Statistiques descriptives

L3 Maths-Eco
Université de Nantes

Frédéric Lavancier

Le vocabulaire est issu de la démographie, domaine d'application initial des statistiques.

Population : ensemble étudié
Exemples : la population française, l'ensemble des entreprises d'une région, un ensemble de sites géographiques, un ensemble de dates,...

Individus : les éléments composant la population
Exemples : une personne, une entreprise, un site géographique, une date,...

Variables : les caractéristiques observées sur chaque individu
Exemples :
- *pour des personnes : sexe, CSP, âge, salaire,...*
- *pour des entreprises : nombre de salariés, secteur d'activité,...*
- *pour des sites géographiques : altitude, type de végétation,...*
- *pour des dates : cours d'une action, température, ventes journalières,...*

Lorsqu'on on observe une seule variable au cours du temps (par exemple la température journalière), on parle d'une **série temporelle**.

Un **jeu de données** est composé de l'observation des variables sur les individus issus de la population. Il se présente généralement sous forme d'un tableau dont les lignes sont les individus et les colonnes les variables.

Exemple de jeu de données : Cours d'indices boursiers nationaux.

	DAX	SMI	CAC	FTSE
02/01/91	1628.75	1678.1	1772.8	2443.6
03/01/91	1613.63	1688.5	1750.5	2460.2
04/01/91	1606.51	1678.6	1718.0	2448.2
07/01/91	1621.04	1684.1	1708.1	2470.4
08/01/91	1618.16	1686.6	1723.1	2484.7
09/01/91	1610.61	1671.6	1714.3	2466.8
⋮	⋮	⋮	⋮	⋮
24/12/98	5598.32	7952.9	4041.9	5680.4
27/12/98	5460.43	7721.3	3939.5	5587.6
28/12/98	5285.78	7447.9	3846.0	5432.8
29/12/98	5386.94	7607.5	3945.7	5462.2
30/12/98	5355.03	7552.6	3951.7	5399.5
31/12/98	5473.72	7676.3	3995.0	5455.0

Les individus : les 1860 jours ouvrés du 01/01/1991 au 31/12/1998
Les variables : les indices allemands (DAX), suisses (SMI), français (CAC) et anglais (FTSE).

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

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

Les individus : les poteries numérotées de 1 à 13
Les variables : le site archéologique et différents composés chimiques.

Type de variables

Deux grandes catégories déclinées en deux types.

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

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

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

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

Que fait-on à partir d'un jeu de données?

Quelques exemples :

- Le "reporting" : résumer de façon efficace un jeu de données afin de rendre l'information lisible facilement et rapidement.
  Outils : résumés numériques, représentations graphiques.
- L'inférence : les observations sont souvent recueillies auprès d'un échantillon et non de toute la population d'intérêt. L'inférence statistique vise à induire certaines caractéristiques de la population à partir de leur observation sur un échantillon.
  Outils : estimation, tests statistiques.
- L'étude du lien entre des variables : quantifier le lien, voire le modéliser.
  Outils : analyse bivariée, analyse multivariée, modèles de régression,...
- La classification des individus : chercher des regroupements entre individus selon leurs profils.
  Outils : analyse multivariée, analyse discriminante,...
- La prévision : prédire des valeurs futures (pour une série temporelle) ou la valeur manquante d'une variable pour un individu.
  Outils : modélisation temporelle, modèles de régression,...

2 Quels logiciels?

Tableurs (Excel ou Open Office) :

- 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 les macros complémentaires (régression linéaire, ANOVA, tests), mais sont peu pratiques à utiliser.

$\Rightarrow$ Les tableurs sont globalement mal adaptés aux études statistiques.

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

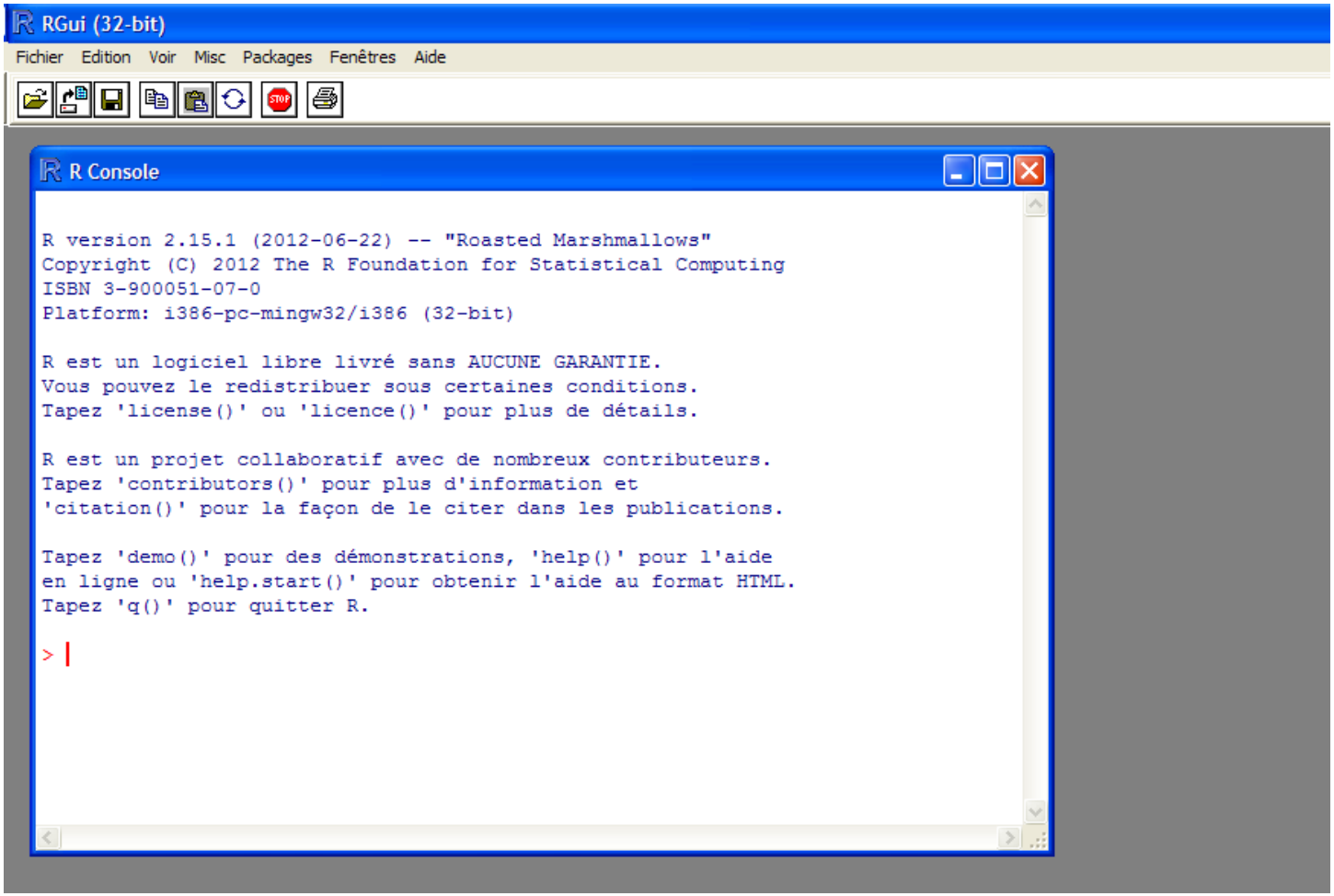
Logiciel spécialisé gratuit: R

On utilisera 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.



R Commander

Il existe une interface graphique à R permettant de travailler avec un jeu de données : **R Commander**

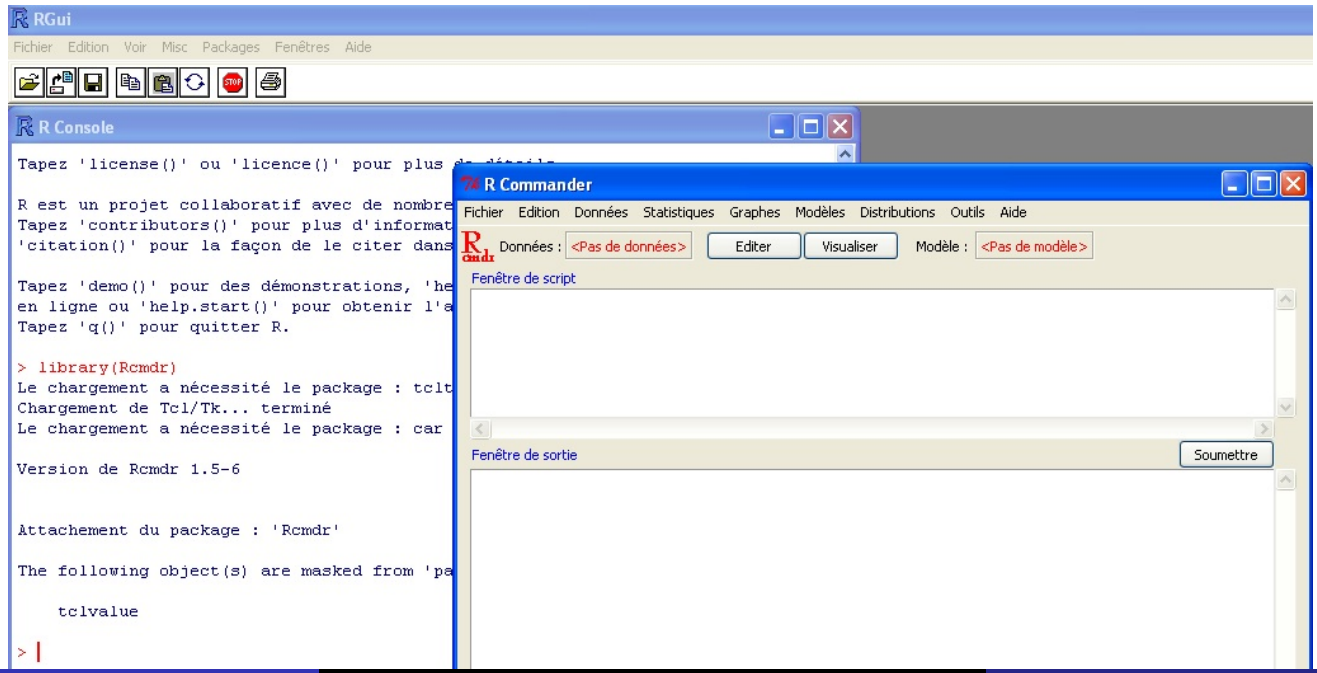
Avantages : utilisation clique-boutons, prise en main rapide.
Inconvénients : fonctionnalités limitées, parfois instable.

On lance R Commander par la commande:

```
library(Rcmdr)
```

(On peut le réouvrir après fermeture avec la commande `Commander()`)

L'importation et l'analyse d'un jeu de données se fait à l'aide des menus déroulants (cf TP).



3 Analyse univariée

- Variable qualitative (ou facteur)
  - Résumés numériques
  - Représentations graphiques
- Variable quantitative
  - Résumés numériques
  - Représentations graphiques

3. Analyse univariée
   - Variable qualitative (ou facteur)
     - Résumés numériques
     - Représentations graphiques
   - Variable quantitative

Variable qualitative : résumés numériques

Soit une variable qualitative A ayant k modalités notées $A_1, \ldots, A_k$.

L'observation de la variable A sur n individus peut se résumer par un **tableau des fréquences**.

Modalités de A	A_1	A_2	$\cdots$	A_k
Effectifs observés	n_1	n_2	$\cdots$	n_k
Fréquences observées	f_1	f_2	$\cdots$	f_k

où n_i est l'effectif dans la modalité A_i et $f_i = n_i/n$, pour tout $i = 1, \ldots, k$.

Propriétés : On a $\sum_{i=1}^{k} n_i = n$ et $\sum_{i=1}^{k} f_i = 1$

Dans R Commander: Statistiques\Résumés\Distributions de fréquences...
renvoie l'effectif et la fréquence de chaque modalité.

Le **mode** est la modalité la plus fréquente dans l'échantillon.

3. Analyse univariée
   - Variable qualitative (ou facteur)
     - Résumés numériques
     - Représentations graphiques
   - Variable quantitative

Variable qualitative : représentations graphiques

Les fréquences de chaque modalité peuvent être résumées par :

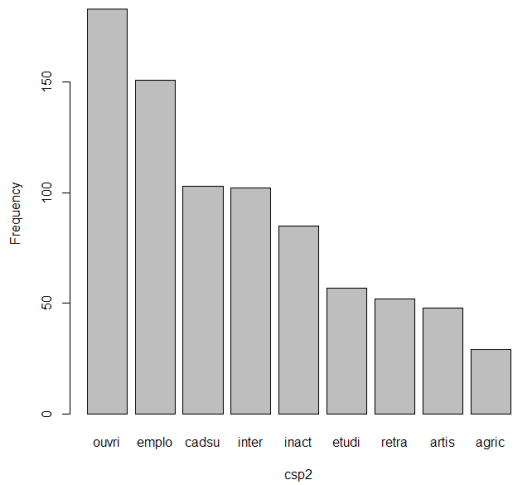
- un graphe en barres : chaque modalité en abscisse est représentée par une barre dont la hauteur est proportionnelle à la fréquence de la modalité. Les barres ne sont pas collées les unes aux autres.
  Dans R Commander : Graphes\Graphe en barres...
- un diagramme circulaire (ou camembert) : chaque modalité est représentée par un secteur dont l'aire est proportionnelle à la fréquence.
  Dans R Commander : Graphes\Graphe en camembert...

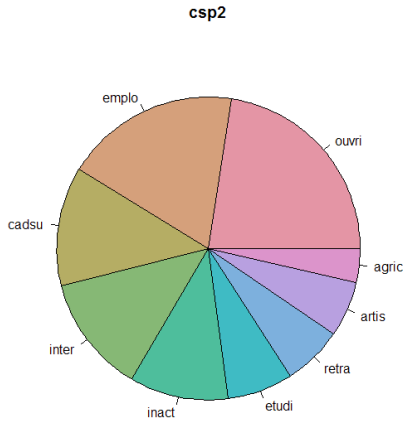
Ordre des modalités. Si la variable est ordinale, les modalités sont présentées dans leur ordre naturel. Sinon, les modalités sont classées par ordre décroissant de leur fréquence d'apparition.

Dans R Commander : l'ordre des modalités peut être modifié dans Données\Gérer les variables dans le jeu de données actif\Réordonner une variable facteur.

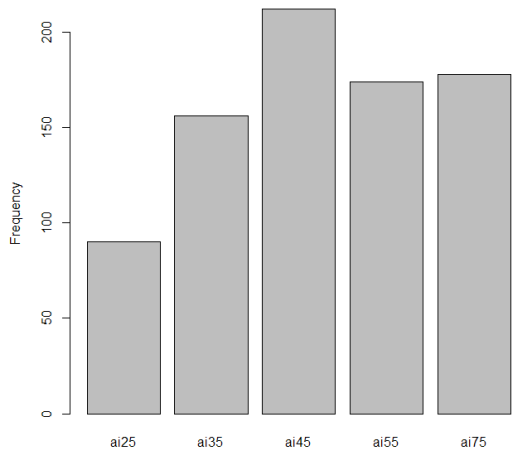
Exemples

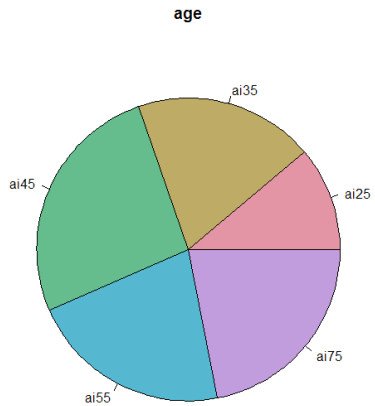
Distribution de la CSP des clients d'une banque (variable qualitative nominale) :





Distribution de la classe d'âge des clients d'une banque (variable qualitative ordinale) :





3 Analyse univariée

- Variable qualitative (ou facteur)
- Variable quantitative
  - Résumés numériques
  - Représentations graphiques

Dans R Commander : Statistiques\Résumés\Statistiques descriptives...

On note $x_1, \ldots, x_n$ l'échantillon des n valeurs numériques.

Les mesures de position d'un échantillon

- Le **mode** est la valeur la plus fréquente dans l'échantillon.
- La **moyenne**, notée $\bar{x}_n$, est $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$
  Propriété (preuve en cours) : $\bar{x}_n = \underset{y}{\text{argmin}} \sum_{i=1}^n (x_i - y)^2$
- La **médiane** est le nombre m séparant l'échantillon ordonné en 2 parties égales. Plus précisément, m vérifie :

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{x_i \leq m\}} \geq \frac{1}{2} \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{x_i \geq m\}} \geq \frac{1}{2}$$

  Propriétés (preuve en cours) :
  On note $x_{(1)}, \ldots, x_{(n)}$ l'échantillon ordonné.

  1. $m \in \underset{y}{\text{argmin}} \sum_{i=1}^n |x_i - y|$
  2. si $n = 2k+1$ (n impair): m est unique et vaut $m = x_{(k+1)}$
     si $n = 2k$ (n pair): m n'est pas unique, l'ensemble des solutions est $[x_{(k)}, x_{(k+1)}]$ et on choisit en pratique $m = \frac{1}{2}(x_{(k)} + x_{(k+1)})$

Remarque : la médiane est plus robuste aux valeurs extrêmes que la moyenne

Les **quantiles.**
Soit $p \in [0,1]$, le **quantile** d'ordre p, noté $Q(p)$, sépare l'échantillon ordonné en deux parties de taille respective np et $n(1-p)$ approximativement.
Plus précisément, $Q(p)$ vérifie

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq Q(p)\}} \geq p \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \geq Q(p)\}} \geq 1-p \tag{1}$$

On note $\lfloor . \rfloor$ la partie entière inférieure et $\lceil . \rceil$ la partie entière supérieure.
Propriétés : $Q(p)$ n'est pas nécessairement unique, l'ensemble des solutions étant $[x_{(\lceil np \rceil)}, x_{(\lfloor np \rfloor + 1)}]$. Pour $p = 0.5$, on retrouve la définition de la médiane.

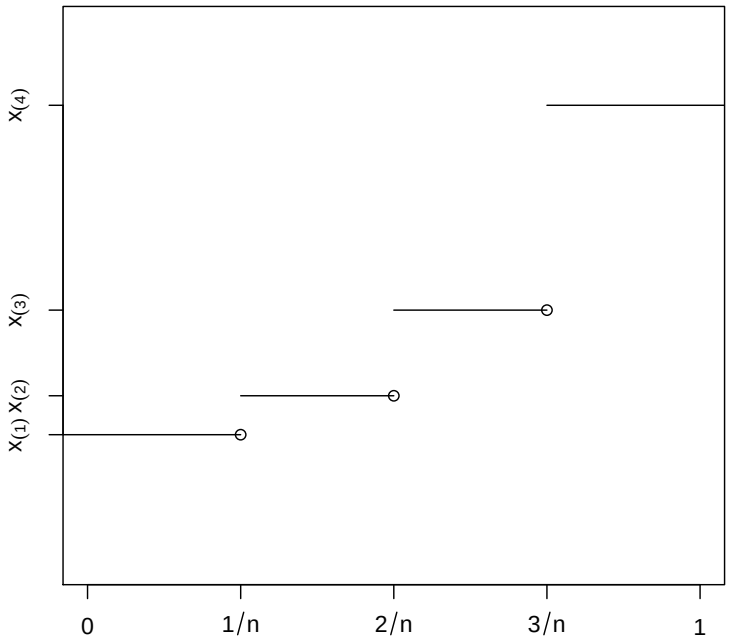
Calcul en pratique :

1. $Q(p) = x_{(\lceil np \rceil)}$: choix qui respecte la définition mais n'est pas cohérent avec la convention adoptée pour la médiane (lorsque $p = 0.5$).
2. $Q(p) = x_{(\lceil np \rceil)}$ si np n'est pas un entier, et $Q(p) = \frac{1}{2}(x_{(np)} + x_{(np+1)})$ sinon. Ce choix respecte la définition et est cohérent avec la convention adoptée pour la médiane.

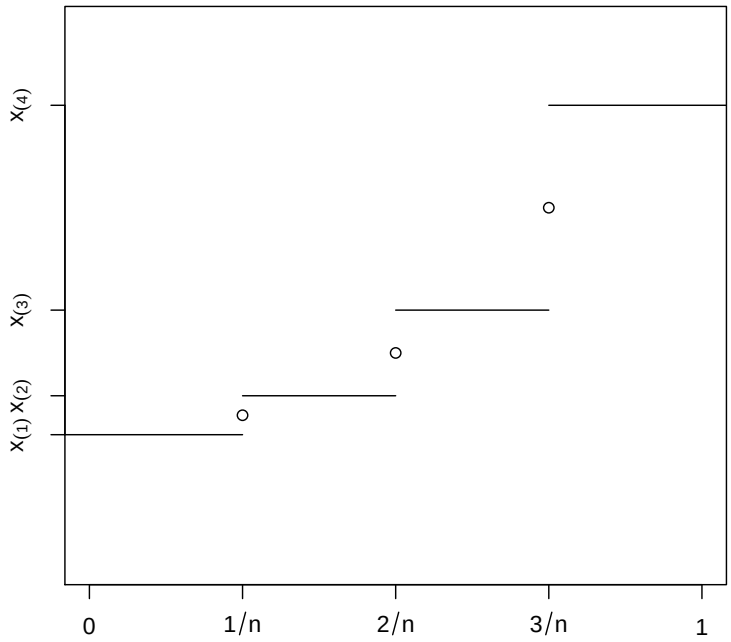
Pour $p = 0.25, 0.5, 0.75$, les quantiles sont appelés **quartiles**.
Pour $p = 0.1, \ldots, 0.9$, ce sont les **déciles**.

Représentation de $Q(p)$ en fonction de p, selon le choix de sa définition.
(Exemple pour $n = 4$)

Choix 1



Choix 2



Autres choix possibles pour $Q(p)$.

R en propose 9 (voir l'aide de la fonction quantile).

- Les deux premiers sont les 2 choix précédents.
- Le troisième est un autre choix conduisant à $Q(p)$ discontinu.
- Les six autres sont des choix rendant $Q(p)$ continu mais dans ce cas la propriété initiale (1) n'est plus vérifiée.

L'idée est d'interpoler linéairement les points $(p_k, x_{(k)})$ où $k = 1, \ldots, n$ et où le choix de p_k diffère selon les versions:

- Le choix $p_k = k/n$ correspond à une interpolation linéaire du choix 1 ci-dessus et correspond au choix 4 dans R.
- Pour les autres choix, la motivation est inférentielle. En supposant que $X_1, \ldots, X_n$ est un échantillon i.i.d dont la loi est de fonction de répartition (f.d.r.) F, le but est d'estimer au mieux le quantile théorique de la loi. En notant F_n la f.d.r. empirique, on a $k/n = F_n(X_{(k)})$. Ainsi au lieu de choisir $p_k = k/n = F_n(X_{(k)})$ on considère $F(X_{(k)})$.
  Quelle que soit F, $F(X_{(k)})$ est distribué comme la $k$ème statistique d'ordre d'un échantillon i.i.d suivant la loi uniforme sur $[0, 1]$, qui est une loi $Beta(k, n + 1 - k)$. On peut alors choisir par exemple :
  $p_k = E(F(X_{(k)}))$ ou $p_k = mode(F(X_{(k)}))$ ou $p_k = mediane(F(X_{(k)}))$.

Le choix par défaut sous R (choix 7) correspond à $p_k = mode(F(X_{(k)})) = \frac{k-1}{n-1}$.

Si le but n'est pas inférentiel mais descriptif, le choix 2 semble plus approprié.

Les mesures de dispersion d'un échantillon

- L'**étendue** vaut $max - min$, c'est à dire $x_{(n)} - x_{(1)}$.
- La **variance** est $V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. On a $V = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$.
- La **variance corrigée** est $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
- L'**écart-type** est $\sigma = \sqrt{V}$
- L'**écart-type corrigé** est $S = \sqrt{S^2}$
- L'**écart moyen absolu** est $EMA = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|$, où m est la médiane.
- L'**écart inter-quartile** est $Q(0.75) - Q(0.25)$

La mesure la plus utilisée est l'écart-type, qui peut être interprétée comme suit:

Proposition (voir preuve en cours)

Soit $\alpha \geq 1$ et $\mathcal{I} = [\bar{x}_n - \alpha\,\sigma,\ \bar{x}_n + \alpha\,\sigma]$, alors

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{x_i \in \mathcal{I}} \geq 1 - \frac{1}{\alpha^2}$$

En particulier (pour $\alpha = 2$), au moins 3/4 des observations appartiennent à l'intervalle $[\bar{x}_n - 2\sigma, \bar{x}_n + 2\sigma]$.

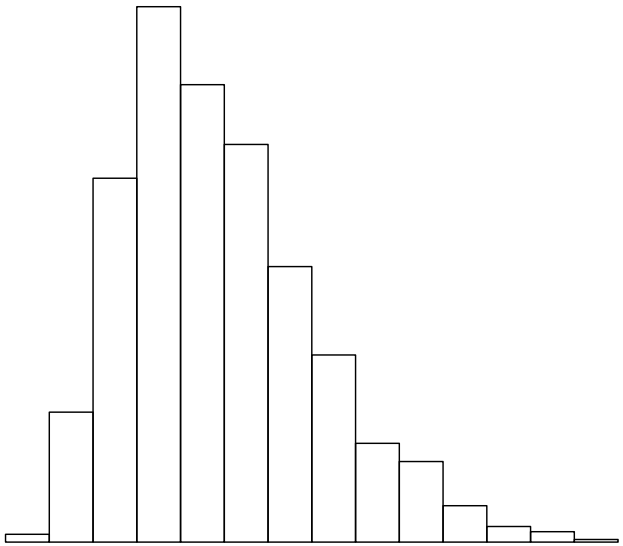
Attention : La variance et l'écart-type calculés dans R correspondent à leur version "corrigée" S^2 et S.

Pour mesurer l'asymétrie de la répartition des valeurs, on utilise le **coefficient d'asymétrie** ou **skewness** en anglais.

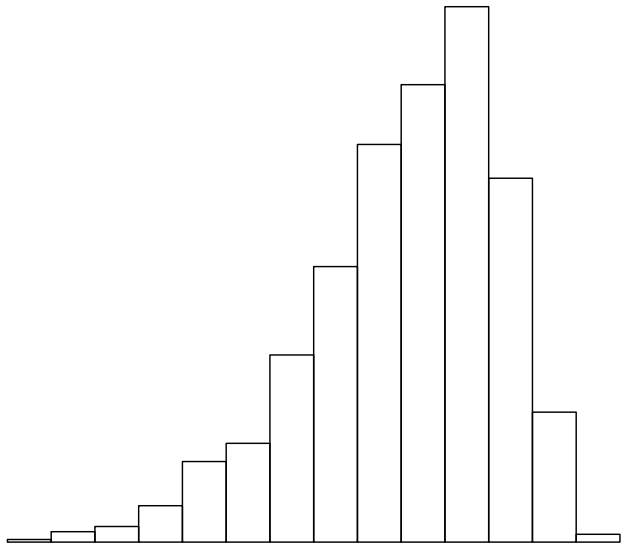
$$\gamma_1 = \frac{m_3}{m_2^{3/2}}$$

où $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k$ est le moment centré d'ordre k (la variance si $k = 2$)
Dans une optique inférentielle, R propose deux autres expressions du skewness (voir l'aide de la fonction skewness du package "e1071")

Skewness positif ($\bar{x}_n > mediane$)



Skewness négatif ($\bar{x}_n < mediane$)



Pour mesurer l'aplatissement de la répartition des valeurs, on utilise l'**excès d'aplatissement** ou **kurtosis normalisé** :

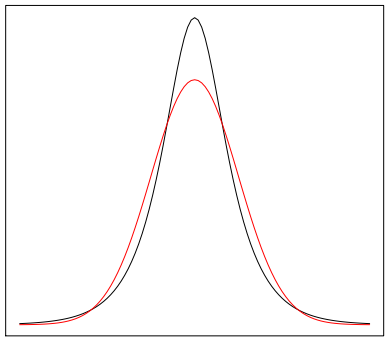
$$\gamma_2 = \frac{m_4}{m_2^2} - 3$$

où $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k$ est le moment centré d'ordre k.

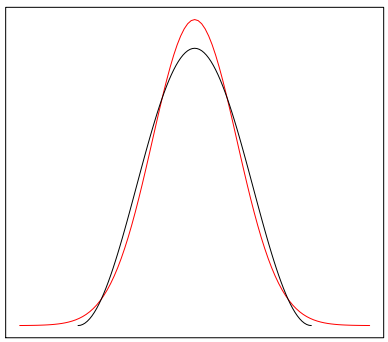
Rem : le "3" correspond au kurtosis (non normalisé) théorique d'une $\mathcal{N}(0, \sigma^2)$. γ_2 compare donc l'aplatissement de la distribution à celui d'une loi normale.

Dans une optique inférentielle, R propose deux autres expressions du kurtosis (voir l'aide de la fonction kurtosis du package "e1071")

$\boxed{\gamma_2 > 0}$: distribution **leptokurtique** (en noir). Profil plus "pointue" qu'une normale de même variance (en rouge). C'est le cas en présence de nombreuses valeurs extrêmes. Le kurtosis est utilisé en finance pour repérer ce phénomène.

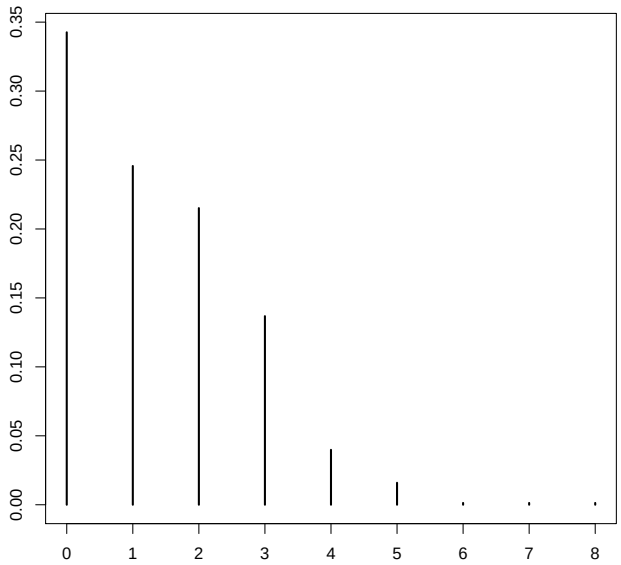


$\boxed{\gamma_2 < 0}$: distribution **platikurtique** (en noir). Profil plus "plat" qu'une normale de même variance (en rouge).



Le diagramme en bâtons (pour une variable quantitative discrète).
Les bâtons sont placés en abscisse au niveau de chaque valeur possible de la variable discrète et leur hauteur est proportionnelle à la fréquence observée.

Exemple : répartition du nombre d'enfants par femme dans un échantillon de femmes actives américaines.

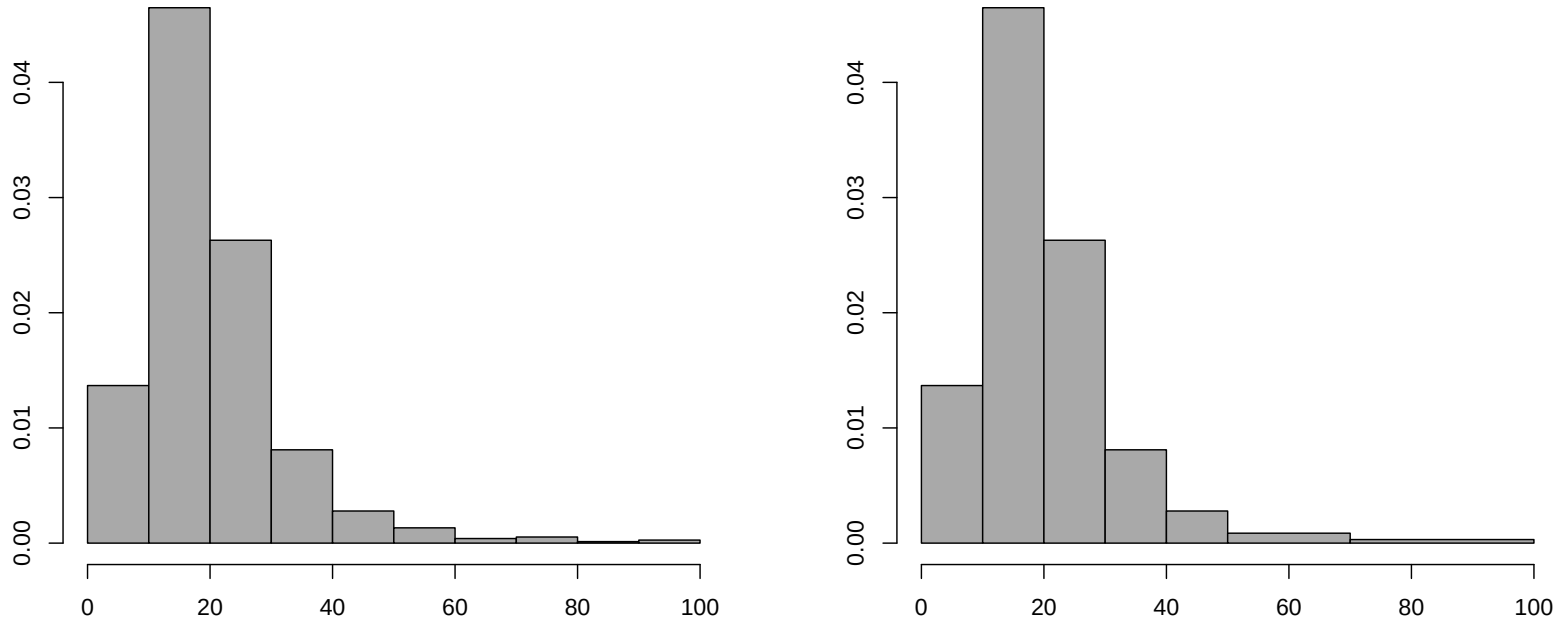


Dans R Commander : le diagramme en bâtons n'est pas proposé.
Dans R : en ligne de commandes, il faut d'abord calculer les fréquences avec la fonction table puis représenter le résultat avec plot.

L'histogramme (pour une variable quantitative continue).
L'ensemble des valeurs possibles est découpé en intervalles disjoints appelés **classes**. Au niveau de chaque classe s'élève un rectangle dont l'aire est égale à la fréquence observée de la classe.

Exemple : répartition du salaire annuel des femmes actives américaines (en milliers de dollars)

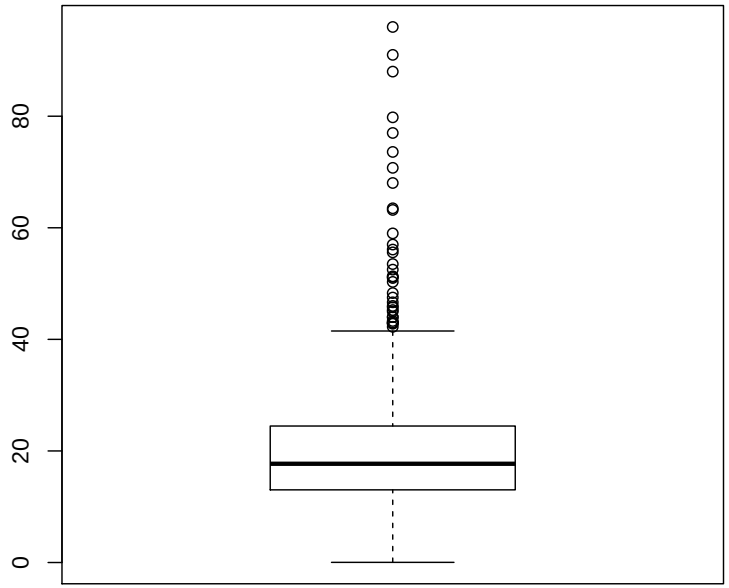
Gauche : classes de même taille. Droite : classes de tailles différentes (les 2 dernières).



Dans R Commander : Graphes\Histogramme...

Le boxplot ou **boîte à moustaches** ou **boite de dispersion**.
Le rectangle central est délimité par le premier et le troisième quartile et la médiane y est symbolisée par un trait.
Les "moustaches" partent de chaque côté jusqu'à la valeur minimale et maximale de l'échantillon, sous réserve que leur longueur ne dépasse pas $1.5(Q(0.75) - Q(0.25))$, soit 1.5 fois la hauteur de la boîte. Sinon, les moustaches s'arrêtent à la dernière valeur avant cette limite et les valeurs restantes sont représentées par des points isolés.

Exemple : répartition du salaire des femmes actives américaines.



Dans R Commander : Graphes\Boite de dispersion...

4 Analyse bivariée

- Variable quantitative/ Variable qualitative
- Variable qualitative/ Variable qualitative
- Variable quantitative/ Variable quantitative

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

On suppose que le facteur admet I modalités contenant chacune n_i individus ($i = 1, \ldots, I$). On a donc $\sum_{i=1}^{I} n_i = n$.
On note x_{ij} la valeur de la variable quantitative pour l'individu j se trouvant dans la modalité i du facteur ($i = 1, \ldots, I$ et $j = 1, \ldots, n_i$).
On note $\bar{x}_i$ la moyenne dans la modalité i et $\bar{x}$ la moyenne totale, i.e.

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

Le lien entre la variable et le facteur est parfois mesuré par le **rapport de corrélation eta**:

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

Propriétés (cf cours) :

- **Formule de décomposition de la variance** : La variance totale est la somme de la variance inter-modalités et de la variance intra-modalités.

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{I} n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

- On a ainsi $0 \leq \eta \leq 1$

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

Exemple : quelques statistiques descriptives du taux de NO2 dans les véhicules en fonction de la fluidité du trafic (jeu de données vu en TP).

	mean	sd	0%	25%	50%	75%	100%	data:n
A	244.1691	143.9747	16.53659	140.3315	203.6442	321.9209	844.3919	79
B	232.5729	123.3186	53.64617	143.6053	199.0793	299.4594	573.0764	68
C	235.1250	136.0099	52.36647	126.5375	202.1544	302.8114	673.3514	65
D	266.9988	149.9037	13.40037	171.7825	220.6241	337.2726	806.6694	74

Avec les notations précédentes, on a ici $I = 4$, $n_1 = 79$, $n_2 = 68$, $n_3 = 65$, $n_4 = 74$, $\bar{x}_1 = 244$, $\bar{x}_2 = 233$, $\bar{x}_3 = 235$, $\bar{x}_4 = 267$.

• Pour calculer η, on peut utiliser le menu Statistiques\Moyennes\ANOVA à un facteur... : η^2 correspond au rapport des "Sum Sq" (somme des carrés) pour le facteur sur ceux des "Residuals" (résidus).

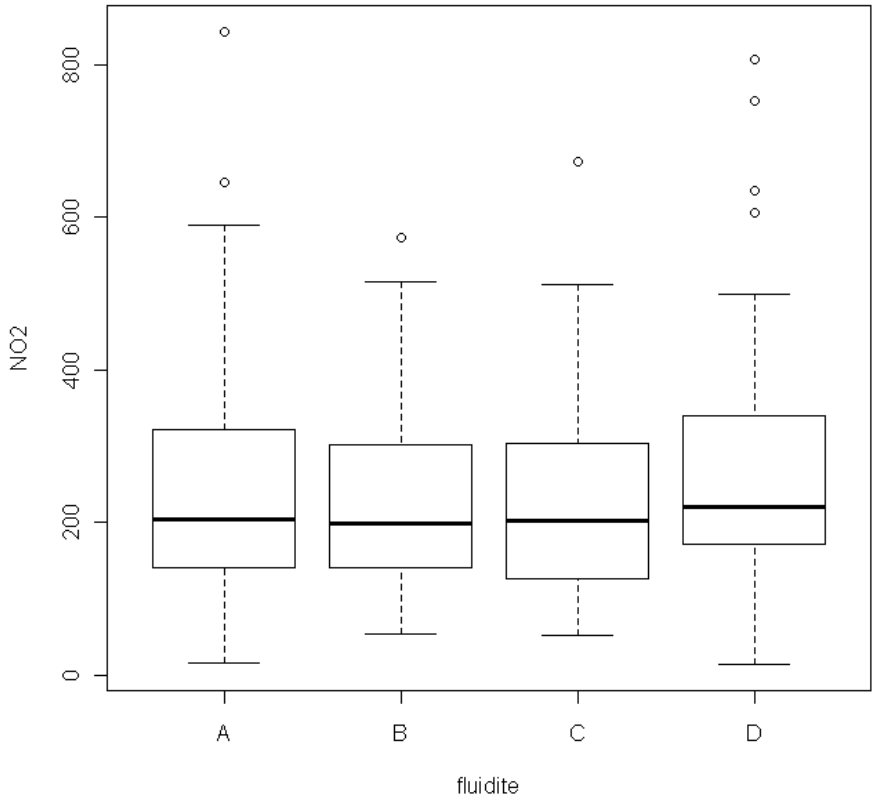
Pour l'exemple ci-dessus, la sortie de l'ANOVA donne:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fluidite	3	52687	17562	0.907	0.438
Residuals	282	5460050	19362		

Ainsi $\eta^2 = \frac{52687}{5460050}$, d'où $\eta \approx 0.1$.

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").



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

On suppose que le premier facteur admet I modalités et le second J modalités.
n_{ij} : nombre d'individus ayant la modalité i pour le premier facteur et j pour le second.
$n_{i.}$: nombre d'individus ayant la modalité i pour le premier facteur
$n_{.j}$: nombre d'individus ayant la modalité j pour le second facteur

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

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

Sous R Commander, on peut le construire automatiquement à l'aide du menu Statistiques\Tables de contingence\Tri croisé...

Exemple : Pour les variables "type" et "fluidite" du jeu de données NO2trafic, le tableau de contingence (avec l'option par défaut "Pas de pourcentages") est :

```
          type
fluidite  P  U  A  T  V
       1 21 21 19  9  9
       2 20 17 16  8  7
       3 17 17 16  8  7
       4 20 20 18  8  8
```

Remarque : On peut remplacer les effectifs n_{ij} par les fréquences n_{ij}/n en choisissant l'option "Pourcentages du total" dans l'onglet "Statistiques".

• La **distribution conditionnelle** du second facteur sachant la modalité i du premier facteur est donnée par les fréquences:

$$\frac{n_{ij}}{n_{i.}}, \quad \text{pour } j = 1, \ldots, J.$$

Pour chaque i, on a évidemment : $\sum_{j=1}^{J} \frac{n_{ij}}{n_{i.}} = 1$.

Les **profils lignes** correspondent à l'ensemble de ces distributions conditionnelles pour $i = 1, \ldots, I$.

L'intéret des profils lignes est de comparer la distribution du second facteur selon les modalités du premier facteur. S'il y a indépendance entre les deux facteurs, les profils lignes doivent être similaires.

Dans R Commander, on obtient les profils lignes en choisissant l'option "Pourcentages des lignes" dans l'onglet "Statistiques" du menu Statistiques\Tables de contingence\Tri croisé...

• De même on peut s'intéresser aux **profils colonnes** qui sont les distributions conditionnelles du premier facteur sachant le second. Elles sont données par les fréquences

$$\frac{n_{ij}}{n_{.j}}, \quad \text{pour } i = 1, \ldots, I.$$

Exemple (suite) :
Les profils lignes du tableau précédent sont :

	type						
fluidite	P	U	A	T	V	Total	Count
1	26.6	26.6	24.1	11.4	11.4	100.1	79
2	29.4	25.0	23.5	11.8	10.3	100.0	68
3	26.2	26.2	24.6	12.3	10.8	100.1	65
4	27.0	27.0	24.3	10.8	10.8	99.9	74

Les profils colonnes sont :

	type				
fluidite	P	U	A	T	V
1	26.9	28.0	27.5	27.3	29.0
2	25.6	22.7	23.2	24.2	22.6
3	21.8	22.7	23.2	24.2	22.6
4	25.6	26.7	26.1	24.2	25.8
Total	99.9	100.1	100.0	99.9	100.0
Count	78.0	75.0	69.0	33.0	31.0

On constate peu de différences entre les profils lignes (de même pour les profils colonnes), il ne semble donc pas y avoir de lien entre les deux facteurs.

Pour mesurer le lien entre les deux facteurs, on calcule la distance du khi-deux.

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

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

Propriétés (voir preuve en cours):

$$0 \leq \chi^2 \leq n \times [min(I, J) - 1]$$

On peut mesurer le lien entre deux variables qualitatives par le **V de Cramer** :

$$V = \sqrt{\frac{\chi^2}{n \times [min(I, J) - 1]}}$$

On a $0 \leq V \leq 1$.

Sous R Commander, la distance du χ^2 est donnée dans la sortie du menu Statistiques\Tables de contingence\Tri croisé... sous l'appellation "`X-squared`", en cochant "Test Chi-deux d'indépendance" dans l'onglet "Statistiques".

Exemple (suite) : on lit dans R Commander `X-squared = 0.3513`.
On en déduit $V = \sqrt{0.3513/(283 \times 3)} = 0.02$. L'absence de lien entre les variables "type" et "fluidite" se confirme.

Pour comprendre plus profondément le lien éventuel entre les deux facteurs :

- L'option "Imprimer les fréquences attendues" donne le tableau des effectifs (et non fréquences...) théoriques $\frac{n_{i.} n_{.j}}{n}$ s'il y avait eu indépendance.
- L'option "Composants de la statistique du chi-deux" donne le tableau des résidus, c'est à dire chaque terme composant la somme du χ^2.

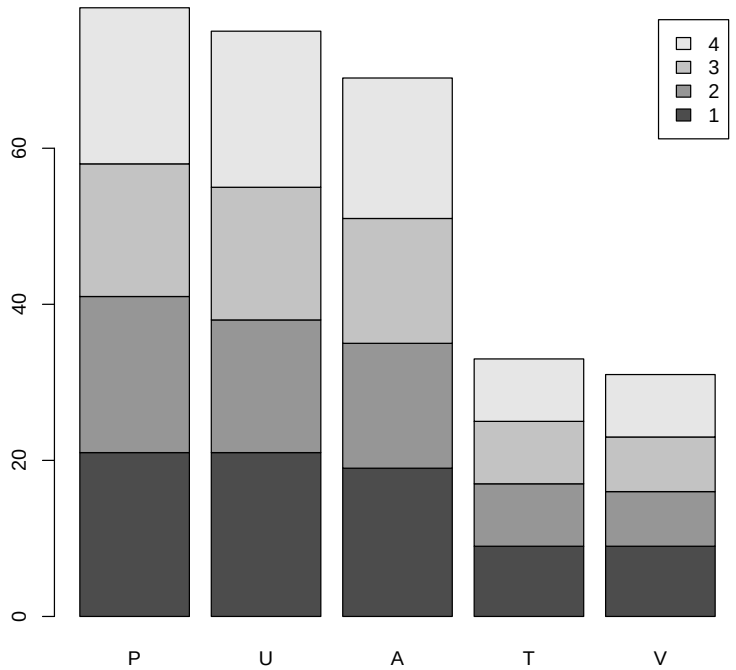
Un résidu élevé témoigne d'une sur-représentation (ou sous-représentation) de la modalité croisée par rapport à une situation d'indépendance.

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

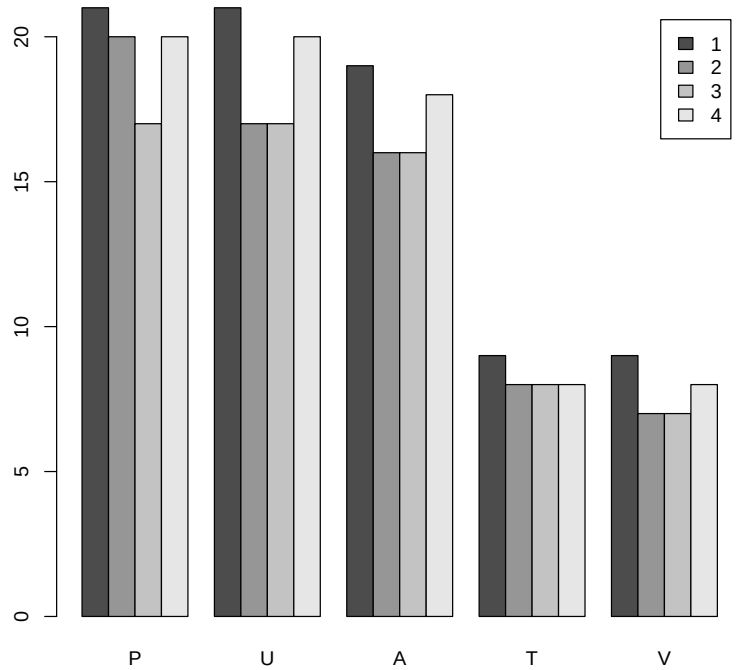
Ces graphes se font en lignes de commande : si le tableau de contingence se nomme `tab`, il suffit de taper barplot(tab) ou barplot(tab,beside=TRUE)

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

barplot(tab,legend.text=TRUE)



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

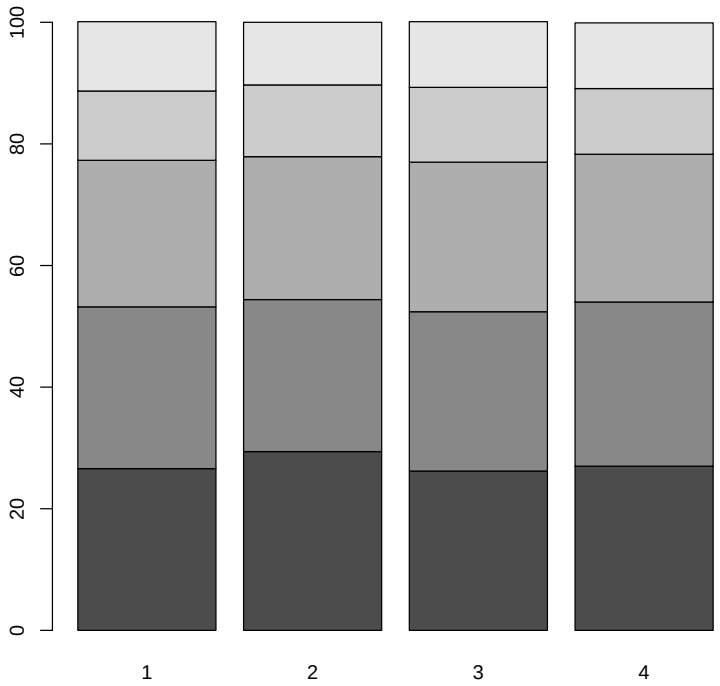


Remarque : si on souhaite représenter les fréquences et non les effectifs, il suffit de diviser `tab` par l'effectif total n, barplot(tab/n).

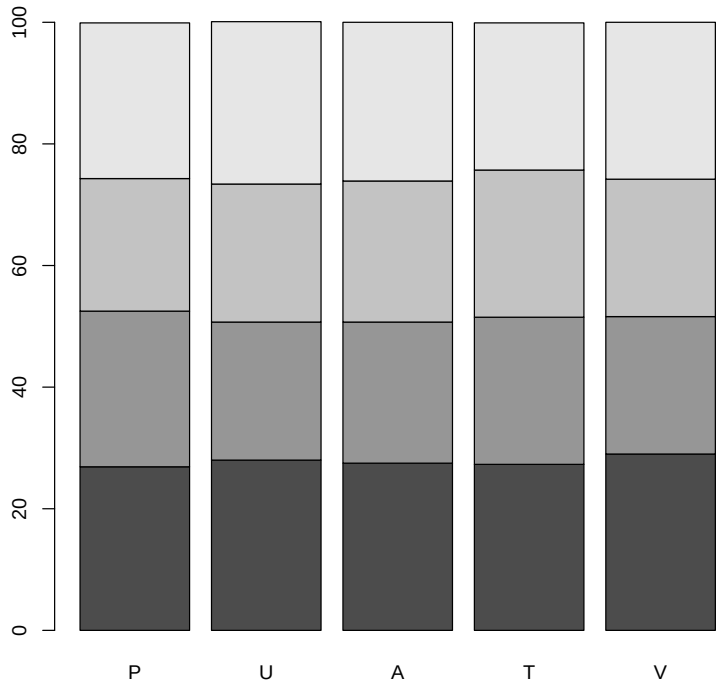
On peut de même représenter les distributions conditionnelles.

On suppose que les tableaux se nomment :
`tablig` : tableau des profils lignes (sans les colonnes `Total` et `Count`)
`tabcol` : tableau des profils colonnes (sans les lignes `Total` et `Count`)

barplot(t(tablig))



barplot(tabcol)



Remarque : il faut représenter la transposée du tableau des profils lignes, ce qui explique le t(tablig).

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

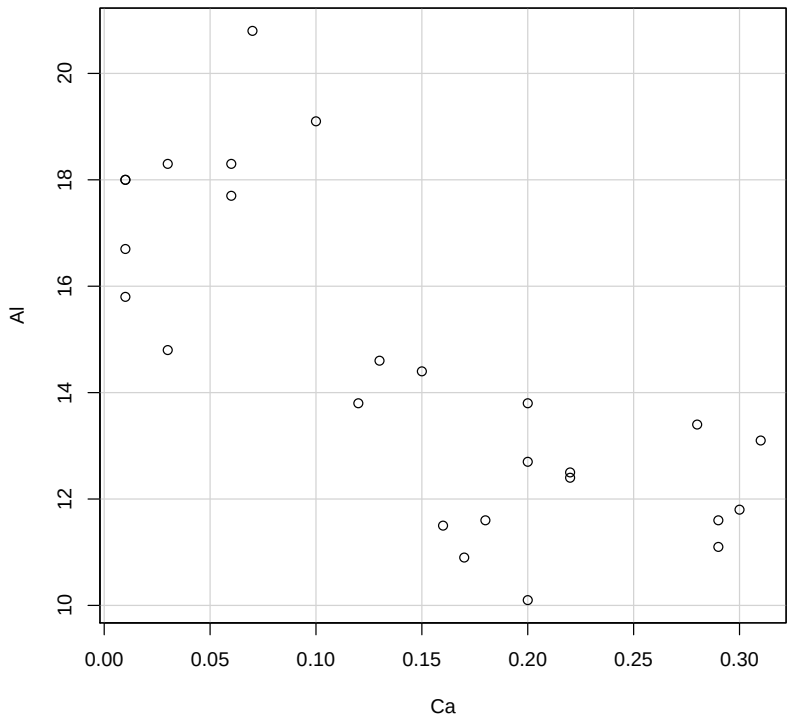
Lien entre deux variables quantitatives : nuage de points

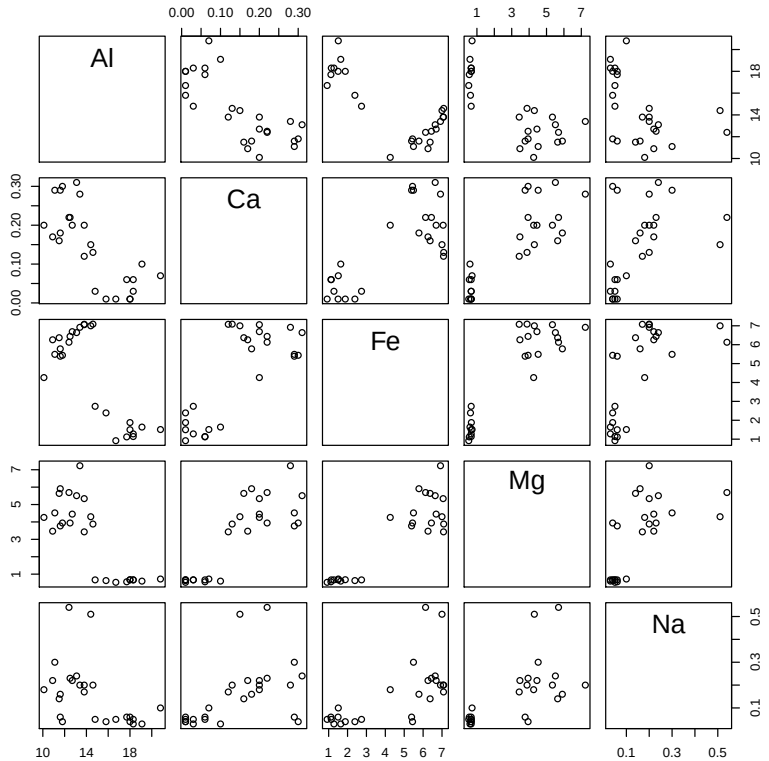
Soit $x_1, \ldots, x_n$ les valeurs de la première variable quantitative X.
Soit $y_1, \ldots, y_n$ les valeurs de la seconde variable quantitative Y.

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

Dans R Commander : Graphe\Nuage de points...ou Graphe\Matrice de nuages de points... si on en souhaite plusieurs.

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





On définit la **covariance** entre X et Y par :

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n$$

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

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

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

où σ_X (resp. σ_Y) désigne l'écart-type de X (resp. Y).

Propriétés (cf cours) : La corrélation r est toujours comprise entre -1 et 1:

- si $r = 1$, il y a un lien linéaire "parfait" positif :

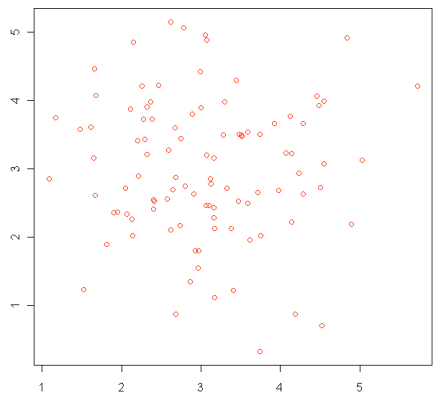
  $r = 1$ ssi il existe $\alpha \geq 0$ et β tel que $y_i = \alpha x_i + \beta$ pour tout $i = 1, \ldots, n$

- si $r = -1$, il y a un lien linéaire "parfait" négatif :

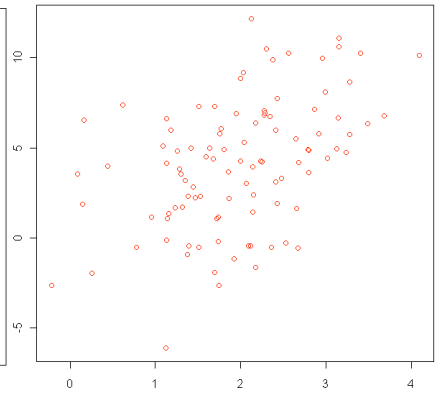
  $r = -1$ ssi il existe $\alpha \leq 0$ et β tel que $y_i = \alpha x_i + \beta$ pour tout $i = 1, \ldots, n$

- si $r = 0$, il n'y a aucun lien linéaire (mais il peut exister un lien non-linéaire).

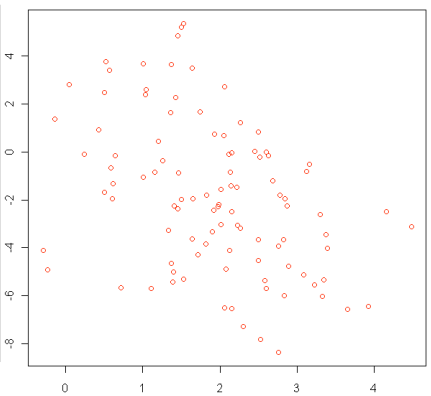
Quelques exemples de nuages de points avec la corrélation correspondante.



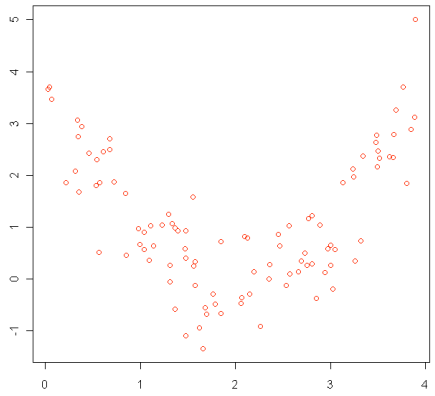
Aucun lien ($r \approx 0$)



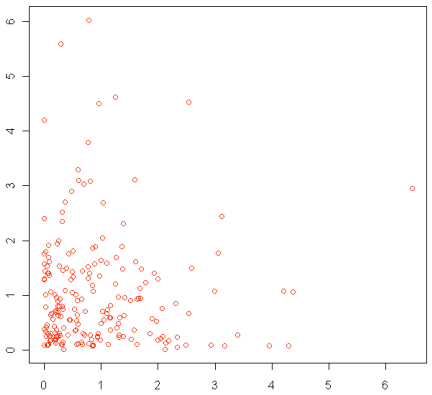
Lien linéaire ($r \approx 0.4$)



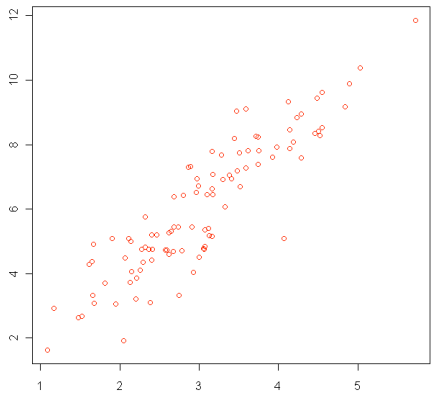
Lien linéaire ($r \approx -0.4$)



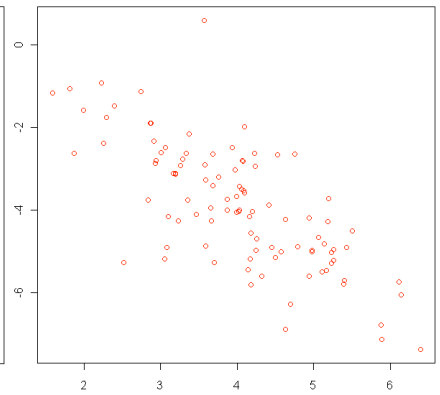
Lien non-linéaire($r \approx 0$)



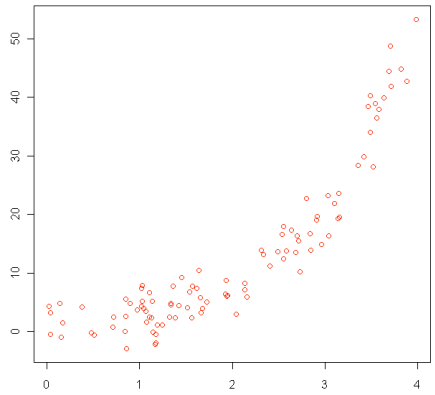
Aucun lien ($r \approx 0$)



Lien linéaire ($r \approx 0.9$)



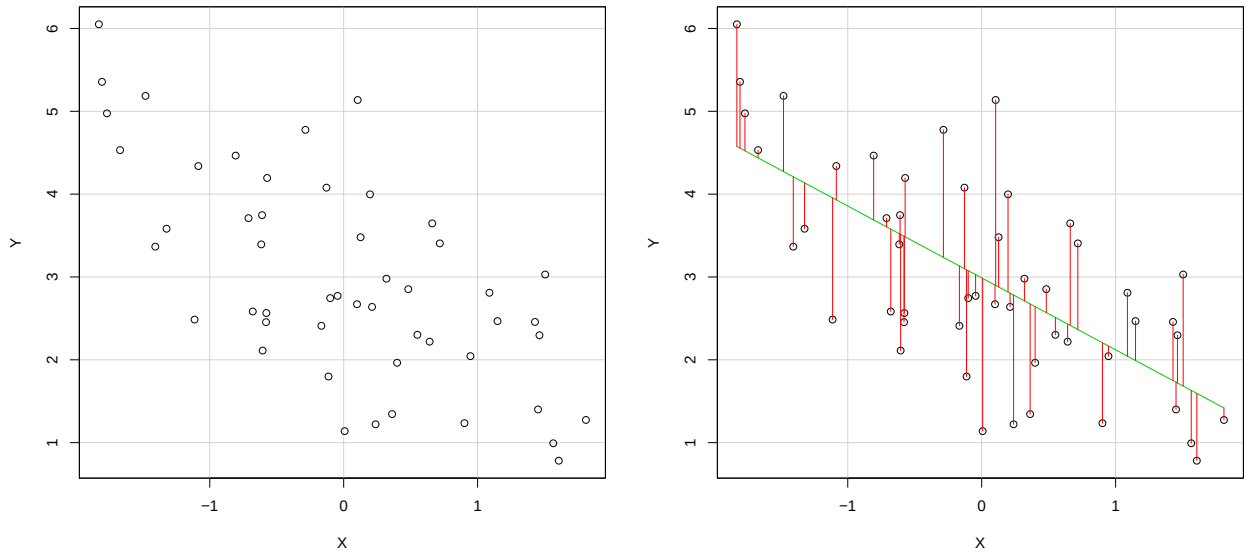
Lien linéaire ($r \approx -0.8$)



Lien non-linéaire ($r \approx 0.8$)

Ajuster la droite des moindres carrés

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



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

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

On trouve, si $var(X) \neq 0$:

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

- Dans R Commander : Statistiques\Ajustement de modèles\Régression linéaire
- La droite des moindres carrés peut servir à prédire une nouvelle valeur de Y : pour $X = x$, Y est prédit par $\hat{y} = \hat{a}x + \hat{b}$.
- Pour mesurer la qualité d'ajustement de la droite on considère les **résidus** : $e_i = y_i - (\hat{a}x_i + \hat{b})$. Il s'agit des segments rouges du graphique précédent. Pour l'ajustement de la droite précédente, on a toujours : $\frac{1}{n}\sum_{i=1}^{n} e_i = 0$.
- Formule de décomposition de la variance:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

que l'on note généralement $SCT = SCE + SCR$ où SCT : "Somme des Carrés Totaux", SCE : "SC Expliqués" et SCR : "SC Résiduels".

- On définit le coefficient de détermination : $R^2 = \frac{SCE}{SCT}$. On a
  - $0 \leq R^2 \leq 1$ d'après la formule de décomposition de la variance
  - $R^2 = r^2$ où r est le coefficient de corrélation linéaire entre Y et X (cf cours)
  - L'ajustement par la droite est d'autant meilleur que R^2 est proche de 1 (car dans ce cas $\sum_{i=1}^{n} e_i^2 \approx 0$)

5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

De nombreuses données sont acquises à intervalles réguliers dans le temps : ce sont des **séries temporelles**.

Des méthodes spécifiques permettent :

- de les représenter graphiquement
- de décrire leurs tendances temporelles et leurs effets périodiques
- d'étudier leur dépendance temporelle
- de les prédire

Représentation graphique :

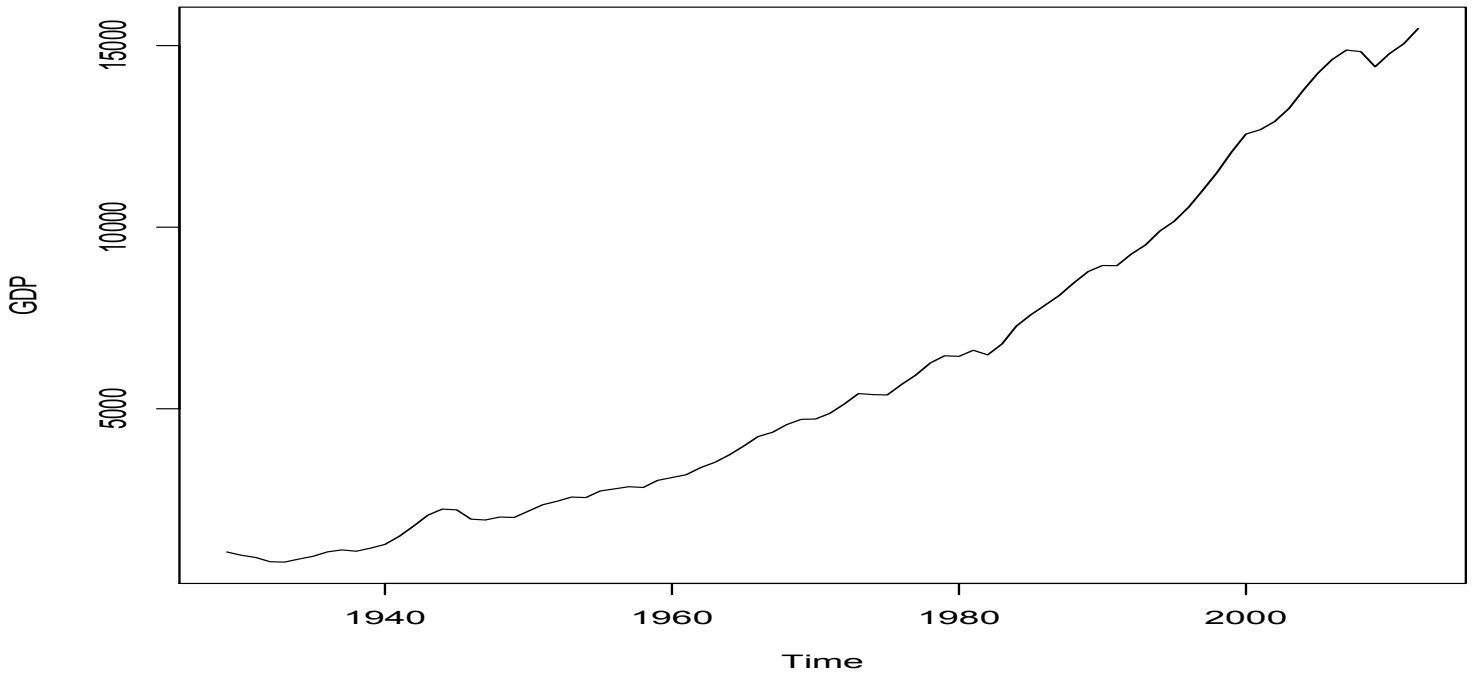
- En abscisse : le temps (secondes, jours, mois, années, etc...);
- En ordonnée : les valeurs des observations.

Exemples :

PIB des Etats-Unis.

Une observation par an.

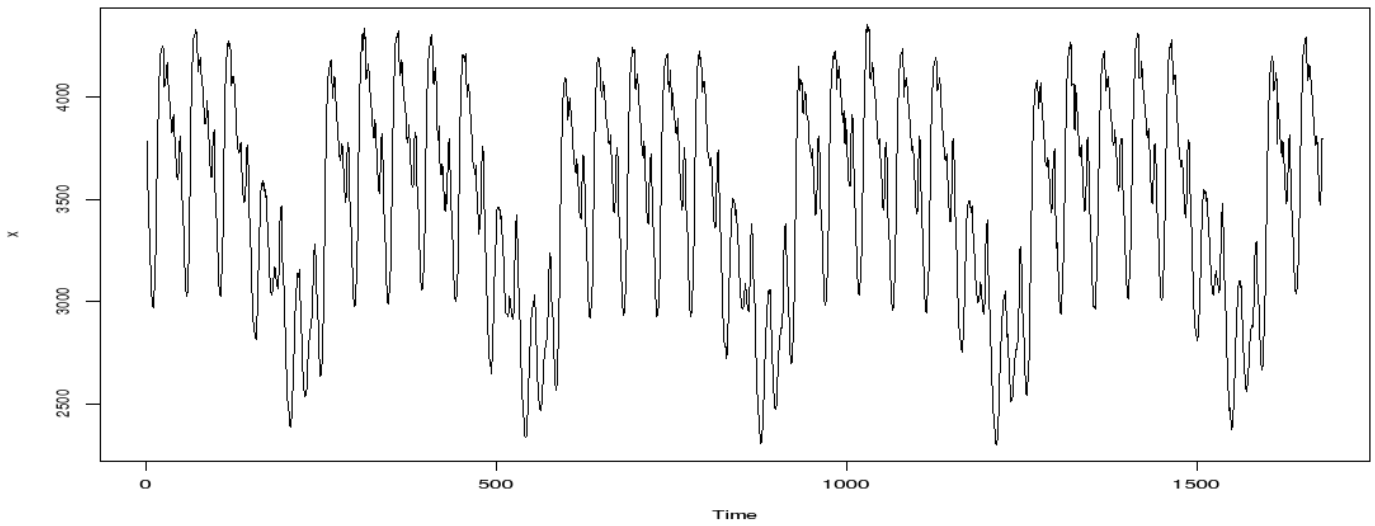
Intervalle d'observation : 1929 - 2012.



Consommation d'électricité en Australie.

Une donnée par demi-heure.

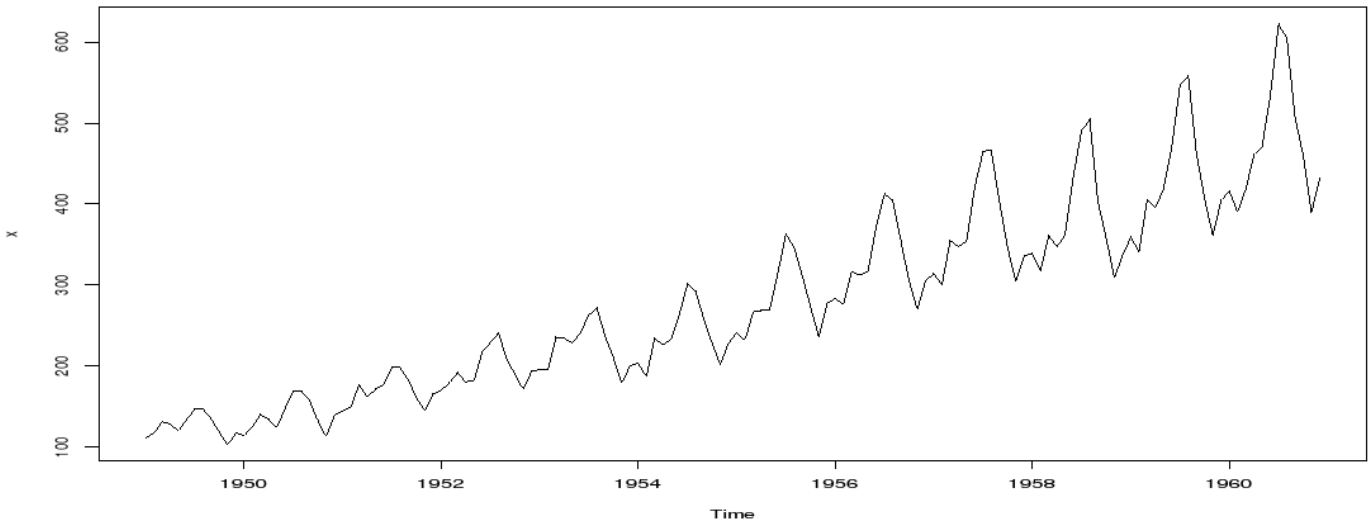
Intervalle d'observation : 35 jours en juin-juillet 1991.



Consommation d'électricité

Trafic aérien.

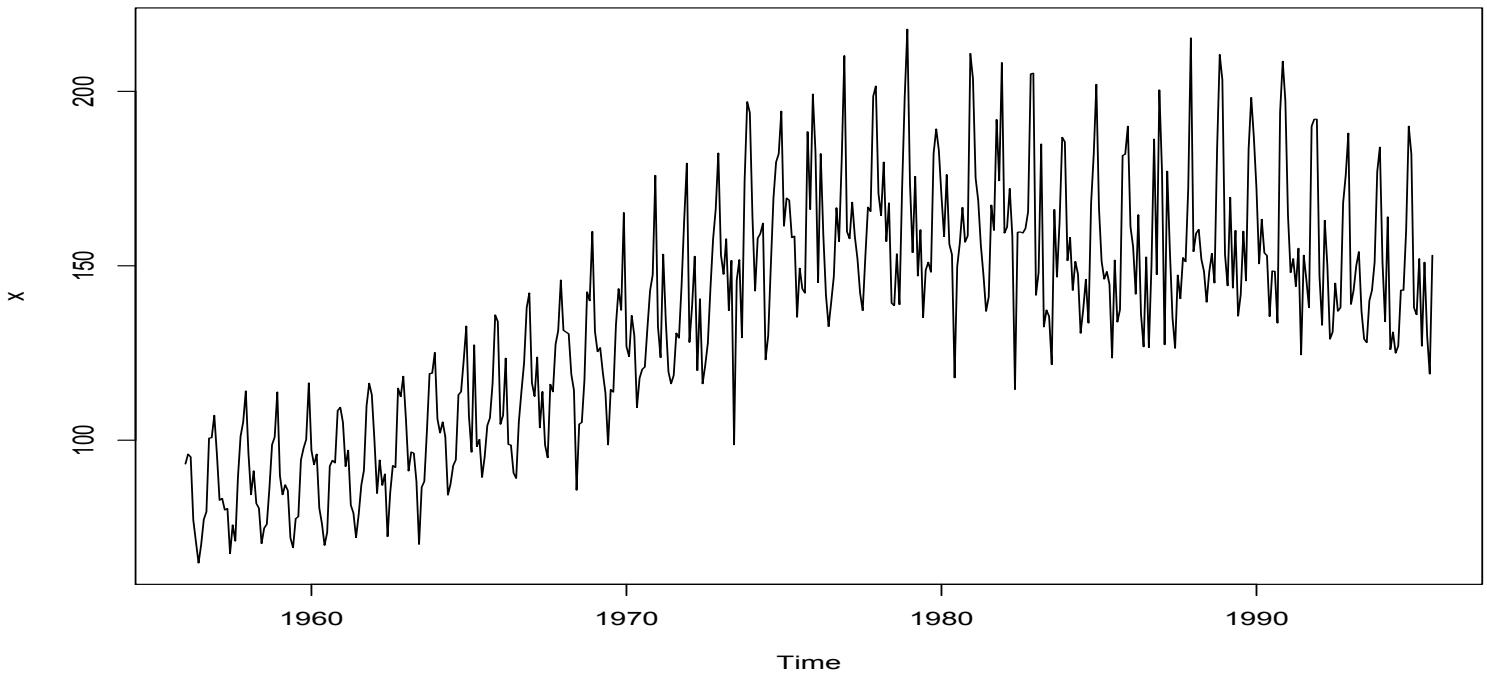
Une donnée par mois de 1949 à 1961.



Trafic aérien

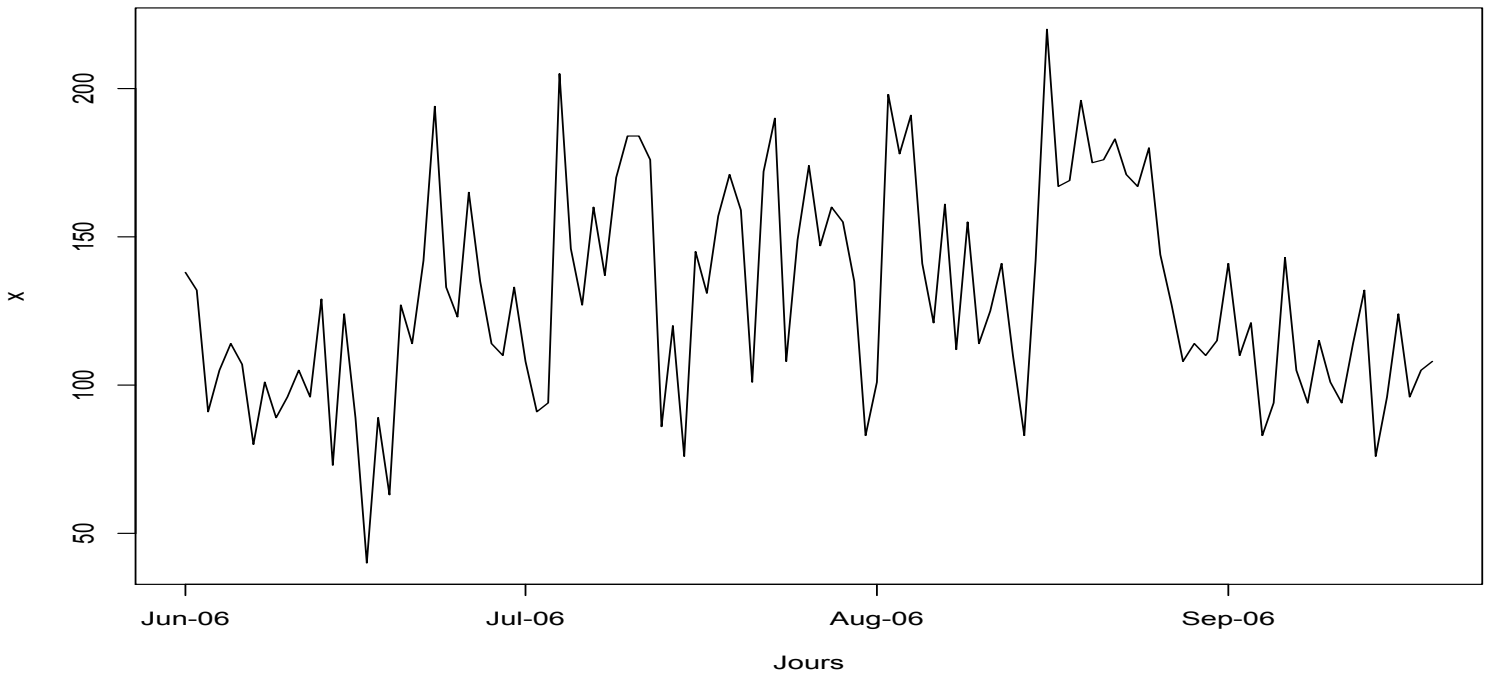
Production de bière en Australie.

Une donnée par mois de 1956 à 1995.



Maximum journalier d'ozone à Rennes.

Une donnée par jour de juin à septembre 2006.



L'étude sous R se fait par lignes de commandes.

Supposons que la série temporelle est la variable `var` du tableau `tab`.

• On commence par définir la série temporelle sous R :

```
x = ts(tab$var, start=1980, freq=12)
```

ts ("time series") : permet à R de voir la variable comme une série temporelle
start (optionnel) : permet de préciser la date de début de la série
freq (optionnel) : précise la période de la série, s'il y en a une (ici on suppose que la période vaut 12, ce qui est typique de données mensuelles).

• On peut alors travailler avec la série sous R, notamment la représenter avec

```
plot(x)
```

• La légende en abscisse se gère en général toute seule. On peut néanmoins imposer une datation personnelle, par exemple comme ceci :

```
dates=seq(as.POSIXlt("2006/6/7"),as.POSIXlt("2006/9/25"),"days")
plot(dates, x, type="l", xaxt="n",xlab='Jours')
r = as.POSIXct(round(range(dates), "days"))
axis.POSIXct(1, at=seq(r[1], r[2], by="month"), format="%b-%y")
```

5 Aspects temporels

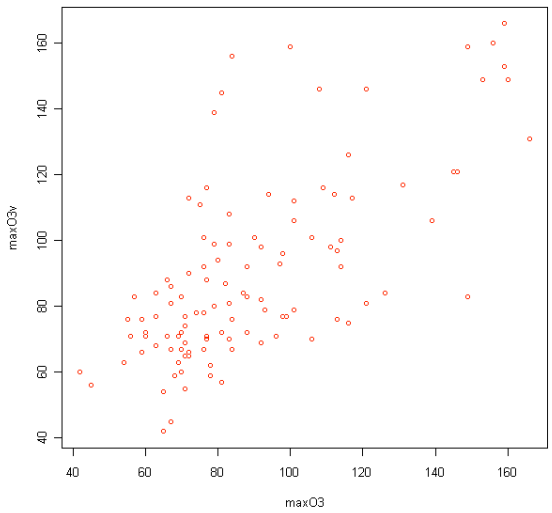
- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

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

Cela peut se vérifier en calculant la corrélation ou en représentant le nuage de points entre les valeurs de la série et ses valeurs au pas de temps précédent.

Exemple :
Pour la série 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 vaut $r = 0.68$.
- La série est donc corrélée positivement à son passé immédiat.



On note x_t la valeur de la série à l'instant t, observée pour $t = 1, \ldots, n$.

- La **fonction d'autocovariance** est définie pour tout entier h de 0 à n par:

$$\sigma(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} (x_t - \bar{x})(x_{t+h} - \bar{x}),$$

où $\bar{x}$ désigne la moyenne empirique de la série : $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$.

Remarque :

  - la sommation s'arrête à $n - h$ car au-delà il n'y a plus de valeurs disponibles pour calculer x_{t+h}.
  - $\sigma(0)$ correspond à la variance de la série.

- La **fonction d'autocorrélation** (ACF en anglais) est définie pour tout entier h de 0 à n par :

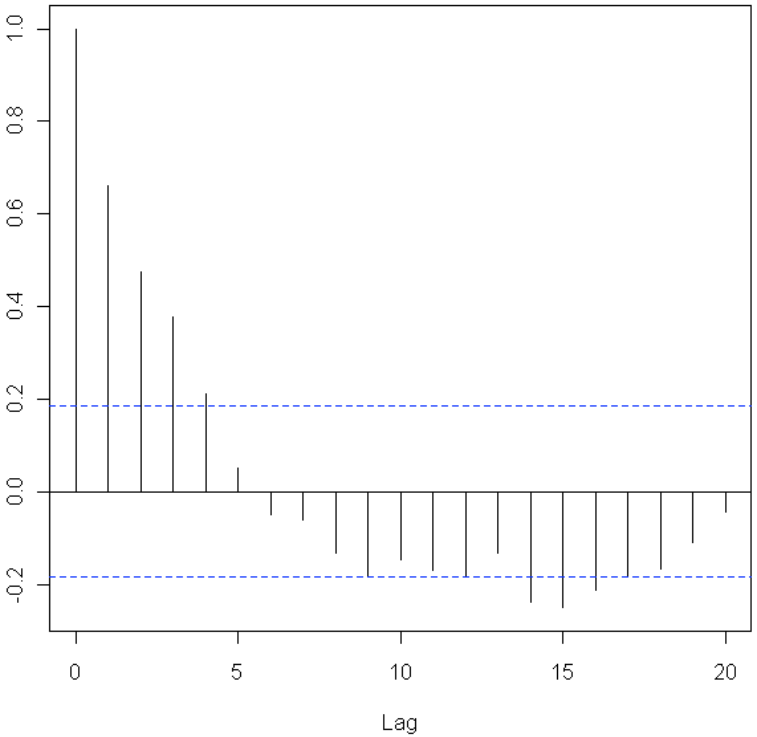
$$\rho(h) = \frac{\sigma(h)}{\sigma(0)}.$$

Remarque :
$\rho(h)$ calcule la corrélation entre les valeurs de la série et les valeurs "h plus tard".

Les ACF : exemple

Sous R : la fonction acf(x) calcule et représente les ACF de la série x.

Exemple : les ACF de la série des max d'ozone.



Chaque baton représente $\rho(h)$ où h ="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.

Remarque : les batons entre les pointillés peuvent être considérés négligeables.

5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

Tendance et saisonnalité

On observe généralement dans une série temporelle :

- une **tendance**, déterministe, qui représente le comportement moyen de la série au cours du temps. On la note m_t.
  Par exemple, pour une tendance linéaire : $m_t = at + b$.

- une **saisonnalité**, déterministe, de période T, qui représente un comportement périodique (ou saisonnier) de la série (par exemple des pics de ventes tous les mois de décembre pour une série mensuelle).
  On la note s_t et on a : $s_{t+T} = s_t$, pour tout t.
  La connaissance de $s_1, \ldots, s_T$ fournit le **profil saisonnier**.

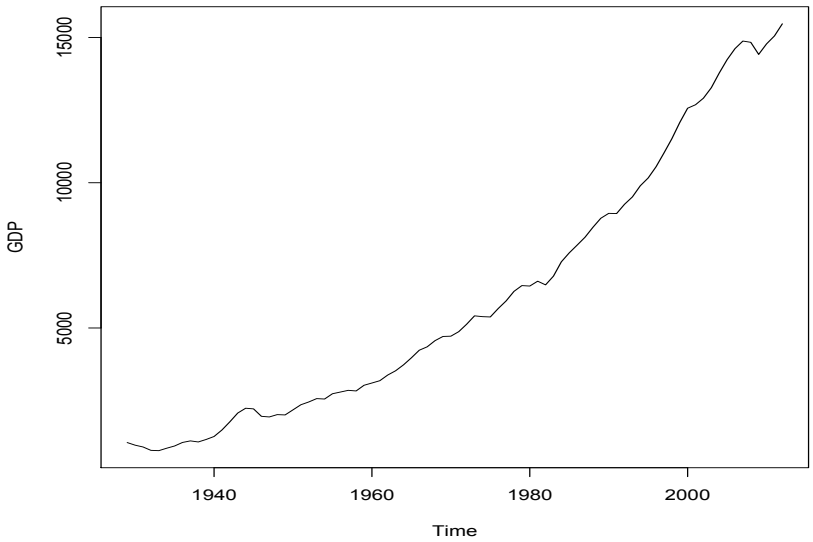
- un **reste aléatoire**, qui contient les variations aléatoires au cours du temps, ces dernières pouvant être liées entre elles dans le temps.

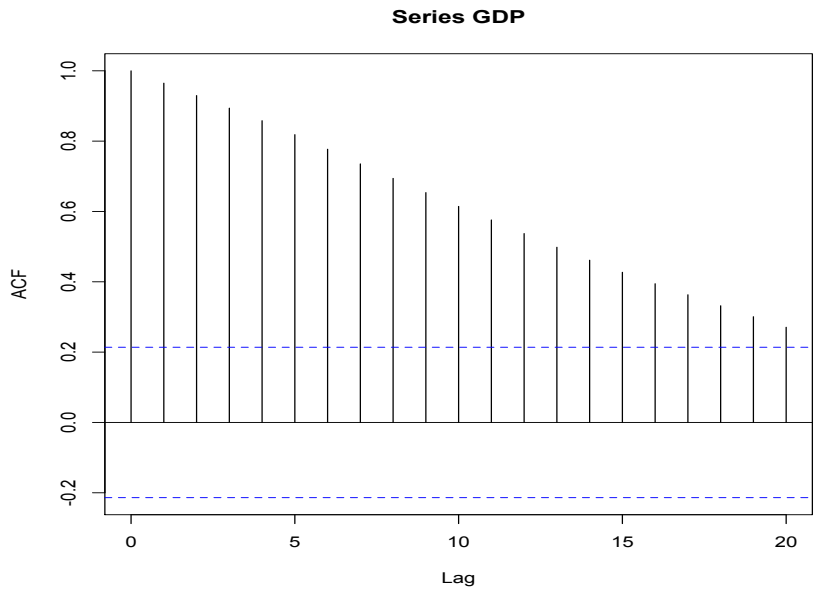
La tendance et la saisonnalité (avec sa période T) sont généralement visibles sur la représentation graphique de la série. On peut également détecter leur présences sur les ACF. Voir les exemples ci-dessous.

PIB des USA de 1929 à 2012

Série avec tendance croissante, non linéaire.

$\longrightarrow$ tendance croissante visible sur la série.
$\longrightarrow$ les ACF décroissent lentement et restent non négligeables

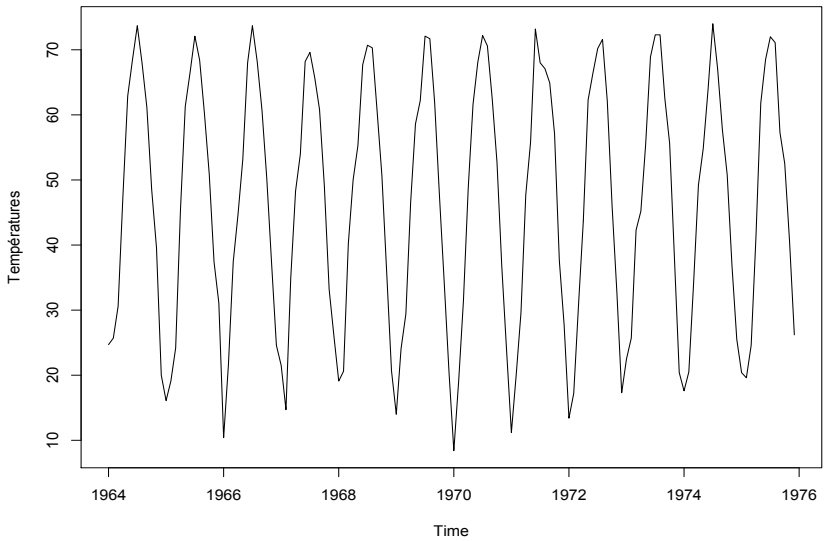


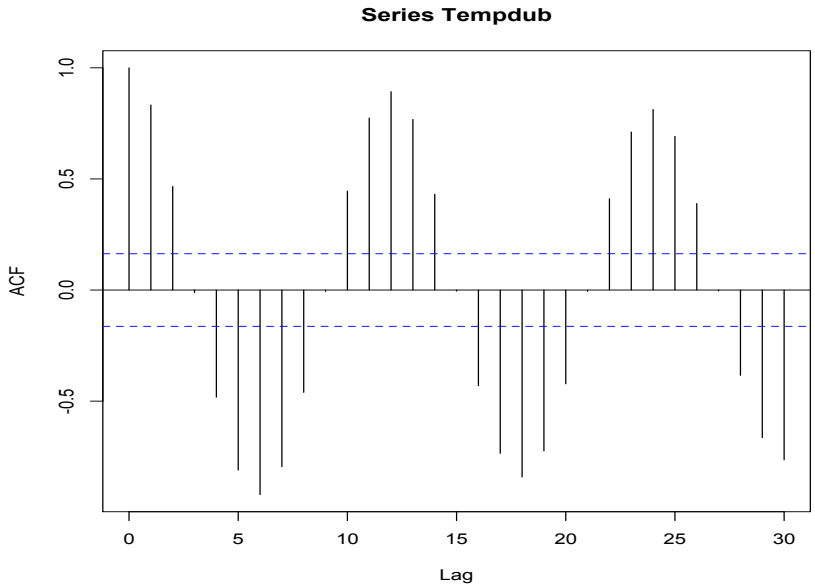


Températures moyennes mensuelles à Dubuque, Iowa, de 1964 à 1975

Série avec saisonnalité de période $T = 12$.

$\longrightarrow$ aspect saisonnier de période 12 visible sur la série.
$\longrightarrow$ les ACF ont également un comportement périodique de période 12

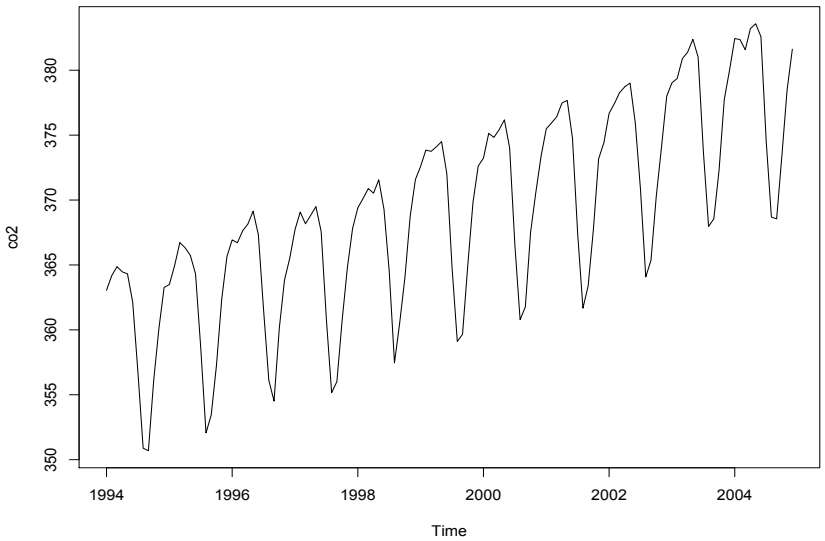


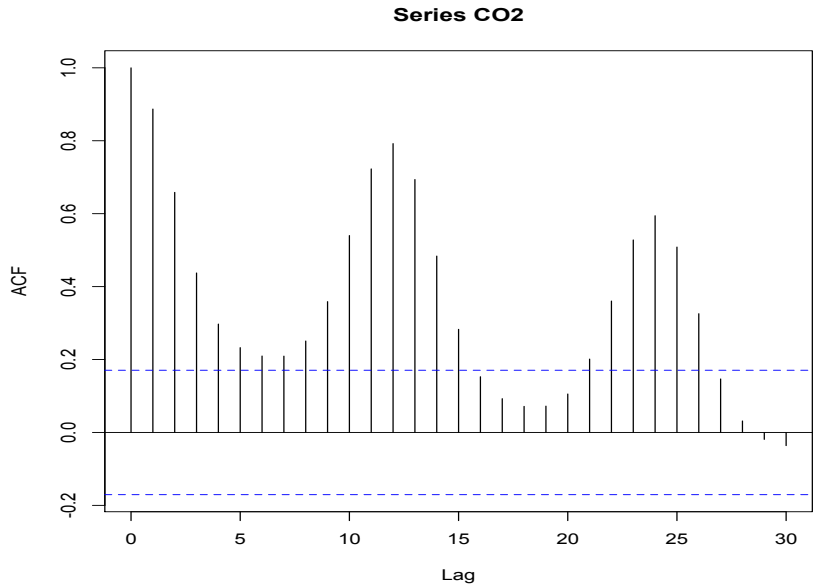


Concentration de CO2 mensuelle à Hawaï de 1994 à 2004

Série avec tendance et saisonnalité de période $T = 12$.

$\longrightarrow$ tendance linéaire et saisonnalité de période 12 visibles sur la série.
$\longrightarrow$ les ACF ont un comportement périodique et décroissent faiblement.

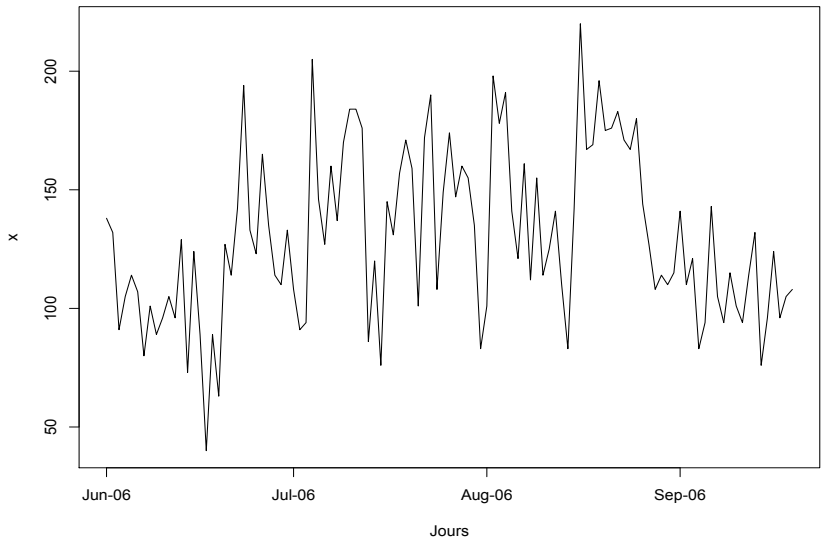


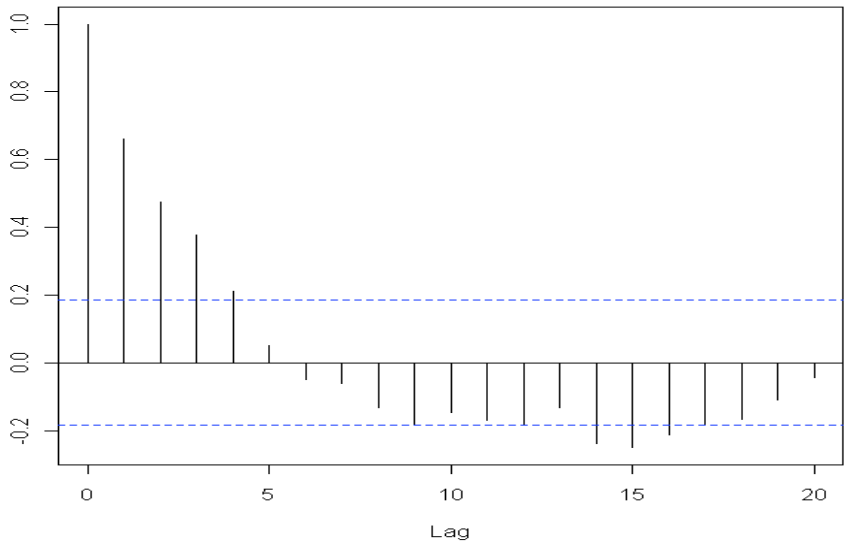


Max d'ozone journalier à Rennes l'été 2006

Série sans tendance ni saisonnalité.

$\longrightarrow$ aucune structure visible sur la série.
$\longrightarrow$ les ACF décroissent rapidement pour devenir négligeables.

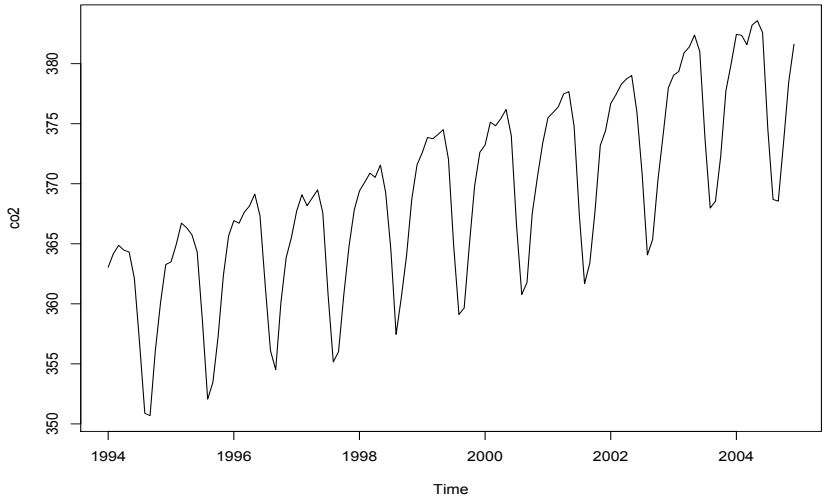


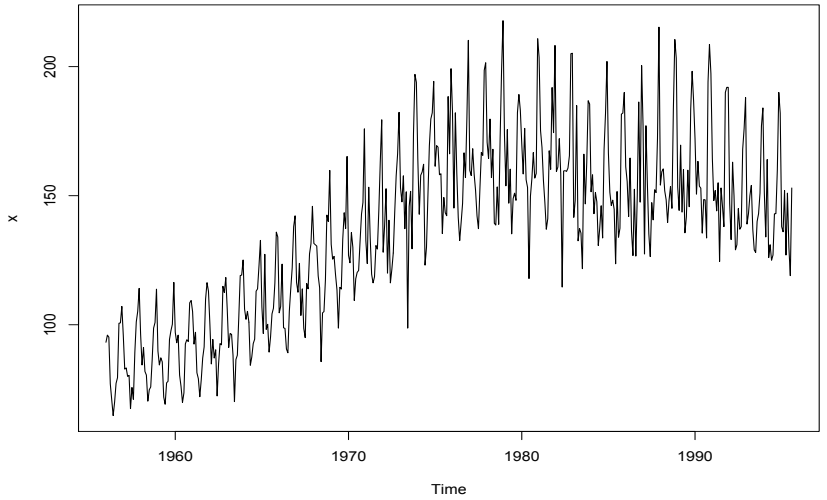


Décomposition additive et multiplicative

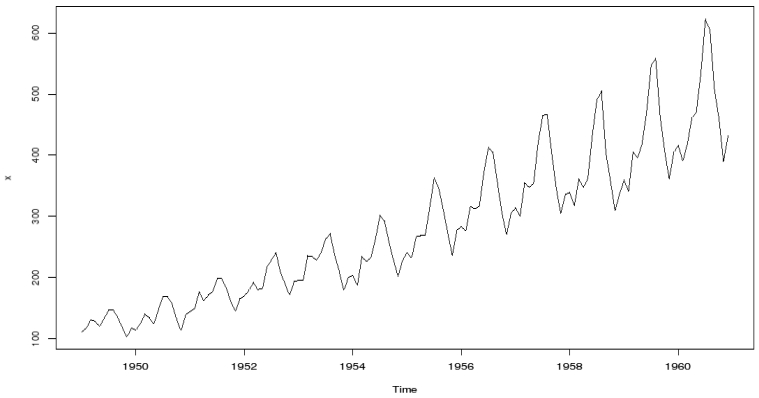
On note x_t la valeur de la série à l'instant t, observée pour différents t.
La plupart des séries observées peuvent se décomposer :

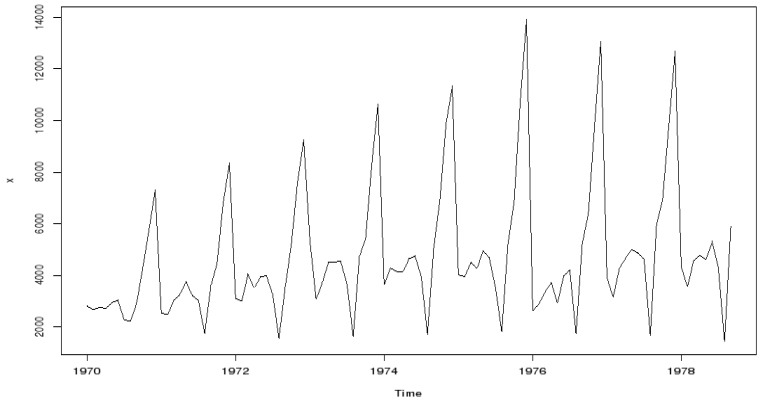
- de façon additive (la plus courante) : $x_t = m_t + s_t + r_t$





- de façon multiplicative : $x_t = m_t \times s_t \times r_t$





5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

Estimation de la tendance

On peut estimer la tendance m_t de deux façons.

- De façon paramétrique.
  On suppose que m_t admet une forme paramétrique spécifique, par exemple $m_t = at + b$ pour une tendance linéaire, et on estime les paramètres par la méthode des moindres carrés.

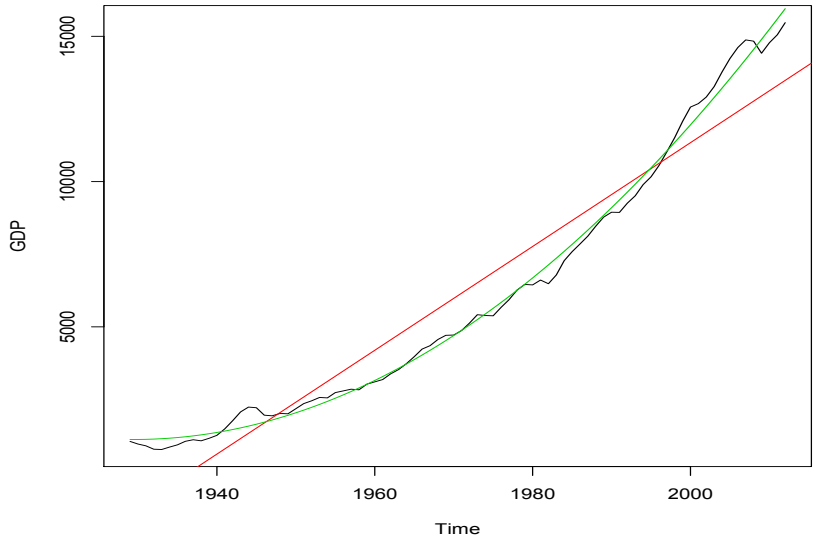
- De façon non-paramétrique.
  On effectue des moyennes mobiles de la série, ce qui équivaut à la lisser : le lissage résultant estime la tendance, mais aucune formule exprimant cette tendance n'est fournie.

Exemple : On considère la série du PIB aux USA.

On décide d'estimer la tendance m_t à l'aide de deux modèles paramétriques:

1. En supposant que m_t est linéaire : $m_t = at + b$. On en déduit une estimation de a et b par les moindres carrés. La droite estimée est représentée en rouge ci-dessous.
2. En supposant que m_t est quadratique : $m_t = at^2 + bt + c$. On estime de même a, b et c par les moindres carrés. La courbe estimée est représentée en vert ci-dessous.
   Le code R pour la tendance quadratique est donné ci-dessous.

```
temps=as.numeric(time(GDP))
temps2=temps^2
reg=lm(GDP ~ temps + temps2)
plot(GDP)
lines(temps,reg$fitted.values)
```



La **moyenne mobile** de x associée aux $k_1 + k_2 + 1$ coefficient $a_{-k_1}, \ldots, a_{k_2}$ est

$$M(t) = \sum_{i=-k_1}^{k_2} a_i x_{t+i}, \qquad t = k_1 + 1, \ldots, n - k_2,$$

où les poids somment à 1 : $\sum_{i=-k_1}^{k_2} a_i = 1$.

- Chaque valeur x_t est donc remplacée par la moyenne pondérée des valeurs autour de x_t : k_1 valeurs avant, k_2 valeurs après, soit en tout $k_1 + k_2 + 1$ valeurs moyennées.
- En pratique, $M(t)$ se calcule pour $t = k_1 + 1$ à $n - k_2$ afin que toutes les valeurs à moyenner soient accessibles.

Quelques moyennes mobiles standards:

- La **moyenne mobile arithmétique** d'ordre p pour laquelle $k_1 = p$, $k_2 = 0$ et $a_i = \frac{1}{p+1}$. Elle n'utilise que le passé de x_t.

$$\bar{M}_p(t) = \frac{1}{p+1} \sum_{i=-p}^{0} x_{t+i} = \frac{1}{p+1} \sum_{i=0}^{p} x_{t-i}, \qquad t = p+1, \ldots, n.$$

Elle est utilisée en finance afin de comparer la valeur présente d'une action x_n (à l'instant $t = n$) à sa tendance passée.

- La **moyenne mobile centrée d'ordre** p, pour laquelle $k_1 = k_2$ et
  - si $p = 2k+1$, $a_i = \frac{1}{p}$ pour tout $i = -k, \ldots, k$

$$M_{2k+1}(t) = \frac{1}{2k+1}(x_{t-k} + \cdots + x_{t+k})$$

  - si $p = 2k$, $a_{-k} = a_k = \frac{1}{2p}$ et $a_i = \frac{1}{p}$ pour $i = -(k-1), \ldots, (k-1)$.

$$M_{2k}(t) = \frac{1}{2k}\left(\frac{x_{t-k}}{2} + x_{t-k+1} + \cdots + x_{t+k-1} + \frac{x_{t+k}}{2}\right)$$

Exemple

M_3 : Moyenne mobile centrée d'ordre 3
M_4 : Moyenne mobile centrée d'ordre 4

On considère le lissage par M_3 et M_4 de la série des 9 valeurs suivantes.

$x(1)$	$x(2)$	$x(3)$	$x(4)$	$x(5)$	$x(6)$	$x(7)$	$x(8)$	$x(9)$
4	6	5	3	7	5	4	3	6

On obtient :

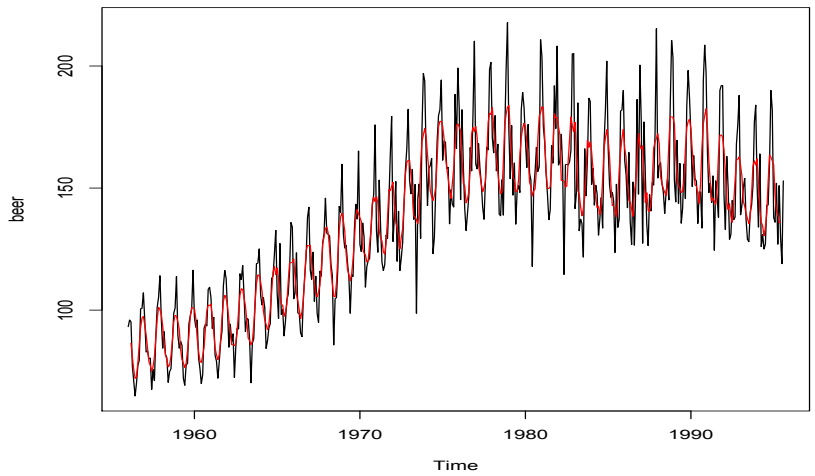
$M_3(1)$	$M_3(2)$	$M_3(3)$	$M_3(4)$	$M_3(5)$	$M_3(6)$	$M_3(7)$	$M_3(8)$	$M_3(9)$
–	5	4.67	5	5	5.33	4	4.33	–

$M_4(1)$	$M_4(2)$	$M_4(3)$	$M_4(4)$	$M_4(5)$	$M_4(6)$	$M_4(7)$	$M_4(8)$	$M_4(9)$
–	–	4.875	5.125	4.875	4.75	4.625	–	–

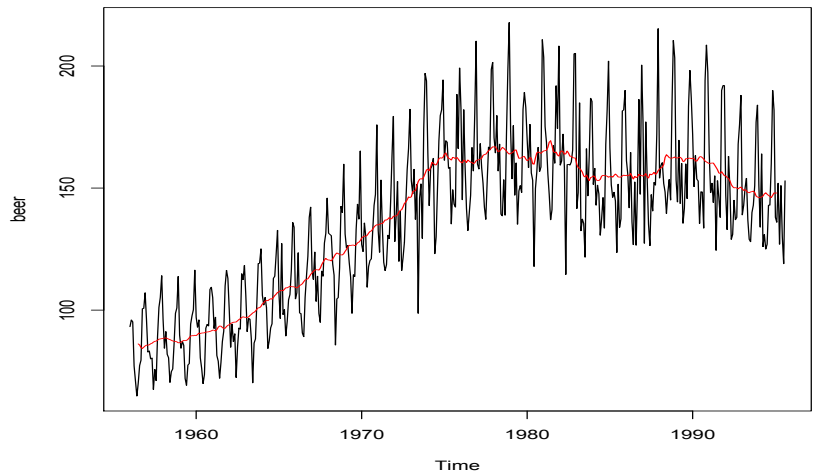
Exemples de calcul :
$M_3(2) = \frac{1}{3}\left(x(1) + x(2) + x(3)\right) = \frac{1}{3}(4 + 6 + 5) = 5$
$M_4(3) = \frac{1}{4}(\frac{x(1)}{2} + x(2) + x(3) + x(4) + \frac{x(5)}{2}) = \frac{1}{4}(\frac{4}{2} + 6 + 5 + 3 + \frac{7}{2}) = 4.875$

Application de M_p aux données de production de bières en Australie.
$\longrightarrow$ Plus l'ordre est grand, plus le lissage est important.

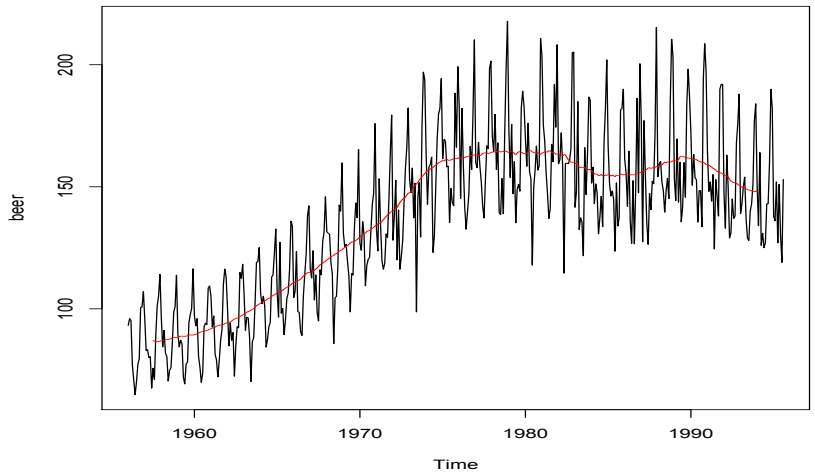
$p = 5$



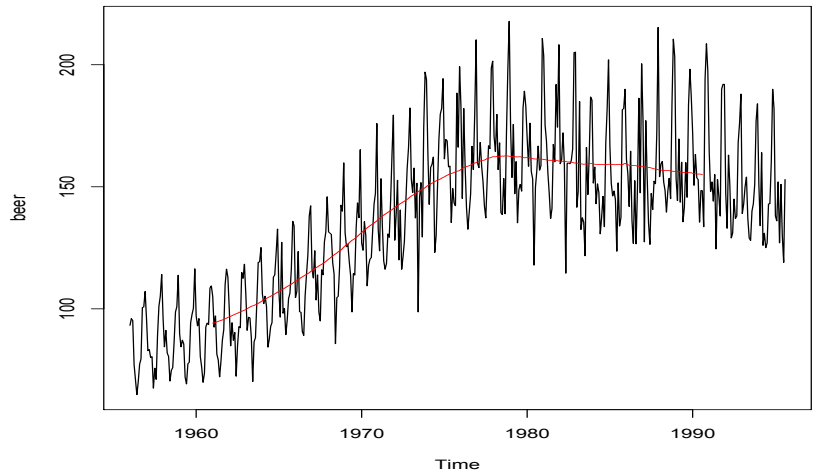
$p = 12$



$p = 36$



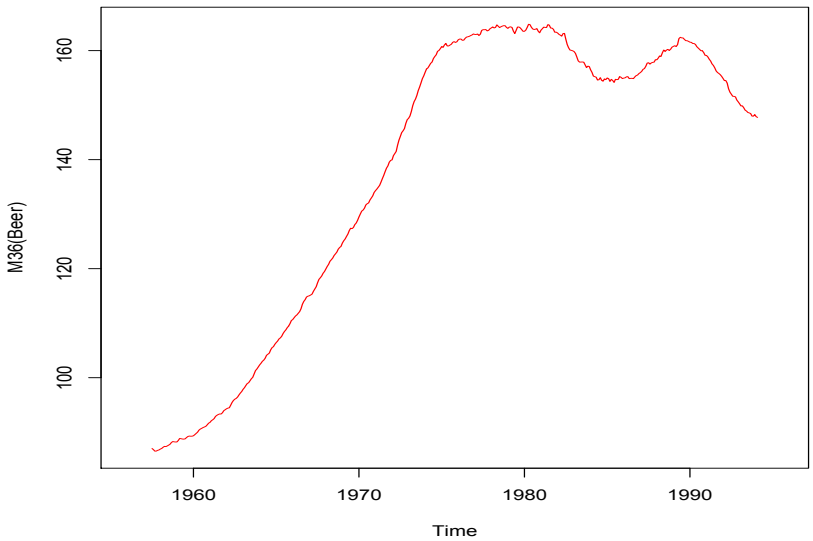
$p = 120$

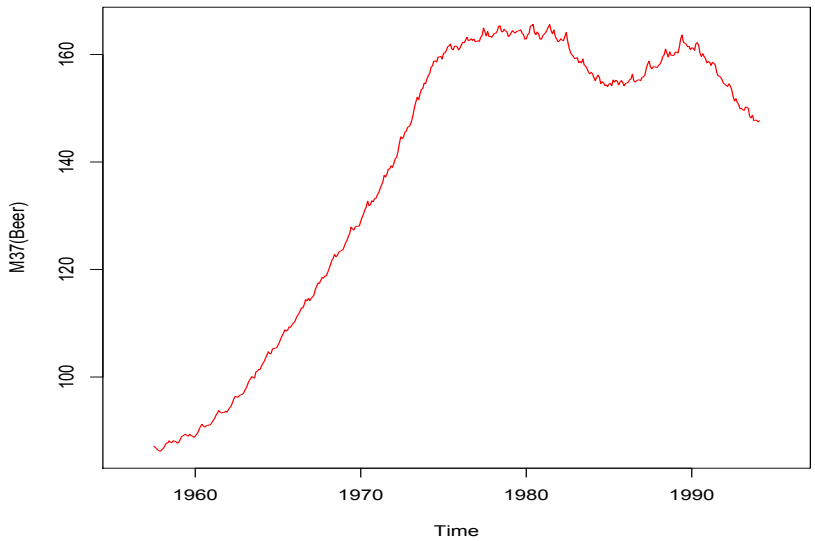


Moyenne mobile en présence d'une saisonnalité

Pour estimer la tendance en présence d'un saisonnalité de période T, il est conseillé d'utiliser une moyenne mobile centrée **d'ordre multiple de** T, pour atténuer l'effet saisonnier.

Exemple : le lissage de la série précédente (pour laquelle $T = 12$) donne pour M_{36} (à gauche) et M_{37} (à droite) :





$\Longrightarrow$ Le lissage par M_{37}, bien que d'un ordre supérieur à M_{36}, est parasité par l'effet saisonnier initial.

Sour R :
Pour appliquer la moyenne mobile centrée M_p à la série x, on utilise la commande filter, en précisant les coefficients de la moyenne mobile.

Plus précisément :

- si $p = 2k + 1$

  filter(x, rep(1/(2*k+1), 2*k+1))

  La commande `rep(a,n)` crée un vecteur contenant n fois la valeur a.

- si $p = 2k$

  filter(x, c(1/(4*k), rep(1/(2*k), 2*k-1), 1/(4*k)))

  La commande `c()` permet de concaténer dans un même vecteur un ensemble de valeurs, ici les $2k + 1$ coefficients : $\frac{1}{4k}$, $(2k - 1)$ fois $\frac{1}{2k}$, et $\frac{1}{4k}$.

5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

Pour estimer la saisonnalité, on élimine au préalable la tendance de la série. On travaille donc avec la nouvelle série $\tilde{x}_t$ où

- $\tilde{x}_t = x_t - \hat{m}_t$ si on a supposé une décomposition additive, avec $\hat{m}_t$ une estimation de la tendance.
- $\tilde{x}_t = x_t / \hat{m}_t$ dans le cas d'une décomposition multiplicative.

Le profil saisonnier $s_1, \ldots, s_T$ s'estime à partir de $\tilde{x}_t$

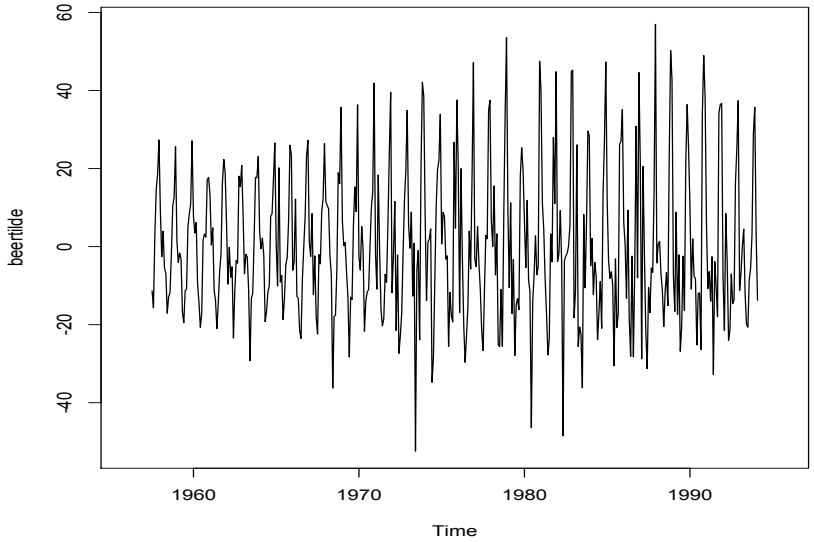
- de manière paramétrique :
  on suppose généralement que $s_t = a + b\cos(2\pi t/T) + c\sin(2\pi t/T)$ et les coefficients a, b et c sont estimés par moindres carrés.
- ou de manière non-paramétrique :
  Soit $k = 1, \ldots, T$ et soit n_k le nombre d'instants multiples de k parmi $1, \ldots, n$. On estime le profil saisonnier de la manière suivante:

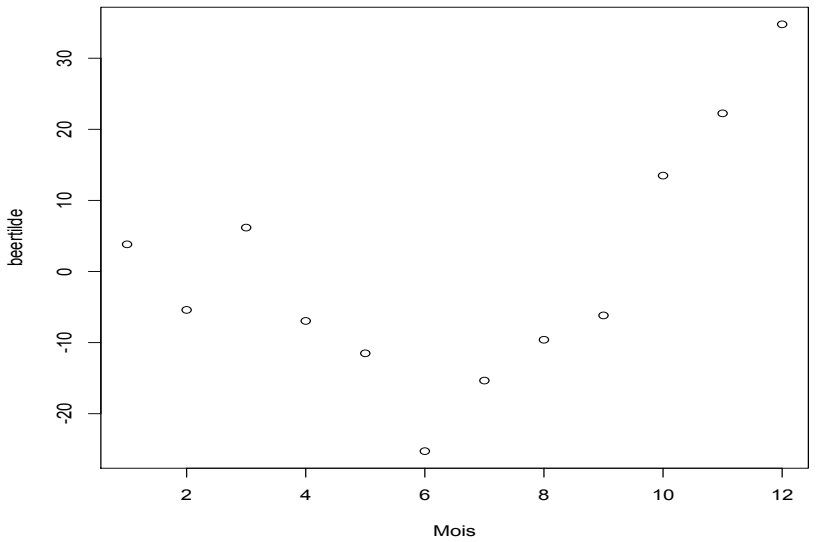
$$\hat{s}_k = \frac{1}{n_k} \sum_{i=0}^{n_k - 1} \tilde{x}_{i+kT}.$$

  La série complète $\hat{s}_1, \ldots, \hat{s}_n$ s'obtient par périodicité en répétant $\hat{s}_1, \ldots, \hat{s}_T$.
  Exemple : pour une série mensuelle de période $T = 12$, $\hat{s}_1$ correspond à la moyenne des mois de janvier, $\hat{s}_2$ à la moyenne de mois de février, etc.

Exemple : Cas d'une décomposition additive

La tendance de la série de production de bières a été estimée avec M_{36}
La série **moins** sa tendance est représentée à gauche.
Le profil saisonnier, à droite, est déduit en moyennant chaque mois.

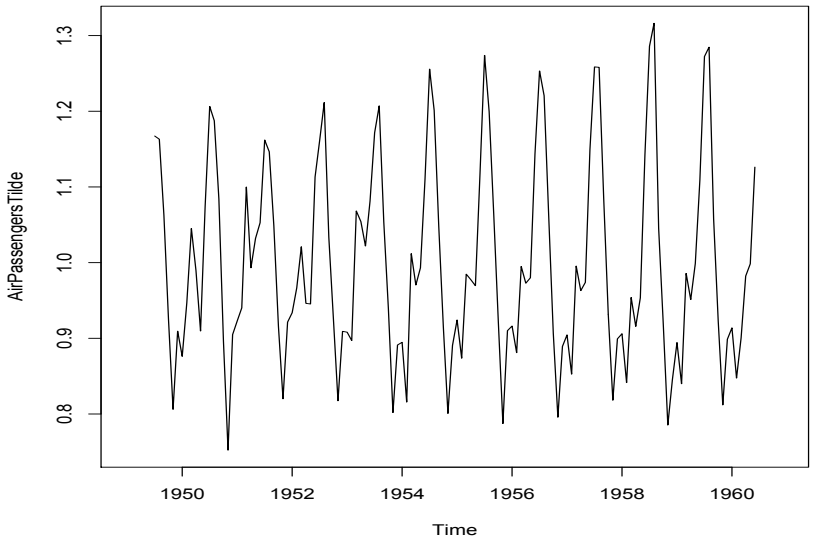


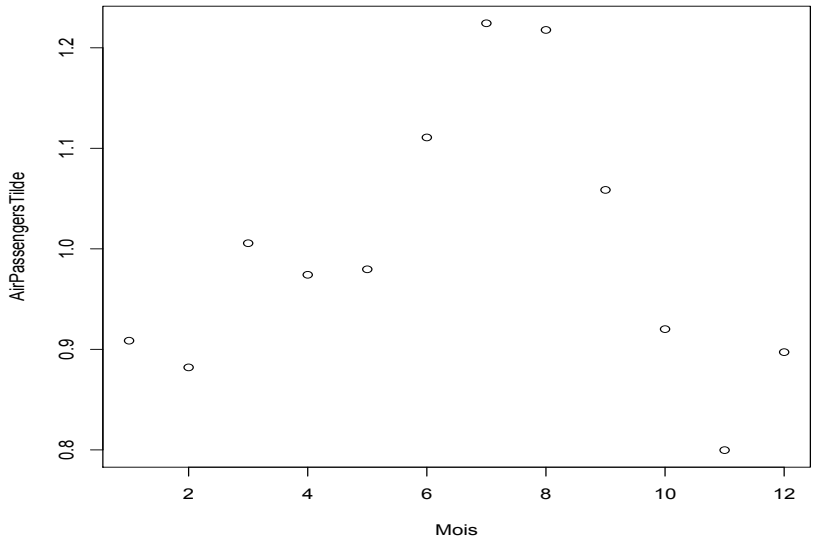


Si x est la série sans tendance : `profil=aggregate(x~cycle(x),FUN=mean)`

Exemple : Cas d'une décomposition multiplicative

La tendance de la série de trafic aérien a été estimée avec M_{12}
La série **divisée** par sa tendance est représentée à gauche.
Le profil saisonnier, à droite, est déduit en moyennant chaque mois.





5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

Décomposition totale d'une série

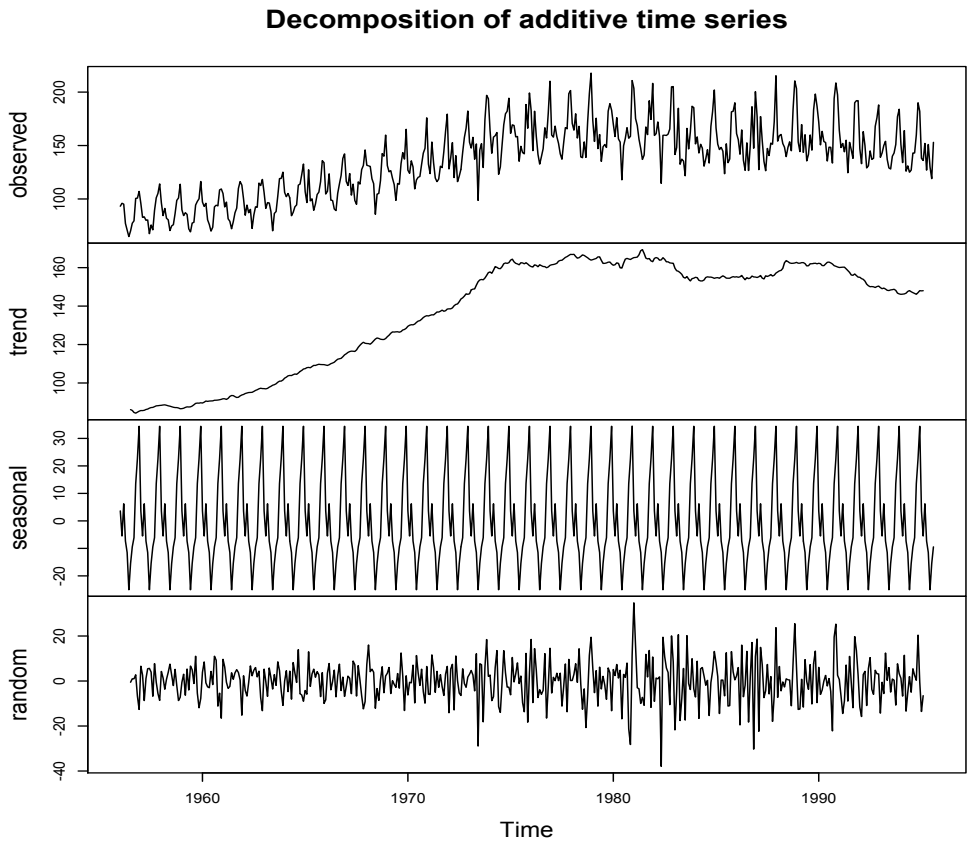
En supposant que la série s'écrit $x_t = m_t + s_t + r_t$, on peut estimer chacune des composantes:

1. On estime m_t par $\hat{m}_t$ (de façon paramétrique ou par moyennes mobiles)
2. On estime s_t à partir de $x_t - \hat{m}_t$, ce qui donne $\hat{s}_t$
3. On en déduit une estimation du reste $\hat{r}_t = x_t - \hat{m}_t - \hat{s}_t$

Pour obtenir le graphe ci-contre:

```
plot(decompose(x))
```

Pour récupérer la tendance :
`decompose(x)$trend`
Pour récupérer la saisonnalité :
`decompose(x)$seasonal`
Pour récupérer le reste :
`decompose(x)$random`
Pour récupérer le profil saisonnier :
`decompose(x)$figure`

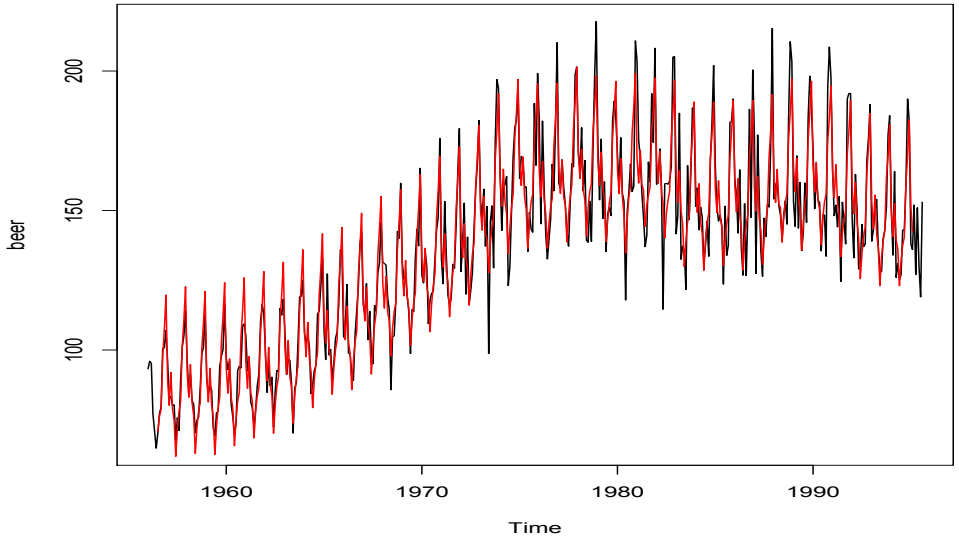


Remarque : le même type de décomposition est possible dans le cas multiplicatif avec

`plot(decompose(x,"multiplicative"))`

Série ajustée et série CVS

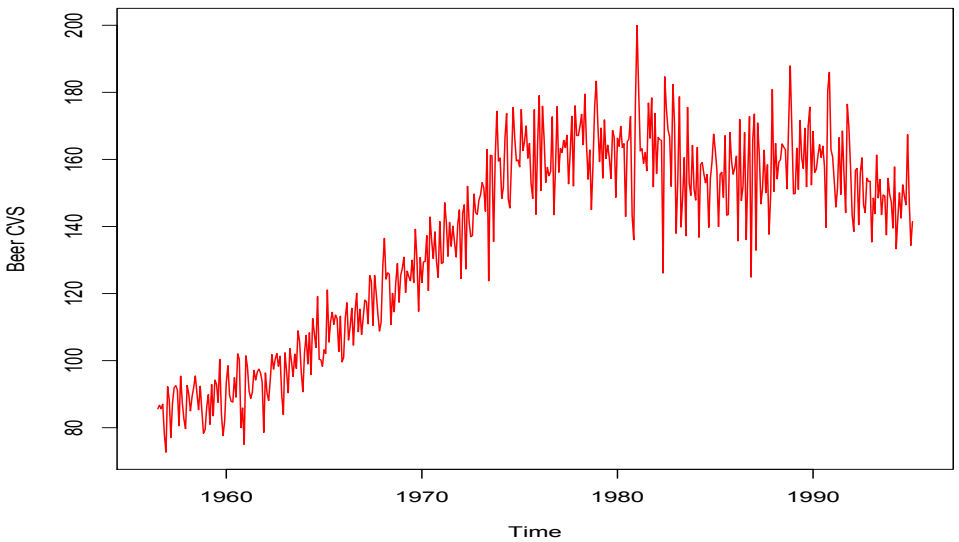
La **série ajustée** $\hat{m}_t + \hat{s}_t$ est un lissage de x_t qui respecte la tendance et la saisonnalité.



La **série CVS** (corrigée des variations saisonnières) correspond à $x_t - \hat{s}_t = \hat{m}_t + \hat{r}_t$.

Elle permet d'analyser les variations de la série sans être influencé par l'aspect saisonnier.

On l'utilise parfois pour réestimer m_t de façon plus fine.



5 Aspects temporels

- Représentation graphique
- Dépendance temporelle, la fonction d'autocorrélation (ACF)
- Tendance et saisonnalité
- Estimation de la tendance
- Estimation de la saisonnalité
- Série ajustée, Série CVS (Corrigée des Variations Saisonnières)
- Pour aller plus loin : prévision et modélisation

On suppose une décomposition additive : $x_t = m_t + s_t + r_t$, où s_t est de période T.

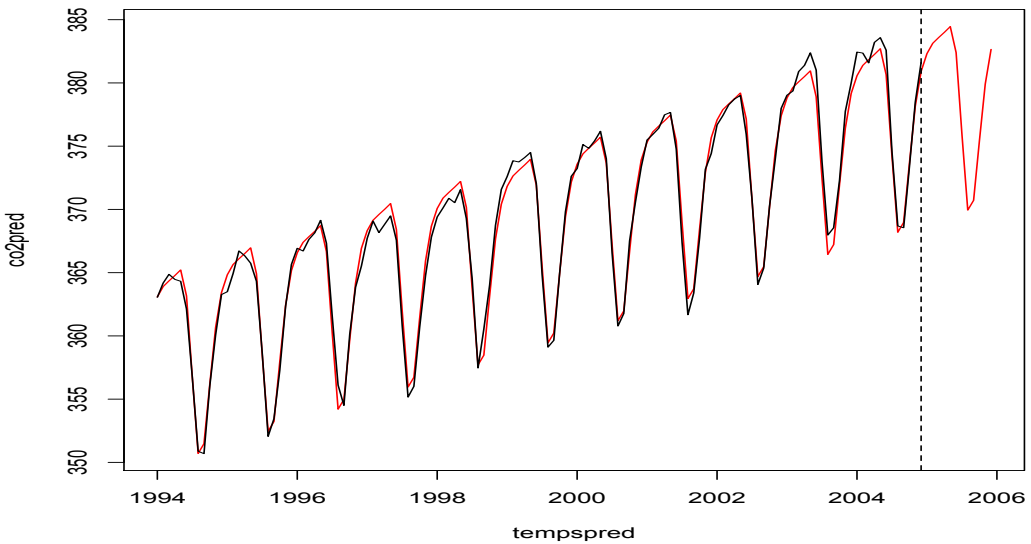
- Une prévision de la série est possible grâce à la série ajustée $\hat{m}_t + \hat{s}_t$, pourvu que $\hat{m}_t$ ait une forme paramétrique.

  Exemple : si $\hat{m}_t = \hat{a}t + \hat{b}$, alors la prévision de la série en $t = n+1$ est

  $$\hat{x}_{n+1} = \hat{m}_{n+1} + \hat{s}_{n+1} = \hat{a}(n+1) + \hat{b} + \hat{s}_{n+1-T}$$

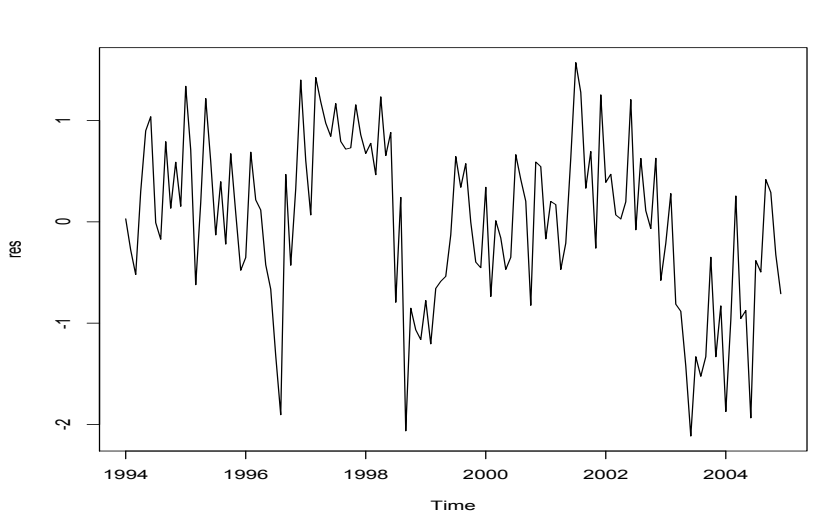
  où la dernière égalité est obtenue par périodicité de $\hat{s}_t$.

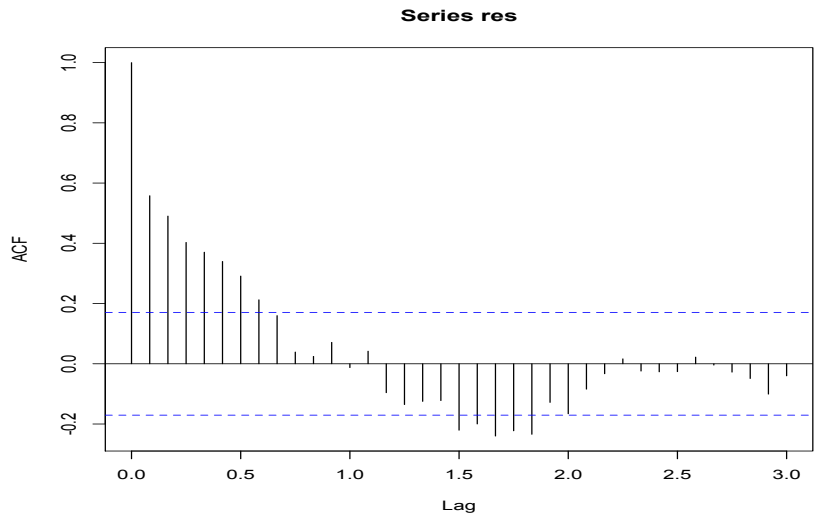
  Application : série de CO2 à Hawaï.
  En noir : la série initiale, observée de 1994 à 2004.
  En rouge : la série ajustée $\hat{a}t + \hat{b} + \hat{s}_t$, prolongée en 2005



- Pour améliorer la prévision, il convient d'essayer de prédire également le reste aléatoire r_t, estimé par $\hat{r}_t = x_t - \hat{m}_t - \hat{s}_t$.

  Exemple : $\hat{r}_t$ pour la série précédente (série noire - série rouge) et ses ACF





  Les ACF montre que $\hat{r}_t$ est corrélé avec son passé.
  $\longrightarrow$ On souhaite tirer parti de cette dépendance pour prédire $\hat{r}_t$ et améliorer la prévision en 2005.
  $\longrightarrow$ Cela se fait en modélisant la dépendance de $\hat{r}_t$, par exemple à l'aide de modèles ARMA : voir le cours d'économétrie.