# Chapitre A

# Série statistique 1 variable

## Sommaire



Dans tout ce chapitre, on considèrera les 3 séries statistiques suivantes :

**Série A :**
Notes obtenues à un contrôle dans une classe de 30 élèves :
2 – 3 – 3 – 4 – 5 – 6 – 6 – 7 – 7 – 7 – 8 – 8 – 8 – 8 – 8 – 9 – 9 – 9 – 9 – 9 – 9 – 10 – 10 – 11 – 11 – 11 – 13 – 13 – 15 – 16

**Série B :**
Salaires en euros des employés d'une entreprise :

| Salaires | [900;1200[ | [1200;1400[ | [1400;1600[ | [1600;1800[ | [1800;2000[ | [2000;2400[ | TOTAL |
|---|---|---|---|---|---|---|---|
| Effectif | 30 | 30 | 60 | 40 | 20 | 20 | 200 |

**Série C :**
Proportion d'adhérents à un club sportif dans différentes sections :

- 17% jouent au handball,
- 25% houent au rugby,
- 58% jouent au tennis.

# I) Méthodes de représentation

## 1) Vocabulaire

**vocabulaire:**

- La **population** est l'ensemble des individus sur lesquels portent l'étude statistique.(Par exemple votre classe de BTS , la population féminine, les fonctionnaires ...dont chaque élément est appelé **individu**.
- Un **échantillon** est une partie de la population considérée.
- Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu :

  Lorsque le caractère ne prend que des valeurs (ou **modalités**) numériques, il est **quantitatif** :
  — **discret** s'il ne peut prendre que des valeurs isolées (notes, âge ...)
  — **continu** dans le cas contraire (poids, taille ...). Dans ce cas on effectue souvent un regroupement des valeurs par **classes**.

  Sinon, on dit qu'il est **qualitatif** (couleur des yeux, sport pratiqué ...) : les caractères ne sont pas des nombres.
- A chaque valeur (ou classe) est associée un **effectif** $n$ : c'est le nombre d'individus associés à cette valeur.

Faire des **statistiques**, c'est recueillir, organiser, synthétiser, représenter et exploiter des données, numériques ou non, dans un but de comparaison, de prévision, de constat ...

Les plus gros "consommateurs" de statistiques sont les assureurs (risques d'accidents, de maladie des assurés), les médecins (épidémiologie), les démographes (populations et leur dynamique), les économistes (emploi, conjoncture économique), les météorologues ...

## 2) Tableaux

**Fréquence:**

On considère une série statistique X à caractère quantitatif, dont les $p$ valeurs sont données par $x_1, x_2, \ldots, x_p$ d'effectifs associés $n_1, n_2, \ldots, n_p$ avec $n_1 + n_2 + ... + n_p = \mathrm{N}$. ?

- A chaque valeur (ou classe) est associée une fréquence $f_i$ : c'est la proportion d'individus associés à cette valeur.
- $f_i = \dfrac{n_i}{\mathrm{N}}$ est un nombre compris entre 0 et 1, que l'on peut écrire sous forme de pourcentage.
- L'ensemble des fréquences de toutes les valeurs du caractère s'appelle la distribution des fréquences de la série statistique.

**Exemple:**

*On peut représenter la **série A** par un tableau d'effectifs, et le compléter par la distribution des fréquences :*

| *Notes* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Eff.* | 0 | 1 | 2 | 1 | 1 | 2 | 3 | 5 | 6 | 2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 |
| *Fréq. en %* | 0 | 3 | 7 | 3 | 3 | 7 | 10 | 17 | 20 | 7 | 10 | 0 | 7 | 0 | 3 | 3 | 0 | 0 | 0 |

**Remarques:**

- *On peut vérifier que la somme des fréquences est égale à 1 (ou à 100 si on les exprime en pourcentages).*
- *On peut aussi faire un regroupement par classe, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.*

**Exemple:**

*Toujours pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient :*

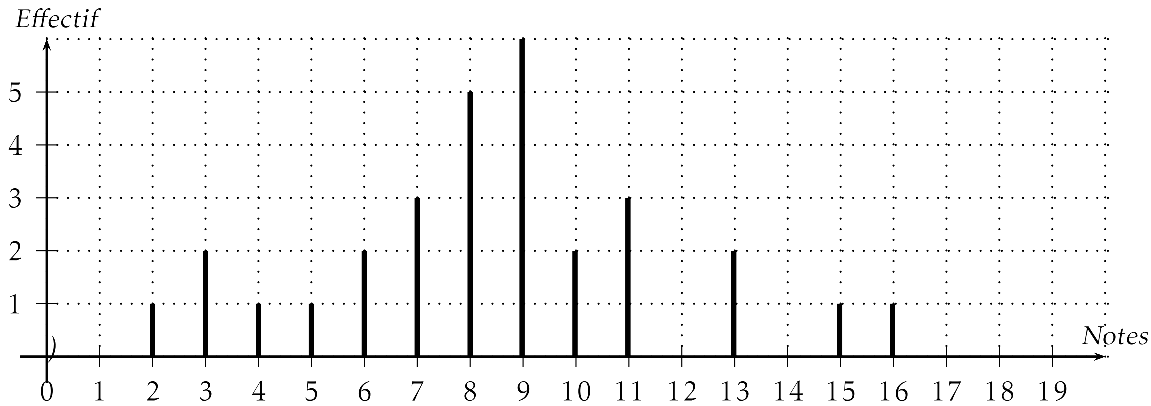
| *Notes* | [ 0 ; 5 [ | [ 5 ; 10 [ | [ 10 ; 15 [ | [ 15 ; 20 [ | *total* |
|---|---|---|---|---|---|
| *Effectif* | 4 | 17 | 7 | 2 | 30 |
| *Fréquence* | 0,13 | 0,57 | 0,23 | 0,07 | 1 |

## 3) Graphiques

Lorsque le caractère étudié est **quantitatif et discret**, on peut représenter la série statistique étudiée par un **diagramme en bâtons** : la hauteur de chaque bâton est alors proportionnelle à l'effectif (ou à la fréquence) associé à chaque valeur.

**Exemple:**

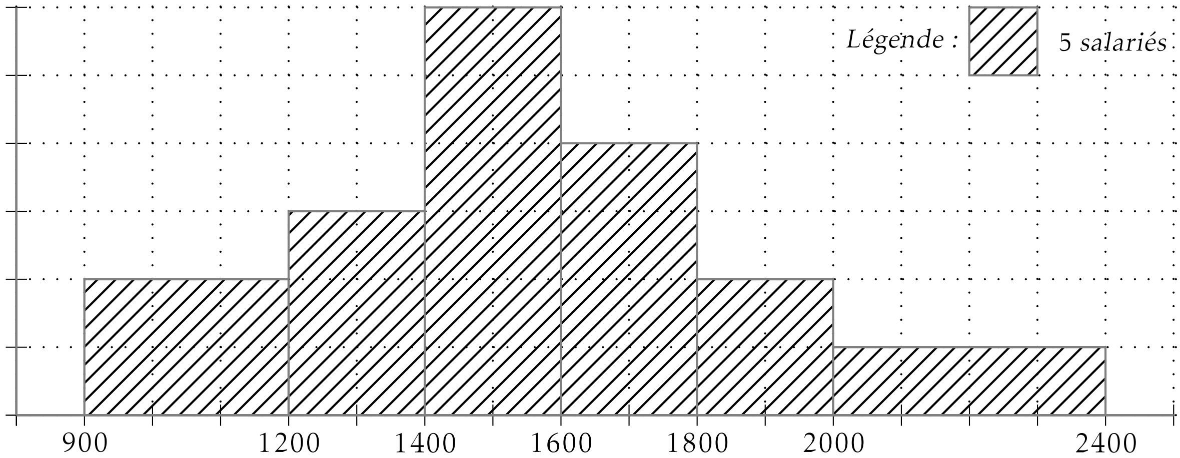
*Voici le diagramme en bâtons représentant la série des notes de la **série A** :*




Lorsque le caractère étudié est **quantitatif et continu**, et lorsque les modalités sont regroupées en classes, on peut représenter la série par un **histogramme** : l'aire de chaque rectangle est alors proportionnelle à l'effectif (ou à la fréquence) associée à chaque classe.
Lorsque les classes ont la même **amplitude**, c'est la hauteur qui est proportionnelle à l'effectif.

**Exemple:**

*Pour la **série B**, on obtient par exemple l'histogramme suivant :*




Enfin, lorsque le caractère est **qualitatif**, on peut représenter la série par :

• **Un diagramme circulaire** (« camemberts ») :
La mesure de chaque secteur angulaire est proportionnelle à l'effectif associé.

• **Un diagramme en tuyaux d'orgue :**
Chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence.

• **un diagramme en bandes :**
Chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence.

**Exemple:**

*Diagrammes de la **série C***

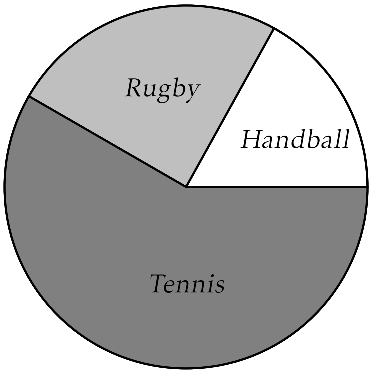



*Diagramme circulaire*

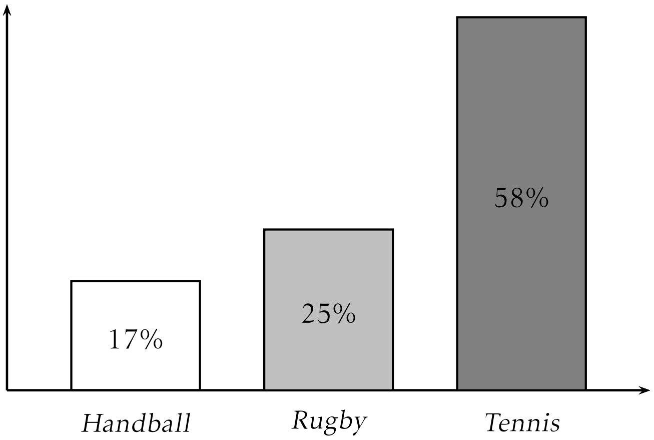



*Diagramme en tuyau d'orgue*

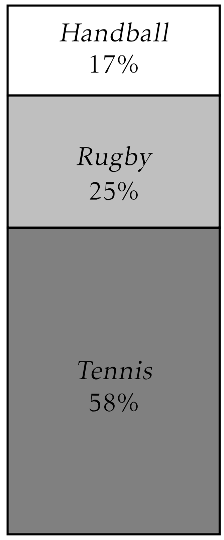



*Diagramme en bandes*

## II) Caractéristiques de position

Dans le premier paragraphe, on a commencé à condenser les informations pour les rendre plus lisibles. Dans ce deuxième paragraphe, on va synthétiser encore davantage l'information pour les caractères quantitatifs en cherchant quelques nombres permettant de décrire au mieux la population observée.

### 1) Moyenne

**Moyenne pondérée:**

Soit une série statistique à caractère quantitatif, dont les $p$ valeurs sont données par $x_1, x_2, \ldots, x_p$ d'effectifs associés $n_1, n_2, \ldots, n_p$ avec $n_1 + n_2 + ... + n_p = \mathrm{N}$.
La moyenne pondérée de cette série est le nombre noté $\overline{x}$ qui vaut

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p} = \frac{1}{\mathrm{N}}\sum_{i=1}^{p} n_ix_i.$$

**Remarque:**

*Lorsque la série est regroupée en classes, on calcule la moyenne en prenant pour valeurs $x_i$ le **centre de chaque classe** ; ce centre est obtenu en faisant la moyenne des deux extrémités de la classe.*

**Exemple:**

- *Dans la **série A**, la moyenne du contrôle est égale à* $\overline{m} = \dfrac{254}{30} \approx 8{,}47$.
- *Dans la **série B**, une estimation du salaire moyen est donné par :* $\overline{S} = \dfrac{460500}{280} \approx 1644{,}64$.

**Remarque:**

*On peut aussi calculer une moyenne à partir de la distribution de fréquences :*

$$\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p = \sum_{i=1}^{p} f_i x_i.$$

**Propriété A1:**

- *Si on ajoute (ou soustrait) un même nombre $k$ à toutes les valeurs d'une série, alors la moyenne de cette série se trouve augmentée (resp. diminuée) de $k$.*
- *Si on multiplie (ou divise) par un même nombre non nul $k$ toutes les valeurs d'une série, alors la moyenne de cette série se trouve multipliée (resp. divisée) par $k$.*

**Exemple:**

*On considère la **série A** :*

- *Si on ajoute $1{,}5$ points à chaque note du contrôle, alors la moyenne de classe devient* $\overline{m} = 8{,}47 + 1{,}5 = 9{,}97$.
- *Si on augmente chaque note de $10\%$, cela revient à multiplier chaque note par $1{,}1$, ce qui donne* $\overline{m} = 8{,}47 \times 1{,}1 = 9{,}32$.

**Propriété A2:**

*Soit une série statistique, d'effectif total $N$, de moyenne $\overline{x}$.*
*Si on divise cette série en deux sous-groupes **disjoints** d'effectifs respectifs $p$ et $q$ (avec $p + q = N$) de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$, alors on a :*

$$\overline{x} = \frac{p}{N} \times \overline{x}_1 + \frac{q}{N} \times \overline{x}_2.$$

**Exemple:**

*On suppose par exemple que les 12 garçons de la classe de la **série A** ont obtenu une moyenne globale de 8 sur 20.*

- *La moyenne du groupe formé par les filles de la classe vérifie :* $9{,}47 = \frac{12}{30} \times 8 + \frac{18}{30} \times \overline{m}_f$ .
- *Soit* $\overline{m}_f = \frac{30}{18}\left(9{,}47 - \frac{12}{30} \times 8\right) = 10{,}45$.

## 2) Médiane

**Médiane:**

Soit une série statistique ordonnée dont les $n$ valeurs sont $x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$.
La médiane est un nombre M qui permet de diviser cette série en deux sous-groupes de même effectif.

- Si $n$ est impair, $n$ est la valeur de cette série qui est située au milieu, à savoir la valeur dont le rang est $\frac{n+1}{2}$, notée $x_{\frac{n+1}{2}}$.
- Si $n$ est pair, $n$ est le centre l'intervalle médian, qui est l'intervalle formé par les deux nombres situés « au milieu » de la série, à savoir $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$.

**Exemple:**

- *La médiane de la série « $2-5-6-8-9-9-10$ » est 8.*
- *La médiane de la série « $2-5-6-8-9-9$ » est 7.*
- *La médiane de la série « $2-5-6-6-9-10$ » est 6.*

**Exemple:**

*On souhaite calculer la médiane de la **série A**.*

- *Pour cela, on commence par remplir le tableau des effectifs cumulés croissants :*

| *Notes* | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Eff.* | 0 | 1 | 2 | 1 | 1 | 2 | 3 | 5 | 6 | 2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 |
| *ECC.* | 0 | 1 | 3 | 4 | 5 | 7 | 10 | 15 | 21 | 23 | 26 | 26 | 28 | 28 | 29 | 30 | 30 | 30 | 30 |

- *Ensuite, l'effectif étant de 30, on chosit la moyenne entre la* $15^{ime}$ *et la* $16^{ime}$ *note.*
  *On obtient* $\text{Med} = \frac{8+9}{2} = 8{,}5$.
- *Ce qui signifie que la moitié des notes est inférieure ou égale à 8,5, et que l'autre moitié des notes est supérieure*
  *ou égale à 8,5.*

**Méthode pratique :**

*Dans le cas de répartition par classes, la médiane peut être évaluée soit graphiquement, soit par interpolation affine à l'aide d'un polygône des effectifs cumulés.*

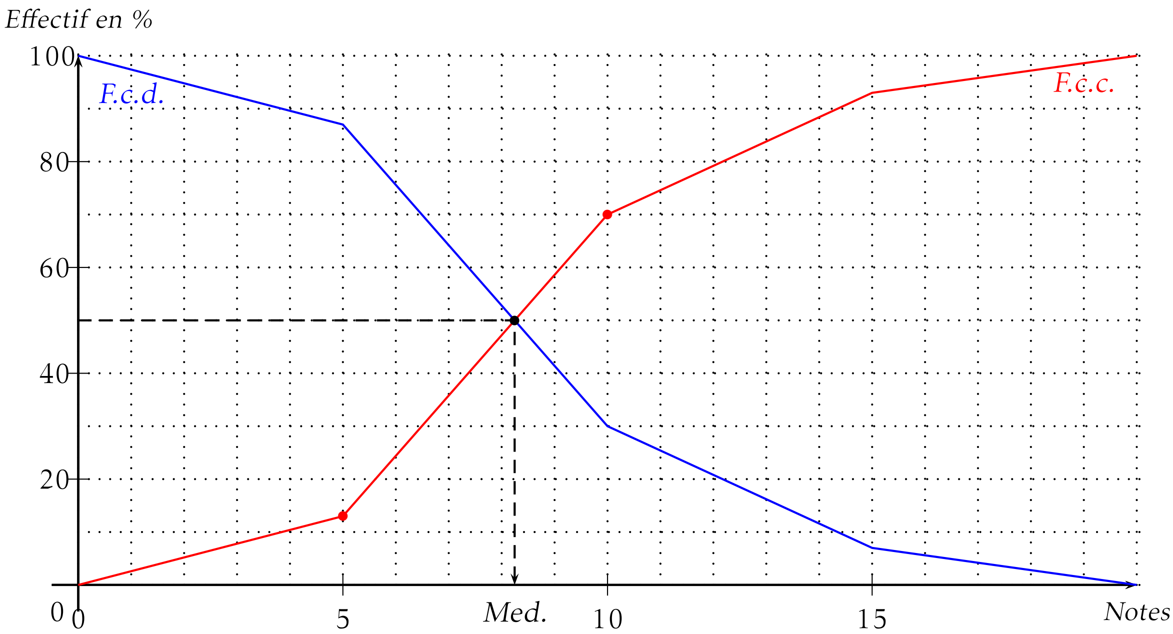
**Exemple:**

*On choisit la répartition par classes de la **série A** :*

- *On commence par créer le tableau des fréquences cumulées croissantes :*
  *(On en profite aussi pour indiquer les fréquences cumulées décroissantes).*

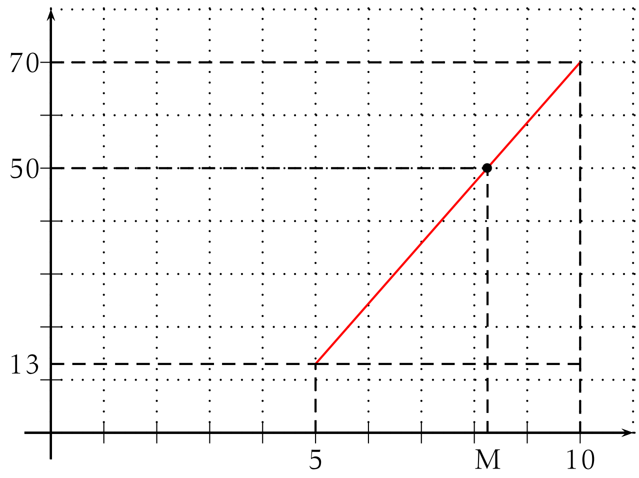
| *Notes* | [ 0 ; 5 [ | [ 5 ; 10 [ | [ 10 ; 15 [ | [ 15 ; 20 [ |
|---|---|---|---|---|
| *Fréq. en %* | 13 | 57 | 23 | 7 |
| *F.c.c.* | 13 | 70 | 93 | 100 |
| *F.c.d.* | 87 | 43 | 7 | 0 |

- *Puis on place les points correspondants aux extrémités de chaque classe sur un graphique :*




- *On détermine le point du polygône d'ordonnée* 50% *et on trouver eniron* 8,2.
- *Pour trouver la médiane, on peut aussi tracer le polygône des fréquences cumulées décroissantes et lire l'abscisse*
  *du point de concours des deux polygônes. On trouve aussi* 8,2.
- *Enfin, par le calcul,* 50% *se situe dans l'intervalle* [ 5 ; 10 [.
  *On fait l'hypothèse que les longeurs des axes sont uniformément réparties dans cette classe.*
  *On peut alors procéder à une interpolation linéaire d'après le théorème de Thalès :*

**Exemple:**




$$\frac{M-5}{10-5} = \frac{50-13}{70-13} \Longleftrightarrow \frac{M-5}{5} = \frac{37}{57} \Longleftrightarrow M = 5\times\frac{37}{57} + 5 = \frac{470}{57} \approx 8{,}25.$$

## 3) Quartiles, déciles ...

**Quartiles-Déciles:**

Soit une série statistique.

- On appelle quartiles de la série un triplet de réels ( $Q_1$ ; $Q_2$ ; $Q_3$ ) qui sépare la série en quatre groupes de même effectif.
- On appelle déciles de la série un 9-uplet de réels ( $D_1$ ; $D_2$ ; ... ; $D_9$ ) qui sépare la série en dix groupes de même effectif.

**Remarque:**

*Par définition, si* $X$ *est une série statistique,* $Q_2 = D_5 = Med(X)$.

**Méthode pratique :**

*Le calcul des valeurs des quartiles ou des déciles se fait en général à partir des graphiques des effectifs (ou fréquences) cumulés croissants, par interpolation linéaire.*
*La calculatrice donne les valeurs de* $Q_1$*, Med et* $Q_3$*.*

**Exemple:**

- *Pour la **série A**, la calculatrice nous donne* $Q_1 = 7$, $Md = 8{,}5$ *et* $Q_3 = 10$.
- *Graphiquement, on trouve* $D_1 \approx 3{,}8$ *et* $D_9 \approx 14{,}2$.
- *Pour la **série B**, on trouve* $Q_1 = 1500$, $Med = 1700$ *et* $Q_3 = 1900$.

# III) Caractéristiques de dispersion

## 1) Étendue

Il s'agit de la première mesure de la dispersion d'une série statistique. Son principal mérite a longtemps été d'exister, et de fournir une information sur la dispersion très simple à obtenir.

**Étendue:**

Soit X une série statistique discrète. On appelle étendue de la série le réel, défini par $Etd(X) = \max(X) - \min(X)$.

**Exemple:**

*L'étendue de la **série A** est de* $16 - 2 = 14$.

## 2) Intervalle inter-quartile

**Intervalle inter-quartile:**

On appelle intervalle inter-quartiles l'intervalle $[\,Q_1\,;\,Q_3\,]$.
L'amplitude de cet intervalle est appelée écart inter-quartiles.

**Exemple:**

- *Dans la **série A**, l'intervalle interquartile est l'intervalle* $[\,7\,;\,10\,]$ *dont l'écart vaut* $10 - 7 = 3$.
- *Cet intervalle comprend donc la moitié des notes de la série située au centre de celle-ci.*

## 3) Variance d'une série statistique

**Variance:**

La variance d'une série statistique est le nombre noté $V(x)$ obtenu comme moyenne des carrés des écarts constatés par rapport à la moyenne de la série :

$$V(X) = \frac{n_1(x_1-\bar{x})^2 + n_2(x_2-\bar{x})^2 + \ldots + n_p(x_p-\bar{x})^2}{n_1+n_2+\ldots+n_p} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\bar{x})^2.$$

**Remarque:**

*Cette formule s'applique bien sûr au cas d'une série statistique sans coefficients : on est ramené à une série pour laquelle tous les coefficients valent* 1*.*

**Propriété A3:**

*On utilise aussi la formule de Koenig :*

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2.$$

## 4) Écart-type d'une série statistique

**Écart-type:**

L'écart-type d'une série statistique X, noté $\sigma(X)$, est la racine carrée de la variance de cette série :

$$\sigma(X) = \sqrt{V(X)}.$$

**Exemple:**

*L'écart-type de la* ***série B*** *vaut :* $\sigma(X) = \sqrt{109561} = 331$.

**Propriété A4:**

*La variance et l'écart-type présentent les propriétés suivantes :?*

- *La variance et l'écart-type sont des nombres positifs ou nuls,*
- *Une variance nulle ou un écart-type nul signifient que toutes les valeurs de la série son égales à sa moyenne,*
- *Plus la variance (ou l'écart-type) d'une série est grande, plus cette série est dispersée autour de sa moyenne,*

# Chapitre B

# Statistiques à 2 variables

## Sommaire

Le problème qui se pose dans les séries statistiques à deux variables est principalement celui du lien qui existe ou non entre chacune de ces deux variables.

# I) Vocabulaire

Ce cours est élaboré à partir de l'exemple suivant :

**Exemple:**

*Le tableau suivant donne l'évolution du nombre d'adhérents d'un club de rugby de* 2001 *à* 2006.

| Année | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|
| Rang $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 |
| Nombre d'adhérents $y_i$ | 70 | 90 | 115 | 140 | 170 | 220 |

*Le but est d'étudier cette série statistique à deux variables (le rang et le nombre d'adhérents) afin de prévoir l'évolution du nombre d'adhérents pour les années suivantes.*

## 1) Nuage de points

La première étape consiste à réaliser un graphique qui traduise les deux séries statistiques ci-dessus.

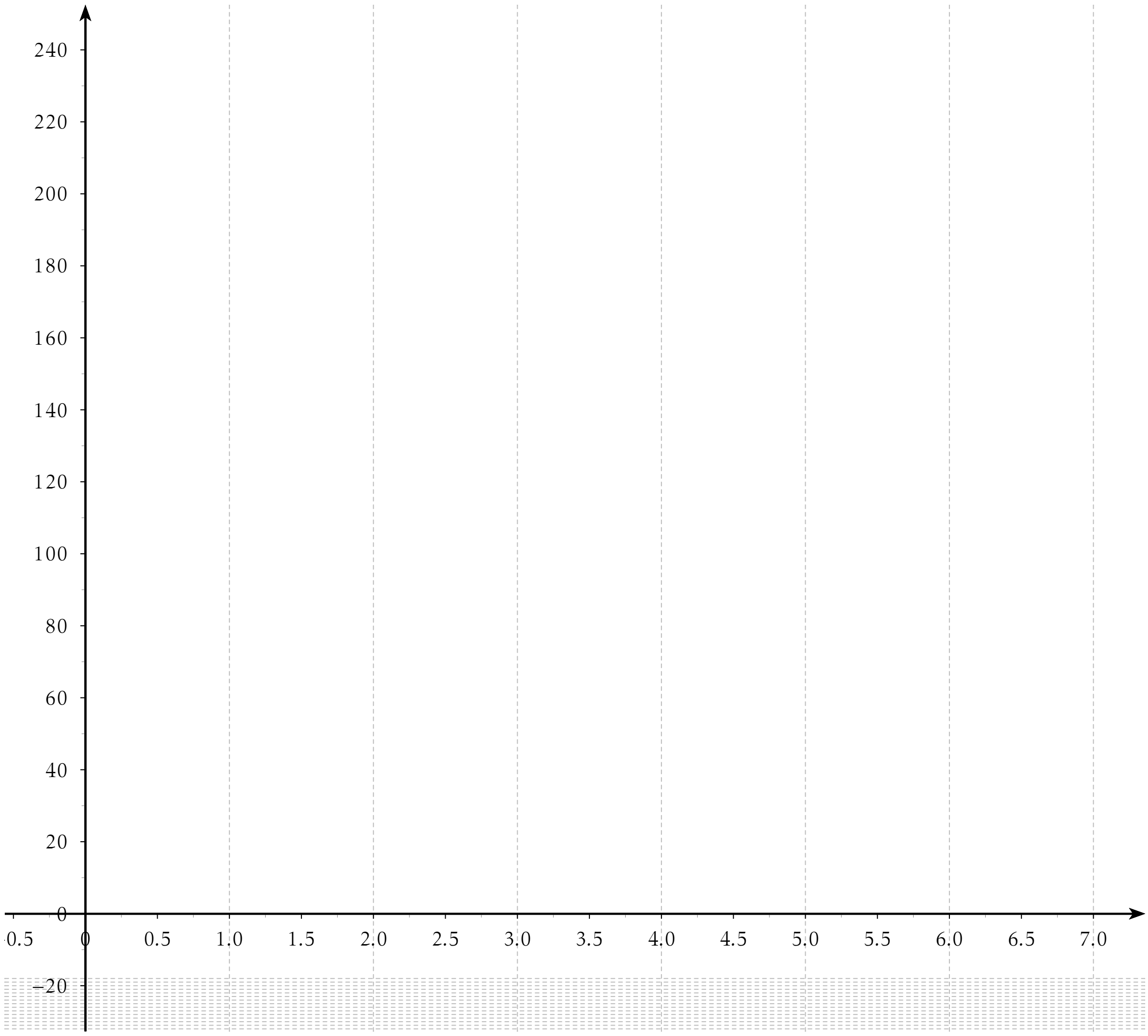
**Nuage de points:**

Soient X et Y deux variables statistiques numériques observées sur $n$ individus. Dans un repère orthogonal, l'ensemble des $n$ points de coordonnées $(x_i, y_i)$ forment le **nuage de points** associé à cette série statistique.

**Exemple:**

*Dans notre exemple, si on place le rang en abscisses, et le nombre d'adhérents en ordonnées, on peut représenter par un point chaque valeur. On obtient ainsi une succession de points, dont les coordonnées sont :* $(1;70)$, $(2;90)$, ... $(6;220)$, *forment un nuage de points.*

**Exercice B1** *Dans le repère ci dessous, représenter le nuage de points associé à la série .*




**Méthode pratique :**

*Pour entrer une série statistique à deux variables sur calculatrice :*

| *T.I.* | *Casio* |
|---|---|
| *Touche* STAT | |
| *Menu EDIT* | *Menu STAT* |
| *Entrer les valeurs $x_i$ dans L1* | *Entrer les valeurs $x_i$ dans List1* |
| *Entrer les valeurs $y_i$ dans L2* | *Entrer les valeurs $y_i$ dans List2* |
| *Régler les valeurs du repère avec la touche* WINDOWS | *Choisir* GRPH |
| | *Régler les paramètres avec* SET |
| *Appuyer sur la touche* TRACE | *Choisir* GPH1 |

## 2) Le problème de l'ajustement

Le tracer met en évidence la possibilité de "reconnaître" graphiquement la possibilité d'une relation fonctionnelle entre les deux grandeurs observées (ici rang et nombre d'adhérent).

Le problème de l'établissement de cette relation fonctionnelle entre les deux séries est un **problème de l'ajustement**.

### 3) Point moyen

**Point moyen:**

Soit une série statistique à deux variables, X et Y , dont les valeurs sont des couples $(x_i ; y_i)_{1\leqslant i\leqslant n}$. On appelle point moyen de la série le point G de coordonnées

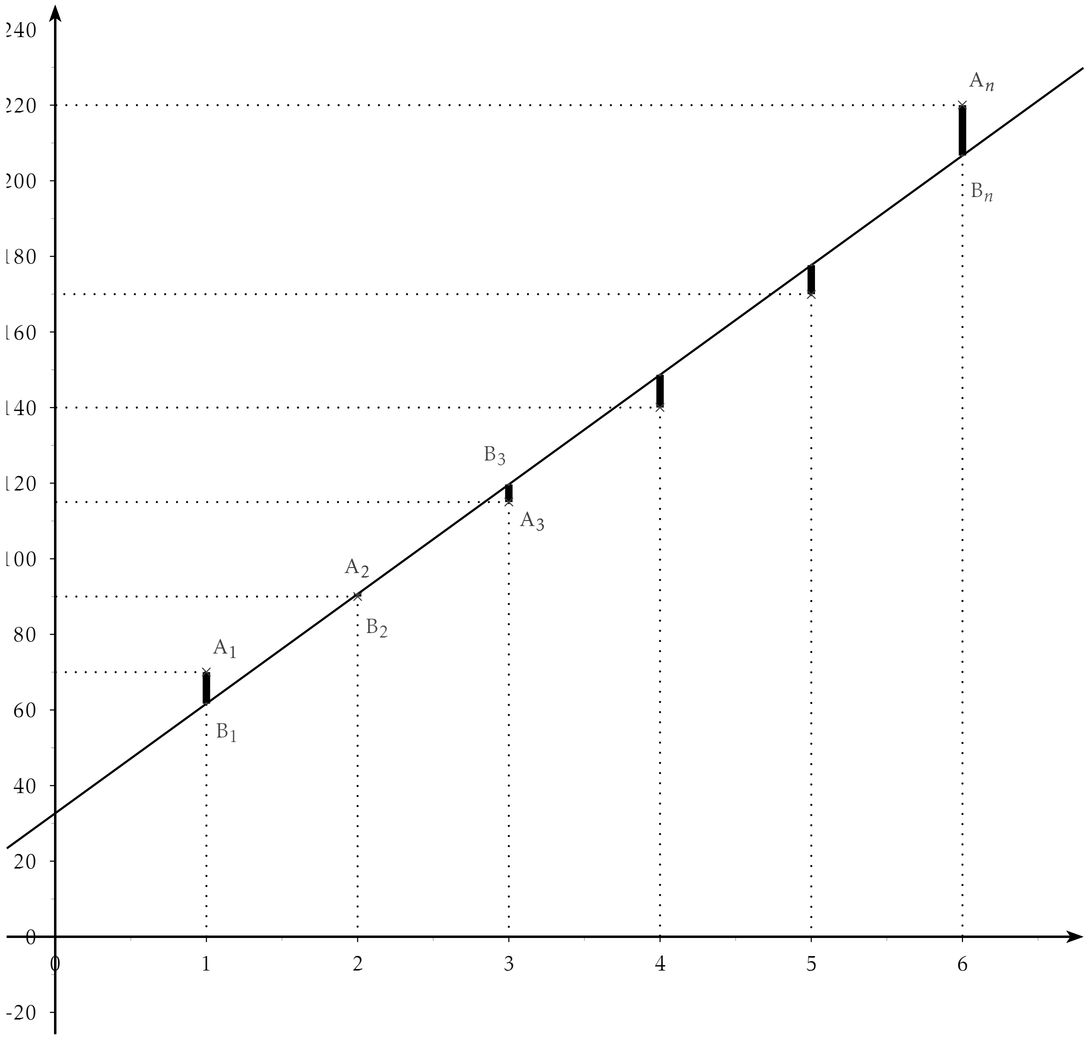
$$G\left(\frac{x_1+\cdots+x_n}{n} ; \frac{y_1+\cdots+y_n}{n}\right)$$

**Exercice B2** *Déterminer les coordonnées des points moyens suivants :*

1. $G_1$ *des années allant de* 2001 *à* 2003
2. $G_2$ *des années allant de* 2004 *à* 2006
3. G *point moyen du nuage de points tout entier.*

## II) Différents ajustements

### 1) Ajustement affine : Méthode des moindres carrés



Pour cette méthode, on va calculer l'erreur commise que l'on va ensuite essayer de minimiser. Pour la méthode des moindres carrés, on va calculer la somme des carrés des longueurs des segments verticaux noirs (on aurait pu s'intéresser aux écarts horizontaux) :

$$E_{a,b} = A_1B_1^2 + A_2B_2^2 + \ldots + A_nB_n^2 = \sum_{i=1}^{n} A_iB_i^2$$

Le but est donc de trouver $a$ et $b$ pour rendre cette erreur minimale.

**Méthode pratique :**

*On estime la relation :* $Y = aT + b$ *via la méthode des moindres carrées, en donnant comme valeurs à la variable T :* $1, 2, \ldots, n$.

$$\begin{cases} a = \dfrac{COV(T;Y)}{V(T)} \\ b = \bar{Y} - a\bar{T} \end{cases}$$

*En BUT, c'est Excel/TI qui va se charger des calculs :*

*Sur T.I.*

*Touche* STAT
*Menu* CALC
4 :LinReg(ax+b) ENTER *L1,L2*

*Sur Casio*

*Menu STAT*
CALC SET
*2VarX List : List1*
*2VarY List : List2*
*2VarX Freq : 1*
*L'affichage des résultats se fait avec* REG X

**Exercice B3** *Déterminer une équation de la droite d'ajustement obtenue par la méthode des moindres carrés.*

**Remarque:**

*Le point moyen G du nuage appartient toujours à la droite de régression de y en x.*
*Ne vous privez pas de le vérifier à chaque fois !*

**Coefficient de corrélation linéaire:**

Afin d'apprécier la qualité d'un ajustement affine, on introduit un nouveau paramètre, le coefficient de corrélation linéaire.
Le coefficient de corrélation linéaire d'une série statistique de variables x et y est le nombre défini par :

$$r = \frac{COV(X,Y)}{\sigma_x\sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y}}{\sigma_x\sigma_y}$$

**Propriété B1:**

- *Plus le coefficient de régression linéaire est proche de 1 en valeur absolue, meilleur est l'ajustement linéaire. En général on considère que la corrélation est forte lorsque* $r \geq 0{,}90$.
- *Lorsque* $r = \pm 1$, *la droite de régression passe par tous les points du nuage, qui sont donc alignés.*
- $-1 \leq r \leq 1$

**Méthode pratique :**

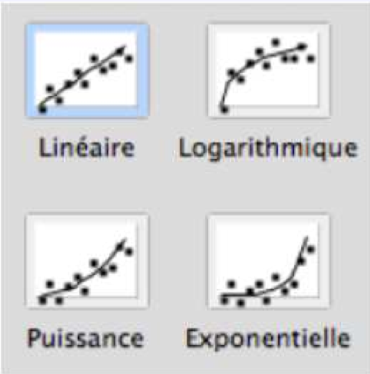
*Encore une fois c'est Excel qui va nous donner le moyen de calculer la valeur de r, en même temps que l'équation de la droite de régression.*

**Exercice B4** *TD TND regression1 Ajustement linéaire*

## 2) Les ajustements non linéaires

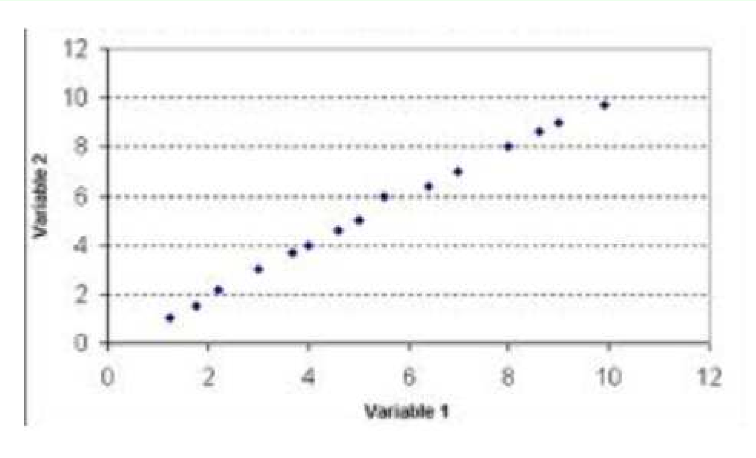
**Ajustement non linéaires:**

Il est possible via des changements de variables d'estimer un nuage de point de deux variables quantitatives via une fonction non linéaire (exponentielles, logarithmique, puissances ...)

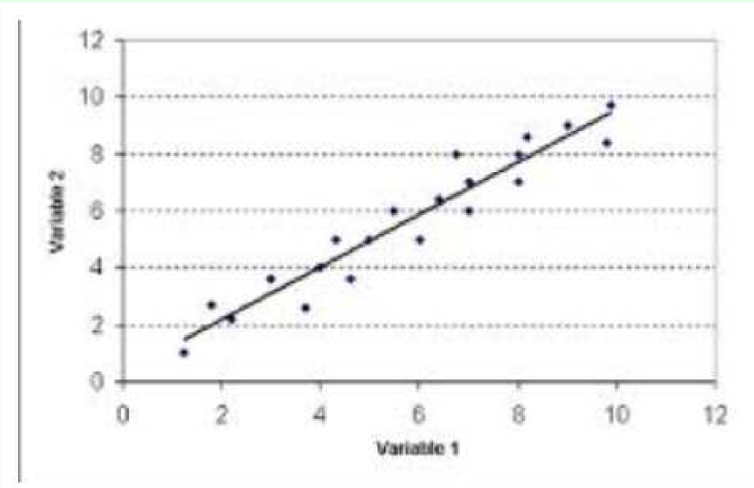



**Méthode pratique :**

1. *Le nuage de points*
   - *Insertion graphique - « Nuage de points »*
   - *Sélection des données - « ordo.=val de Y »et « abs.=val de X »*
   - *Mise en forme adéquate - utiliser les options*




2. *L'ajustement*
   - *Sélectionner les points - « Courbe de tendance »*
   - *Choisir un modèle - « linéaire, exponentiel, logarithmique ou puissance »*
   - *Mise en forme adéquate - utiliser les options*




*Bien entendu, tout document* ***nécessite un titre****, chaque axe doit indiquer ce qu'il représente ET il faut afficher l'****équation de la courbe de tendance*** *ainsi que le* ***coefficient*** $R^2$

**Remarque:**

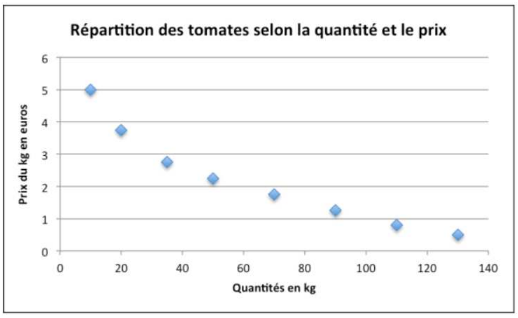
*L'intérêt d'avoir un outil aussi puissant qu'Excel permet ici de tester en quelques secondes les différents ajustement possibles et de retenir celui qui permet le meilleur coefficient de corrélation.*

**Exemple:**

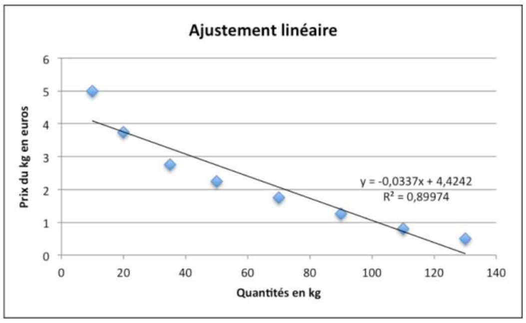
*On cherche à étudier la relation entre les prix et des quantité sur le marché des tomates. On a observé les variations suivantes :*

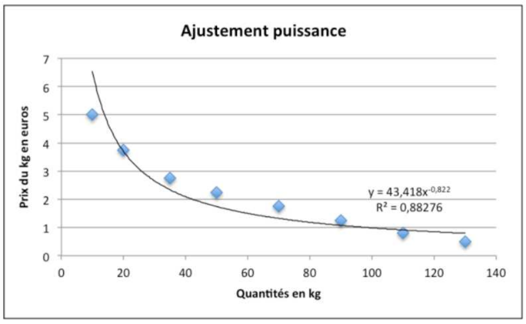
| | A | B |
|---|---|---|
| 1 | Quantités en kg | Prix du kg en euros |
| 2 | 10 | 5 |
| 3 | 20 | 3,75 |
| 4 | 35 | 2,75 |
| 5 | 50 | 2,25 |
| 6 | 70 | 1,75 |
| 7 | 90 | 1,25 |
| 8 | 110 | 0,8 |
| 9 | 130 | 0,5 |

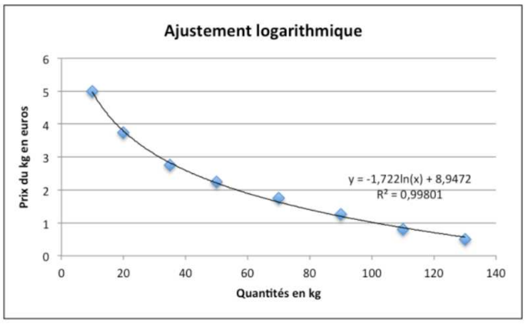
*Ce qui nous donne le nuage de point suivant :*

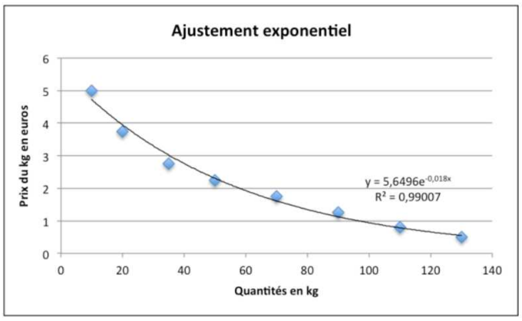



*Testons les différents ajustements que nous propose Excel :*




***Interprétation :***

- *Le meilleur ajustement est l'ajustement logarithmique.*
- *Si la quantité est de 140 kg, le prix sera de :* $-1{,}722\ln(140) + 8{,}9472 = 0{,}44$€

**Exercice B5** *TD TND Ajustement non linéaire1*

# Chapitre C

# Série chronologique

**Série chronologique:**

**Une série chronologique** ou **temporelle** décrit l'évolution d'une variable quantitative par rapport au temps.

Nous verrons dans ce chapitre, plusieurs façons de s'intéresser à ces séries chronologiques et de les étudier.

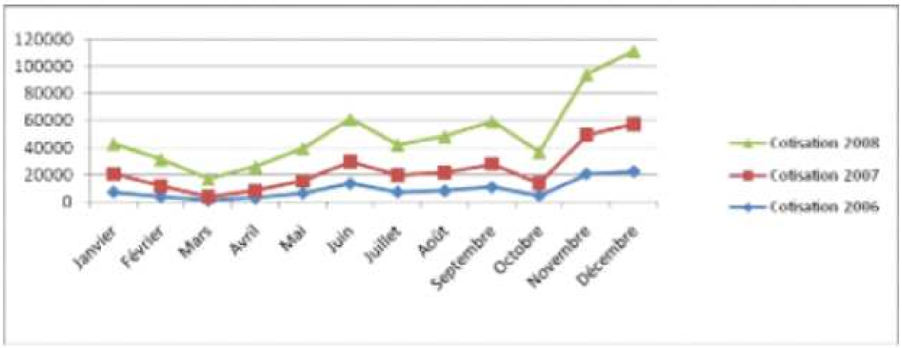
## I) Présentation des données temporelles

En premier lieu nous allons étudier et apprendre à représenter ces séries chronologiques via Excel.

**Méthode pratique :**

*Nuages de points périodiques*

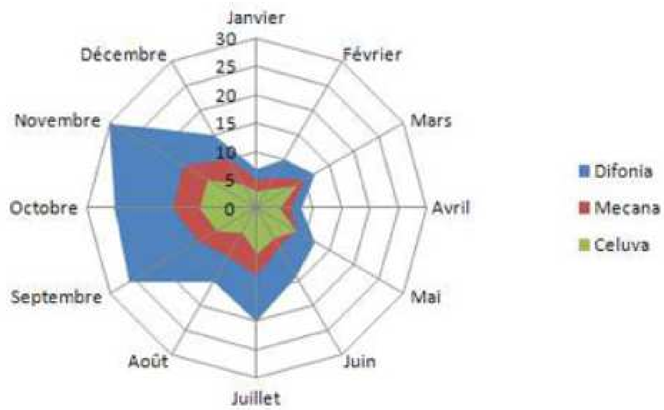
- *Insertion graphique puis "Nuage de points"*
- *Sélection des données puis "ordo.=val.de la Variable quantitative" et "abs.=les sous périodes"*
- *"Mise en forma adéquate puis utiliser les options*



**Méthode pratique :**

*Graphique polaire*

- *Insertion graphique puis "Radar"*
- *Sélection des données puis "ordo.=val.de la Variable quantitative" et "abs.=les sous périodes"*
- *"Mise en forma adéquate puis utiliser les options*




**Méthode pratique :**

*Nuage de points annuel*

- *Insertion graphique puis "Nuage de points"*
- *Sélection des données puis "ordo.=val.de la Variable quantitative" et "abs.=les sous périodes"*
- *"Mise en forma adéquate puis utiliser les options*

**Remarque:**

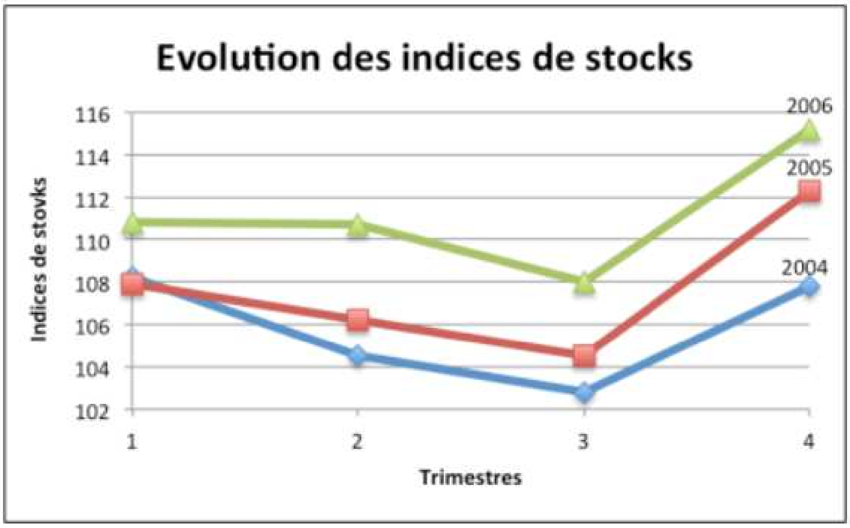
- *Les représentations graphiques doivent être surmontées d'un titre.*
- *Il faut indiquer ce que chaque axe représente.*

**Exemple:**

*Le tableau suivant donne les indices trimestriels de stocks de matières en valeur des industries agricoles et alimentaires (IAA) :*

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | 1er trimestre | 2nd trimestre | 3ème trimestre | 4ème trimestre |
| 2 | 2004 | 108,2 | 104,5 | 102,8 | 107,8 |
| 3 | 2005 | 107,9 | 106,2 | 104,5 | 112,3 |
| 4 | 2006 | 110,8 | 110,7 | 108 | 115,2 |

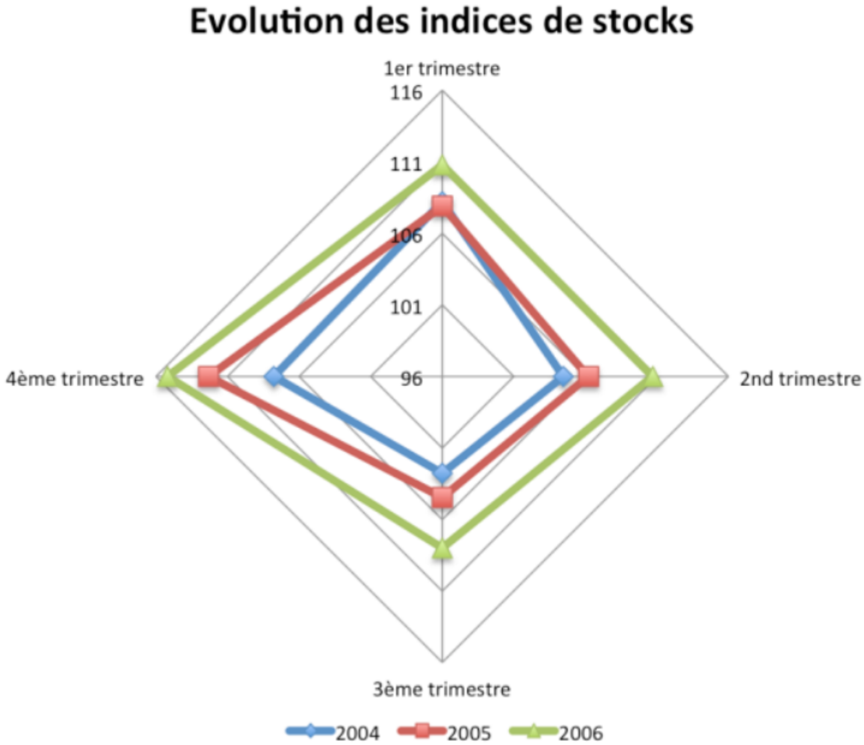
*On obtient alors les représentations suivantes :*




***Interprétation :***

- *Les indices de stocks ont une évolution saisonnière :*

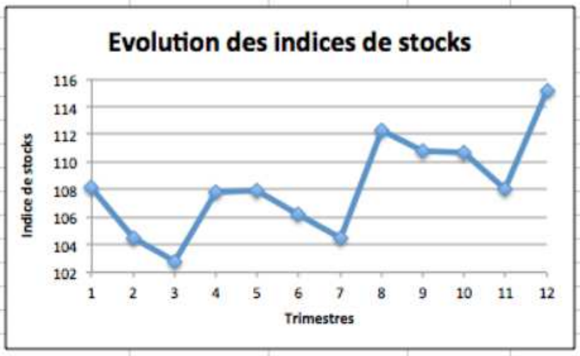
  *ils baissent entre le 1er trimestre et le 3ème trimestre mais augmente de façon importante au 4ème trimestre.*
- *Par ailleurs ils ont tendance à augmenter entre 2004 et 2006.*




***Interprétation :***

- *Les indices de stocks ont tendance à être systématiquement plus élevé au 4ème trimestre*
- *Par ailleurs ils ont tendance à augmenter entre 2004 et 2006.*

| Périodes | Indices de Stocks |
|---|---|
| 1 | 108,2 |
| 2 | 104,5 |
| 3 | 102,8 |
| 4 | 107,8 |
| 5 | 107,9 |
| 6 | 106,2 |
| 7 | 104,5 |
| 8 | 112,3 |
| 9 | 110,8 |
| 10 | 110,7 |
| 11 | 108 |
| 12 | 115,2 |




***Interprétation :***

- *Les indices de stocks sont marqués par une évolution saisonnière et une tendance croissante.*

**Exercice C6** *TD TND seriechrono1 Présentation des données*

## II) Composante tendancielle

### composante tendancielle:

On appelle la **composante tendancielle** ($F_t$) d'une série chronologique ou temporelle son orientation générale :

- Tendance durable à la croissance
- Tendance durable à la décroissance

On parle encore de *composante générale*, ou *extra-saisonnière* ou encore de *trend*

### Méthode pratique :

*On estime la relation :* $Y = aT + b$ *via la méthode des moindres carrées, en donnant comme valeurs à la variable T :* $1, 2, \ldots, n$.

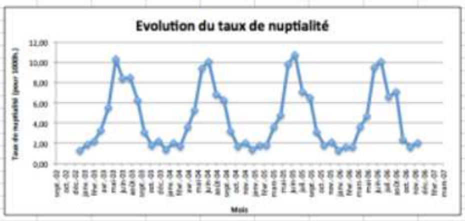
$$\begin{cases} a = \dfrac{COV(T;Y)}{V(T)} \\ b = \quad \bar{Y} - a\bar{T} \end{cases}$$

### Exemple:

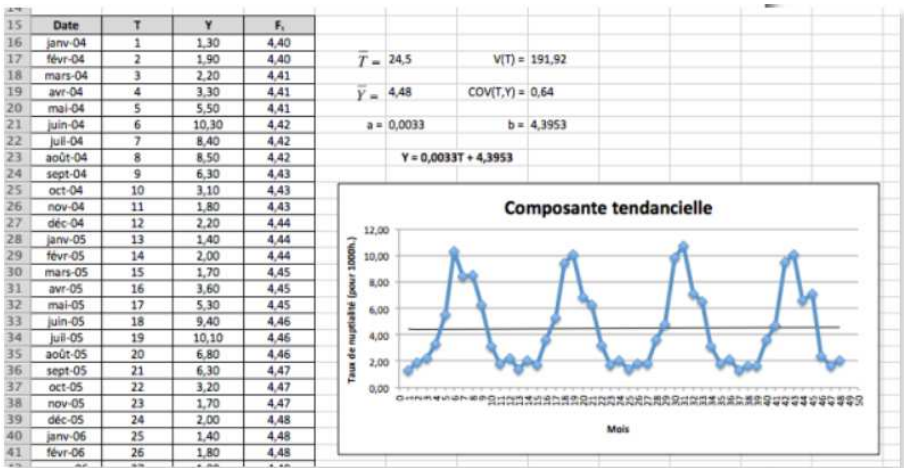
*Considérons la série suivante qui donne le taux mensuel de nuptialité (nombre de mariages pour 1000 habitants) en France métropolitaine :*

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Mois | 2003 | 2004 | 2005 | 2006 |
| 2 | Janvier | 1,30 | 1,40 | 1,40 | 1,30 |
| 3 | Février | 1,90 | 2,00 | 1,80 | 1,60 |
| 4 | Mars | 2,20 | 1,70 | 1,80 | 1,60 |
| 5 | Avril | 3,30 | 3,60 | 3,60 | 3,60 |
| 6 | Mai | 5,50 | 5,30 | 4,80 | 4,70 |
| 7 | Juin | 10,30 | 9,40 | 9,80 | 9,50 |
| 8 | Juillet | 8,40 | 10,10 | 10,70 | 10,10 |
| 9 | Août | 8,50 | 6,80 | 7,10 | 6,60 |
| 10 | Septembre | 6,30 | 6,30 | 6,50 | 7,10 |
| 11 | Octobre | 3,10 | 3,20 | 3,10 | 2,40 |
| 12 | Novembre | 1,80 | 1,70 | 1,80 | 1,60 |
| 13 | Décembre | 2,20 | 2,00 | 2,10 | 2,00 |

| | Date | T | Y |
|---|---|---|---|
| 15 | Date | T | Y |
| 16 | janv-03 | 1 | 1,30 |
| 17 | févr-03 | 2 | 1,90 |
| 18 | mars-03 | 3 | 2,20 |
| 19 | avr-03 | 4 | 3,30 |
| 20 | mai-03 | 5 | 5,50 |
| 21 | juin-03 | 6 | 10,30 |
| 22 | juil-03 | 7 | 8,40 |
| 23 | août-03 | 8 | 8,50 |
| 24 | sept-03 | 9 | 6,30 |
| 25 | oct-03 | 10 | 3,10 |
| 26 | nov-03 | 11 | 1,80 |
| 27 | déc-03 | 12 | 2,20 |
| 28 | janv-04 | 13 | 1,40 |
| 29 | févr-04 | 14 | 2,00 |
| 30 | mars-04 | 15 | 1,70 |
| 31 | avr-04 | 16 | 3,60 |
| 32 | mai-04 | 17 | 5,30 |
| 33 | juin-04 | 18 | 9,40 |
| 34 | juil-04 | 19 | 10,10 |
| 35 | août-04 | 20 | 6,80 |




*L'estimation de la composante tendancielle par les MCO donne alors les résultats suivants, qu'on peut alors représenter graphiquement :*



| | Date | T | Y | $F_t$ |
|---|---|---|---|---|
| 16 | janv-04 | 1 | 1,30 | 4,40 |
| 17 | févr-04 | 2 | 1,90 | 4,40 |
| 18 | mars-04 | 3 | 2,20 | 4,41 |
| 19 | avr-04 | 4 | 3,30 | 4,41 |
| 20 | mai-04 | 5 | 5,50 | 4,41 |
| 21 | juin-04 | 6 | 10,30 | 4,42 |
| 22 | juil-04 | 7 | 8,40 | 4,42 |
| 23 | août-04 | 8 | 8,50 | 4,42 |
| 24 | sept-04 | 9 | 6,30 | 4,43 |
| 25 | oct-04 | 10 | 3,10 | 4,43 |
| 26 | nov-04 | 11 | 1,80 | 4,43 |
| 27 | déc-04 | 12 | 2,20 | 4,44 |
| 28 | janv-05 | 13 | 1,40 | 4,44 |
| 29 | févr-05 | 14 | 2,00 | 4,44 |
| 30 | mars-05 | 15 | 1,70 | 4,45 |
| 31 | avr-05 | 16 | 3,60 | 4,45 |
| 32 | mai-05 | 17 | 5,30 | 4,45 |
| 33 | juin-05 | 18 | 9,40 | 4,46 |
| 34 | juil-05 | 19 | 10,10 | 4,46 |
| 35 | août-05 | 20 | 6,80 | 4,46 |
| 36 | sept-05 | 21 | 6,30 | 4,47 |
| 37 | oct-05 | 22 | 3,20 | 4,47 |
| 38 | nov-05 | 23 | 1,70 | 4,47 |
| 39 | déc-05 | 24 | 2,00 | 4,48 |
| 40 | janv-06 | 25 | 1,40 | 4,48 |
| 41 | févr-06 | 26 | 1,80 | 4,48 |

$\overline{T} =$ 24,5 V(T) = 191,92

$\overline{Y} =$ 4,48 COV(T,Y) = 0,64

a = 0,0033 b = 4,3953

**Y = 0,0033T + 4,3953**



**Exercice C7** *TD seriechrono2 Composante tendancielle*

## III) Composante saisonnière

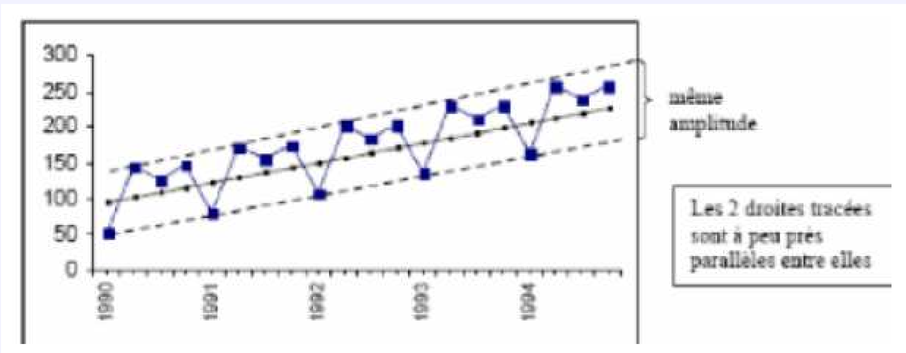
**Composante saisonnière:**

On appelle la composante saisonnière ($S_t$) d'une série chronologique ou temporelle ses évolutions périodiques sur l'année :

- dues aux saisons (printemps, été, ...)
- dues aux usages (fêtes, vacances, ...) On parle également de composante périodique.

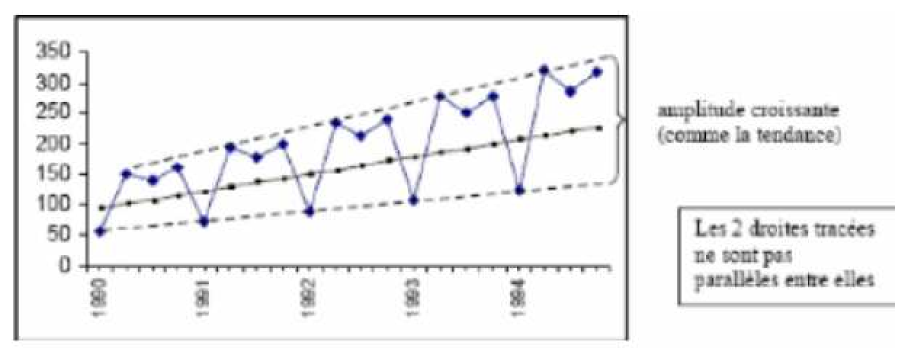
Deux situations peuvent exister dans les séries temporelles :

- **Le modèle additif :**
  1. Les fluctuations sont d'amplitudes constantes autour du trend
  2. Le modèle est tel que : $Y_t = T_t + S_t + \epsilon_t$





- **Le modèle multiplicatif :**
  1. Les fluctuations sont d'amplitudes liées à la valeur du trend.
  2. Le modèle est tel que $Y_t = T_t \times S_t \times \epsilon_t$




**Remarque:**

*Il arrive que le choix entre les deux modèles ne soit pas aussi clair, mais dans ce cas les résultats , quelque soit le modèle choisi seront très proches.*

**Méthode pratique :**

***Pour le modèle additif***

1. *Écart saisonniers à chaque date :* $S_t = Y_t - F_t$
2. *Coefficients saisonniers de la période :* $S_i$ = *Moyenne des écarts saisonniers de la même saison*
3. *Somme des coefficients saisonniers : La somme des* $S_i$ *sur une période doit être égale à 0.*
4. *Coefficients saisonniers corrigés :* $S'_i = S_i - m$ *lorsque la somme des* $S_i$ *n'est pas nulle, et où* $m$ *est la moyenne des* $S_i i$

***Pour le modèle multiplicatif***

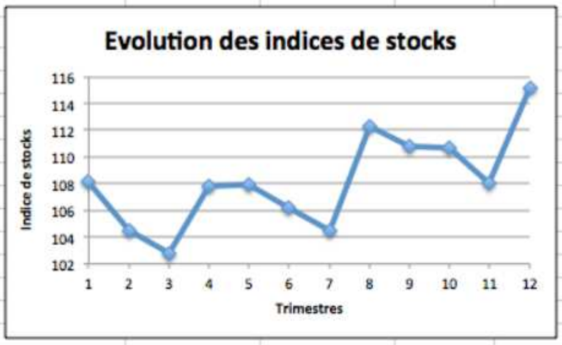
1. *Rapports saisonniers à chaque date :* $S_t = \dfrac{Y_t}{F_t}$
2. *Coefficients saisonniers de la période* $S_i$ = *Moyenne des rapports saisonniers de la même saison.*
3. *Produit des coefficients saisonniers : La somme des* $S_i$ *sur une période doit être égale au nombre de saisons.*
4. *Coefficients saisonniers corrigés :* $S'_i = \dfrac{S_i}{m}$ *où* $m$ *est la moyenne des* $S_i$

**Exemple:**

*Reprenons les indices trimestriels de stocks de matières en valeur des industries agricoles et alimentaires (IAA) :*

| | 1er trimestre | 2nd trimestre | 3ème trimestre | 4ème trimestre |
|---|---|---|---|---|
| 2004 | 108,2 | 104,5 | 102,8 | 107,8 |
| 2005 | 107,9 | 106,2 | 104,5 | 112,3 |
| 2006 | 110,8 | 110,7 | 108 | 115,2 |

| Périodes | Indices de Stocks |
|---|---|
| 1 | 108,2 |
| 2 | 104,5 |
| 3 | 102,8 |
| 4 | 107,8 |
| 5 | 107,9 |
| 6 | 106,2 |
| 7 | 104,5 |
| 8 | 112,3 |
| 9 | 110,8 |
| 10 | 110,7 |
| 11 | 108 |
| 12 | 115,2 |




***Interprétation :***

- *Les indices de stocks sont marqués par une évolution saisonnière et une tendance croissante.*
- *On peut alors estimer la composante tendancielle via la méthode des MCO et la composante saisonnière via la méthode du schéma additif*

| T | Y | $F_t$ | $s_t$ | $S_t$ | $Y_{cvs}$ | $\hat{y}$ | $\varepsilon_t$ |
|---|---|---|---|---|---|---|---|
| 1 | 108,2 | 104,58 | 3,62 | 1,72 | 106,48 | 106,31 | 1,89 |
| 2 | 104,5 | 105,25 | -0,75 | -0,78 | 105,28 | 104,47 | 0,03 |
| 3 | 102,8 | 105,91 | -3,11 | -3,47 | 106,27 | 102,44 | 0,36 |
| 4 | 107,8 | 106,58 | 1,22 | 2,53 | 105,27 | 109,11 | -1,31 |
| 5 | 107,9 | 107,24 | 0,66 | 1,72 | 106,18 | 108,97 | -1,07 |
| 6 | 106,2 | 107,91 | -1,71 | -0,78 | 106,98 | 107,13 | -0,93 |
| 7 | 104,5 | 108,57 | -4,07 | -3,47 | 107,97 | 105,10 | -0,60 |
| 8 | 112,3 | 109,24 | 3,06 | 2,53 | 109,77 | 111,77 | 0,53 |
| 9 | 110,8 | 109,91 | 0,89 | 1,72 | 109,08 | 111,63 | -0,83 |
| 10 | 110,7 | 110,57 | 0,13 | -0,78 | 111,48 | 109,79 | 0,91 |
| 11 | 108 | 111,24 | -3,24 | -3,47 | 111,47 | 107,77 | 0,23 |
| 12 | 115,2 | 111,90 | 3,30 | 2,53 | 112,67 | 114,43 | 0,77 |

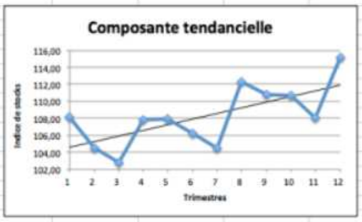
$\bar{T}$ = 6,5 V(T) = 11,92

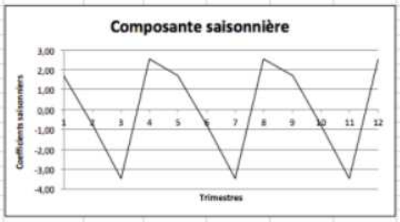
$\bar{Y}$ = 108,24 COV(T,Y) = 7,93

a = 0,6654 b = 103,92

Y = 0,6654 + 103,92

| | |
|---|---|
| $S_1$ | 1,72 |
| $S_2$ | -0,78 |
| $S_3$ | -3,47 |
| $S_4$ | 2,53 |
| Total | 0,00 |




***Interprétation :***

- *La composante saisonnière est donnée par la colonne $S_t$*
- *On peut alors la représenter graphiquement*

**Exercice C8** *TD seriechrono2 Composante saisonnière*

# Bibliographie


- Cours Statistiques appliqués IUT-Amiens-TC *(C.DOLIGER)*