$ dans le cas bilatéral (comme dans l'exemple précédent p55). Ces données supplémentaires renforceraient probablement la décision du dernier test en fournissant une p-value plus petite. Seule une nouvelle étude peut confirmer cela.

Remarque 2 : On peut refaire les tests en cochant Oui dans Variances égales ? car leur égalité, vraisemblable au vu des boxplots, est confirmée par un test. Les résultats confirment notre décision : les moyennes semblent significativement différentes.

7 Test de comparaison de 2 moyennes

- t-test indépendant
- t-test apparié

t-test apparié

Sous R Commander : Statistiques\Moyennes\t-test apparié...

Il suffit de sélectionner les deux séries dont on veut comparer les valeurs.
Notons x la première série et y la seconde série.
Le test porte sur la moyenne des valeurs de $x - y$.
En notant m cette moyenne, on peut tester :
$H_0 : m = 0$ contre $H_1 : m \neq 0$ (cas Bilatéral)
$H_0 : m \geq 0$ contre $H_1 : m < 0$ (cas Différence < 0)
$H_0 : m \leq 0$ contre $H_1 : m > 0$ (cas Différence > 0)

En ligne de commandes : si x et y sont les deux séries appariées à comparer, on utilise

```
t.test(x,y,paired=TRUE)
```

On peut préciser le type de test comme ci-dessus (bilatéral ou non), voir ?t.test

t-test apparié : un exemple

On considère les deux mesures de NO2 en doublon au site "JUS" à Rouen (cf p37).

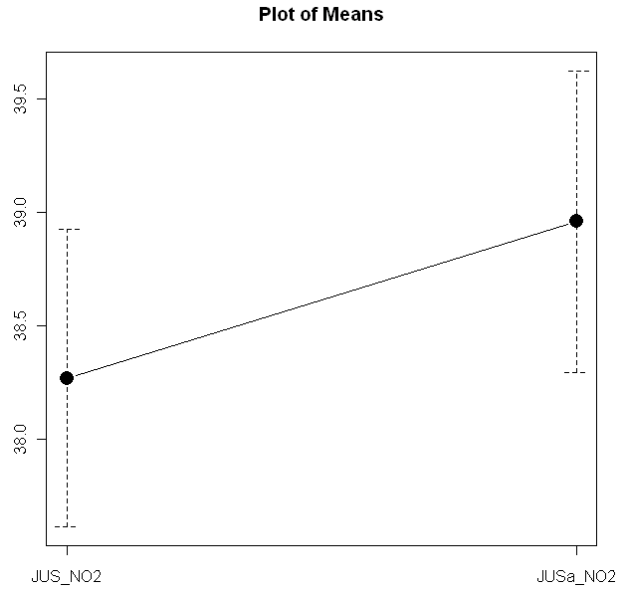
En chaque instant, les deux mesures devraient être similaires (aux fluctuations et erreurs de mesure près). Il est donc naturel de les soumettre à un t-test apparié pour le vérifier :

Au vu de la p-value, on refuse le test d'égalité. On conclut donc qu'en moyenne, la différence des deux mesures n'est pas nulle.

```
        Paired t-test

data:  tab$JUS_NO2 and tab$JUSa_NO2
t = -28.7655, df = 3623, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6958947 -0.6070854
sample estimates:
mean of the differences
             -0.6514901
```

Par contre, l'égalité des deux moyennes par un t-test indépendant est acceptée :

```
        Welch Two Sample t-test

data:  variable by facteur
t = -1.4488, df = 7257.114, p-value = 0.1474
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6208029  0.2431655
sample estimates:
 mean in group JUS_NO2 mean in group JUSa_NO2
              38.26980                38.95862
```

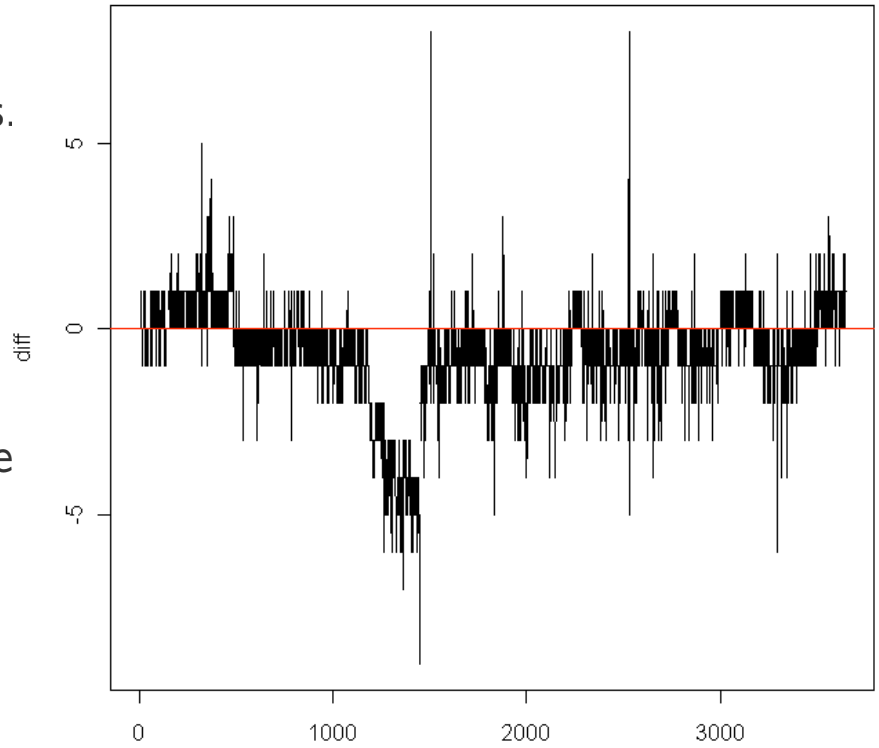


t-test apparié : un exemple (suite)

Ce résultat surprenant se comprend en visualisant la série des différences entre les mesures.

Il y a une succession de biais positifs puis négatifs. Le changement de signe survient lors d'un nouvel étalonnage effectué par le technicien. Ainsi

- La décision H_1 du t-test apparié s'explique : il y a toujours un biais entre les 2 mesures.
- Comme ces biais se compensent sur la durée de la série, l'égalité des moyennes globales par le t-test indépendant est acceptée.



Conclusion :

L'utilisation d'un t-test apparié n'a de sens que si l'on souhaite comparer les valeurs de 2 variables une à une et non leur moyenne globale.

8 Généralisation du t-test indépendant : l'ANOVA

ANOVA à un facteur

L'ANOVA (à un facteur) permet de tester si une variable quantitative admet la même moyenne sur tous les groupes définis par un facteur au **nombre quelconque de modalités**. Le t-test indépendant est donc un cas particulier de l'ANOVA à un facteur lorsque le facteur n'a que 2 modalités.

Sous R Commander : Statistiques\Moyennes\ANOVA à un facteur...
Le test suivant est effectué :

$\mathbf{H_0}$: la moyenne est la même pour toutes les modalités
contre $\mathbf{H_1}$: la moyenne est différente des autres pour au moins une modalité.

La p-value de ce test est retournée ainsi que les moyennes par modalités.

Exemple : Sur les données de NO2 en fonction du type de route :

```
> summary(AnovaModel.1)
             Df  Sum Sq Mean Sq F value    Pr(>F)
type          4  437719  109430   6.059 0.0001088 ***
Residuals   281 5075018   18061
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

> numSummary(tab$NO2 , groups=tab$type, statistics=c("mea
      mean       sd  n
A 215.5343 126.6993 69
P 246.9322 126.2990 78
T 348.7938 209.7387 33
U 233.7392 115.6702 75
V 224.9077 109.1470 31
```

La p-value est très faible donc on conclut à $\mathbf{H_1}$ en prenant un risque très faible.

Légende :
"A" : Autoroute, "P" : Périurbain, "T" : Tunnel, "U" : Urbain, "V" : Voie rapide urbaine

ANOVA à un facteur (suite)

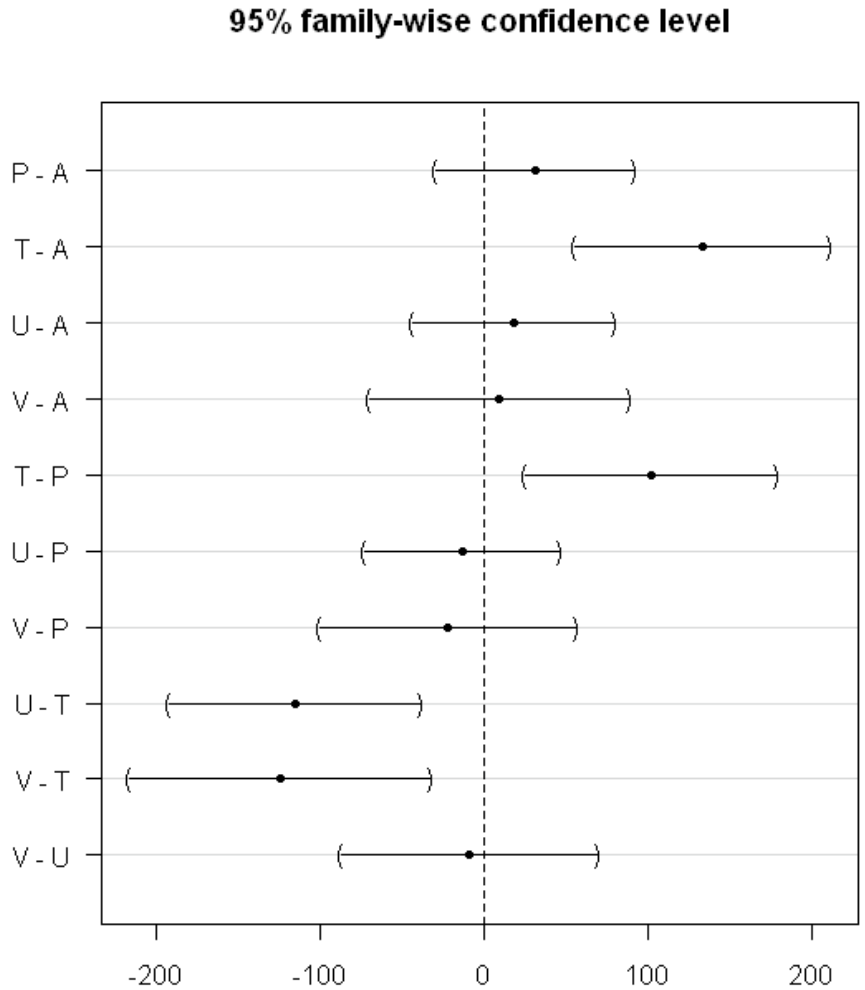
Pour en savoir plus sur le détail des moyennes dans chaque modalité et analyser leurs différences, on peut cocher Comparaison multiple de moyennes.

Exemple (suite) :

La différence des moyennes entre chaque paire de modalités est représentée avec son intervalle de confiance : s'il contient 0, les 2 moyennes ne sont pas significativement différentes, sinon elles le sont (c'est équivalent au t-test bilatéral).

Ici, le facteur type semble significatif dans l'ANOVA uniquement à cause de la modalité "T" (tunnel).

On pourrait effectuer une ANOVA sur la sous-table composée des modalités autres que "T" pour confirmer la non-significativité du facteur type sans "T".



ANOVA à un facteur : compléments

Une fois une ANOVA effectuée, les résultats sont stockés dans le modèle dont le nom apparaît à droite du bouton Visualiser.
On peut analyser plus en détails le modèle grâce au menu Modèles. La démarche est similaire à l'analyse des modèles de régression (voir la partie suivante).

En ligne de commandes : L'ANOVA se fait à l'aide de la fonction aov puis summary.
On peut également utiliser la fonction lm (voir partie suivante) : l'ANOVA est en effet un cas particulier de régression linéaire.

Attention :
Le facteur doit être vu comme un facteur. A défaut : Données\Gérer les variables dans le jeu de données actif\Convertir des variables numériques en facteurs... ou as.factor().
L'estimation des effets de chaque modalité d'un facteur dans l'ANOVA (visible dans Modèles\Intervalles de confiance... ou avec lm) dépend du codage du facteur.
On peut vérifier le codage utilisé avec model.matrix(nomdumodele). Le codage du facteur peut être changé dans Données\Gérer les variables dans le jeu de données actif\Définir les contrastes pour un facteur. Le plus naturel est le choix "dummy". Dans ce cas :

- le coefficient moyen correspond à l'effet de la première modalité,
- les autres coefficients correspondent aux effets respectifs des autres modalités par rapport à la première.

On peut changer l'ordre des modalités d'un facteur dans Données\Gérer les variables dans le jeu de données actif\Réordonner une variable facteur.

Principe général

Une **régression** permet de quantifier à l'aide d'une formule la relation moyenne entre une variable quantitative (dite variable réponse) et d'autres variables (dites variables explicatives).

Exemple : $Ozone \approx fonction(Température)$?
Si oui, quelle est cette fonction ?

Régression linéaire : cas particulier où la fonction est supposée linéaire.

Exemple : $Ozone \approx \beta_0 + \beta_1 Température$?
Si oui, que valent les paramètres β_0 et β_1 ?

Le "$\approx$" précédent doit se lire "$=$ en moyenne". Le modèle précis auquel on s'intéresse est en fait :

$$Ozone = \beta_0 + \beta_1 Température + \epsilon$$

où ϵ est un terme d'erreurs de moyenne nulle, contenant les erreurs de mesures, l'incertitude des données, l'erreur de modélisation. Lors de la mise en oeuvre d'un modèle de régression linéaire, on espère que ce bruit ϵ soit le plus "petit" possible.

On peut considérer **plusieurs** variables explicatives, **quantitatives** ou de type **facteur** (bien vérifier dans ce cas que la variable est vue comme un facteur, cf p70) :

Exemple : $Ozone = \beta_0 + \beta_1 Température + \beta_2\ Présence\_pluie + \epsilon$
où la variable $Présence\_pluie$ vaut 1 s'il a plu, 0 sinon.

Principe général

A quoi sert une régression linéaire ?

- à confirmer le lien observé graphiquement (cf parties 3 et 4) entre la variable réponse et les variables explicatives ;
- à comprendre l'impact marginal d'une variable explicative sur la variable réponse ;

  Ex : d'après le modèle précédent, la hausse de 1 degré de la température engendre une hausse de la concentration en ozone de β_1 (ou une baisse selon le signe de β_1)
- à prédire certaines situations

  Ex : d'après le modèle précédent, à température fixée un jour sec, on en déduit la concentration en ozone.

Pourquoi se limiter à la régression linéaire ?
L'estimation de quelques paramètres (β_0, β_1, etc) est plus aisée que l'estimation d'une fonction complète inconnue. De plus, l'interprétation du modèle est simple (voir ci-dessus).
$\Rightarrow$ Dans un problème de régression, on se ramène si possible à une régression linéaire. Il convient toutefois de bien vérifier que le lien linéaire est légitime.

Comment juger de la qualité du modèle ? Les points suivants seront à considérer :

- les variables explicatives choisies sont-elles pertinentes ?
- le lien avec la variable réponse est-il linéaire ?
- le terme d'erreur ϵ se comporte-t-il de façon "raisonnable" ?

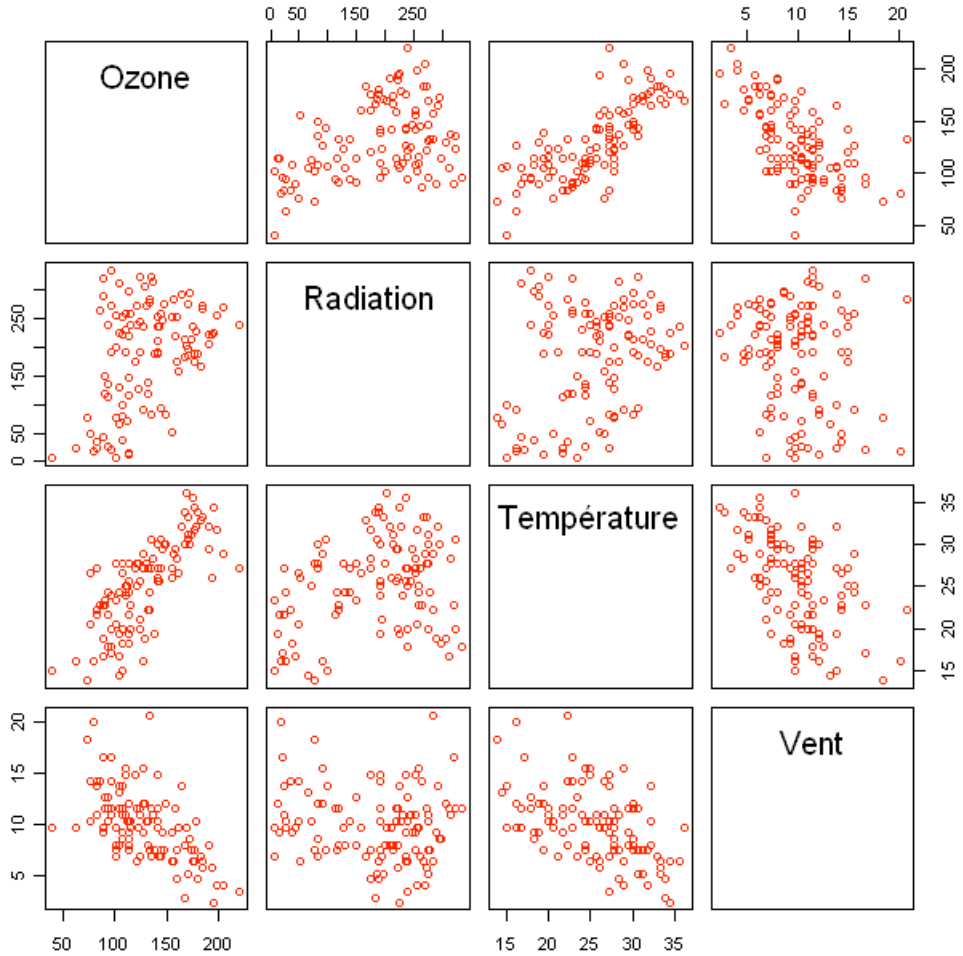
Mise en oeuvre sur un exemple

Exemple : le jeu de données air contient des mesures journalières d'ozone, de température, de radiation solaire et de vitesse du vent. On désire expliquer la concentration d'ozone.

On étudie dans un premier temps les relations entre l'ozone et les autres variables (cf partie 3) :

- par les nuages de points ci-contre
- par le calcul des corrélations :

	Ozone	Radiation	Température	Vent
Ozone	1.0000000	0.4212425	0.7521317	-0.5996008
Radiation	0.4212425	1.0000000	0.2943825	-0.1273656
Température	0.7521317	0.2943825	1.0000000	-0.4971196
Vent	-0.5996008	-0.1273656	-0.4971196	1.0000000



Il existe donc un lien entre l'ozone et les autres variables. Au vu des nuages de points, ce lien peut-être considéré **linéaire** (les points des trois nuages de gauche semblent alignés le long d'une droite).

Mise en oeuvre sur un exemple

Sous R Commander : Statistiques\Ajustement de modèles\Régression linéaire...
ou Statistiques\Ajustement de modèles\Modèle linéaire...

En commandes : fonction lm, voir son utilisation dans les sorties de R Commander.

Dans l'exemple précédent : si on choisit en variable réponse ozone et en variables explicatives les 3 autres variables, on obtient le résumé suivant :

- Call : rappelle le modèle proposé.
- Residuals : résume les valeurs des résidus, c'est à dire du ϵ
- Coefficients : donne une estimation des valeurs des coefficients β_0, β_1, β_2, β_3 devant chaque variable explicative, ainsi que le résultat du test de Student (voir p78)
- des statistiques sur la qualité globale du modèle sont enfin données : le R^2, le R^2 ajusté, le résultat du test de Fisher (voir p78)

```
Call:
lm(formula = Ozone ~ Radiation + Température + Vent, data = a

Residuals:
    Min      1Q  Median      3Q     Max
-44.538 -15.002  -1.305  13.604  59.321

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.55226   15.07531   3.486 0.000712 ***
Radiation    0.08786    0.02235   3.931 0.000150 ***
Température  3.58809    0.43988   8.157 7.13e-13 ***
Vent        -3.05278    0.63041  -4.843 4.35e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 

Residual standard error: 20.41 on 107 degrees of freedom
Multiple R-squared: 0.6796, Adjusted R-squared: 0.6707
F-statistic: 75.66 on 3 and 107 DF,  p-value: < 2.2e-16
```

Dans cet exemple la formule de régression estimée serait :

$$Ozone \approx 52.55 + 0.09\ Radiation + 3.59\ Température - 3.05\ Vent$$

Diagnostics

Il existe de nombreux outils pour évaluer la qualité d'un modèle et sa sensibilité aux données. Certains sont disponibles dans le menu Modèles. Il servent à évaluer le modèle sélectionné à droite du bouton Visualiser.
Tous ces outils ne sont pas détaillés ici. Nous renvoyons aux ouvrages spécialisés (voir les références). Néanmoins, les éléments principaux à considérer sont les suivants :

- Les variables sont-elles pertinentes dans le modèle ? Cela revient à savoir si le coefficient β devant une variable est significativement non-nul. On effectue pour cela le **test de Student** $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$. Sa p-value est donnée pour chaque variable dans le paragraphe Coefficients de la sortie du modèle.
- Le modèle est-il globalement bien explicatif ? Cela revient à savoir si le bruit du modèle ϵ est "petit". Le **coefficient de détermination R^2** répond à cette question : il calcule la part de variabilité de la variable réponse reproduite par le modèle (sans le bruit). Le R^2 est toujours compris entre 0 et 1. Plus R^2 est proche de 1, plus le modèle est explicatif. Mais le R^2 a le défaut d'augmenter avec le nombre de variables même si celles-ci ne sont pas pertinentes. Le **R^2 ajusté** corrige ce défaut.
- Les liens sont-ils vraiment linéaires ? Cela se voit dans les nuages de points initiaux. Si ce n'est pas le cas, une transformation des variables peut être envisagée.

D'autres éléments sont à inspecter comme le comportement précis des résidus, l'influence des points extrêmes, etc. Ils peuvent conduire à reconsidérer la modélisation.

Diagnostics : exemple

Suite de l'exemple précédent :

- Dans la sortie, on observe que pour chaque variable explicative (y compris la constante "(intercept)") la p-value associée au test de Student est très faible. On conclut donc que chacun des coefficients est significativement non-nul et donc que les variables explicatives sont pertinentes.
- Par ailleurs, l'écart-type des résidus vaut 20.41. Cela n'est pas informatif en soi, il faut le comparer à l'écart-type "initial" de la variable réponse pour savoir si les résidus sont "petits" ou non. C'est l'objet du R^2, ou mieux du R^2 ajusté. Ici il vaut $R^2_{ajusted} = 67\%$. Il y a donc 67% de la variabilité de l'ozone qui est expliquée par le modèle estimé.
- Enfin le test de Fisher est effectué. Il teste

  H_0 : aucune variable n'est significative (tous les coefficients sont nuls)

  contre H_1 : au moins un coefficient est non nul.

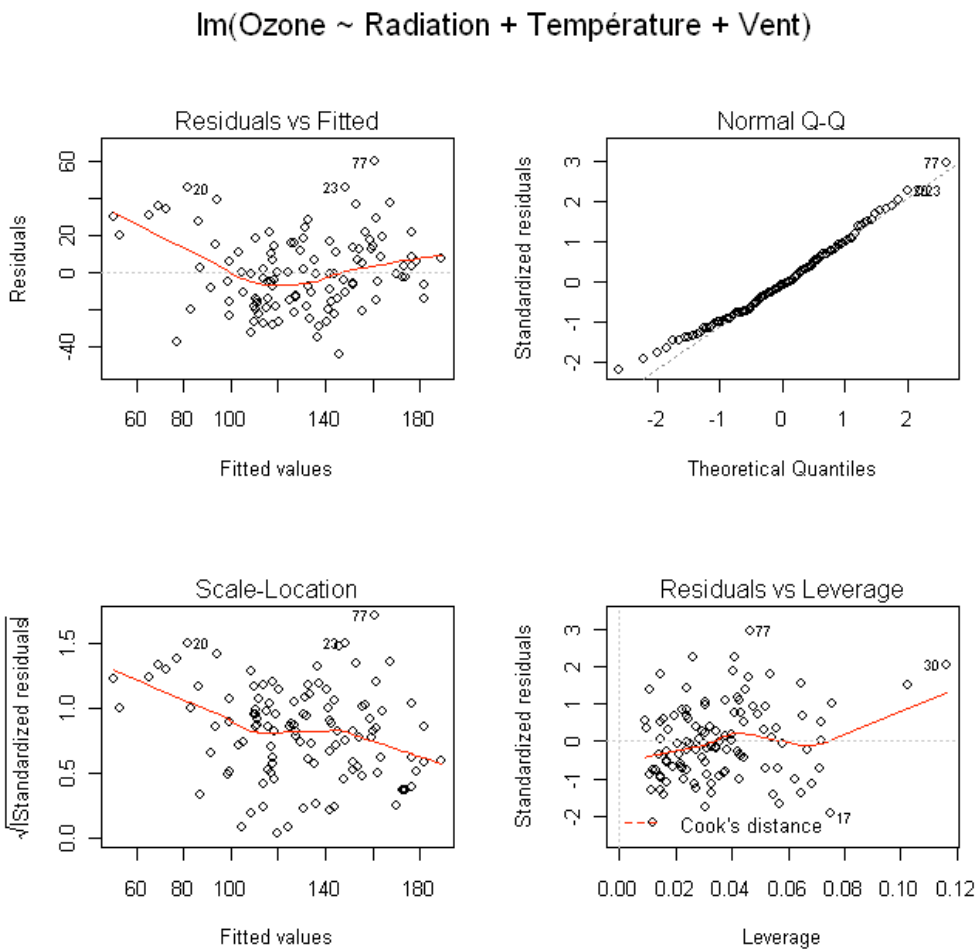
  Sa p-value est négligeable donc on rejette H_0, ce qui n'est pas surprenant vu les résultats des tests de Student précédents.

Diagnostics : analyse des résidus, exemple

Suite de l'exemple précédent :

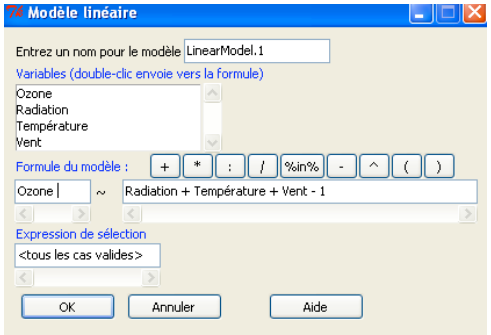
Modèles\Graphes\Diagnostics graphiques propose une analyse rapide des résidus.

- Les deux graphes de gauche représentent les résidus (normalisés pour le graphe du bas) de chaque observation.
  Si le modèle est correctement spécifié, les résidus doivent être à peu près distribués de façon équivalente autour de 0 tout au long de l'axe des abscisses. C'est le cas ici.
- Le graphe en bas à droite représente "l'effet levier" (leverage) de chaque observation. Cet effet mesure l'influence d'un point dans l'estimation du modèle. On repère les observations influentes à droite du graphe.
- Dans le graphe en haut à droite, si les points sont alignés sur la diagonale (comme ici), on accepte la normalité des résidus. Cette propriété n'est pas indispensable si le jeu de données contient suffisamment d'observations.



Compléments

- On peut enlever la constante (intercept) d'un modèle, c'est à dire l'imposer comme étant nulle. Pour cela il faut ajouter "-1" à l'équation du modèle.
  Attention : dans ce cas, le R^2 n'a plus de sens.



- Il faut éviter de mettre deux variables explicatives fortement corrélées dans le même modèle . Cela constitue une redondance d'informations qui perturbe l'estimation du modèle. Pour vérifier si ce problème a lieu dans le modèle estimé Modèles\Diagnostics numériques\Inflation de variance des facteurs calcule le "vif" pour chaque variable explicative. Si le "vif" est grand (> 10) cela témoigne d'un problème de redondance entre deux variables : il faut alors supprimer une des deux variables du modèle.
- Il est possible de choisir automatiquement le meilleur sous-modèle du dernier modèle estimé dans Modèles\Sous-ensemble....
  Chaque ligne correspond à un modèle testé. Chaque modèle repose sur les variables grisées sur la ligne. Le modèle retenu est celui du haut.
  Attention : la sélection n'est faite que selon un seul critère (par ex le BIC). Aucune autre analyse n'est prise en compte (analyse des résidus, "vif", etc.).
- Si LinearModel.1 est le nom du modèle estimé, LinearModel.1$fitted.values contient les valeurs du modèle estimé pour chaque observation, LinearModel.1$residuals est le vecteur des résidus (uniquement en ligne de commandes).

10 Aspects temporels

- Séries temporelles et dépendance temporelle
- Tirer profit de la dépendance
- Analyser les cycles et la tendance
- Précautions en présence de dépendance temporelle

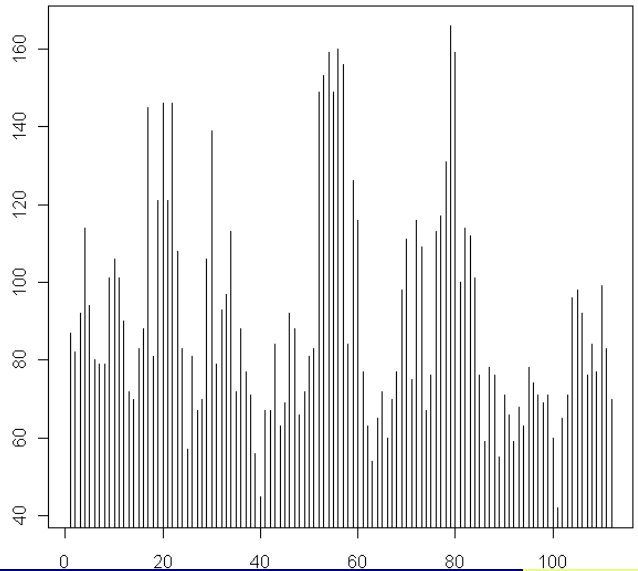
Représentation en temporel

Lorsque les données sont temporelles, par exemple pour des mesures de concentration d'un polluant effectuées à intervalles de temps réguliers, il est pratique de les représenter sous forme de courbe :

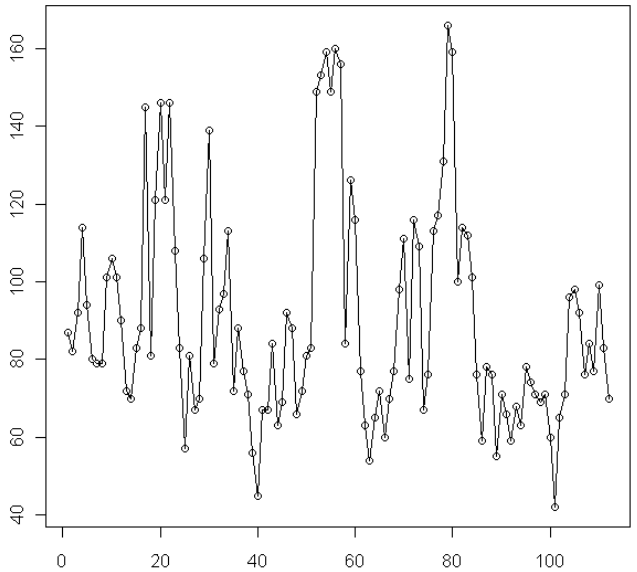
- facile sous Excel
- peu pratique avec R Commander, éventuellement Graphes\Graphe indexé...
- en ligne de commandes sous R : pour représenter une série x : plot(x,type='l')
  pour ajouter les points à la courbe : points(x)

Exemple : Ozone max journalier à Rennes durant l'été 2001 (issu de "Statistiques avec R")

Avec Graphes\Graphe indexé... :



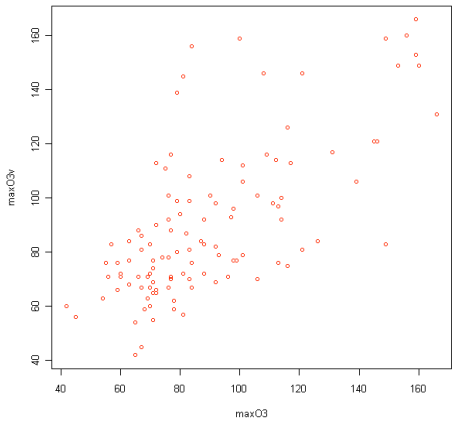
En ligne de commandes :



La dépendance temporelle

Dans une série temporelle, il est courant d'observer une dépendance entre les différentes valeurs de la série, notamment entre les valeurs voisines.

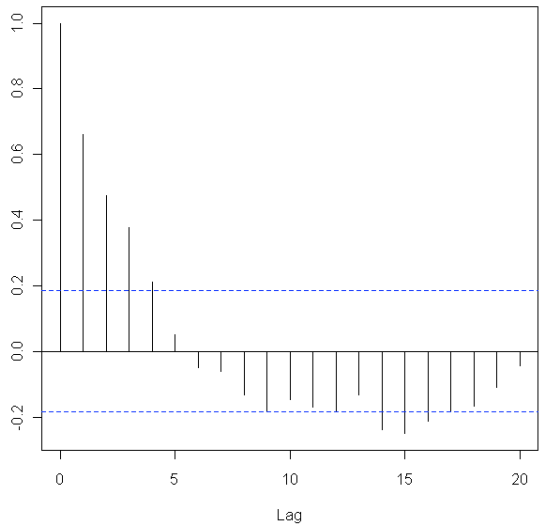
Exemple : pour la série précédente des max d'ozone, on construit la série des valeurs de la veille. Le nuage de points entre la série initiale et la série des valeurs de la veille est donné ci-contre. La corrélation entre les deux séries vaut $r = 0.68$
La série est donc corrélée positivement à son passé immédiat.



Compléments :
acf(x,na.action=na.pass) calcule les corrélations de la série x avec ses passés à tout ordre.

Exemple : les ACF de la série précédente sont données ci-contre.
Chaque baton représente la corrélation de la série avec son passé d'ordre "Lag" :
Lag=0 corrélation de la série avec elle-même (elle vaut toujours 1).
Lag=1 corrélation de la série avec son passé immédiat (on retrouve $r = 0.68$).
Lag=2 corrélation de la série avec la série des valeurs de l'avant-veille.
Lag=... etc.

Rem : Les batons entre les pointillés peuvent être considérés nuls.



Inclure le passé dans la régression

Une régression peut être améliorée si on tient compte des dépendances temporelles.

Exemple : On régresse les pics d'ozone précédents sur la température à 12h, la nébulosité à 9h et la vitesse du vent à 9h. A droite, on a ajouté au modèle les pics d'ozone de la veille.

```
Call:
lm(formula = maxO3 ~ Ne9 + T12 + Vx9, data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-33.7679  -9.6841  -0.4943   8.1700  47.5676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.0584    12.7275   0.947  0.34553
Ne9          -1.9703     0.7751  -2.542  0.01245 *
T12           4.1938     0.4781   8.772 2.84e-14 ***
Vx9           1.9130     0.6868   2.785  0.00632 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.2 on 108 degrees of freedom
Multiple R-squared: 0.6787, Adjusted R-squared: 0.6698
F-statistic: 76.05 on 3 and 108 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = maxO3 ~ maxO3v + Ne9 + T12 + Vx9, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
maxO3v       0.35483    0.05789   6.130 1.50e-08 ***
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
T12          2.76409    0.47450   5.825 6.07e-08 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 107 degrees of freedom
Multiple R-squared: 0.7622, Adjusted R-squared: 0.7533
F-statistic: 85.75 on 4 and 107 DF,  p-value: < 2.2e-16
```

Modèles\Diagnostics numériques\Test Durbin Watson d'autocorrélation...permet de tester si les résidus d'un modèle sont non-corrélés (H_0). Si H_0 est refusée, le modèle ne tient pas assez compte de la dépendance initiale et peut donc être amélioré. Pour les modèles ci-dessus :

```
        Durbin-Watson test

data:  maxO3 ~ Ne9 + T12 + Vx9
DW = 1.3153, p-value = 0.0001684
alternative hypothesis: true autocorelation is not 0
```

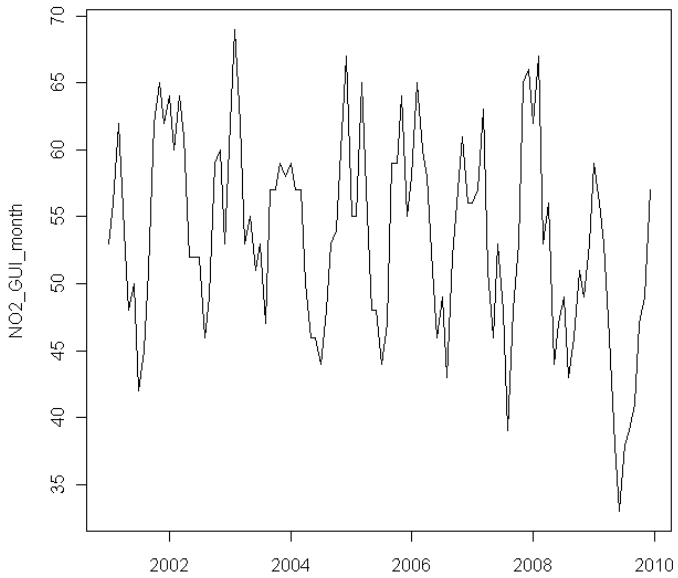
```
        Durbin-Watson test

data:  maxO3 ~ maxO3v + Ne9 + T12 + Vx9
DW = 1.9439, p-value = 0.6682
alternative hypothesis: true autocorelation is not 0
```

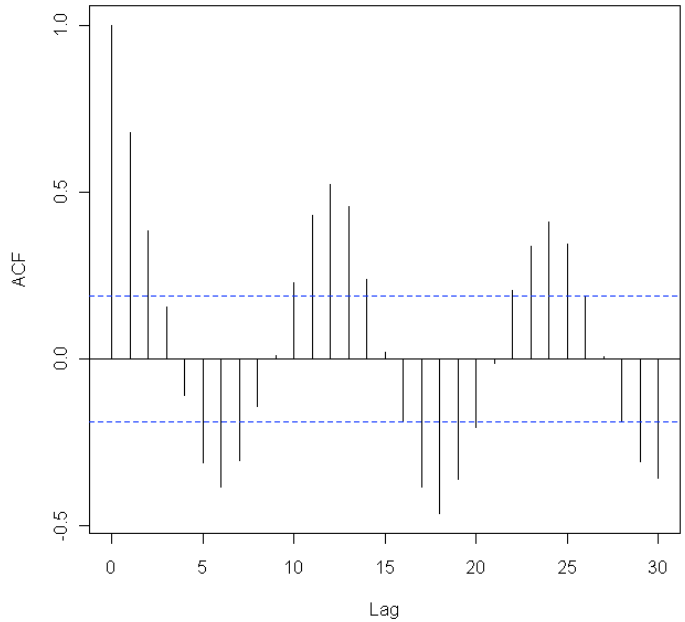
Mettre en valeur les cycles

Les séries horaires ont généralement une période (ou un cycle) de 24h.
Les séries mensuelles ont généralement une période de 12 mois.

On peut confirmer cette périodicité en visualisant les ACF : ils héritent de la période de la série.

Exemple : NO2 mesuré à la station GUI (centre ville de Rouen)





Mettre en valeur les cycles

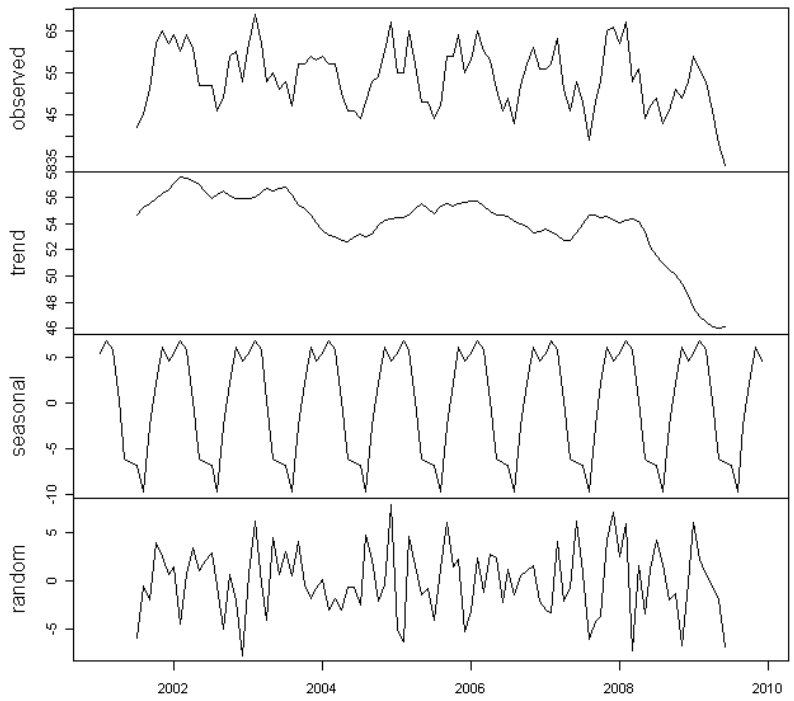
Il est pratique de définir une série au format ts ("time series") en précisant sa période et sa date de départ. Ex : x=ts(NO2_GUI_month, freq=12, start=2001).

On peut ainsi décomposer la série selon sa tendance, sa composante cyclique et un reste, grâce à la fonction decompose ou stl .

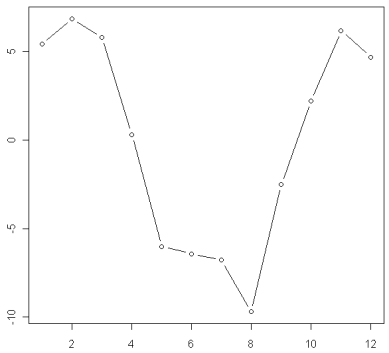
Ex : pour la série ci-dessus plot(decompose(x)) ou plot(stl(x, s.window='periodic'))

La série initiale ("observed" en haut) est la somme :

- d'une tendance ("trend"),
- d'une composante cyclique ("seasonal"),
- d'un reste ("random").



Il est possible de visualiser le profil typique d'un cycle avec plot(decompose(x)$figure)



Détecter une tendance décroissante (ou croissante)

Si la série est annuelle (1 valeur par an) : deux possibilités

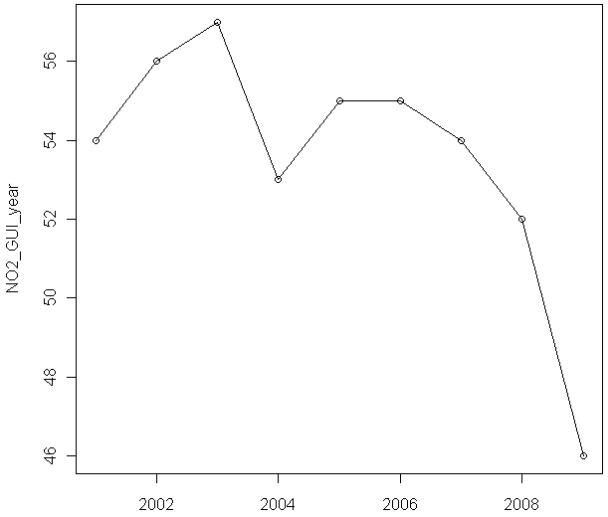
- **test de MannKendall** : H_0 : la série ne présente pas de tendance monotone,
  contre H_1 : la série présente une tendance monotone.
  Sous R : fonction MannKendall dans la librairie Kendall.
- A partir de la **corrélation de Spearman** entre la série et les numéros d'observations, on teste : H_0 : la corrélation est nulle, contre H_1 : elle est non nulle.
  Sous R : cor.test(x,1:length(x),method="spearman",use="complete.obs")
  Sous R Commander : Statistiques\Résumés\Test de corrélation... : choisir la série d'intérêt et la série des numéros d'observation et cocher Coefficient de Spearman.

Si la série est mensuelle (1 valeur par mois) : deux possibilités

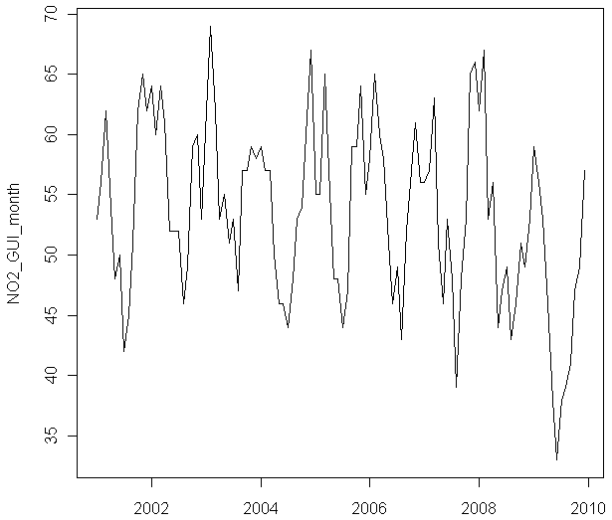
- **test de MannKendall saisonnier**.
  Sous R : la série x doit être au format ts ("time series") en précisant sa période : x=ts(x,freq=12), puis SeasonalMannKendall(x) dans la librairie Kendall.
- Construire la **série corrigée des variations saisonnières** puis appliquer les tests classiques (MannKendall et Spearman) :
  Sous R : x=ts(x,freq=12) comme ci-dessus. La série corrigée des variations saisonnières est alors : x_cvs=decompose(x)$trend+decompose(x)$random.

Détecter une tendance : un exemple

NO2 mesuré à la station GUI (centre ville de Rouen) par ans puis par mois :




```
> library(Kendall)
> MannKendall(x)
tau = -0.514, 2-sided pvalue =0.07314
> cor.test(x,1:length(x),method="spearman",use="complete.obs")

        Spearman's rank correlation rho

data:  x and 1:length(x)
S = 195.6329, p-value = 0.06883
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.6302744
```




```
> library(Kendall)
> x=ts(x,freq=12,start=2001)
> SeasonalMannKendall(x)
tau = -0.313, 2-sided pvalue =6.0921e-05
> x_cvs=decompose(x)$trend+decompose(x)$random
> MannKendall(x_cvs)
tau = -0.29, 2-sided pvalue =2.9354e-05
> cor.test(x_cvs,1:length(x_cvs),method="spearman",use="complete.obs")

        Spearman's rank correlation rho

data:  x_cvs and 1:length(x_cvs)
S = 206657, p-value = 4.996e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4016346
```

$\Longrightarrow$ Mieux vaut privilégier l'étude des séries mensuelles qui contiennent plus de données.

Estimer la tendance par une droite

Les tests précédents testent la présence d'une tendance, **éventuellement non-linéaire**.

On peut néanmoins souhaiter ajuster une droite linéaire à cette tendance.

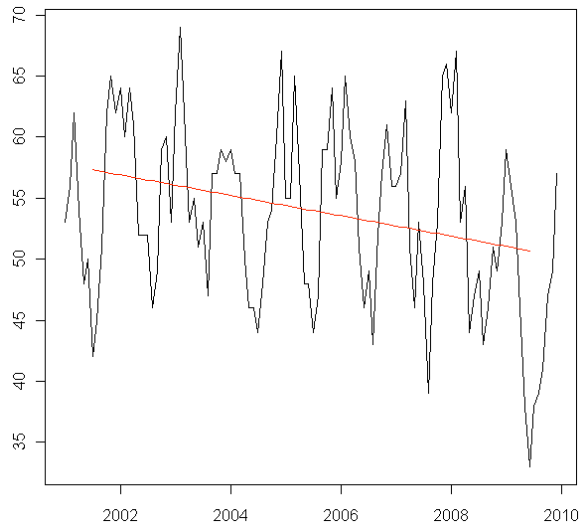
Pour cela, il suffit de faire une régression linéaire de la série corrigée des variations saisonnières sur la série des numéros d'observations (ou des dates). Dans l'ordre :

1. On enlève les valeurs manquantes à x_cvs
2. On récupère la variable des dates (temps)
3. On effectue la régression
4. On superpose le tracé de la série et de la droite estimée.

```
> x_cvs2=na.omit(x_cvs)
> temps=as.numeric(time(x_cvs2))
> reg=lm(x_cvs2~temps)
> plot(x)
> lines(temps,reg$fitted.values,col=2)
```

La représentation est donnée ci-contre.

Il est possible d'ajouter des bornes de confiance à cette droite. On les obtient par
pred=predict(reg,interval='confidence')
puis matplot(temps,pred,type='l',col=2,add=T)
pour les ajouter au graphe ci-contre.



Dépendance et t-tests

Conséquence de la dépendance pour les t-tests :
Les t-tests sous R (cf partie 7) ne tiennent pas compte de la dépendance. En présence d'une dépendance temporelle positive (i.e. la somme des batons dans les ACF est positive), **la p-value des t-tests est en réalité supérieure à celle annoncée** par R.
De même, pour les tests de Mann Kendall et de Spearman programmés sous R, les p-value dans les cas "positivement dépendants" sont en réalité supérieures à celles annoncées.

Gestion des valeurs manquantes :
Il ne faut pas supprimer les valeurs manquantes dans les séries temporelles, cela briserait la structure temporelle de la série.
La plupart des fonctions ont des options qui permettent de gérer les valeurs manquantes (comme mean, cor).

D'autres nécessitent la totalité des valeurs de la série (comme decompose) : on peut dans ce cas remplacer les valeurs manquantes par une interpolation de leurs valeurs voisines avec la fonction na.approx de la librairie zoo.

Références

- Films Tutoriaux en ligne (avec R Commander) :
  www.screencast.com/users/agrocampus/folders/FilmR
- Poly sur l'utilisation de R : www.math.sciences.univ-nantes.fr/∼philippe/R_freeware_files/Anne-Philippe-cours-R.pdf
- Statistiques générales avec R :
  "Statistiques avec R" de PA. Cornillon, A. Guyader, F. Husson, N. Jégou, J. Josse, M. Kloareg, E. Matzner-Lober, L. Rouvière.
- Langage R, Statistiques avec R :
  "Le logiciel R" de P. Lafaye de Micheaux, R. Drouilhet, B. Liquet.
- Analyse de données (ACP, etc.), exemples avec R :
  "Analyse de données avec R" de F. Husson, S. Lê et J. Pagès.
- Poly sur la régression et l'ANOVA avec R :
  "Practical Regression and Anova using R" de J.J. Faraway.
  cran.r-project.org/doc/contrib/Faraway-PRA.pdf
- Régression, exemples avec R :
  "Régression. Théorie et applications." de PA. Cornillon et E. Matzner-Lober.