groupe esaip
ENSEIGNEMENT SUPÉRIEUR
ÉCOLE D'INGÉNIEURS

Modélisation probabiliste et statistique

1. Variables aléatoires (rappels)
2. Echantillonage
3. Estimation statistique
4. Tests statistiques paramétriques
5. Analyse de la variance
6. Tests statistiques non paramétriques
7. Fiabilité
8. Régression et corrélation
9. Séries chronologiques

1. VARIABLES ALÉATOIRES

Définition de la variable aléatoire

Variable aléatoire : application X : $\Omega \rightarrow \mathbf{R}$

Espace fondamental Ω :
l'ensemble de tous les événements élémentaires

X(Ω) : l'ensemble des valeurs possibles prises par **X** (l'image de Ω par **X**)

Variable aléatoire: **discrète ou continue**

Exemple d'une variable discrète

On lance trois fois une pièce de monnaie.

A chaque événement élémentaire, on peut associer un réel x qui est le nombre de « pile » obtenu.

Variable aléatoire :
application X : $\Omega \rightarrow \mathbf{R}$

ω	X(ω) = x
(F, F, F)	0
(F, F, P)	1
(F, P, F)	1
(F, P, P)	2
(P, F, F)	1
(P, F, P)	2
(P, P, F)	2
(P, P, P)	3

-la probabilité d'avoir une fois pile : $P(X(\omega)=1)=\frac{3}{8}$

La loi de d'une variable aléatoire discrète

La **loi de probabilité** associe à chacune des valeurs possibles x_i de la variable aléatoire discrète X la probabilité de l'événement correspondant, c'est-à-dire

$$p_i = P(X = x_i)$$

Notation : $P(X = x_i) = P\{\omega \in \Omega ; X(\omega) = x_i\}$

La loi de probabilité est déterminée si l'on connaît l'ensemble des couples (x_i, p_i)

$$P(X = x_i) = p_i \geq 0 \quad ; \quad \sum p_i = 1$$

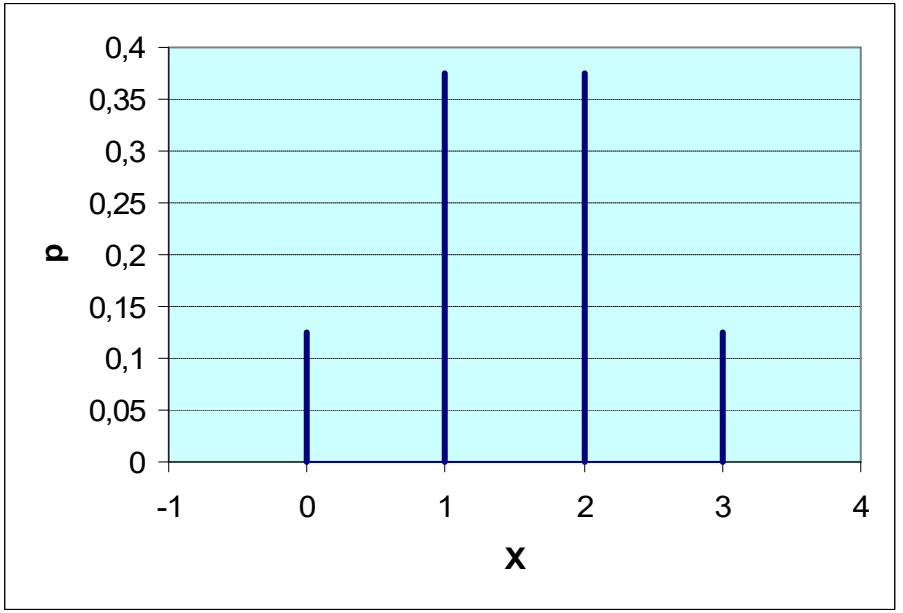
Exemple

On lance trois fois une pièce de monnaie

X : $\Omega \rightarrow$ R,
$X(\omega) = x_i$ le nombre de « pile » obtenu

x_i	0	1	2	3
p_i	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Diagramme en bâtons associé à la variable aléatoire :

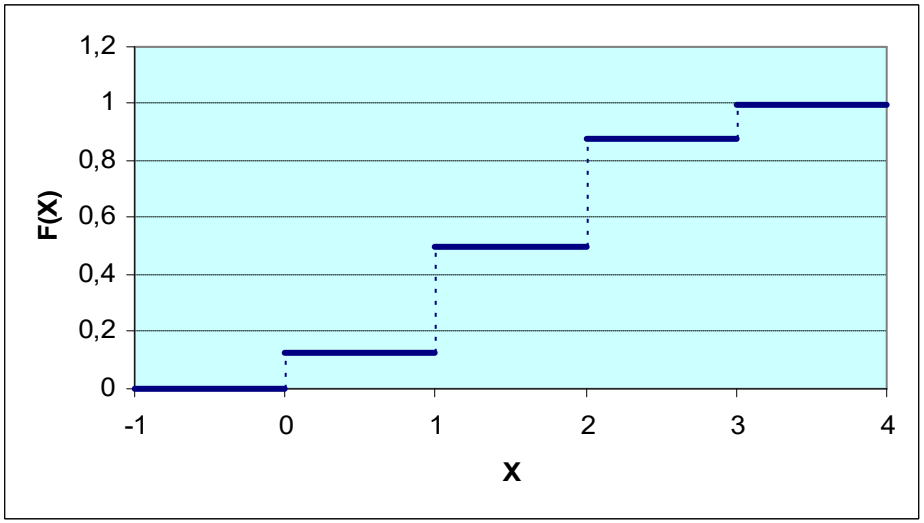


Fonction de répartition de la variable aléatoire discrète X :

$$F : \mathbf{R} \rightarrow [\,0,\ 1\,]$$

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i) = \sum_{x_i < x} p_i = p_1 + p_2 + ... + p_k$$

Représentation graphique :
courbe de probabilités cumulées.



Exemple

On lance trois fois une pièce de monnaie
X : $\Omega \rightarrow$ **R,**
X(ω) = x_i **le nombre de « pile » obtenu**

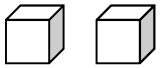
Caractéristiques numériques d'une variable aléatoire discrète :

- **Espérance mathématique :** $E(X) = \sum_i p_i x_i$

- **Variance :** $Var(X) = E(X^2) - (E(X))^2 = \sum_i p_i x_i^2 - \left(\sum_i p_i x_i \right)^2$

- **L'écart-type** : $\sigma(X) = \sqrt{Var(X)}$

Exemple On lance deux dés cubiques.



événement élémentaire : $\omega=(\omega_1, \omega_2)$

X: variable aléatoire donnant la somme des points obtenus

ω_1 \ ω_2	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Loi de probabilité:

$X(\omega)=x_i$	2	3	4	5	6	7	8	9	10	11	12
p_i	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Espérance mathématique (le gain moyen): $E(X)=\sum_i p_i x_i =$

$$=\tfrac{1}{36}\times 2+\tfrac{2}{36}\times 3+\tfrac{3}{36}\times 4+\tfrac{4}{36}\times 5+\tfrac{5}{36}\times 6+\tfrac{6}{36}\times 7+\tfrac{5}{36}\times 8+\tfrac{4}{36}\times 9+\tfrac{3}{36}\times 10+\tfrac{2}{36}\times 11+\tfrac{1}{36}\times 12=7$$

Variance : $Var(X)=\sum_i p_i x_i^2-\left(\sum_i p_i x_i\right)^2=(\tfrac{1}{36}\times 2^2+\tfrac{2}{36}\times 3^2+\tfrac{3}{36}\times 4^2+\tfrac{4}{36}\times 5^2+\tfrac{5}{36}\times 6^2+$

$$+\tfrac{6}{36}\times 7^2+\tfrac{5}{36}\times 8^2+\tfrac{4}{36}\times 9^2+\tfrac{3}{36}\times 10^2+\tfrac{2}{36}\times 11^2+\tfrac{1}{36}\times 12^2)-7^2=\tfrac{1974}{36}-49\approx 54{,}8-49=5{,}8$$

L'écart-type : $\sigma(X)=\sqrt{Var(X)}\approx 2{,}4$

Quelques lois usuelles discrètes

Loi binomiale de paramètres n et p notée B(n, p)

Soit n épreuves de satisfaisant aux trois conditions suivantes :

- chaque épreuve ne peut conduire qu'à deux résultats complémentaires succès ou échec avec les probabilités respectives *p* et 1-*p*.
- La probabilité de succès est la même dans chaque épreuve.
- Les épreuves sont indépendantes

Exemple : lancer de n pièces de monnaie.

X : nombre de succès observés sur n épreuves de Bernouilli

X suit la **loi binomiale** de paramètres n et p.

La probabilité **de k succès observés sur n épreuves de Bernouilli**:

$$P(X = k) = C_n^k \, p^k \, (1 - p)^{n-k}$$

Rappel:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1)...(n-k+1)}{k \cdot (k-1) \cdot 3 \cdot 2 \cdot 1}$$

Propriétés :

$$E(X) = np \qquad Var(X) = npq$$

Loi de Poisson de paramètre $\lambda>0$ notée $P(\lambda)$

On dit qu'une variable aléatoire X suit une loi de Poisson de paramètre $\lambda>0$ lorsque X prend les valeurs entières 0, 1, 2, …, k, …

et $$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$$

- ***La loi de Poisson décrie la probabilité de réalisation de k succès dans unité de mesure (par exemple pendant unité de temps).***

- ***La loi de Poisson est utilisée pour décrire des événements peu fréquents.***

Exemple X : nombre d'accidents de travail par jours dans une entreprise

Propriétés :

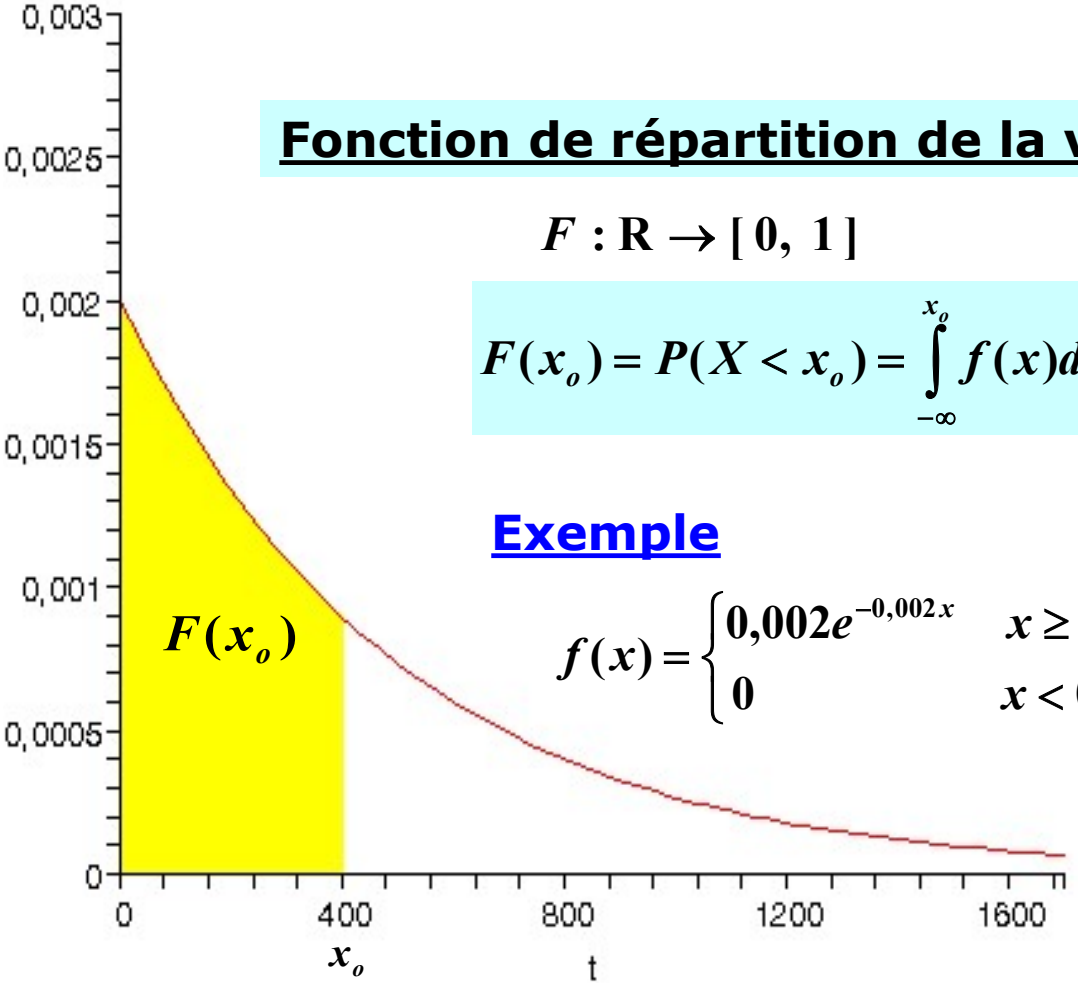
$$E(X)=\lambda \quad , \quad Var(X)=\lambda$$

λ représente le taux moyen de succès dans unité de mesure (par exemple par unités de temps)

Variables aléatoires continues

X(Ω) infini non dénombrable

Exemple X mesurant la durée de bon fonctionnement, en jours (pas nécessairement un nombre entier), d'un équipement particulier fabriqué en grande série; X(Ω) =[0,+∞]



Fonction de répartition de la variable aléatoire continue

$$F : \mathrm{R} \rightarrow [\,0,\ 1\,]$$

$$F(x_o) = P(X < x_o) = \int_{-\infty}^{x_o} f(x)dx$$

f : densité de probabilité

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}$$

Exemple

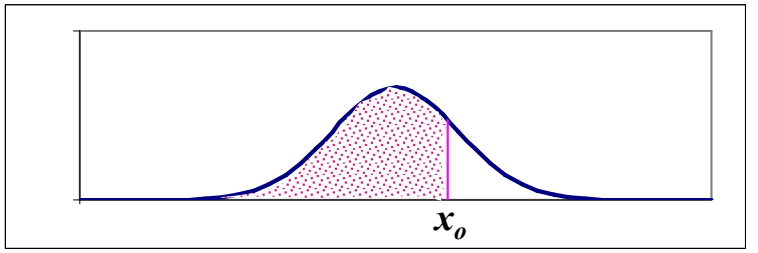
$$f(x) = \begin{cases} 0{,}002e^{-0{,}002x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$F(x_o) = P(X < x_o) = \int_{0}^{x_o} f(x)dx$$

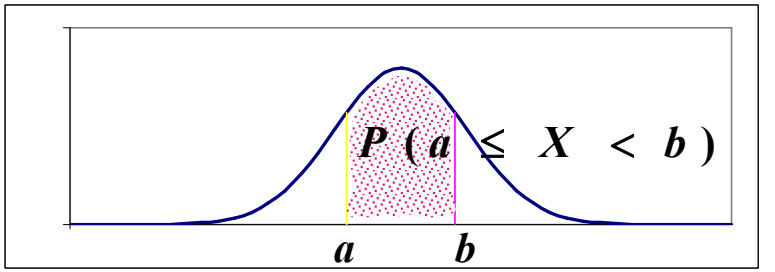
$$P(X \leq 400) = F(400) \approx 0{,}55$$

Propriétés de la fonction de répartition

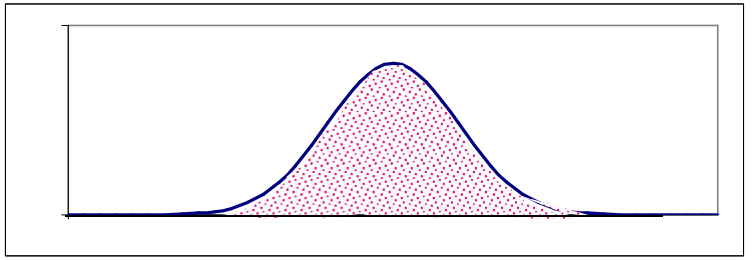
$$P(X < x_o) = F(x_o) = \int_{-\infty}^{x_o} f(x)dx$$



$$P(a \le X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$



$$P(-\infty < X < \infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$



Caractéristiques d'une variable aléatoire continue

- **Espérance mathématique** $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$
- **Variance** $Var(X) = E(X^2) - (E(X))^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} x\ f(x) \right)^2$
- **L'écart-type** : $\sigma(X) = \sqrt{Var(X)}$

Lois normales des variables aléatoires continues

Loi normale centrée réduite

Z~N(0,1)

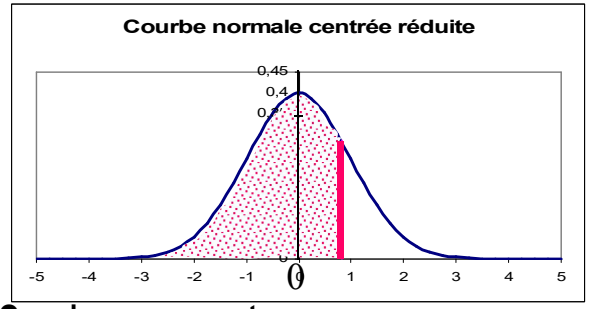
Z(Ω)=R

Densité de probabilité de Z :

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

- Moyenne **E(Z) = 0** , **Var(Z) = 1**, **σ = 1**
- Si $z_o \geq 0$ alors on trouve $P(Z<z_0)=\Pi(z_0)$

dans la table de loi normale



- Si $z_o < 0$ alors pour trouver

$$P(Z<z_0)=1-P(Z<-z_0)$$

on utilise la symétrie de la courbe.

- Probabilité associée à un intervalle $[z_1,\ z_2]$:

$$P(z_1<Z<z_2)=P(Z<z_2)-P(Z<z_1)$$

Loi normale de paramètres m et σ

X~N(m, σ)

X(Ω)=R

Densité de probabilité de X :

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} \quad , x \in \mathbf{R}$$

- Moyenne $E(X)=m$ et $Var(X)=\sigma^2$

X ~ N(m, σ)

⇓ **changement de variable** $Z=\frac{X-m}{\sigma}$

Z ~ N(0, 1)

- $P(X<x_0)=P(Z<\frac{x_0-m}{\sigma})$

le résultat est lu dans la table de N(0, 1)

- $P(x_0<X<x_1)=P(\frac{x_0-m}{\sigma}<Z<\frac{x_1-m}{\sigma})$

changement de variable $Z=\frac{X-m}{\sigma}$

2.THEORIE DE L'ECHANTILLONAGE

Notation:

Paramètres caractérisants la population parente :

$\mathbf{N_p}$: nombre des éléments de la population (N_p peut être l'infinie)

m : la moyenne de la population $\mathbf{m} = \frac{x_1 + ... + x_{N_p}}{\mathbf{N_p}} = \frac{\sum_{i=1}^{N_p} x_i}{\mathbf{N_p}}$

σ : écart type de la distribution de la population $\sigma = \sqrt{\frac{\sum_{i=1}^{N_p} (x_i - \mathbf{m})^2}{\mathbf{N_p}}}$

σ^2 : variance de la distribution de la population

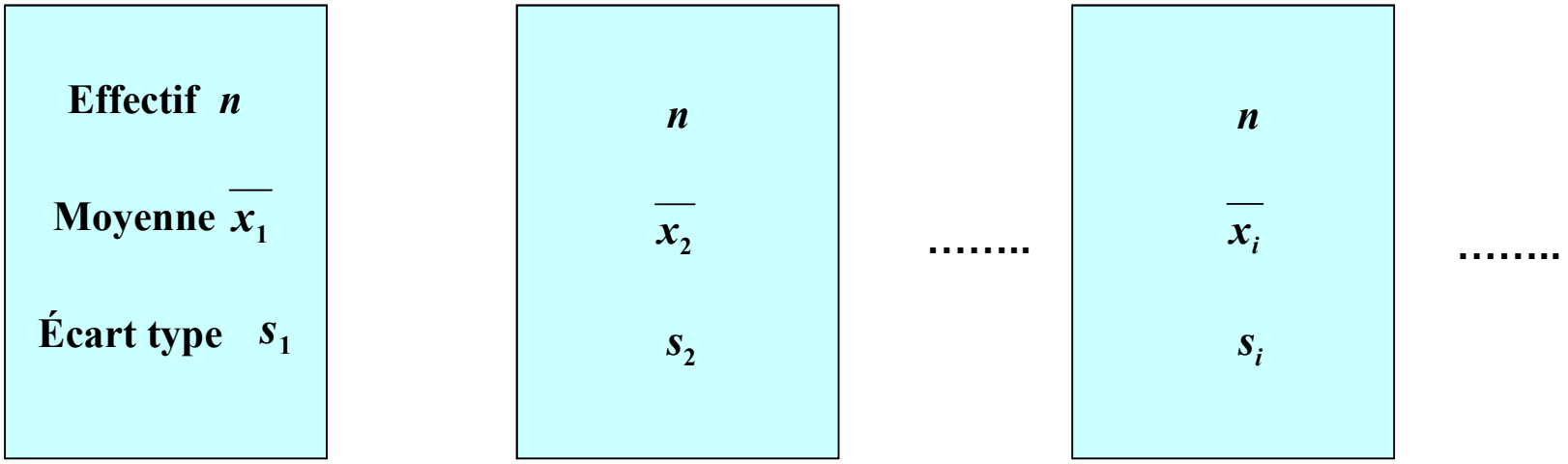
Paramètres caractérisants un échantillon:

n : nombre des éléments dans un échantillon, $n \leq N_p$

$\bar{x}$: la moyenne de l'échantillon $\bar{x} = \frac{x_1 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

s : écart type d'échantillon $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$, $\mathbf{s}^2$ - variance d'échantillon

- On a choisi de façon aléatoire des échantillons de taille *n*.



<u>Deux types d'échantillonnage</u>:

- cas de la population infinie et (ou) d'échantillonnage non-exhaustif (avec remise)
- cas de la population finie et d'échantillonnage exhaustif (sans remise)

Distribution d'échantillonnage de la moyenne :

La distribution des diverses valeurs que peut prendre la moyenne d'échantillon $\overline{X}$ obtenue de tous les échantillons possibles de même taille d'une population donnée.

Théorème

Soit une population de moyenne *m* et d'écart type σ.
Soit $\overline{X}$ la variable aléatoire qui, à tout échantillon aléatoire prélevé **avec remise** et d'effectif *n* fixé, associe la moyenne de cet échantillon.
Alors, pour *n suffisamment grand*, $\overline{X}$ suit approximativement la loi normale

$$N\left(\mathbf{m}, \frac{\sigma}{\sqrt{n}}\right)$$

Remarques

- En pratique, on peut appliquer le théorème quand $n \geq 30$ et $N_p \geq 2n$.

- Lorsque les échantillons de taille *n* sont prélevés **sans remise** dans la population d'effectif N_p, alors, pour n suffisamment grand, $\overline{X}$ suit approximativement la loi normale $N\left(\mathbf{m}, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}\right)$

Distribution d'échantillonnage de la moyenne est caractérisée par:

$m_{\overline{X}}$:la moyenne de la distribution d'échantillonnage de la moyenne

$\sigma_{\overline{X}}$: écart type de la distribution d'échantillonnage de la moyenne

Population infinie et (ou) échantillonnage non- exhaustif (avec remise)	Population finie et échantillonnage exhaustif (sans remise)
$m_{\overline{X}} = m \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$	$m_{\overline{X}} = m \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N_p - n}{N_p - 1}}$

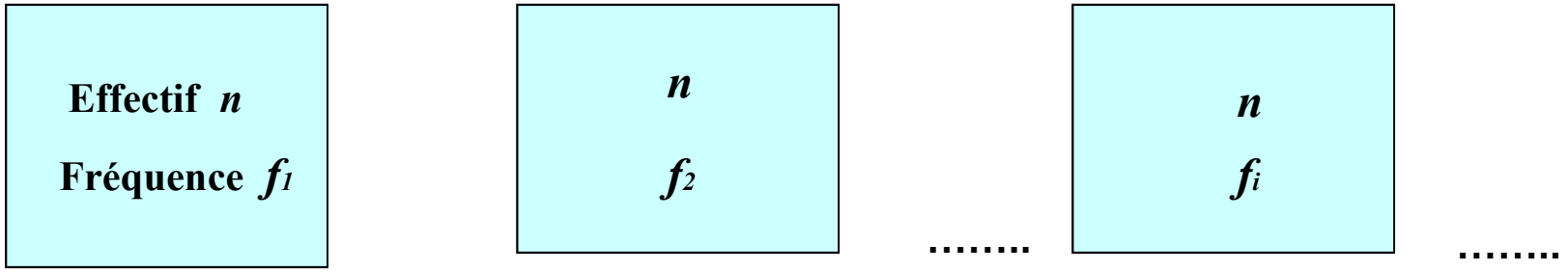
- La distribution d'échantillonnage de la moyenne suit la loi normale si $n \geq 30$ et $N_p \geq 2n$.
- En pratique, on peut négliger le facteur $\sqrt{\frac{N_p - n}{N_p - 1}}$ si $\frac{n}{N_p} \leq 0{,}1$.
- La variable $Z = \frac{\overline{X} - m}{\sigma_{\overline{X}}}$ suit la loi normale centrée réduite N(0,1)

Dans le cas d' échantillonnage non-exhaustif c'est la variable $Z = \frac{\sqrt{n}}{\sigma}(\overline{X} - m)$ qui suit la loi normale centrée réduite N(0,1)

Distribution d'échantillonnage de la fréquence (de la proportion) :

Supposons que les éléments d'une population étudiée possèdent une certaine propriété avec une fréquence ***p***.

• Prélevons avec remise dans cette population des échantillons aléatoires de même effectif n et mesurons pour chacun d'eux la fréquence ***f*** avec laquelle les éléments possèdent cette même propriété:



Échantillon 1 Échantillon 2 Échantillon *i*

Théorème

Soit une population dont les éléments possède une certaine propriété avec une fréquence ***p***.

Soit ***F*** la variable aléatoire qui, à tout échantillon aléatoire prélevé **avec remise** et d'effectif *n* fixé, associe la fréquence avec laquelle les éléments de cet échantillon possèdent cette propriété.

Alors, pour *n suffisamment grand*, ***F*** suit approximativement la loi normale

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Distribution d'échantillonnage de la fréquence est caractérisée par:

$\mathbf{m}_F$:la moyenne de la distribution d'échantillonnage de la fréquence

σ_F : écart type de la distribution d'échantillonnage de la fréquence

Population infinie et (ou) échantillonnage non-exhaustif (avec remise)	Population finie et échantillonnage exhaustif (sans remise)
$\mathbf{m}_F = p \qquad \sigma_F = \sqrt{\frac{p(1-p)}{n}}$	$\mathbf{m}_F = p \qquad \sigma_F = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{\mathbf{N}_p - n}{\mathbf{N}_p - 1}}$

- La distribution d'échantillonnage de la fréquence suit la loi normale si $n \cdot p \geq 5$ et $n \cdot (1-p) \geq 5$

  Habituellement, si p n'est pas trop voisin de 0 ou 1, $n \geq 30$ est suffisant.

- On peut négliger le facteur $\sqrt{\frac{\mathbf{N}_p - n}{\mathbf{N}_p - 1}}$ si $\frac{n}{\mathbf{N}_p} \leq 0{,}1$

- La variable $Z = \frac{F - p}{\sigma_F}$ suit la loi normale centrée réduite N(0,1)

Distribution d'échantillonnage de la différence et de la somme de deux statistiques

Cas général:

la différence: $\mathbf{m}_{(s_1-s_2)} = \mathbf{m}_{s_1} - \mathbf{m}_{s_2}$ et $\sigma_{(s_1-s_2)} = \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$

la somme: $\mathbf{m}_{(s_1+s_2)} = \mathbf{m}_{s_1} + \mathbf{m}_{s_2}$ et $\sigma_{(s_1+s_2)} = \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$

- Supposons qu'on a deux distributions d'échantillonnage de la moyenne pour des échantillons de taille respectivement n_1 et n_2. Elles sont caractérisées respectivement par $\mathbf{m}_{\overline{X}_1}$, $\mathbf{m}_{\overline{X}_2}$, $\sigma_{\overline{X}_1}$ et $\sigma_{\overline{X}_2}$. Si les échantillons sont considérés comme grands ($n_1 \geq 30$ et $n_2 \geq 30$) alors **la distribution d'échantillonnage de la différence des moyennes** suit la loi normale $\mathbf{N}(\mathbf{m}_{\overline{X}_1-\overline{X}_2}, \sigma_{\overline{X}_1-\overline{X}_2})$
- On a des résultats correspondants pour **la distribution d'échantillonnage de la différence des fréquences**.

distribution d'échantillonnage de la différence des moyennes	$\mathbf{m}_{\overline{X}_1-\overline{X}_2} = \mathbf{m}_{\overline{X}_1} - \mathbf{m}_{\overline{X}_2} = \mathbf{m}_1 - \mathbf{m}_2$ $\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
distribution d'échantillonnage de la différence des fréquences.	$\mathbf{m}_{F_1-F_2} = \mathbf{m}_{F_1} - \mathbf{m}_{F_2} = p_1 - p_2$ $\sigma_{F_1-F_2} = \sqrt{\sigma_{F_1}^2 + \sigma_{F_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

3. THÉORIE D'ESTIMATION

Exemple introductif

Une société s'approvisionne en pièces brutes qui, conformément aux conditions fixées par le fournisseur, doivent avoir une masse moyenne de 780 grammes. Au moment où **500 pièces** sont réceptionnées, on en prélève au hasard un échantillon de **36 pièces** dont on mesure la masse. On obtient les résultats suivants :

Masse des pièces (en grammes)	Nombre de pièces
[745, 755[	2
[755, 765[	6
[765, 775[	10
[775, 785[	11
[785, 795[	5
[795, 805[	2

Question 1:

À combien peut-on estimer **la moyenne** et **l'écart type des masses** pour la population constituée des 500 pièces à l'aide des résultats obtenus sur cet échantillon ?

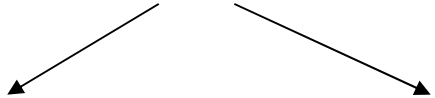
Question 2:

À combien peut-on estimer **la proportion** des pièces dont la masse est strictement inférieure à 765 g parmi la population constituée de l'ensemble de 500 pièces réceptionnées?

Nature du problème

On cherche des informations sur une population à partir d'un échantillon.

Estimation d'un paramètre de population donnée



Estimation ponctuelle
(par une valeur unique)

Estimation par intervalle de confiance

Estimations ponctuelles

- Moyenne inconnue **m** d'une population ⟶ Moyenne $\overline{x}$ d'un échantillon prélevé au hasard

- Proportion inconnue ***p*** des éléments d'une population possédant une certaine propriété ⟶ Proportion ***f*** des éléments possédant une certaine propriété dans un échantillon prélevé au hasard

- Variance inconnue σ^2 d'une population ⟶ $\hat{s}^2 = \frac{n}{n-1} s^2$, où s^2 est la variance d'un échantillon prélevé au hasard

- Ecart type inconnue σ d'une population ⟶ $\hat{s} = \sqrt{\frac{n}{n-1}}\, s$, où ***s*** est l'écart type d'un échantillon prélevé au hasard

Estimation de la moyenne par intervalle de confiance

$\overline{X}$: la variable aléatoire qui, à tout échantillon aléatoire prélevé **avec remise** et d'effectif $\boldsymbol{n}$ f ixé$(\boldsymbol{n \geq 30})$, associe la moyenne de cet échantillon.

$\overline{\boldsymbol{X}}$ suit la loi normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$; $Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite N(0,1).

$\boldsymbol{m}$: moyenne de la population;

$\boldsymbol{\alpha = 1 - niveau\ \ de\ \ confiance}$ (risque fixé *a priori*; par exemple α=0,05)

On cherche un intervalle $\left[\overline{X} - c, \overline{X} + c\right]$ tel que $\boldsymbol{P\left(m \in \left[\overline{X} - c, \overline{X} + c\right]\right) = 1 - \alpha.}$

Alors, $P(\overline{X} - c \leq m \leq \overline{X} + c) = 1 - \alpha$

$$\Leftrightarrow \quad P\left(\frac{-c}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \alpha$$

$$\Leftrightarrow \quad P\left(\frac{-c\sqrt{n}}{\sigma} \leq Z \leq \frac{c\sqrt{n}}{\sigma}\right) = 1 - \alpha \quad \Leftrightarrow \quad 2 \cdot P\left(Z \leq \frac{c\sqrt{n}}{\sigma}\right) - 1 = 1 - \alpha$$

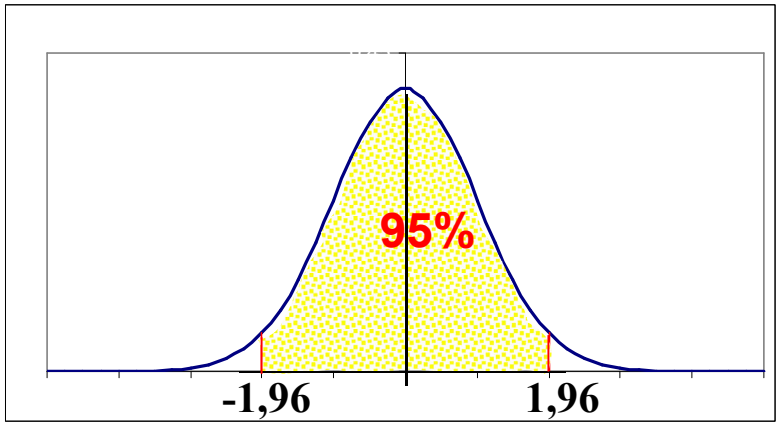
$$\Leftrightarrow \quad P\left(Z \leq \frac{c\sqrt{n}}{\sigma}\right) = 1 - \frac{\alpha}{2}$$

D'où, à l'aide de la table N(0,1) on peut déterminer $\frac{c\sqrt{n}}{\sigma}$ et en déduire c.

Par la suite, nous allons noter $\frac{c\sqrt{n}}{\sigma}$ par z_c.

Exemple

Si on choisi le niveau de confiance de **95%** (la probabilité de 0,95), le risque α**=0,05**.



$$2P(Z \le z_c) - 1 = 0{,}95$$ **(niveau de confiance 95%; risque 5%)**

$$\Downarrow$$

$$P(Z \le z_c) = 0{,}975$$

$$\Downarrow$$ (d'après la table de N(0,1)

$$z_c = 1{,}96$$ **(coefficient de confiance; valeur critique)**

$$\Downarrow$$

$$P(-1{,}96 \le \frac{\sqrt{n}}{\sigma}(\overline{X} - m) \le 1{,}96) = 0{,}95$$

$$\Downarrow$$

$$P\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \le \mathrm{m} \le \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

L'intervalle $\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}};\ \bar{x} + z_c \frac{\sigma}{\sqrt{n}}\right]$ est **l'intervalle de confiance** de la moyenne **m** de la population avec le niveau de confiance $2\Pi(z_c) - 1$

ayant pour centre la moyenne $\bar{x}$ de l'échantillon considéré (de taille $n \geq 30$)

Remarques

- Si σ est inconnue on l'estime par $\hat{s} = \sqrt{\frac{n}{n-1}}\, s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$; dans ce cas l'intervalle de confiance est donc donné par : $\left[\bar{x} - z_c \frac{s}{\sqrt{n-1}},\ \bar{x} + z_c \frac{s}{\sqrt{n-1}}\right]$

- Dans le cas de tirage **sans remise** l'intervalle de confiance de la moyenne est

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}};\ \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}\right]$$

Niveau de confiance : cas particuliers usuels

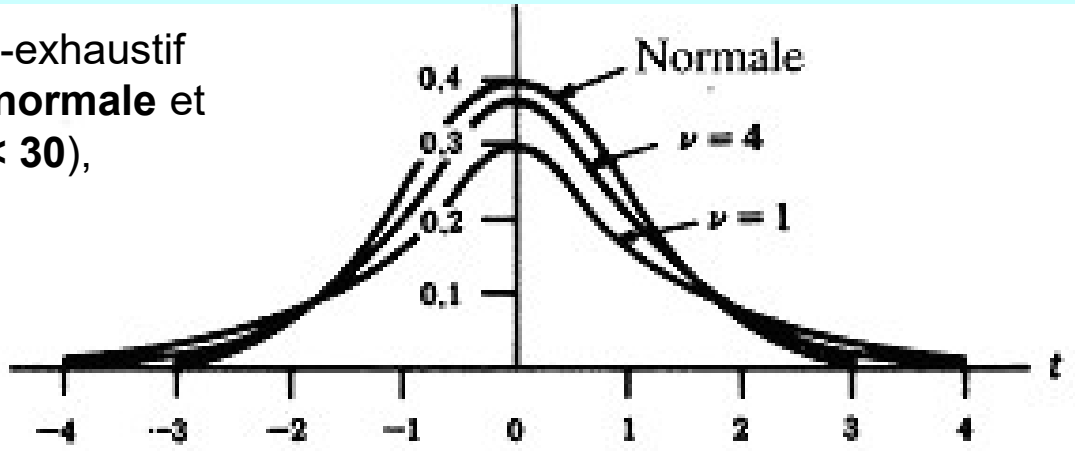
Niveau de confiance	99,73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
Valeur critique z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

Estimation de la moyenne par intervalle de confiance: **cas des petits échantillons**

Dans le cas où l'échantillonnage non-exhaustif s'effectue à partir d'une **population normale** et la taille de l'échantillon est petite (**n < 30**), la statistique

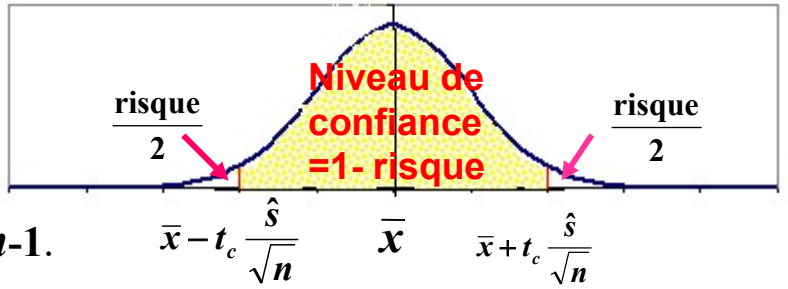
$$t = \frac{\sqrt{n}}{\hat{s}}(\overline{X} - \mathrm{m})$$

suit la **distribution de Student.**



L'intervalle de confiance:

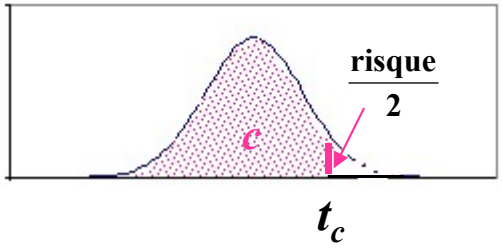
$$\left[\overline{x} - t_c \frac{\hat{s}}{\sqrt{n}};\ \overline{x} + t_c \frac{\hat{s}}{\sqrt{n}}\right]$$



Le coefficient de confiance t_c est calculé à partir de la distribution de Student; il dépend de du niveau de confiance choisi et du nombre de degré de liberté **ν = *n*-1**.

Extrait de la table de Student:

Coefficient de confiance t_c $\left(c = 1 - \frac{\text{risque.}}{2}\right)$	$t_{0.995}$	$t_{0.99}$	$t_{0.975}$	$t_{0.95}$	$t_{0.90}$
ν = *n*-1 = 10	3.17	2.76	2.23	1.81	1.37
ν = *n*-1 = 15	2.95	2.60	2.13	1.75	1.34
ν = *n*-1 = 20	2.84	2.53	2.09	1.72	1.32



Intervalle de confiance de la moyenne m :

	Population infinie et (ou) échantillonnage non- exhaustif (avec remise)	Population finie et échantillonnage exhaustif (sans remise)
$n \geq 30$	$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$
$n < 30$	$\bar{x} \pm t_c \frac{\hat{s}}{\sqrt{n}}$	$\bar{x} \pm t_c \frac{\hat{s}}{\sqrt{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$

- Si σ est inconnue on estime par $\hat{s} = \sqrt{\frac{n}{n-1}}\, s$, où **s** est l'écart type de l' échantillon.
- En pratique, on néglige le facteur $\sqrt{\frac{N_p - n}{N_p - 1}}$ si $\frac{n}{N_p} \leq 0{,}1$.

Estimation de la fréquence par intervalle de confiance

p **:** proportion d'éléments de la population possédant une certaine propriété (fréquence de « succès » dans la population)

f **:** proportion d'éléments de l'échantillon possédant une certaine propriété (fréquence de « succès » dans l'échantillon)

n≥ 30

Intervalle de confiance: $\left[f - z_c \sigma_F \, , \; f + z_c \sigma_F \right]$

Population infinie et (ou) échantillonnage non- exhaustif (avec remise)	population finie et échantillonnage exhaustif (sans remise)
$f \pm z_c \sqrt{\frac{p(1-p)}{n}}$	$f \pm z_c \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$

Remarques

- Si σ_F n'est connu, on peut l'estimer par :

$$\hat{s}_F = \sqrt{\frac{f(1-f)}{n}}$$ (cas d'échantillonage non-exhausfif)

$$\hat{s}_F = \sqrt{\frac{f(1-f)}{n}} \sqrt{\frac{N_p - n}{N_p - 1}}$$ (cas d'échantillonage exhausfif)

4. TESTS STATISTIQUES

Tests de validité d'hypothèse relatifs à une moyenne

Exemple introductif

A partir d'un échantillon de 36 pièces nous avons obtenu la masse moyenne 774,7 g et l'intervalle [770,61 ; 778,79] comme intervalle de confiance de la moyenne m de la population de 500 pièces (avec niveau de confiance 95%).

Question :

- Peut-on considérer que les 500 pièces de la population ont une masse moyenne de 780 g, comme le prévoient les conditions fixées par le fournisseur ?

- Autrement dit, doit-on accepter ou refuser la livraison de ces 500 pièces au vu du résultat obtenu sur l'échantillon ?

Hypothèse nulle (notée H_o) :

La moyenne de la population est **m =** 780.

- Une erreur de première espèce: rejet d'une hypothèse alors qu'elle est vraie.
- Une erreur de deuxième espèce : acceptation d'une hypothèse alors qu'elle est fausse.

Hypothèse nulle (notée Ho) : La moyenne de la population est m = m_o.

Supposons H_o vraie $\Rightarrow$ $\overline{X}$ suit la loi normale $N\left(m_o, \frac{\sigma}{\sqrt{n}}\right)$

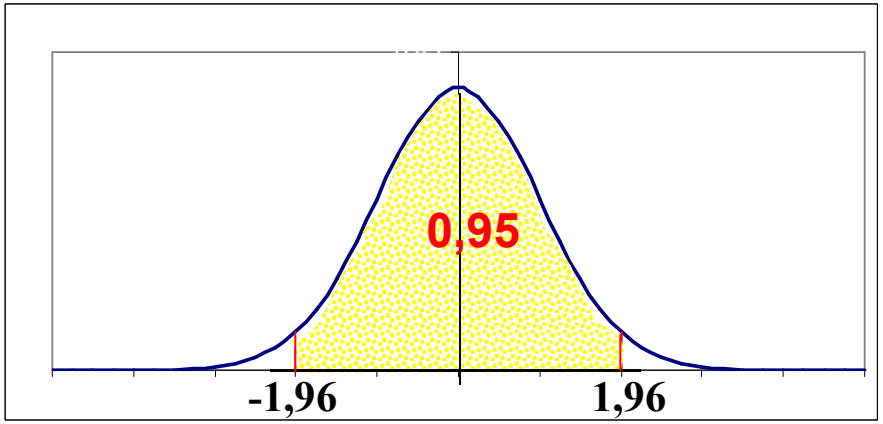
$\Downarrow$

$Z = \frac{\sqrt{n}}{\sigma}(\overline{X} - m_o)$ suit la loi normale centrée réduite ***N*(0,1)**

$\Downarrow$

$$P(-z_c < Z < z_c) = 2\Pi(z_c) - 1$$

Par exemple: Si on choisi le niveau de confiance **0,95** alors



$$2\Pi(z_c) - 1 = 0{,}95 \Rightarrow z_c = 1{,}96$$

$\Downarrow$

$$P(-1{,}96 \le \frac{\sqrt{n}}{\sigma}(\overline{X} - m_o) \le 1{,}96) = 0{,}95$$

$\Downarrow$

$$P\left(m_o - 1{,}96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le m_o + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

En supposant que **m = m_o**, on sait, avant de prélever un échantillon aléatoire de taille *n*, que sa moyenne appartient à l'intervalle $\left[m_o - 1{,}96\frac{\sigma}{\sqrt{n}},\ m_o + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$ avec la probabilité 0,95.

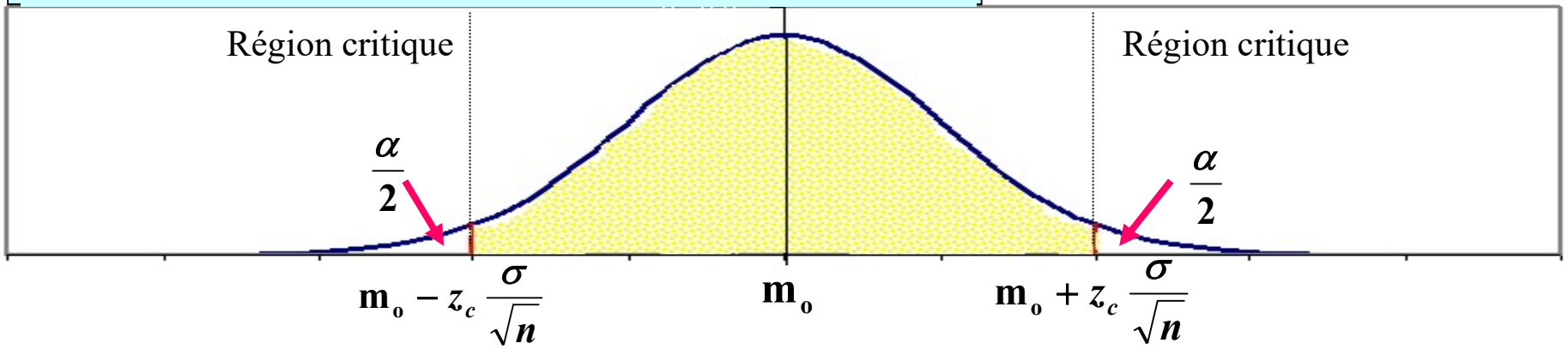
Si l'hypothèse **H₀** est vraie, la moyenne d'un échantillon aléatoire de taille *n* *(n ≥ 30)* appartient à l'intervalle $\left[\mathrm{m}_0 - z_c \frac{\sigma}{\sqrt{n}}, \mathrm{m}_0 + z_c \frac{\sigma}{\sqrt{n}}\right]$ avec la probabilité correspondant au niveau de confiance choisi.

Notons $\alpha = 1 - \textit{niveau de confiance}$ (fixé *a priori*; valeurs habituelles α=0,05 ou α=0,01)

Règle de décision

On prélève un échantillon aléatoire non exhaustif de taille *n* et on calcule sa moyenne $\bar{x}$.

- Si $\bar{x} \notin \left[\mathrm{m}_0 - z_c \frac{\sigma}{\sqrt{n}}, \mathrm{m}_0 + z_c \frac{\sigma}{\sqrt{n}}\right]$, on rejette **H₀**.
- Si $\bar{x} \in \left[\mathrm{m}_0 - z_c \frac{\sigma}{\sqrt{n}}, \mathrm{m}_0 + z_c \frac{\sigma}{\sqrt{n}}\right]$, on accepte **H₀** .



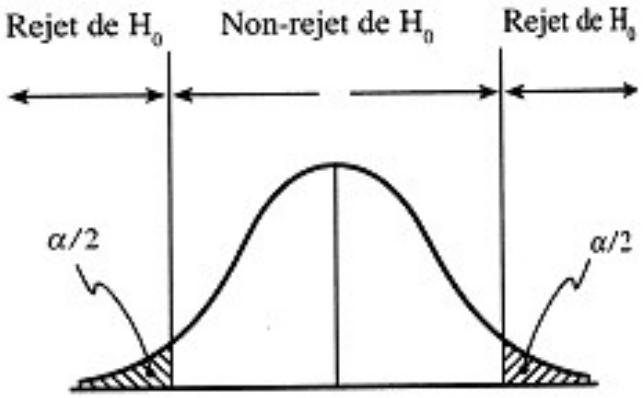
- **Le seuil de signification α définit la probabilité de rejeter à tort H₀**
  (il correspond à la probabilité de l'erreur de première espèce)

- Test bilatéral: la région critique est située des deux côtés de la région où on l'accepte H₀.

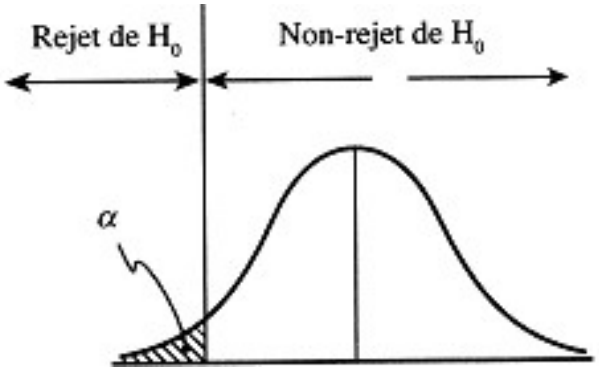
Hypothèse H_o: $m = m_o$

Hypothèse alternative H_1:

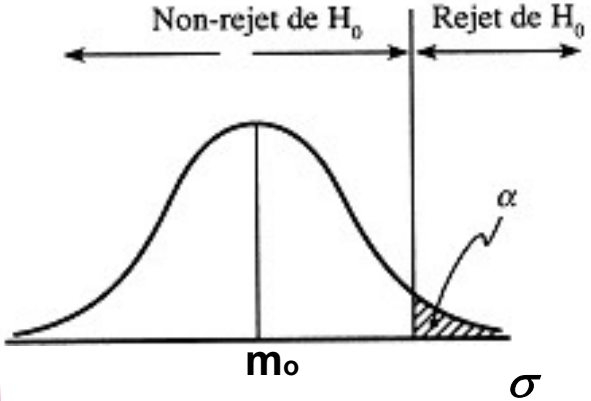
$m \neq m_o$	$m < m_o$	$m > m_o$
$\Downarrow$	$\Downarrow$	$\Downarrow$
Test bilatéral	**Test unilatéral à gauche**	**Test unilatéral à droite**



$$m_o - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad m_o + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



$$m_o - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$



$$m_o + z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Rejeter H_o si

$$\bar{x} \notin \left[m_o - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, m_o + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

Rejeter H_o si

$$\bar{x} < m_o - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Rejeter H_o si

$$\bar{x} > m_o + z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

$$\alpha = 0{,}05 \Rightarrow z_{\alpha/2} = 1{,}96$$

$$\alpha = 0{,}01 \Rightarrow z_{\alpha/2} = 2{,}58$$

$$\alpha = 0{,}05 \Rightarrow z_{\alpha} = 1{,}645$$

$$\alpha = 0{,}01 \Rightarrow z_{\alpha} = 2{,}33$$

Tests de validité d'hypothèse relatifs à une moyenne: **cas des petits échantillons**

- Dans le cas où l'échantillonnage non-exhaustif s'effectue à partir d'une **population normale** et la taille de l'échantillon est petite (**n < 30**),la statistique suit la **distribution de Student.** $t = \frac{\sqrt{n}}{\hat{s}}(\bar{X} - \mathrm{m})$

- Dans ce cas valeur critique $\mathbf{t_c}$ dépend du nombre de degré de liberté $\nu = n - 1$.

Pour un test bilatéral:

$$t_c = t_{n-1,1-\alpha/2}$$

Pour un test unilatéral:

$$t_c = t_{n-1,1-\alpha}$$

Test de comparaisons des moyennes de deux populations

Exemple introductif

Un second fournisseur B livre 800 pièces du même modèle que le fournisseur A. On prélève au hasard et avec remise un échantillon de 50 pièces dont on mesure la masse.

Population B

Moyenne $\mathbf{m_B}$ inconnue

Écart type:

$\sigma_B \rightarrow \hat{s}_B = 12{,}1$

Échantillon

$n = 50$

$\overline{x_B} = 779{,}6$

$s_B = 11{,}99$

Population A

Moyenne $\mathbf{m_A}$ inconnue

Écart type:

$\sigma_A \rightarrow \hat{s}_A = 12{,}5$

Échantillon

$n = 36$

$\overline{x_A} = 774{,}7$

$s_A = 12{,}36$

Question :

- La différence entre ces moyennes provient-elle d'une différence entre les productions des deux fournisseurs ou du choix des échantillons ?
- Autrement dit, comment construire un test permettant de décider, à partir de deux échantillons, s'il y a une différence significative, au seuil de α **%** , entre les moyennes des masses des pièces livrées par les deux fournisseurs.

$\overline{X_A}$: la variable aléatoire qui, à tout échantillon de taille n_A associe la moyenne de l'échantillon $\overline{x_A}$

($n_A \geq 30$) ; elle suit approximativement la loi normale $N\left(\mathbf{m}_A, \frac{\sigma_A}{\sqrt{n_A}}\right)$

$\overline{X_B}$: la variable aléatoire qui, à tout échantillon de taille n_B associe la moyenne de l'échantillon $\overline{x_B}$

($n_B \geq 30$) ; elle suit approximativement la loi normale $N\left(\mathbf{m}_B, \frac{\sigma_B}{\sqrt{n_B}}\right)$

- Supposons que les variables $\overline{X_A}$ et $\overline{X_B}$ sont indépendantes.

La variable $D = \overline{X_B} - \overline{X_A}$ suit une loi normale

$$\mathbf{m}_D = E(\overline{X_B} - \overline{X_A}) = E(\overline{X_B}) - E(\overline{X_A}) = \mathbf{m}_B - \mathbf{m}_A$$

$$\sigma_D^2 = Var(\overline{X_B} - \overline{X_A}) = Var(\overline{X_B}) + Var(\overline{X_A}) = \frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}$$

$$\Rightarrow Z = \frac{D - \mathbf{m}_D}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \text{ suit la loi N(0,1)}$$

Hypothèse nulle H_o : $\mathbf{m}_A = \mathbf{m}_B \quad (\mathbf{m}_D = 0)$

Il n'y a pas différence des moyennes entre les deux populations

Soit d la différence de moyennes de deux échantillons $d = \overline{x_B} - \overline{x_A}$.

On peut tester H_o relativement à des hypothèses alternatives H_1:

$\mathbf{m}_D \neq 0$	$\mathbf{m}_D < 0$	$\mathbf{m}_D > 0$
Rejeter H_o si	**Rejeter H_o si**	**Rejeter H_o si**
$d \notin \left[-z_{\alpha/2}\sigma_D,\ z_{\alpha/2}\sigma_D\right]$	$d < -z_\alpha \sigma_D$	$d > z_\alpha \sigma_D$

Tests de validité d'hypothèse relatifs à une proportion

F : la variable aléatoire qui, à tout échantillon aléatoire d'effectif *n* fixé, associe la fréquence avec laquelle les éléments de cet échantillon possèdent certaine propriété.

p **:** proportion d'éléments de la population possédant une certaine propriété

n≥ 30 **np>5 et n(1-p)>5**.

F suit approximativement la loi normale $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$; $\sigma_F = \sqrt{\frac{p(1-p)}{n}}$

Hypothèse Ho : **p = p_o.**

La proportion d'éléments de la population possédant une certaine propriété est égale p_o.

Si H_o est vraie, alors $Z = \frac{F - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$ suit la loi normale N(0,1).

f **:** proportion d'éléments de l'échantillon possédant une certaine propriété

On peut tester $\mathbf{H_o}$ relativement à des hypothèses alternatives $\mathbf{H_1}$:

$p \neq p_o$.	$p < p_o$.	$p > p_o$.
Test bilatéral	**Test unilatéral à gauche**	**Test unilatéral à droite**
Rejeter H_o si	**Rejeter H_o si**	**Rejeter H_o si**
$f \notin [p_o - z_{\alpha/2}\sigma_F, p_o + z_{\alpha/2}\sigma_F]$	$f < p_o - z_\alpha \sigma_F$	$f > p_o + z_\alpha \sigma_F$

Test de comparaisons des proportions (des fréquences)

Population A

proportion p_A inconnue

F_A suit $\mathrm{N}\left(p_A, \sqrt{\frac{p_A(1-p_A)}{n_A}}\right)$

Échantillon

$n_A \geq 30$

f_A

Population B

proportion p_B inconnue

F_B suit $\mathrm{N}\left(p_B, \sqrt{\frac{p_B(1-p_B)}{n_B}}\right)$

Échantillon

$n_B \geq 30$

f_B

Hypothèse nulle H₀ :

$$p_A = p_B = p \qquad (p_B - p_A = 0)$$

Il n'y a pas différence entre les proportions des éléments possédant un certain caractère dans les deux populations

$$F_B - F_A \sim \mathrm{N}\left(0, \sqrt{\frac{p(1-p)}{n_A} + \frac{p(1-p)}{n_B}}\right)$$ alors $$Z = \frac{F_B - F_A}{\sigma_{F_B - F_A}}$$ suit N(0,1).

Estimation de **p**: $$p = \frac{n_A f_A + n_B f_B}{n_A + n_B}$$

où $$\sigma_{F_B - F_A} = \sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

On peut tester **H₀** relativement à des hypothèses alternatives **H₁**:

$p_B \neq p_A$

Rejeter H₀ si

$f_B - f_A \notin \left]-z_{\alpha/2}\sigma_{F_B-F_A},\ z_{\alpha/2}\sigma_{F_B-F_A}\right[$

$p_B < p_A$

Rejeter H₀ si

$f_B - f_A < -z_\alpha \sigma_{F_B-F_A}$

$p_B > p_A$

Rejeter H₀ si

$f_B - f_A > z_\alpha \sigma_{F_B-F_A}$

5. ANALYSE DE LA VARIANCE (À 1 FACTEUR)

Objectif : **Etude de la liaison entre une variable quantitative X et une variable qualitative Y à *k* modalités.**

Notations

- **La variable Y induit *k* classes (« sous-populations »);**
- **L'échantillon *i* a un effectif n_i.**
- **Effectif total:** $n = \sum_{i=1}^{k} n_i$
- $\overline{x_i}$ **est une estimation de** m_i
- **Moyenne empirique générale :** $\overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$

		Moyenne d'échantillon	Moyenne de sous-population
Y_1	$x_{11}, x_{12}, \ldots x_{1n_1}$	$\overline{x}_1$	m_1
Y_2	$x_{21}, x_{22}, \ldots\ldots x_{2n_2}$	$\overline{x}_2$	m_2
:	:	:	:
Y_k	$x_{k1}, x_{k2}, \ldots x_{kn_k}$	$\overline{x}_k$	m_k

On admet que les valeurs de la variable X sont distribuées normalement dans chaque «sous-populations».

Exemple

Un échantillonnage aléatoire de sol est réalisé sur trois sites A, B et C d'une usine polluante.

variable quantitative X :

Teneur en plomb exprimé en mg/kg .

***variable qualitative* Y :**

Les sites de la usine polluante (modalités : A, B, C)

Site A	**128**	**142**	**98**
Site B	**101**	**85**	
Site C	**67**	**86**	

Question: Est-ce qu'il y a une différence significative entre les teneurs moyennes en plomb de sols de trois sites A,B,C?

Hypothèse nulle à tester (Ho) : égalité des moyennes: $m_1 = m_2 = ... = m_k$

Mise en œuvre du Test

Equation d'analyse de la variance:

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

SCT = SCI + SCE

SCT Somme des Carrés Totale
(écarts de chaque mesure par rapport à la moyenne générale).

SCI Somme des Carrés Interne aux classes.

SCE Somme des Carrés Entre les classes.

Source de variation	Degrés de liberté	Somme des carrés	Carrés moyens	Statistique F
Entre les classes	$\nu_1 = k-1$	SCE	$s_1^2 = \frac{SCE}{k-1}$	$F = \frac{s_1^2}{s_2^2}$
Interne aux classes	$\nu_2 = n-k$	SCI	$s_2^2 = \frac{SCI}{n-k}$	*avec k-1 et n-k degrés de liberté*
Totale	n-1	SCT		

On lit sur la table de Fisher à *k-1* et *n-k* degrés de liberté et au seuil de signification α , la valeur « critique » (théorique):

$$F_\alpha(k-1, n-k)$$

Remarque. Le seuil de signification α, fixé par l'utilisateur, définit la probabilité de rejeter à tort H_0

La règle de décision :

Si $F > F_\alpha(k-1, n-k)$ alors on rejette Ho
sinon l'hypothèse est acceptable.

Exemple

Effectif total:

$$n = \sum_{i=1}^{3} n_i = 3 + 2 + 2 = 7$$

Moyenne générale :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = 101$$

				Somme	Moyenne $\bar{x}_i$	$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$
Site A	128	142	98	368	122,67	1010,67
Site B	101	85		186	93	128
Site C	67	86		153	76,5	180,5
					SCI :	1319,17

Source de variation	Degrés de liberté	Somme des carrés	Carrés moyens	Statistique F
Entre les classes	k-1=3 -1= =2	SCE= 2736,83	$s_1^2 = \frac{SCE}{k-1}$ =1368,42	$F = \frac{s_1^2}{s_2^2}$ =4,15
Interne aux classes	n-k=7 - 3 =4	SCI= 1319,17	$s_2^2 = \frac{SCI}{n-k}$ = 329,79	*avec 2 et 4 degrés de liberté*
Totale	n-1=6	SCT		

$n_i(\bar{x}_i - \bar{x})^2$	
1408,33	
128	
1200,5	
SCE :	2736,83

$$F_c = F_{0.05}(2,4) = 6,94$$

$F < F_c$ **Décision : on ne rejette pas H₀.**

Au risque de 5% on peut admettre l'hypothèse que les teneurs moyennes en plomb sont les mêmes pour les trois sites en question (la différence constatée entre les moyennes empiriques est due au hasard d'échantillonnage).

Remarque. *Les résultas de l'analyse de la variance ne changent pas si on soustrait la même valeur à tous les éléments du tableau de données.*

6. Tests statistiques non paramétriques

Tests indépendantes de la distribution de la population et de ses paramètres.

Test khi-deux d'indépendance des deux variables

Données : **2 variables qualitatives X et Y ;**
X est à k modalités ; Y est à m modalités

Tableau de « contingence »
donnant les effectifs n_{ij}
(n_{ij} est le nombre d'individus possédant la modalité i de X et la modalité j de Y):

X \ Y	1	…	j	…	m	Total
1	n_{11}	…	n_{1j}	…	n_{1m}	L_1
…	…	…	…	…	…	…
i	n_{i1}	…	n_{ij}	…	n_{im}	L_i
…	…	…	…	…	…	…
k	n_{k1}	…	n_{kj}	…	n_{km}	L_k
Total	C_1	…	C_j	…	C_m	n

$$C_j = \sum_{i=1}^{k} n_{ij}$$

$$L_i = \sum_{j=1}^{m} n_{ij}$$

$$\text{effectif total} \quad n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$$

Exemple:

Tableau de satisfaction vis-à-vis du travail de 200 salariés d'une grande entreprise en fonction de la catégorie des salaires

Tableau de « contingence » (Tableau des effectifs observés):

Catégorie salariale Y / X Niveau de satisfaction	Niveau bas (moins de 20 000 €)	Niveau moyen (20 000 à 30 000 €)	Niveau élevé (plus que 30 000 €)	Total
élevé	13	19	25	57
moyen	28	29	28	85
faible	24	18	16	58
Total	65	66	69	200

Effectifs observés:

Catégorie salariale / X Niveau de satisfaction / Y	Moins de 20 000 €	20 000 à 30 000 €	Plus que 30 000 €	Total
élevé	13	19	25	57
moyen	28	29	28	85
faible	24	18	16	58
Total	65	66	69	200

♦**Hypothèse nulle H_0 :**
« X et Y sont indépendantes »

Sous l'hypothèse nulle H_0
l'effectif théorique est :

$$n_{Tij} = \frac{L_i \times C_j}{n}$$

Pour pouvoir appliquer le test du Khi Deux il faut que $n_{Tij} \geq 5$

$$n_{T11} = \frac{57 \times 65}{200} = 18{,}525 \qquad n_{T21} = \frac{85 \times 65}{200} = 27{,}625$$

Effectifs théoriques:

Catégorie salariale / Niveau de satisfaction	Moins de 20 000 €	20 000 à 30 000 €	Plus que 30 000 €	Total
élevé	18.525	18.81	19.665	57
moyen	27.625	28.05	29.325	85
faible	18.85	19.14	20.01	58
Total	65	66	69	200

♦**L'indicateur d'écart**
(la valeur observée de Khi-Deux) :

$$\chi_o^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{Oij} - n_{Tij}\right)^2}{n_{Tij}}$$

Dans notre exemple:

$$\chi_o^2 = \frac{(13-18{,}525)^2}{18{,}525} + \frac{(28-27{,}625)^2}{27{,}625} + \frac{(24-18{,}85)^2}{18{,}85} + \ldots + \frac{(16-20{,}01)^2}{20{,}01} = 5{,}47$$

♦ **On lit sur la table du Khi-Deux à *(k-1)(m-1)* degrés de liberté et au seuil de signification α , *la valeur « critique » ou théorique***

Dans notre exemple:

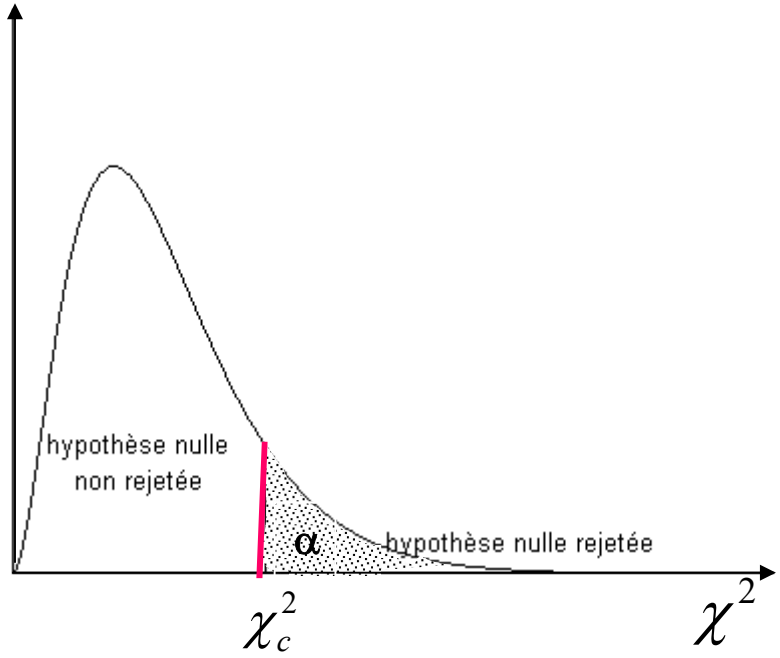
- **nombre de degrés de liberté: *(k-1)(m-1) = (3-1)(3-1) = 4***
- **Le seuil de signification α (la probabilité de rejeter à tort H_0) est fixé par l'utilisateur: α = *0.05***

$$\chi_c^2 = 9{,}49$$

Règle de décision au seuil de signification α:

Si $\chi_o^2 > \chi_c^2$ **alors** ***on rejette l'hypothèse H_0***

sinon *l'hypothèse est vraisemblable*



Dans notre exemple:

$$\chi_o^2 = 5{,}47 \quad \text{et} \quad \chi_c^2 = 9{,}49$$

$\chi_o^2 < \chi_c^2$ **alors l'hypothèse H_o est vraisemblable (les deux variables sont indépendantes)**

Dans la grande entreprise étudiée le niveau de satisfaction du travail des salariés ne dépend pas du niveau de salaire.

Mesure d'association entre deux variables: coefficient de Cramer

Dans le cas où l'hypothèse d'indépendance des deux variables est rejetée, on peut quantifier l'association entre deux variables à l'aide de **coefficient de Cramer** définie par:

$$V = \sqrt{\frac{\chi_0^2}{n(k-1)}}$$

où k est le minimum entre le nombre de lignes et le nombre de colonnes et n est la taille de l'échantillon.

Le coefficient V de Cramer varie entre 0 (aucune association entre les variables) et 1 (parfaite association entre les variables)

Recherche d'un modèle probabiliste de la distribution de donnés

- On cherche à déterminer si un modèle théorique est susceptible de représenter adéquatement le comportement probabiliste de la variable observée et ceci basé sur les effectifs obtenus sur l'échantillon.

Etapes à suivre:

1) Représentation des données obtenues sur l'échantillon sous forme de tableau et/ ou sous forme graphique

2) Choix d'un modèle probabiliste en fonction des observations obtenues

3) Estimation des paramètres de la distribution théorique

4) Vérification de la qualité d'ajustement à l'aide du test khi deux de conformité

Estimation des paramètres de la distribution théorique

Les modèles probabilistes fréquemment employés en statistique sont

-dans le cas de variables discrètes : la distribution binomiale, la distribution de Poisson

-dans cas de variables continues : la distribution normale, la distribution exponentielle

Distribution du caractère X	Paramètres nécessaires pour calculer probabilités	Estimation
Poisson $P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$	λ	$\hat{\lambda}=\bar{x}$ taux moyen de succès dans unité de mesure
binomiale $P(X=k)=C_n^k p^k (1-p)^{n-k}$	p (pour une taille n fixée)	$\hat{p}=\frac{\bar{x}}{n}$ $\bar{x}$ taux moyenne de sucés dans une suite d'échantillons de taille n
normale $P(X<x_0)=\int_{-\infty}^{x_0}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}dx$	m, σ	$\bar{x}$, $\hat{s}$ estimation de la moyenne et de l'écart type à partir de l''échantillon
exponentielle $P(X>x_o)=e^{-\lambda x_o}$	λ	$\lambda=1/E(X)$

Test khi deux de conformité (ou d'ajustement)

Comparaison d'une distribution observée à k classes à une distribution théorique

tableau de effectifs observées, k classes

1	2	3	…	..	k
n_{o1}	n_{o2}	n_{o3}			n_{ok}

tableau de effectifs théoriques, k classes

1	2	3	…	..	k
n_{T1}	n_{T2}	n_{T3}	…	…	n_{Tk}

il faut que $n_{T_i} \geq 5$

- **L'indicateur d'écart (la valeur observée de Khi-Deux) :** $\chi_o^2 = \sum_{i=1}^{k} \frac{(n_{oi} - n_{Ti})^2}{n_{Ti}}$

- Nombre de degrés de liberté υ **= k- 1- r** **r**: nombre de paramètres inconnues;

  Si la distribution théorique est complètement spécifiée υ= k-1; pour la distribution selon la loi normale avec m et σ inconnues υ= k-1-2=k-3

- On lit sur la table du Khi-Deux à ν degrés de liberté et au seuil de signification α , *la valeur « critique » ou théorique* χ_c^2

<u>Règle de décision au seuil de signification α:</u>

Si $\chi_o^2 < \chi_c^2$ la différence entre la distribution observée et la distribution théorique n'est pas significative au seuil α.

Si $\chi_o^2 > \chi_c^2$ la différence est significative.

Exemple:

Lors d'un choix portant sur 3 possibilité A, B et C, les préférences de 120 sujets sont réparties comme suit : 34 pour A, 44 pour B, 42 pour C.

Question:

Peut-on rejeter l'hypothèse d'équiprobabilité des choix (risques 5%) ?

tableau de effectifs observées

A	B	C
34	**44**	**42**

tableau de effectifs théoriques

A	B	C
40	**40**	**40**

Nombre degrés de liberté ν **= k-1=3-1=2**

α **= 5%** $\Rightarrow$ $\chi_c^2 = 5{,}99$

- L'indicateur d'écart

(la valeur observée de Khi-Deux) :

$$\chi_o^2 = \sum_{i=1}^{k} \frac{(n_{oi} - n_{Ti})^2}{n_{Ti}} = \frac{(34-40)^2}{40} + \frac{(44-40)^2}{40} + \frac{(42-40)^2}{40} = 1{,}4$$

$\chi_o^2 < \chi_c^2$ $\Rightarrow$ **Au seuil de 5% on ne peut pas rejeté l'hypothèse d'équiprobabilité des choix ; elle vraisemblable.**

7. FIABILITE

Dans le langage courant :

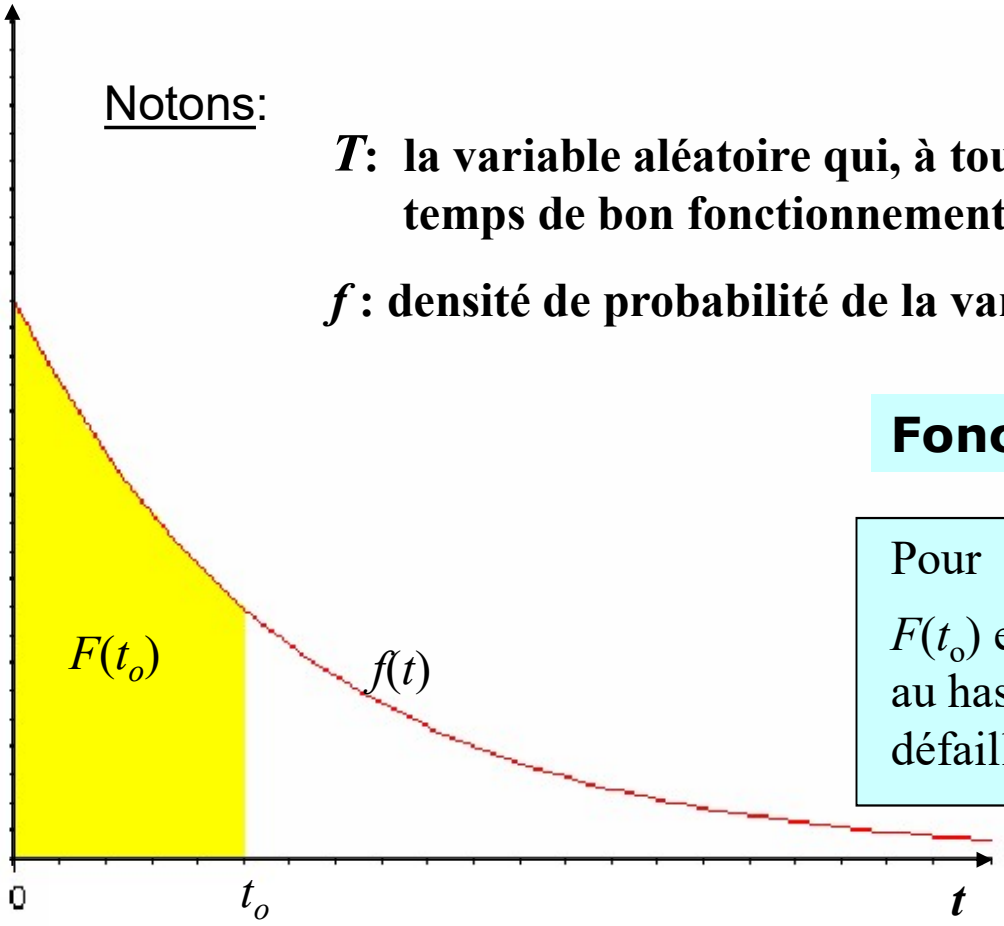
• Un modèle est plus fiable qu'un autre signifie que, en général, qu'un appareil de ce modèle fonctionne correctement plus longtemps qu'un appareil de l'autre modèle ; il s'agit évidemment d'une tendance, non d'une certitude.

• Une défaillance peut être une panne, une avarie, un fonctionnement incorrect,...

Notons:

***T*: la variable aléatoire qui, à tout dispositif tiré au hasard, associe son temps de bon fonctionnement (sa durée de vie avant une défaillance).**

***f* : densité de probabilité de la variable T**



Fonction de défaillance

Pour tout $t_o \geq 0$ $\quad F(t_o) = P(T < t_o)$

$F(t_o)$ est la probabilité qu'un dispositif prélevé au hasard dans la population considérée ait une défaillance avant l'instant t_o.

$T \geq t_o$ **est l'événement contraire de** $T < t_o$ **.**

$$\Downarrow$$

$$P(T \geq t_o) = 1 - P(T < t_o)$$

$$\Downarrow$$

$$P(T \geq t_o) = 1 - F(t_o)$$

Fonction de fiabilité

Pour tout $t_o \geq 0$ $\quad R(t_o) = 1 - F(t_o)$

$R(t_o)$ est la probabilité qu'un dispositif prélevé au hasard dans la population considérée n'ait pas de défaillance avant l'instant t_o.

Estimation de $F(t)$ et $R(t)$

On peut estimer $F(t)$ et $R(t)$ à partir de valeurs observées sur un échantillon.

Exemple:

On a mesuré pour 20 éléments du même type la durée de vie, en heures, avant la première défaillance.

Intervalle de temps	Nombre d'éléments défaillants dans cet intervalle	Instant t_i (en heures)	Nombre total n_i d'éléments défaillants à l'instant t_i	Estimation de $F(t_i)$ par $\frac{n_i}{20}$	Estimation de $R(t_i) = 1 - F(t_i)$
[0,500]	7	500	7	0,35	0,65
]500,1000]	4	1000	11	0,55	0,45
]1000,1500]	3	1500	14	0,70	0,30
]1500,2000]	2	2000	16	0,80	0,20
]2000,2500]	2	2500	18	0,90	0,10
]2500,3000]	1	3000	19	0,95	0,05
]3000,4000]	1	4000	20	1,00	0

n : taille d'échantillon

n_i : nombre d'éléments défaillants à l'instant t_i

Estimation de $F(t_i)$:

- **Méthode des rangs bruts :** $F(t_i) = \dfrac{n_i}{n}$ (si $n \geq 50$)

- **Méthode des rangs moyens :** $F(t_i) = \dfrac{n_i}{n+1}$ (si $20 \leq n < 50$)

- **Méthode des rangs médians :** $F(t_i) = \dfrac{n_i - 0{,}3}{n + 0{,}4}$ (si $n < 20$)

Taux d'avarie

Taux d'avarie moyen par unité de temps entre les instants t_{i-1} et t_i :

$$\frac{F(t_i)-F(t_{i-1})}{R(t_{i-1})(t_i-t_{i-1})}$$

Exemple:

Instant t_i (en heures)	Estimation de $F(t_i)$ par $\frac{n_i}{n}$	Estimation de $R(t_i)=1-F(t_i)$
2000	0,80	0,20
2500	0,90	0,10

Fréquence de défaillances entre les instants t_{i-1} =2000 et t_i =2500 : $\frac{F(t_i)-F(t_{i-1})}{R(t_{i-1})}=\frac{0,90-0,80}{0,20}=50\%$

Taux d'avarie moyen par unité de temps entre les instants t_{i-1} =2000 et t_i =2500 :

$$\frac{F(t_i)-F(t_{i-1})}{R(t_{i-1})(t_i-t_{i-1})}=\frac{0,90-0,80}{0,20(2500-2000)}=0,1\%$$

Interprétation: Si nous avons par exemple 5000 appareils d'un même modèle, au cours chaque heure séparant les instants t_{i-1} =2000 et t_i =2500 il y a 5 appareils qui tombent en panne.

Notons $t_i - t_{i-1} = h$

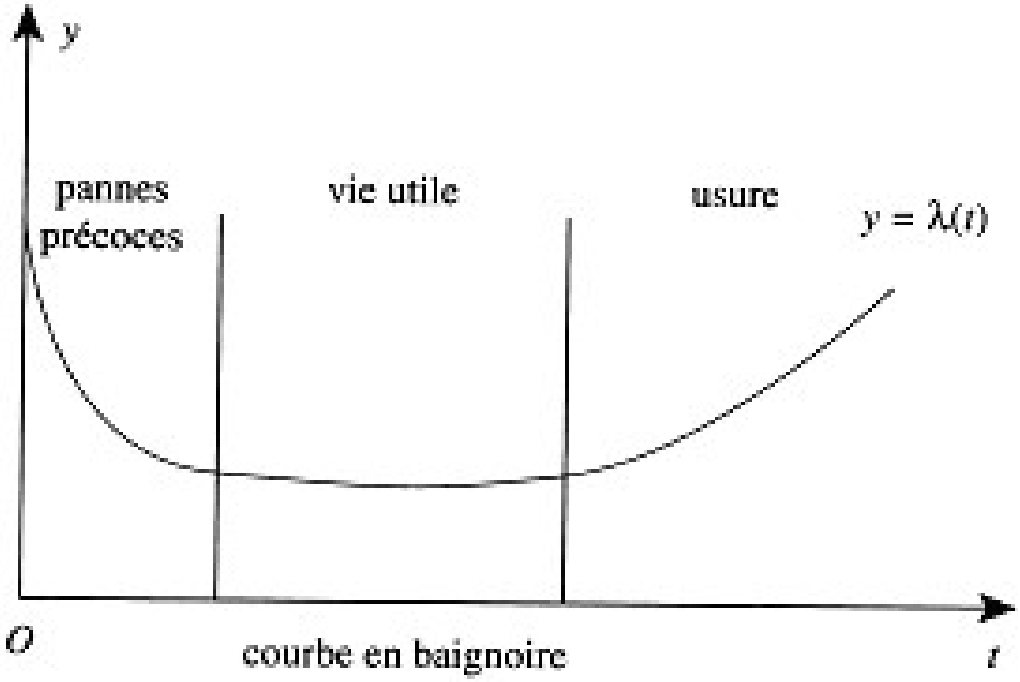
Taux d'avarie moyen: $\dfrac{F(t+h)-F(t)}{h\,R(t\ )}$

Alors le taux d'avarie instantané: $\lim\limits_{h\to 0} \dfrac{F(t+h)-F(t)}{h\,R(t)} = \dfrac{f(t)}{R(t)}$

Taux d'avarie (ou de défaillance) instantané à l'instant *t* : $\lambda(t) = \dfrac{f(t)}{R(t)}$

Allure de la courbe $\lambda(t)$:

- période des pannes précoces dues à défaut de fabrication
- « vie utile » où le taux d'avarie reste stable; les pannes dues au hasard
- Période d'usure où le taux d'avarie augmente avec le temps; les pannes dues à l'usure croissante du matériel.



$$\left.\begin{aligned} \lambda(t) &= \frac{f(t)}{R(t)} \\ R(t) &= 1 - F(t) \\ R'(t) &= -F'(t) = -f(t) \end{aligned}\right\} \Rightarrow \lambda(t) = \frac{f(t)}{1-F(t)} = \frac{F'(t)}{1-F(t)} \quad (1)$$

$$\lambda(t) = -\frac{R'(t)}{R(t)} \quad (2)$$

Si $\lambda(t)$ est connue on peut obtenir $F(t)$ et $R(t)$ à l'aide des équations différentielles (1) et (2):

$$R(t_o) = \exp\left[-\int_0^{t_o} \lambda(t)dt\right] \qquad F(t_o) = 1 - \exp\left[-\int_0^{t_o} \lambda(t)dt\right]$$

Moyenne des Temps de Bon Fonctionnement (MTBF: Mean Time Before Failure)

***T*: la variable aléatoire (de densité de probabilité *f*) qui, à tout dispositif associe son temps de bon fonctionnement.**

Espérance mathématique de *T*: $\mathbf{MTBF} = E(T) = \int_0^{+\infty} t\ f(t)\,dt$

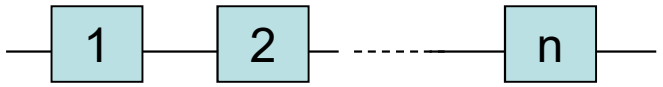
- La durée de vie moyenne d'un élément du type considéré avant sa première défaillance peut être calculer à partir d'un **très grand nombre** d'observations portant sur des éléments prélevés au hasard.

Fiabilité d'un système de n composantes

T_1, T_2, ..., T_n mesurant le temps de bon fonctionnement respectif de chacun des n composants

Nous supposons que les variables aléatoires T_1, T_2, ..., T_n sont indépendantes.

Montage en série



Un système est du type série pour fiabilité lorsqu'il ne fonctionne correctement que si tous ses composantes fonctionnent eux-mêmes correctement.

$$P(T \geq t) = P(T_1 \geq t) \cdot P(T_2 \geq t) \cdot \ldots \cdot P(T_n \geq t)$$

$$\Downarrow$$

Fonction de fiabilité du système:

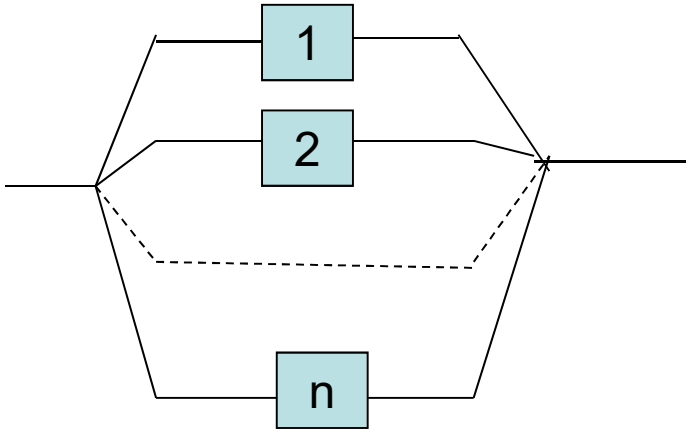
$$R(t) = R_1(t) \cdot R_2(t) \cdot \ldots \cdot R_n(t)$$

où R_i est la fonction de fiabilité de composant i

Fonction de défaillance du système:

$$F(t) = 1 - R_1(t) \cdot R_2(t) \cdot \ldots \cdot R_n(t) = 1 - \left(1 - F_1(t)\right) \cdot \ldots \cdot \left(1 - F_n(t)\right)$$

Montage en parallèle



Un système est du type parallèle pour la fiabilité lorsqu'il n'est défaillant que si tout ses composantes sont eux-mêmes défaillantes.

$$P(T < t) = P(T_1 < t) \cdot P(T_2 < t) \cdot \ldots \cdot P(T_n < t)$$

$$\Downarrow$$

Fonction de défaillance du système:

$$F(t) = F_1(t) \cdot F_2(t) \cdot \ldots \cdot F_n(t)$$

Fonction de fiabilité du système:

$$R(t) = 1 - F_1(t) \cdot F_2(t) \cdot \ldots \cdot F_n(t) = 1 - \left(1 - R_1(t)\right) \cdot \ldots \cdot \left(1 - R_n(t)\right)$$

Loi exponentielle

La loi exponentielle est la loi suivie par la variable *T* lorsque le taux d'avarie est constant; pour t ≥ 0 $\lambda(t) = \lambda$ constante strictement positive.

Remarque

Cette loi concerne tous les matériels pendant une partie de leur vie (vie utile) et les matériels électroniques pendant presque toute leur vie.

- **Fonction de fiabilité :**

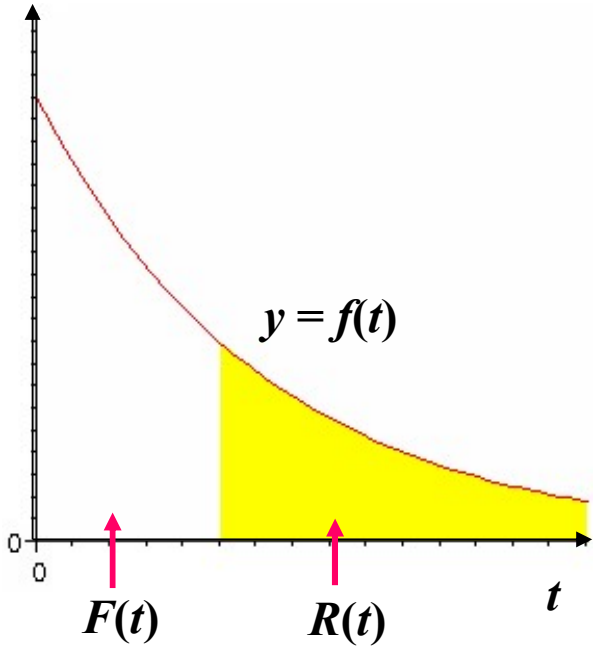
$$R(t) = e^{-\lambda t}$$

- **Fonction de défaillance:**

$$F(t) = 1 - e^{-\lambda t}$$

- **Densité de probabilité de la variable aléatoire *T*:**

$$f(t) = \lambda e^{-\lambda t}$$



MTBF (Espérance mathématique) dans le cas de la loi exponentielle

$$E(T) = \int_0^{+\infty} t\, f(t)dt = \int_0^{+\infty} t\, \lambda\, e^{-\lambda t} dt$$

$\Downarrow$ Intégration par parties

Moyenne des Temps de Bon Fonctionnement: $E(T) = \frac{1}{\lambda}$

Remarque $R(t) = e^{-\lambda t} \Rightarrow R\left(t = \frac{1}{\lambda}\right) = e^{-1} \approx 0{,}368$

Variance

$$V(T) = E(T^2) - \left(E(T)\right)^2 = \int_0^{+\infty} t^2 \lambda\, e^{-\lambda t} dt - \frac{1}{\lambda^2}$$

$\Downarrow$ Intégration par parties

$$V(T) = \frac{1}{\lambda^2}$$

Ecart type $\sigma(T) = \frac{1}{\lambda}$

8. Régression de Y en X

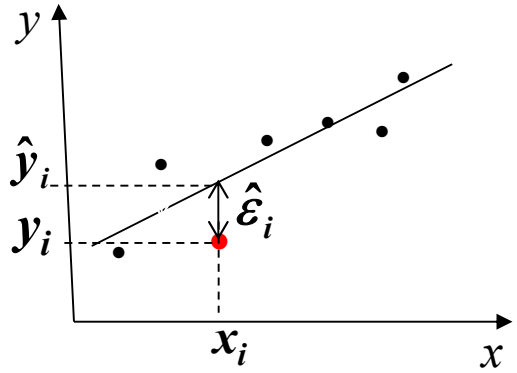
Ajustement par la méthode de moindres carrés ordinaires (MCO)

Données : n couples de points (x_i, y_i), i =1, 2, 3…n

Droite des moindres carrés ordinaires : $y = ax + b$

Valeur ajustée de y_i : $\hat{y}_i = ax_i + b$

Le résidu en *i* (erreur) : $\hat{\varepsilon}_i = y_i - \hat{y}_i$



Critère des MCO :
choix des paramètres (a,b) minimisant la somme des carrés des erreurs :

$$s(a,b) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Solution :

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - a\bar{x}$$

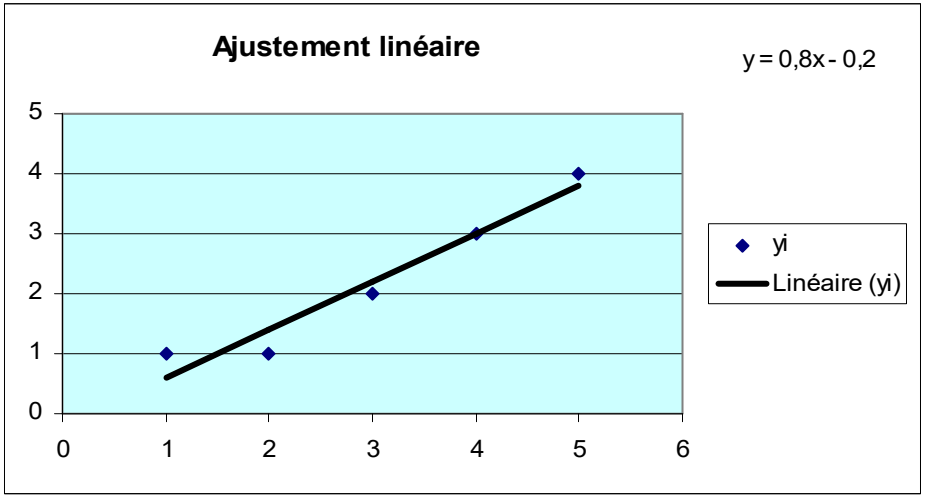
$\bar{x}$, $\bar{y}$, $\overline{xy}$, $\overline{x^2}$ correspondent aux moyennes simples respectivement sur x_i , y_i, x_iy_i, x_i^2.

Exemple

n=5

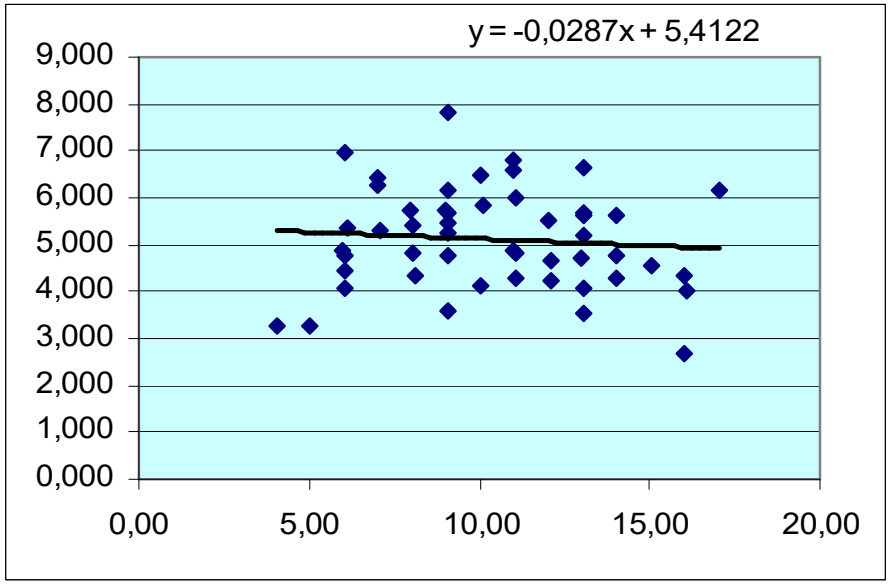
x_i	y_i	$x_i y_i$	x_i^2
1	1	1	1
2	1	2	4
3	2	6	9
4	3	12	16
5	4	20	25
15	11	41	55

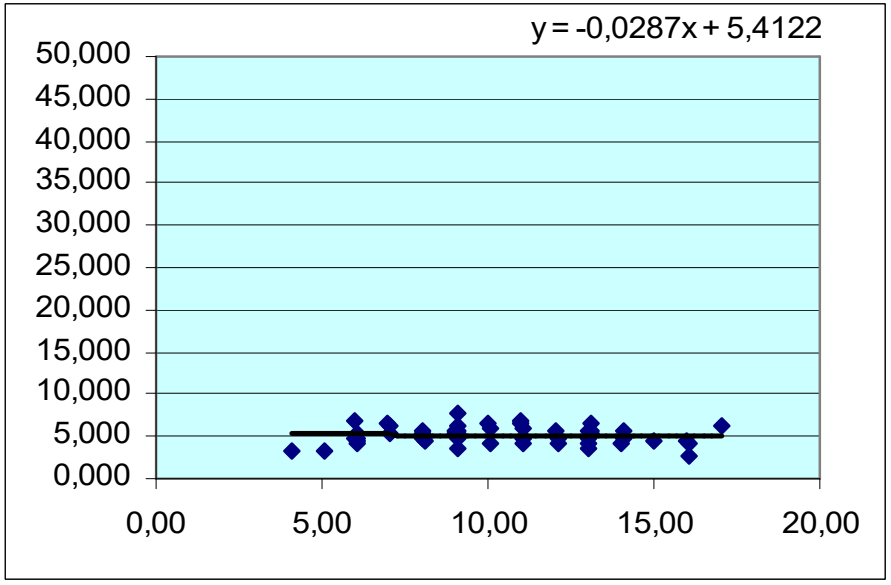
somme



$$\bar{x} = \frac{15}{5} = 3 \qquad \bar{y} = \frac{11}{5} = 2{,}2 \qquad a = \frac{41 - 5 \times 3 \times 2{,}2}{55 - 5 \times 3^2} = 0{,}8 \qquad b = 2{,}2 - 0{,}8 \times 3 = -0{,}2$$

Exemple



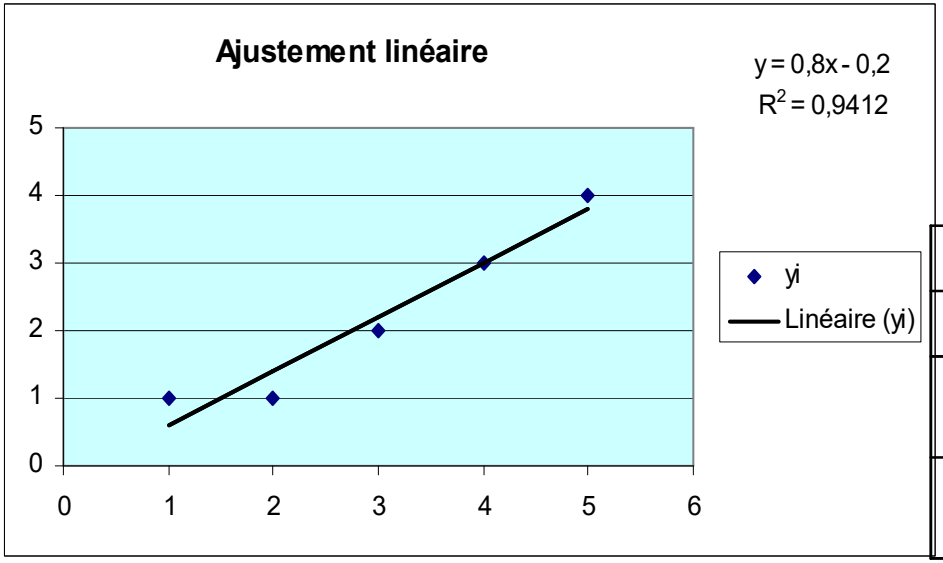


Coefficient de corrélation linéaire r ($\hat{\rho}$)

(mesure de l'efficacité de l'ajustement de la droite MCO aux données) : $r = \dfrac{\mathrm{Cov}_{\text{éch.}}(x,y)}{\sigma(x)\sigma(y)}$

La covariance de l'échantillon $\mathrm{Cov}_{\text{éch}}(x,y)$ est la moyenne empirique des produits, pour les deux variables, des écarts entre les valeurs observées et les moyennes d'échantillon.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\sqrt{\overline{y^2}-\bar{y}^2}}$$



$$a = \frac{\sigma(y)}{\sigma(x)} \cdot r \quad , \qquad b = \bar{y} - a\bar{x}$$

Droite de régression de Y en X:

$$\hat{y} - \bar{y} = \frac{\sigma(y)}{\sigma(x)} r(x - \bar{x})$$

a	0,8
b	-0,2
$r = \hat{\rho}$	0,9701425
$R^2 = \hat{\rho}^2$	0,9411765

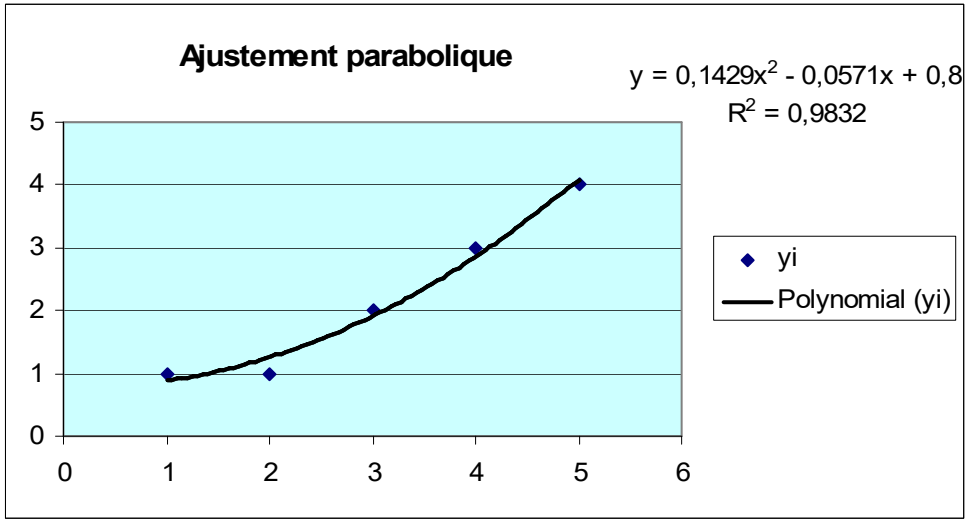
Le coefficient r, calculé à partir d'échantillon d'un taille n, donne une estimation ponctuelle du coefficient de corrélation linéaire de la population noté ρ.

Cas général : coefficient de détermination R^2

(mesure de l'efficacité de l'ajustement de la courbe aux données) :

$$R^2 = \frac{\text{variance de y expliquée par } x}{\text{variance totale de } y} = \frac{\sigma^2(\hat{y})}{\sigma^2(y)}$$

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



Remarque 1 Dans le cas d'ajustement linéaire $R^2 = \hat{\rho}^2$

Remarque 2

$$\sigma^2(y) = R^2\sigma^2(y) + \left(1 - R^2\right)\sigma^2(y)$$

Variance totale de y	Variance due à la régression	Variance résiduelle
$\sigma^2(y)$	$\sigma^2(\hat{y})$	$\sigma^2(\hat{\varepsilon})$

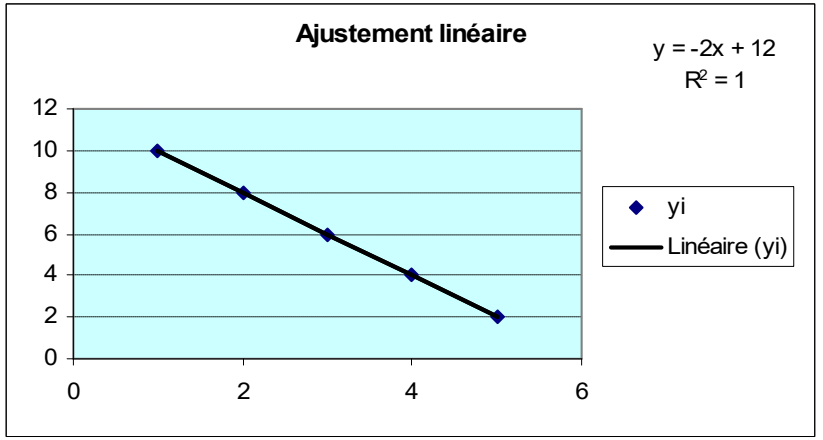
Propriétés du coefficient de corrélation linéaire

- r invariant par changement d'échelle et d'origine
- $-1 \leq r \leq 1$
- $r = \pm 1 \Leftrightarrow$ touts les points sont sur la droite $y = ax + b$ où signe de a = signe de r

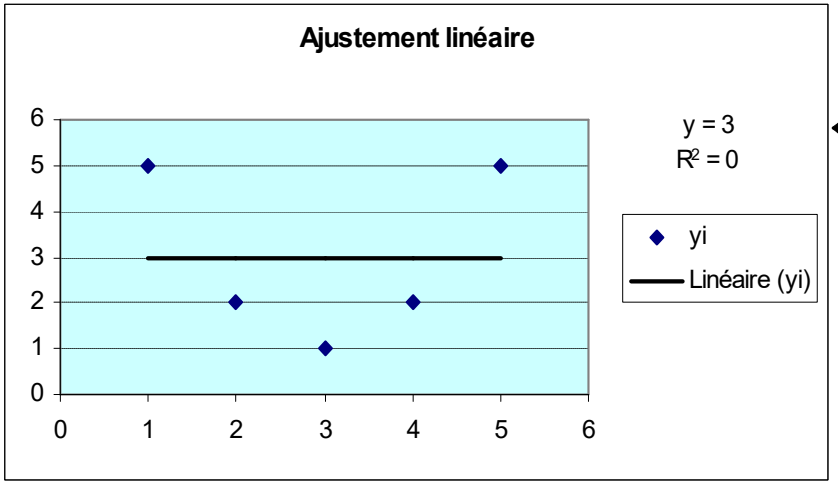
Exemple

$r = -1$

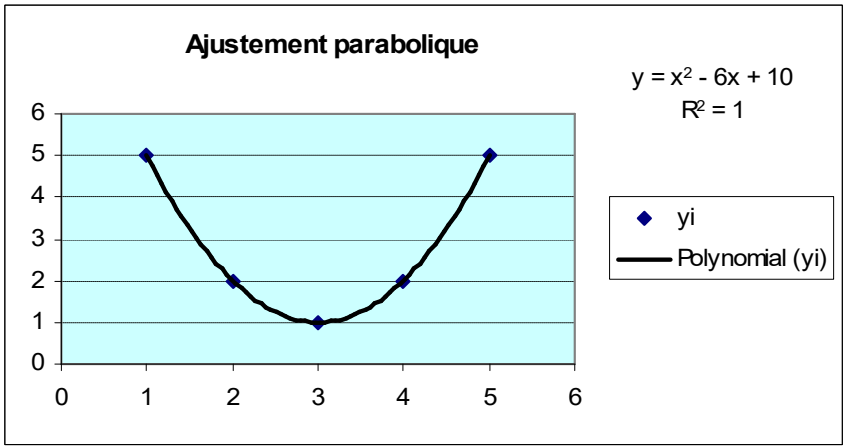
Corrélation linéaire parfaite négative →



← • $r = 0 \Leftrightarrow x_i$ et y_i non corrélées linéairement

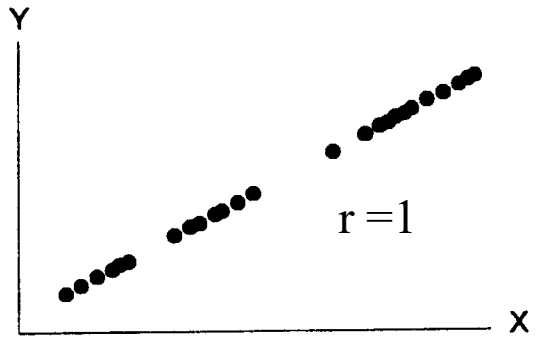


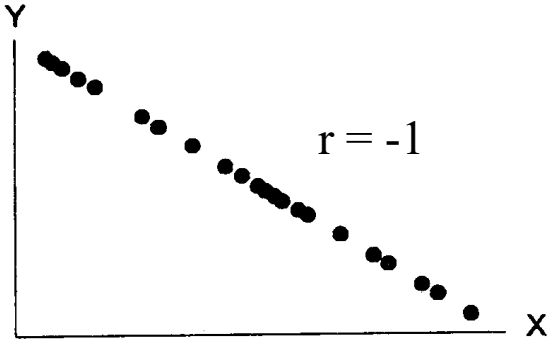
Corrélation parabolique parfaite →

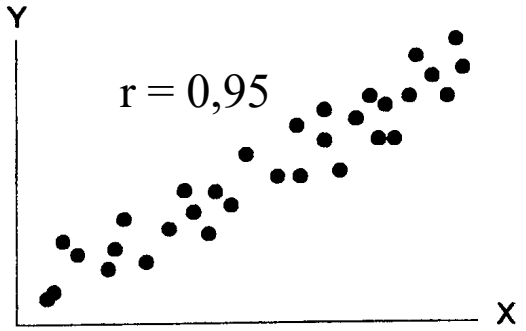


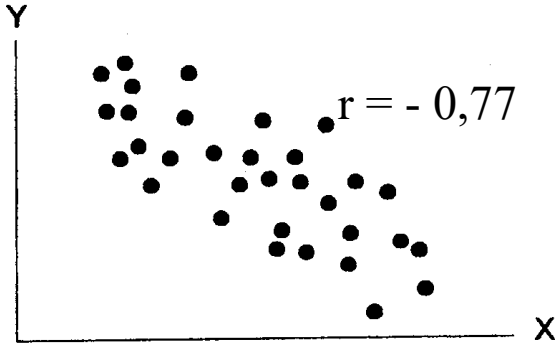
- x et y indépendantes (absence de relation) $\Rightarrow$ $r = 0$ (la réciproque est fausse)

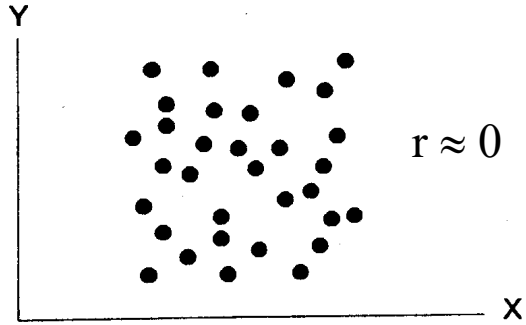
Exemples des nuages de points

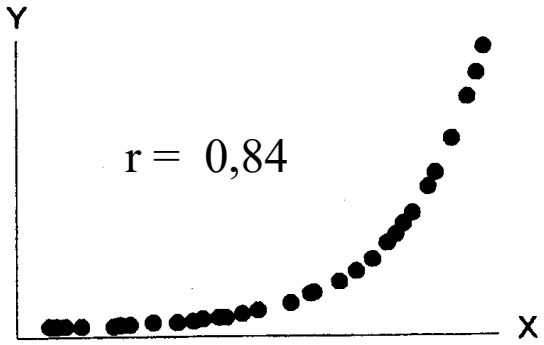


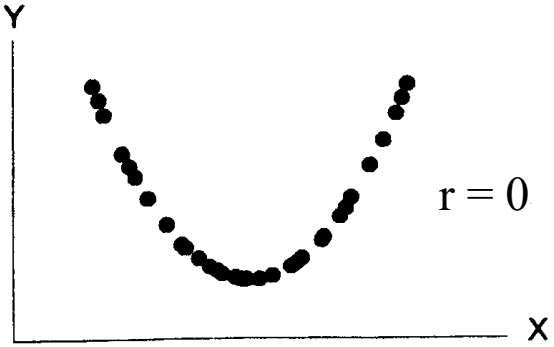












Test de signification r

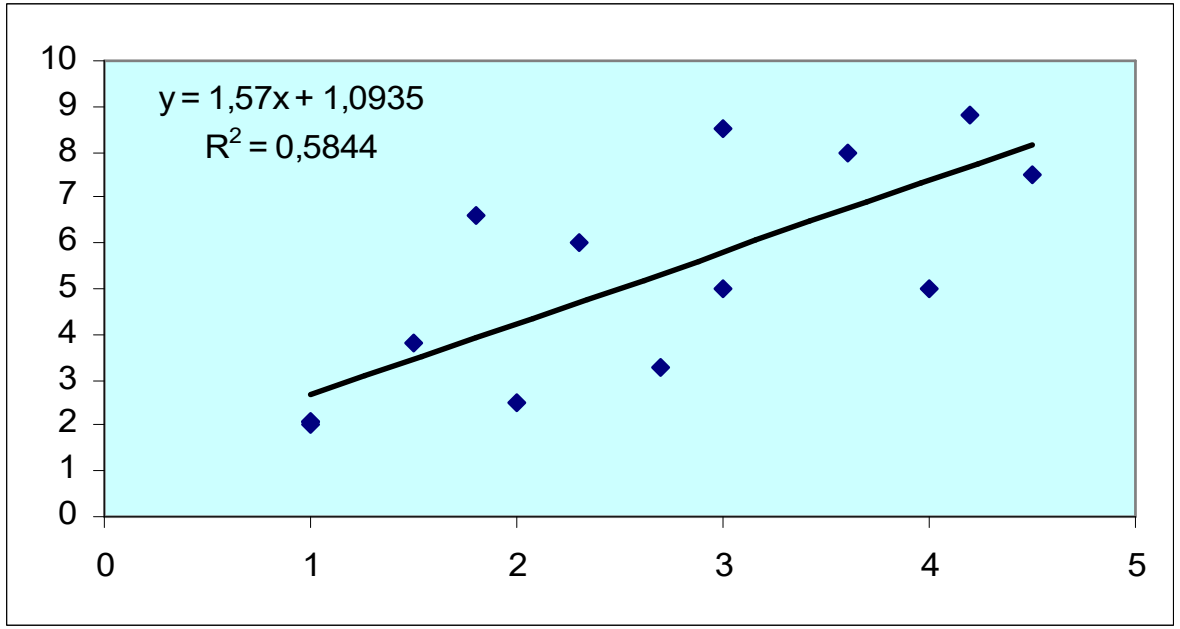
Question Le coefficient r (calculé à partir d'un échantillon) prouve-t-il (au risque α choisi) l'existence d'une corrélation dans la population ?

Exemple

un échantillon de taille $n = 14$

Le nuage de points montre une allure croissante, mais celle-ci est peut-être fortuite, due aux hasard de l'échantillonnage ?

Un nuage d'un petit nombre points peux avoir une allure croissante ou décroissante, alors qu'il n'existe pas de corrélation dans la population parente.



- Un coefficient de corrélation empirique égal à $r = \sqrt{0{,}5844} \approx 0{,}76$

- La corrélation observée sur l'échantillon a-t-il un caractère significatif ?

Hypothèse nulle H_o: X et Y ne sont pas corrélés ($\rho = 0$)

- Soit α le seuil de signification (la probabilité de rejeter à tort H_0).
- **On lit sur la table de *r* (coefficient de corrélation de Bravais – Pearson) à $\upsilon = n-2$ degrés de liberté et au seuil de signification α, *la valeur « critique » de* r_c**

Règle de décision au seuil de signification α:

Si $|r| > r_c$ alors *on rejette l'hypothèse H_o*

sinon *l'hypothèse est vraisemblable*

Dans notre exemple: $n = 14 \Rightarrow \upsilon = 12$; $|r| = 0{,}76$

On lit dans la table de r_c, à la ligne $\upsilon = 12$:

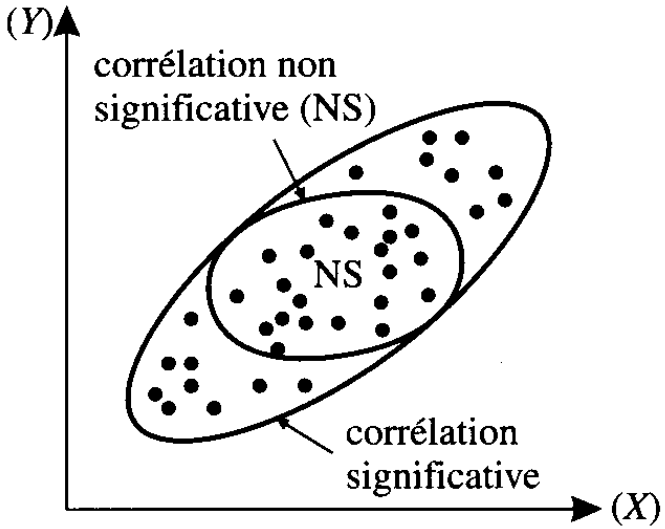
Risque α	5%	1%	0,1%
Valeurs critique r_c	0,5324	0,6614	0,7800

On rejette H_o au seuil de 1% , mais elle est vraisemblable au seuil de 0,1%.

L'allure croissante du nuage de points a une probabilité < 1%, mais >0,1%, d'être due au hasard.

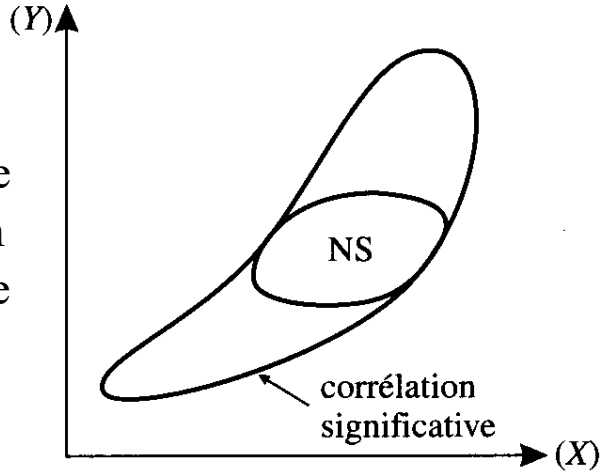
Remarque Plus le nombre υ est grand, plus est petite la valeur de $|r|$ permettant d'affirmer (au risque choisi) qu'une corrélation existe dans la population.

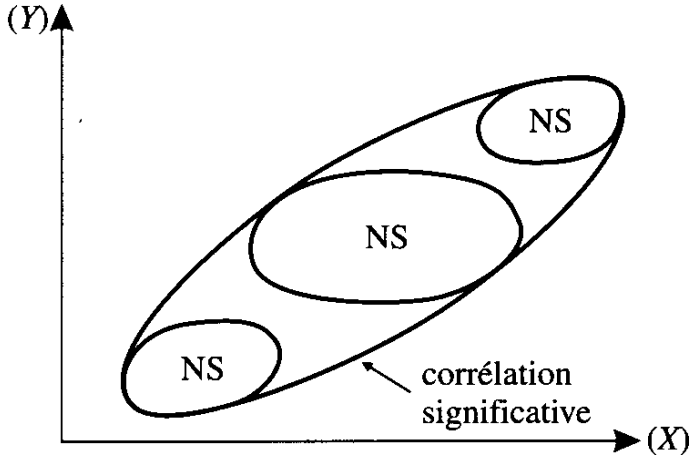
Effet des amplitudes de variation observées sur X et sur Y.



En se restreignant à la région centrale du nuage de points, nous risquons de trouver une corrélation non significative, alors qu'elle l'est fortement si les variables aléatoires sont explorées sur de plus grands intervalles

Cette situation se présente souvent lorsque la liaison statistique est non linéaire

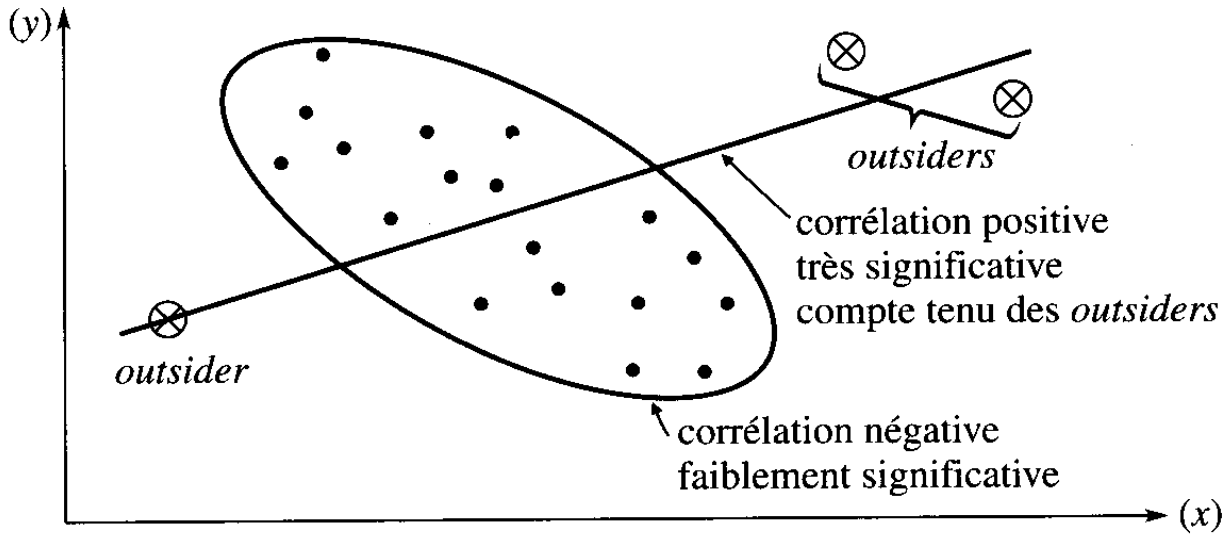




Un problème d'échelle d'observation:
l'amplitude de l'intervalle de variation des deux variables aléatoires peut varier suivant qu'elles sont échantillonnées sur un grand ou un petit espace statistique; la population parente n'est pas la même dans les deux cas.

<u>Conclusion:</u> Les conditions de l'observation ou de l'échantillonnage devront être bien explicitées, et rester bien présentes à l'esprit de l'analyste lors du commentaire et de l'interprétation de toute corrélation observée.

Élimination des « outsiders » ou « intrus »



Un petit nombre de points isolés ou « outsiders » , qui accompagnent un nuage de points, peut être l'origine de covariances anormalement élevées.

Il convient alors de ne pas les faire intervenir dans le calcul.

Ils relèvent soit de l'erreur de mesure soit d'un autre phénomène, ou d'une autre échelle d'observation, insuffisamment représentés dans l'échantillon.

Remarque Il n'existe pas de règle objective permettant de décider si un point est ou non un outsider. Ce sera, dans chaque cas, une question d'appréciation, aidée parfois par la comparaison des corrélations obtenue avec et sans ces couples particuliers.

Corrélation et relation de causalité

Une corrélation significative ne prouve pas forcément une relation de causalité entre les deux variables aléatoires observées.

Les deux variables aléatoires peuvent très bien ne pas être cause l'une de l'autre, mais simplement avoir une cause commune (éventuellement insoupçonnée).

Exemple

On pourrais montrer une forte corrélation entre les ventes d'huile pour bronzer et les ventes de crème glacée.

Il n'y a évidement aucune relation de causalité mais les variations de ces variables sont plutôt attribuables à une cause commune d'ordre climatique.

Régression de X en Y

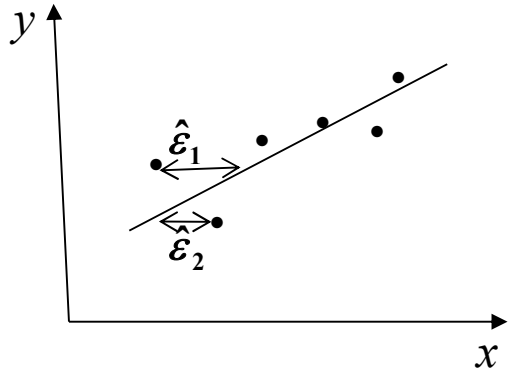
Ajustement par la méthode de moindres carrés ordinaires (MCO)

Données : n couples de points (x_i, y_i), i =1, 2, 3…n

Droite des moindres carrés ordinaires $x = a' y + b'$

Valeur ajustée de x_i : $\hat{x}_i = a' y_i + b'$

Le résidu en *i* (erreur) : $\hat{\varepsilon}_i = x_i - \hat{x}_i$



Critère des MCO :
choix des paramètres (a',b') minimisant la somme des carrés des erreurs :

$$s(a,b) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (x_i - a' y_i - b')^2$$

Solution :

$$a' = \frac{\sigma(x)}{\sigma(y)} \cdot r \quad , b' = \bar{x} - a' \bar{y}$$

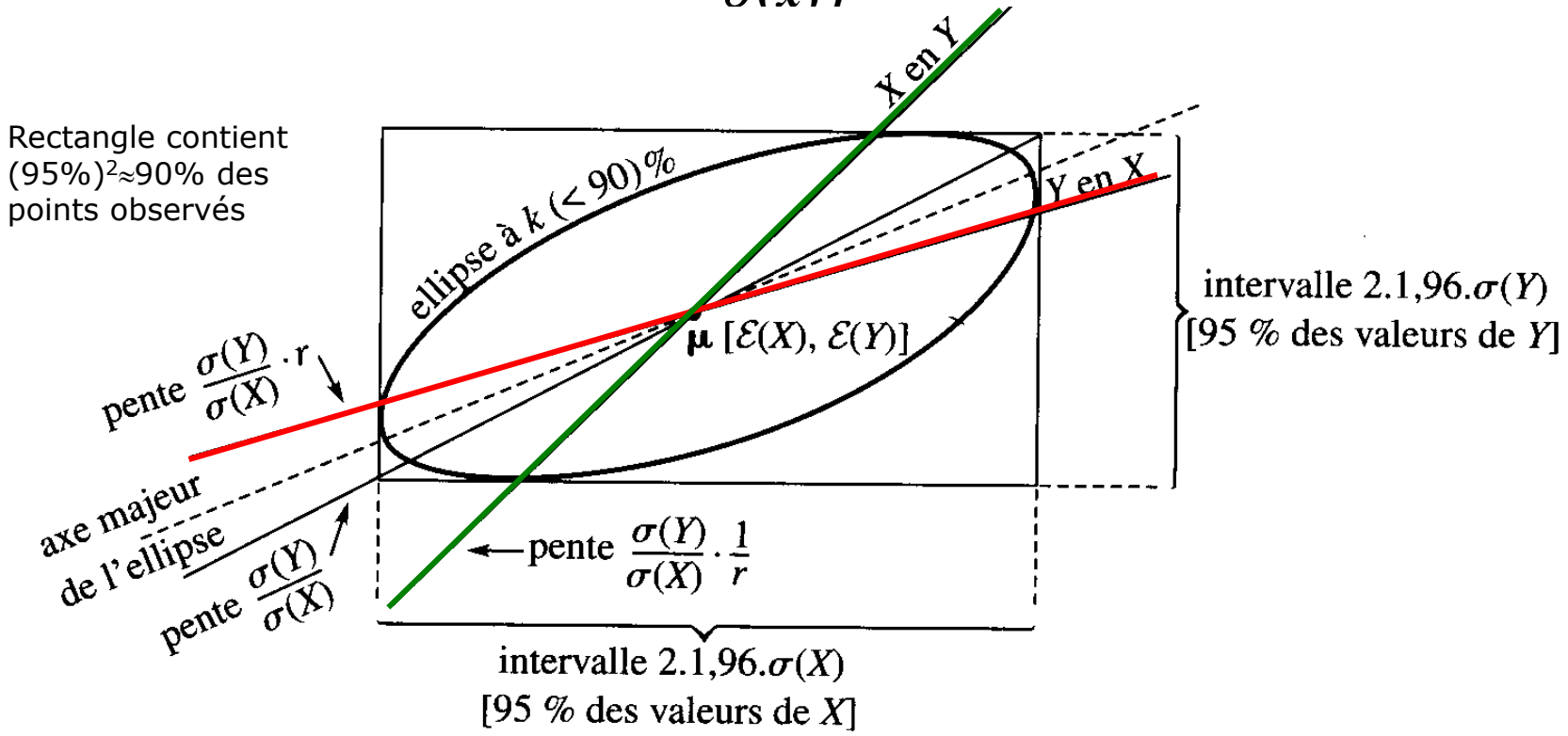
$$\hat{x} = a' y + b' \Rightarrow y = \frac{1}{a'} \hat{x} - \frac{b'}{a'} \Rightarrow y = \frac{\sigma(y)}{\sigma(x)} \frac{1}{r} \hat{x} + (\bar{y} - \frac{\sigma(y)}{\sigma(x)} \frac{1}{r} \bar{x})$$

D'où La droite de régression de X en Y: $y - \bar{y} = \frac{\sigma(y)}{\sigma(x)} \frac{1}{r} (\hat{x} - \bar{x})$

Les deux droites de régression

Droite de régression de Y en X: $\hat{y} - \bar{y} = \frac{\sigma(y)}{\sigma(x)} r(x - \bar{x})$

Droite de régression de X en Y: $y - \bar{y} = \frac{\sigma(y)}{\sigma(x)} \frac{1}{r} (\hat{x} - \bar{x})$



- Les deux droites de régression passent par le point moyen $\mu(\bar{x}, \bar{y})$ de l'échantillon.
- Puisque $|r| < 1$, la première droite a une pente inférieure en valeur absolue au rapport des écarts types $\frac{\sigma(y)}{\sigma(x)}$ la seconde, supérieure.

9. SÉRIES CHRONOLOGIQUES

- **Série chronologique : une suite de valeurs ordonnées dans le temps.**

(y_t) **où** $t = 1, 2, 3, \ldots n$

Hypothèse: Les observations sont effectuées à intervalle de temps constant

- **But de l'étude: décrire, expliquer et prévoir un phénomène évoluant au cours du temps**

Eléments constitutifs d'une série chronologique :

♦ **la composante tendancielle T_t (ou trend)** qui est souvent linéaire de la forme

$$T_t = at + b$$

Si les fluctuations de la série sont trop importantes, on utilise la méthode de moyennes mobiles pour obtenir une série plus lisse, plus régulière.

♦ **la composante saisonnière S_t de période p** représente les fluctuations périodiques (mensuelles (p = 12), trimestrielles (p = 4)…);

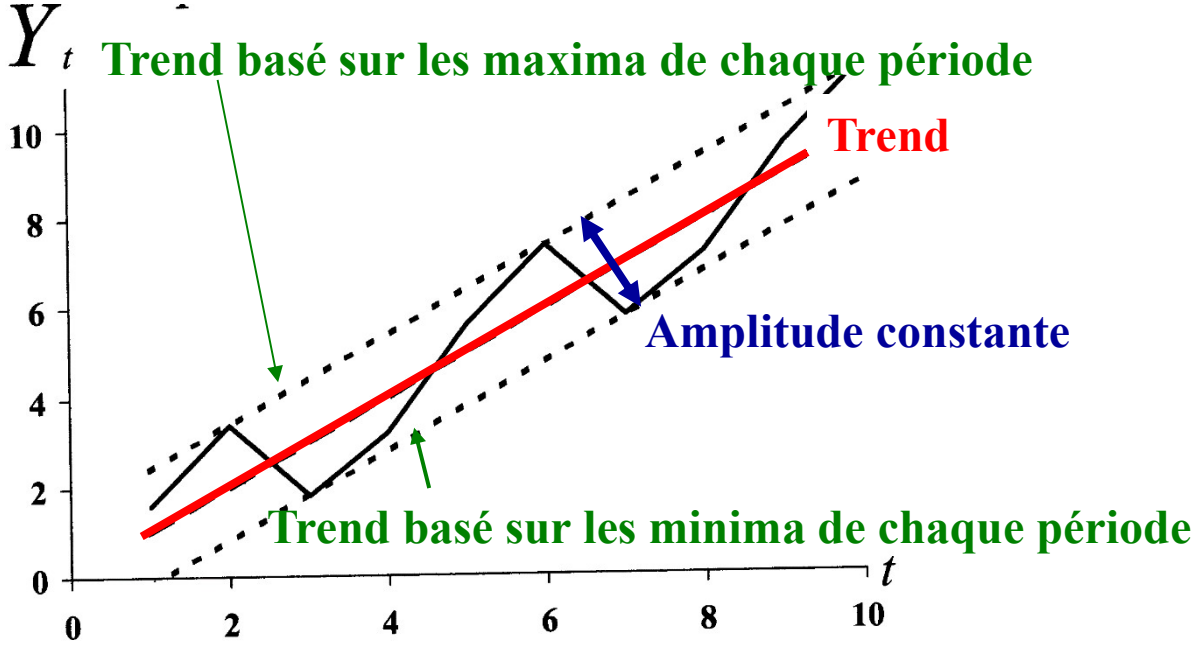
♦ **la composante accidentelle (ou aléa)** ε_t correspond aux fluctuations aléatoires des données et est imprévisible; dans la suite on la suppose de faible amplitude.

Remarque On observe parfois des variations cycliques d'une série chronologique lorsque elle comporte des mouvements oscillatoires liés à l'alternance des différentes phase d'un cycle (économique, solaire…). Dans ce cas le variations sont périodiques avec p >12 mois.

Les modèles de décomposition d'une série chronologique

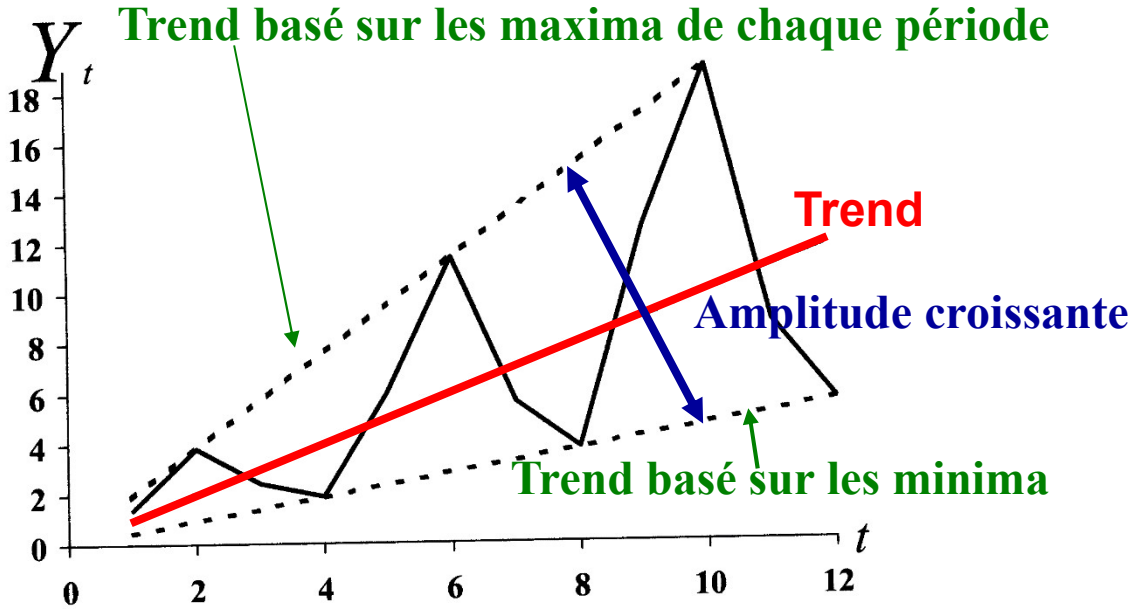
Modèle additif :

$$y_t = T_t + S_t + \varepsilon_t$$



Modèle multiplicatif :

$$y_t = T_t \cdot S_t \cdot \varepsilon_t$$



Choix de modèle

- En fonction de la nature des données
- Graphiquement

Modèle additif :

- **Estimation du trend T_t :**

méthode des MCO appliquée sur la série initiale

$T_t = at + b$ si le trend est linéaire

- **Estimation de composantes saisonnières $\hat{s}_1, \hat{s}_2, \hat{s}_3, ..., \hat{s}_p$:**

- on calcule la série se différences : $S_t = y_t - T_t$

- on estime $\hat{s}_1$ en calculant la moyenne arithmétique des valeurs :

$$S_1, S_{1+p}, S_{1+2p,....}$$

- on estime $\hat{s}_2$ de même façon à partir des valeurs $S_2, S_{2+p}, S_{2+2p,....}$

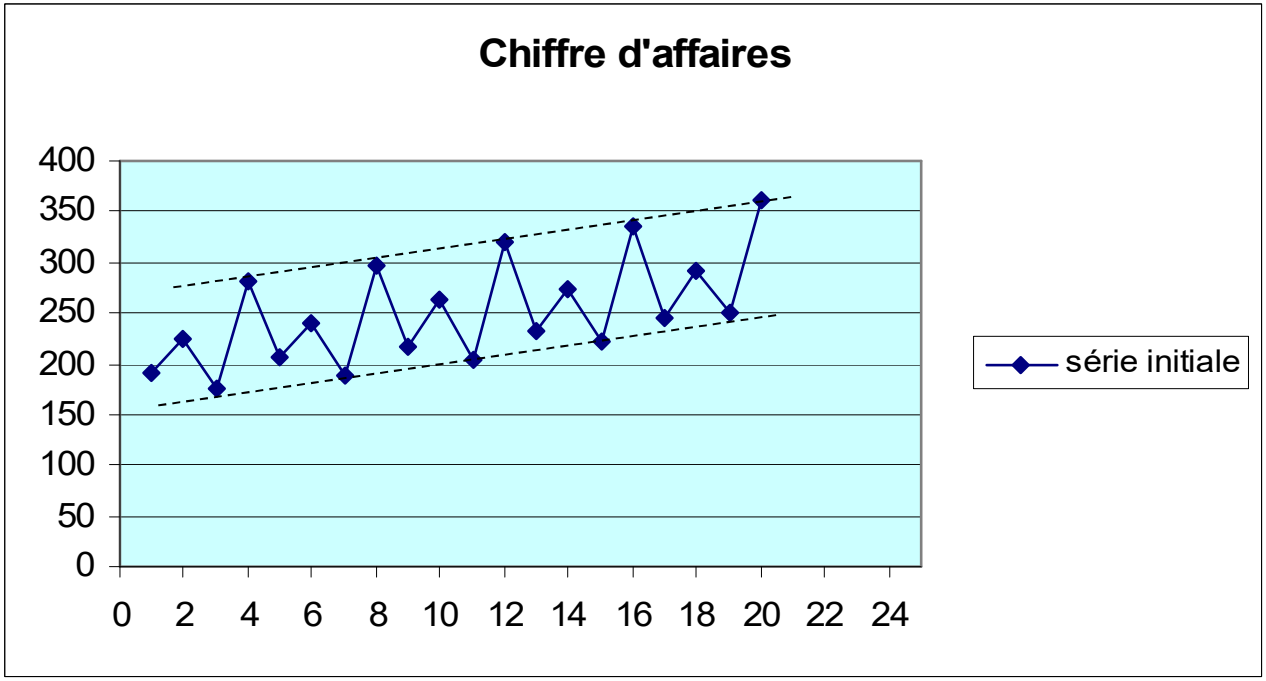
- **Prévision :** $\hat{y}_t = at + b + \hat{s}_t$

Exemple ***Chiffre d'affaires trimestriel d'une entreprise*** :

	année 1	année 2	année 3	année 4	année 5
trimestre 1	190	205,4	215,8	232,2	245,6
trimestre 2	223,7	240,5	262,3	273,1	291,9
trimestre 3	174,4	188,2	203	220,8	249,6
trimestre 4	280,5	295,5	319,5	336,5	360,5

- **Objectif:** prévision du chiffre d'affaires de l'entreprise pour l'année 6

Les droites reliant les extrema sont parallèles



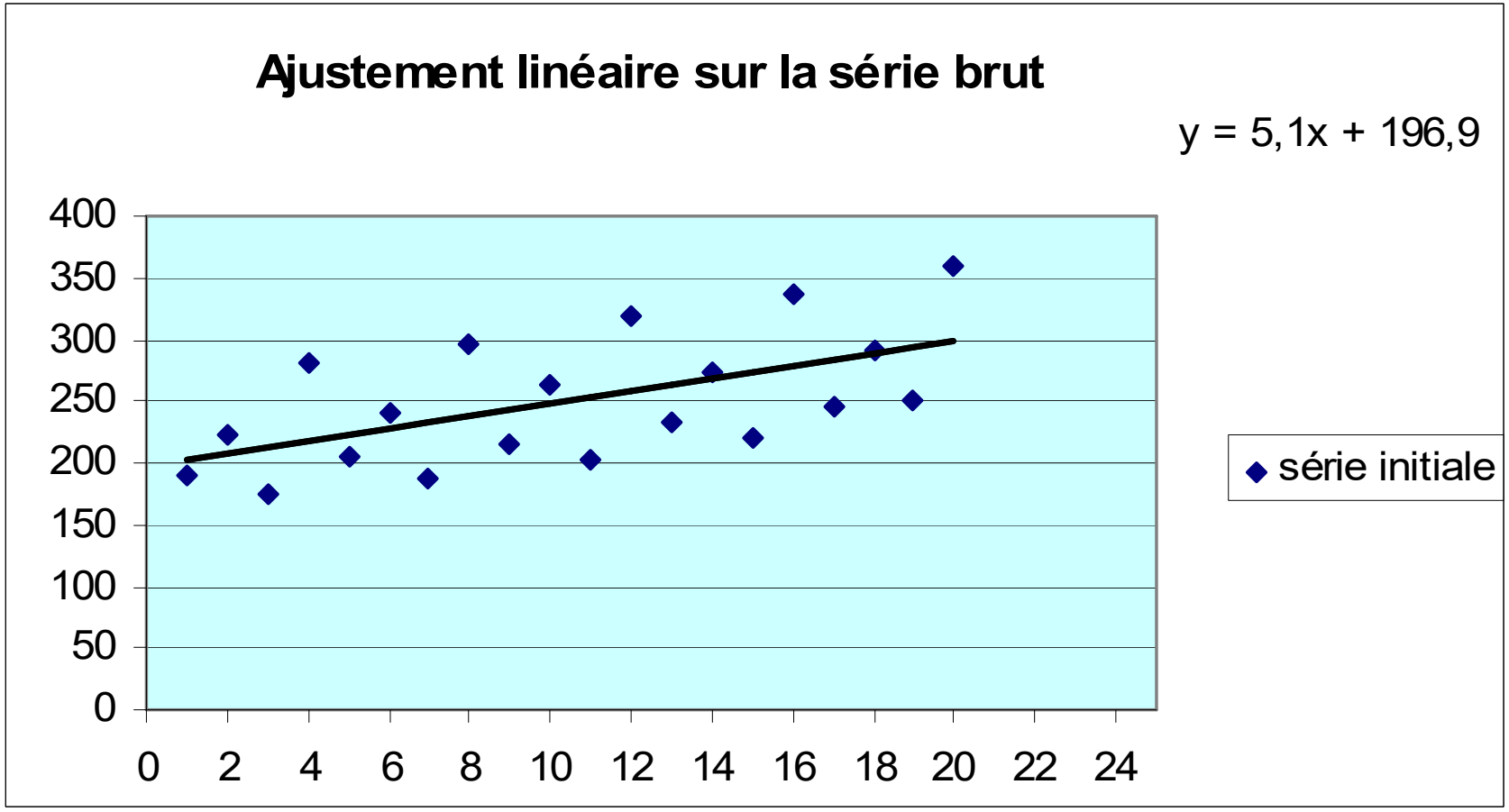
$\Downarrow$

- **Modèle additif :**

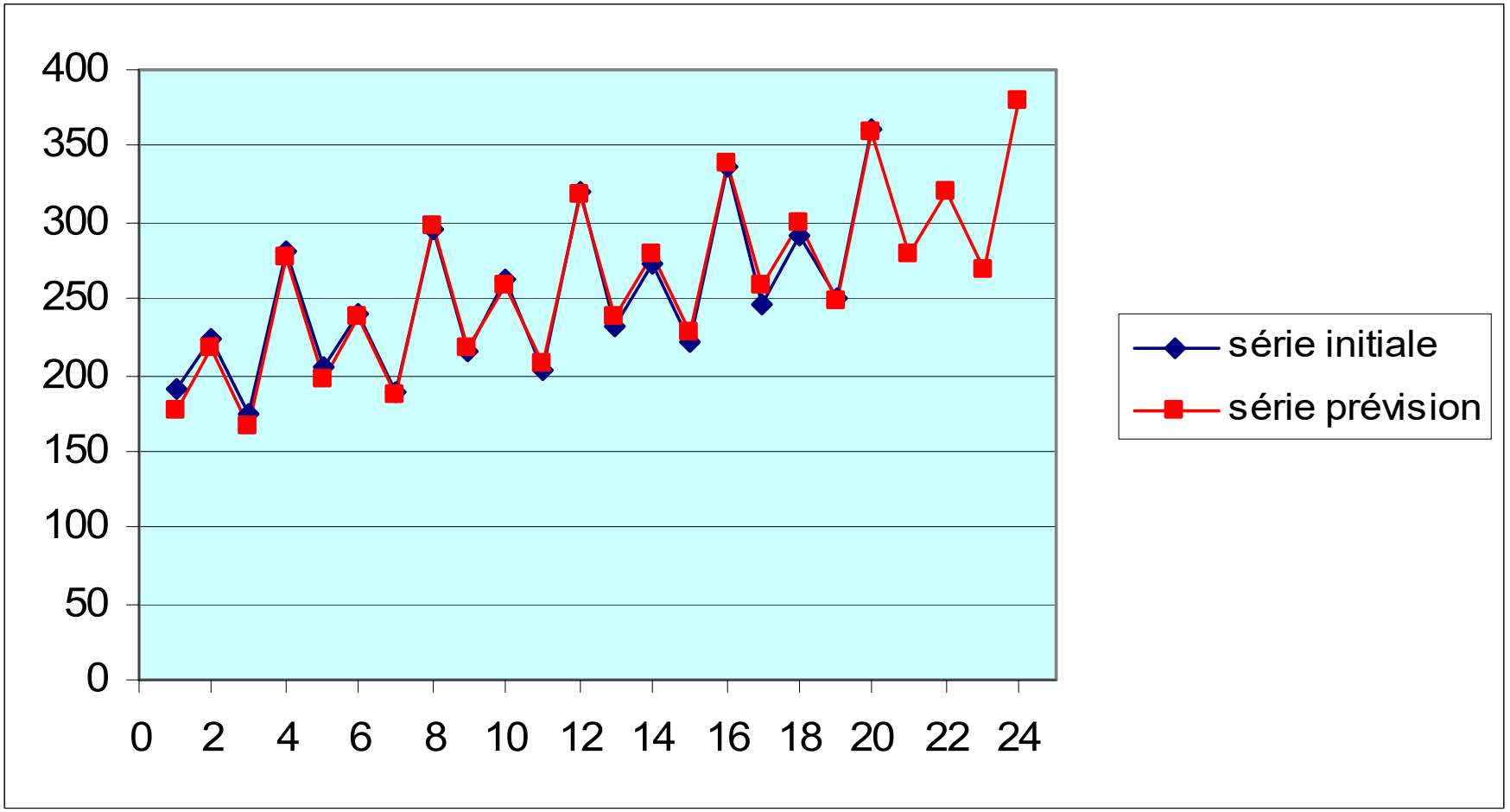
$$y_t = T_t + S_t + \varepsilon_t$$

$$S_t \approx y_t - T_t$$

- **Estimation du trend T_t : méthode des MCO appliquée sur la série initiale**



	t	*Série initiale* y_t	*Trend* $T_t = 5{,}1t+196{,}9$	*Différence* $y_t - y$	*Composante saisonnière* $\hat{s}_t$	*Prévision* $5{,}1t + 196{,}9 + \hat{s}_t$
Année 1	1	190	202	-12	-25	177
	2	223,7	207,1	16,6	10,4	217,5
	3	174,4	212,2	-37,8	-45,8	166,4
	4	280,5	217,3	63,2	60,4	277,7
Année 2	5	205,4	222,4	-17	-25	197,4
	6	240,5	227,5	13	10,4	237,9
	7	188,2	232,6	-44,4	-45,8	186,8
	8	295,5	237,7	57,8	60,4	298,1
Année 3	9	215,8	242,8	-27	-25	217,8
	10	262,3	247,9	14,4	10,4	258,3
	11	203	253	-50	-45,8	207,2
	12	319,5	258,1	61,4	60,4	318,5
Année 4	13	232,2	263,2	-31	-25	238,2
	14	273,1	268,3	4,8	10,4	278,7
	15	220,8	273,4	52,6	-45,8	227,6
	16	336,5	278,5	58	60,4	338,9
Année 5	17	245,6	283,6	-38	-25	258,6
	18	291,9	288,7	3,2	10,4	299,1
	19	249,6	293,8	-44,2	-45,8	248
	20	360,5	298,9	61,6	60,4	359,3
Année 6	21		304		-25	279
	22		309,1		10,4	319,5
	23		314,2		-45,8	268,4
	24		319,3		60,4	379,7


400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16
18
20
22
24
série initiale
série prévision

Modèle multiplicatif :

- **Estimation du trend T_t :**

méthode des MCO appliquée sur la série initiale

$T_t = at + b$ si le trend est linéaire

- **Estimation de coefficients saisonniers $\hat{s}_1, \hat{s}_2, \hat{s}_3, ..., \hat{s}_p$:**

- on calcule la série les rapports saisonniers : $S_t = y_t / T_t$

- on estime $\hat{s}_1$ en calculant la moyenne arithmétique des valeurs :

$$S_1, S_{1+p}, S_{1+2p,....}$$

- on estime $\hat{s}_2$ de même façon à partir des valeurs $S_2, S_{2+p}, S_{2+2p,....}$

- **Prévision :** $\hat{y}_t = (at + b)\hat{s}_t$

Exemple ***Nouvelles immatriculations de voitures particulières de 1996 à 2000 au Luxemburg (source Statec):***

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Aoûte	Sept.	Oct.	Nov.	Déc.
1996	2006	3224	3789	4153	3100	2527	3015	1504	1847	2314	1673	1602
1997	2247	3862	3586	4047	2838	2727	2730	1648	2007	2450	1966	1695
1998	2433	3723	4325	4493	3399	3083	3247	1928	2377	2831	2388	2126
1999	3127	4437	5478	4384	3552	3678	3611	2260	2699	3071	2510	2182
2000	3016	4671	5218	4746	4814	3545	3341	2439	2637	3085	2737	2055

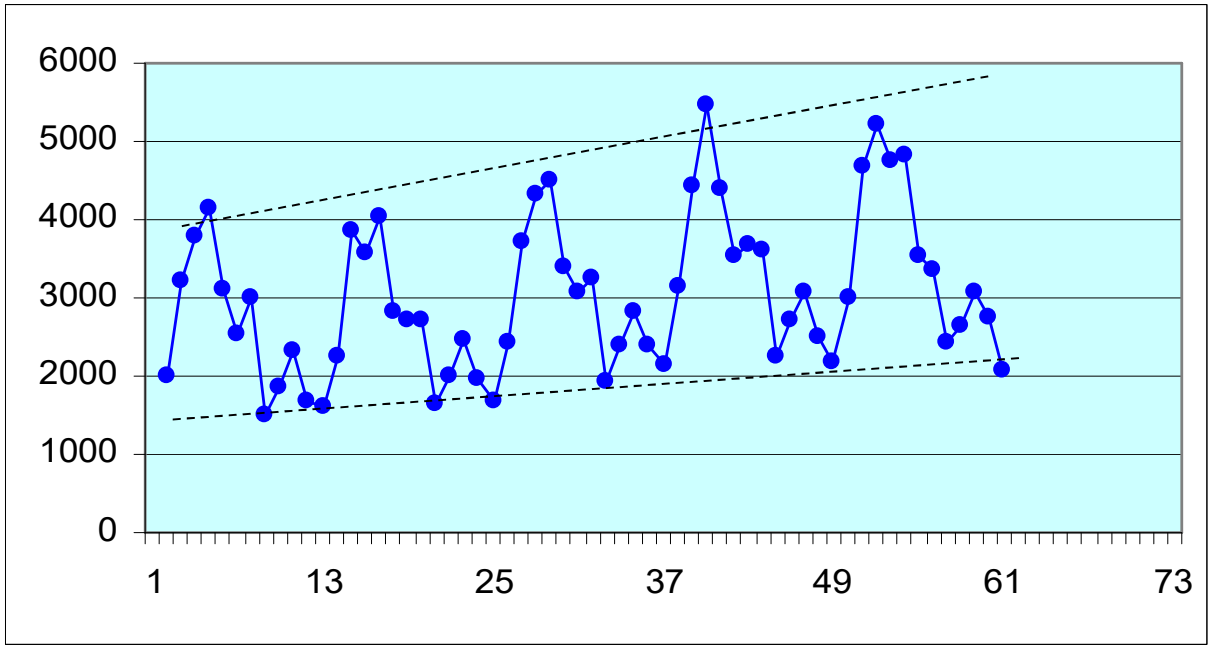
Les droites des trends basés sur des extrema de chaque période ne sont pas parallèles

$\Downarrow$

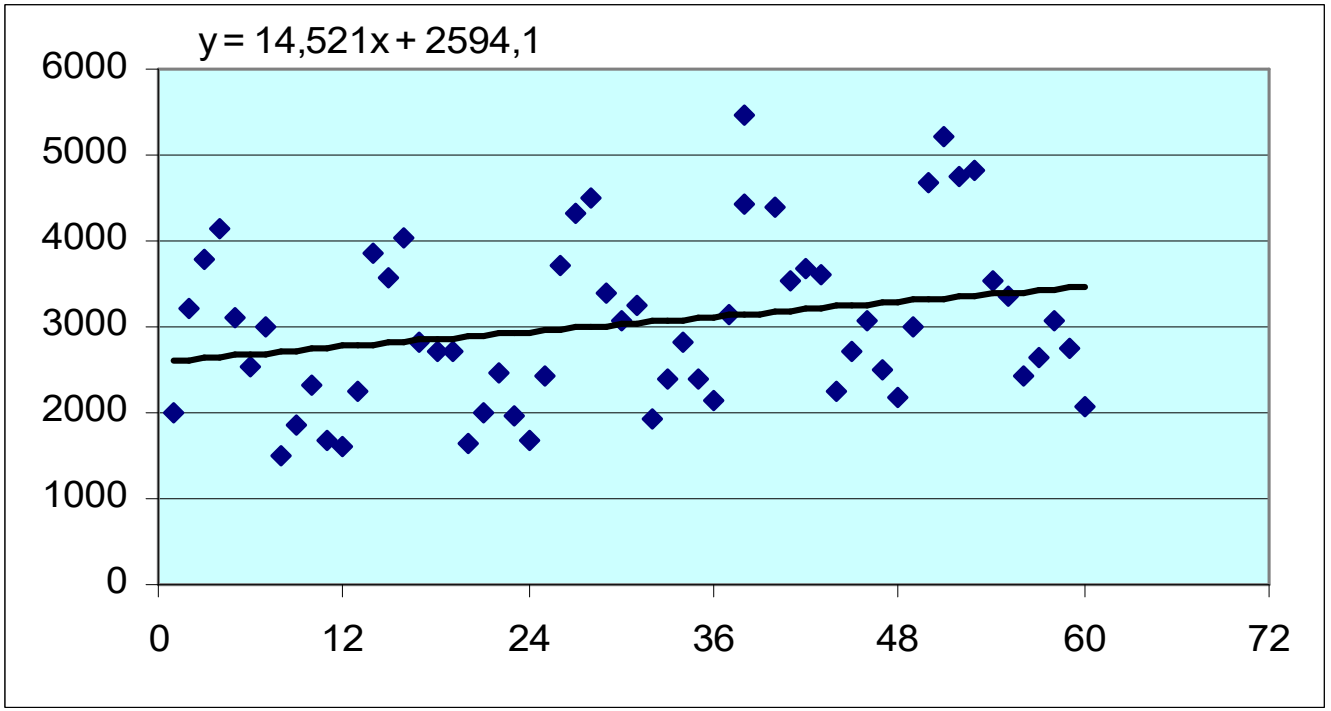
Modèle multiplicatif :

$$y_t = T_t \cdot S_t \cdot \varepsilon_t$$

$$S_t \approx \frac{y_t}{T_t}$$



La droite de régression:
y=14,521x+2594,1



- Les valeurs obtenues à l'aide de **la droite de régression**:

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Aoûte	Sept.	Oct.	Nov.	Déc.
1996	2608,62	2623,14	2637,66	2652,18	2666,71	2681,23	2695,75	2710,27	2724,79	2739,31	2753,83	2768,35
1997	2782,87	2797,39	2811,92	2826,44	2840,96	2855,48	2870,00	2884,52	2899,04	2913,56	2928,08	2942,60
1998	2957,13	2971,65	2986,17	3000,69	3015,21	3029,73	3044,25	3058,77	3073,29	3087,81	3102,34	3116,86
1999	3131,38	3145,90	3145,90	3174,94	3189,46	3203,98	3218,50	3233,02	3247,55	3262,07	3276,59	3291,11
2000	3305,63	3320,15	3334,67	3349,19	3363,71	3378,23	3392,76	3407,28	3421,80	3436,32	3450,84	3465,36

- Pour obtenir **les rapports saisonniers**, il suffit de diviser les valeurs originales par celles obtenues en se basant sur la droite de régression:

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Aoûte	Sept.	Oct.	Nov.	Déc.
1996	0,77	1,23	1,44	1,57	1,16	0,94	1,12	0,55	0,68	0,84	0,61	0,58
1997	0,81	1,38	1,28	1,43	1,00	0,96	0,95	0,57	0,69	0,84	0,67	0,58
1998	0,82	1,25	1,45	1,50	1,13	1,02	1,07	0,63	0,77	0,92	0,77	0,68
1999	1,00	1,41	1,74	1,38	1,11	1,15	1,12	0,70	0,83	0,94	0,77	0,66
2000	0,91	1,41	1,56	1,42	1,43	1,05	0,98	0,72	0,77	0,90	0,79	0,59

- Ensuite il faut calculer la moyenne pour chaque période obtenant ainsi **les coefficients saisonniers:**

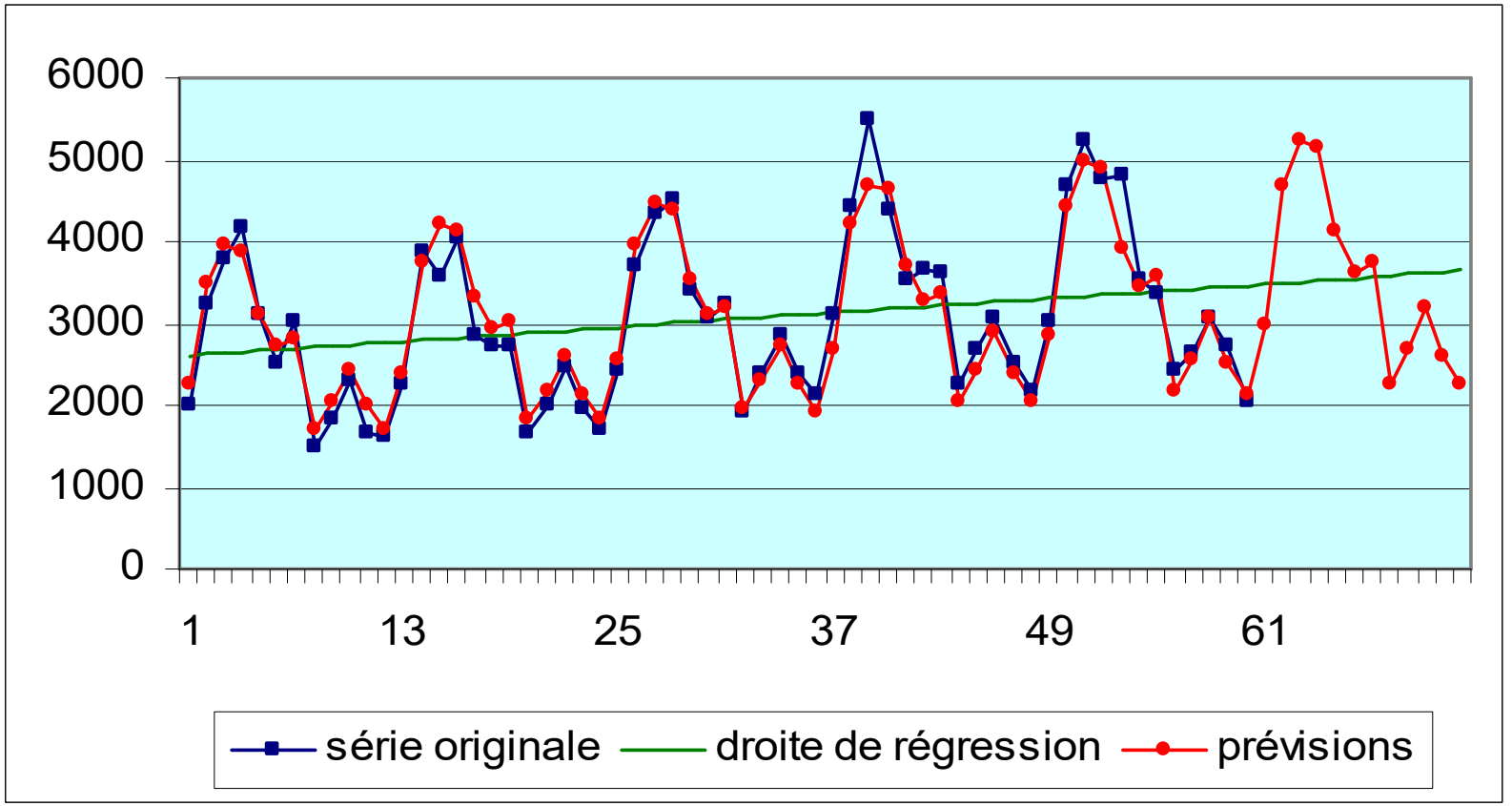
Janv. $\hat{s}_1$	Fév. $\hat{s}_2$	Mars $\hat{s}_3$	Avril $\hat{s}_4$	Mai $\hat{s}_5$	Juin $\hat{s}_6$	Juillet $\hat{s}_7$	Aoûte $\hat{s}_8$	Sept. $\hat{s}_9$	Oct. $\hat{s}_{10}$	Nov. $\hat{s}_{11}$	Déc. $\hat{s}_{12}$
0,86	1,34	1,49	1,46	1,17	1,02	1,05	0,63	0,75	0,89	0,72	0,62

Dans notre exemple, nous n'avons pas besoin d'ajuster ces coefficients puisque leurs somme vaut p =12 (nombre de saisons par période). $\sum_{i=1}^{12} \hat{s}_i = \mathbf{12}$

En général, si $\sum_{i=1}^{p} \hat{s}_i = p'$ ($p' \neq p$ mais $p' \approx p$) alors il faut corriger les coefficients saisonniers en les divisant par ***p'/p.***

Prévisions

$$\hat{\hat{y}}_t = T_t \cdot \hat{s}_t$$



	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Aoûte	Sept.	Oct.	Nov.	Déc.
1996	2248,7	3504,4	3938,7	3868,4	3111,3	2741,5	2826,7	1719,1	2041,0	2433,4	1987,1	1712,4
1997	2398,9	3737,2	4198,9	4122,6	3314,6	2919,7	3009,4	1829,6	2171,6	2588,2	2112,8	1820,2
1998	2549,1	3970,0	4459,1	4376,8	3517,9	3097,8	3192,2	1940,1	2302,1	2743,0	2238,6	1928,0
1999	2699,4	4202,8	4697,6	4630,9	3721,2	3276,0	3374,9	2050,7	2432,6	2897,8	2364,3	2035,8
2000	2849,6	4435,6	4979,5	4885,1	3924,5	3454,2	3557,6	2161,2	2563,1	3052,6	2490,0	2143,5
2001	**2999,8**	**4668,3**	**5239,7**	**5139,2**	**4127,8**	**3632,3**	**3740,3**	**2271,7**	**2693,7**	**3207,4**	**2615,8**	**2251,3**

Moyennes mobiles

■ Si les fluctuations de la série sont importantes, on peut d'abord utiliser la méthode de moyennes mobiles pour obtenir une série plus régulière.

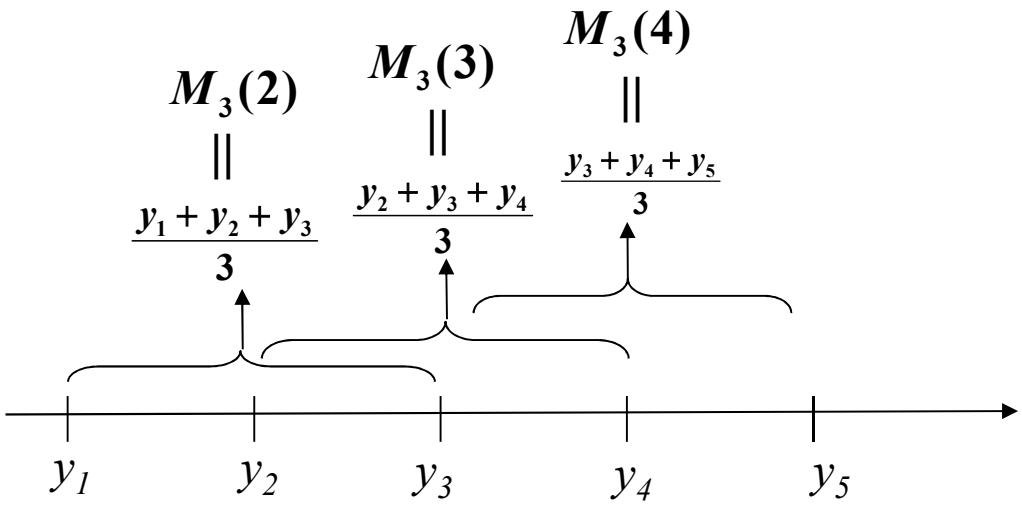
On peut estimer la tendance par lissage selon les moyennes mobiles

■ On appelle moyennes mobiles centrées d'ordre k de la série des $(y_t)_{t=1..n}$ les moyennes arithmétiques calculées sur k valeurs successives et rapportées à la date du « milieu » de la période:

- **Si k impair:** $k=2m+1$ $\quad M_k(t) = \frac{1}{k}\sum_{i=-m}^{i=m} y_{t+i}$

Exemple k = 3

$$M_3(2) = \frac{y_1 + y_2 + y_3}{3} \qquad M_3(3) = \frac{y_2 + y_3 + y_4}{3} \qquad M_3(4) = \frac{y_3 + y_4 + y_5}{3}$$



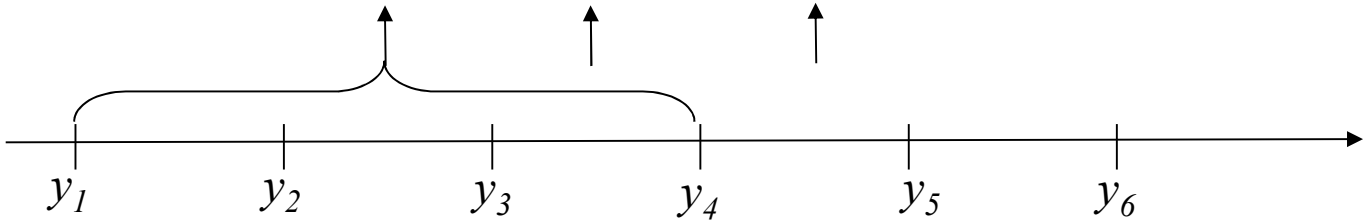
Remarque On ne peut pas déterminer la moyenne mobile pour $2 \cdot m$ valeurs extrêmes du temps t, c'est à dire m de chaque côté (dans notre exemple pour t =1 et t =5)

- **Si *k* pair:** $k=2m$

$$M_k(t) = \frac{1}{k}\left[\frac{y_{t-m}}{2} + \sum_{i=-m+1}^{i=m-1} y_{t+i} + \frac{y_{t+m}}{2}\right]$$

Exemple k = 4

$$M_4(2.5) = \frac{y_1 + y_2 + y_3 + y_4}{4} \qquad M_4(3.5) = \frac{y_2 + y_3 + y_4 + y_5}{4} \qquad M_3(4.5) = \frac{y_3 + y_4 + y_5 + y_6}{4}$$



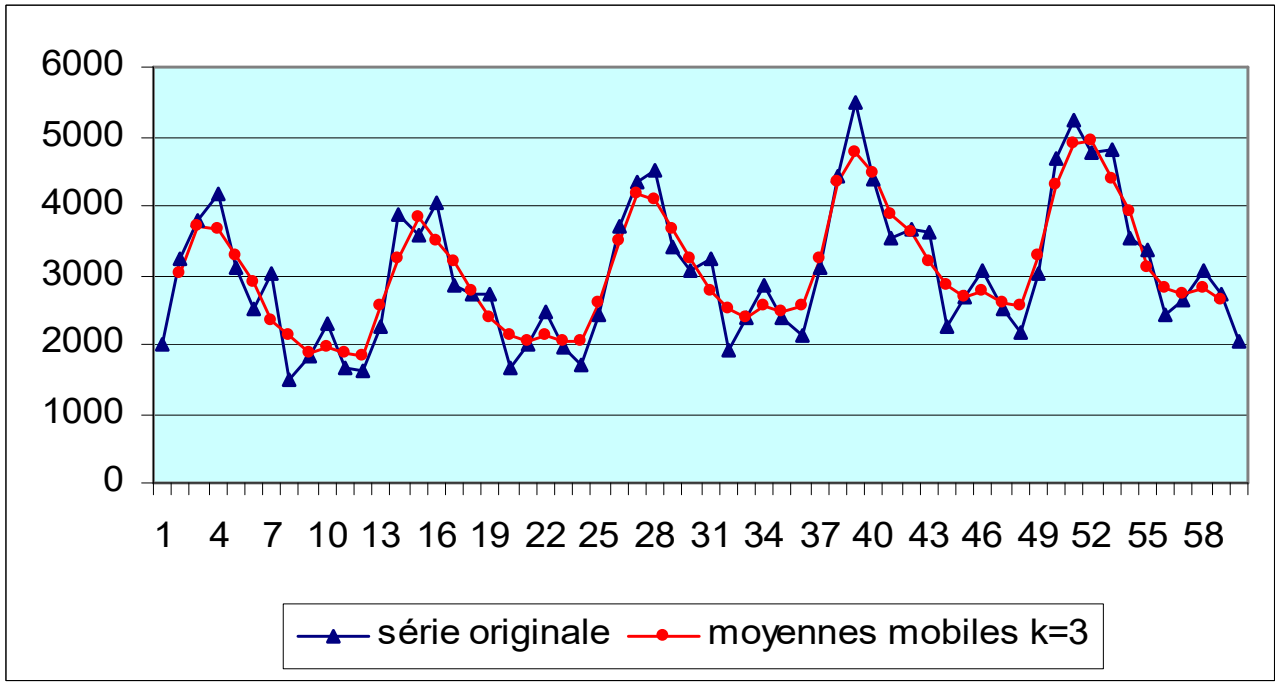
Les moyennes obtenues ne correspondent pas à des abscisses existantes.

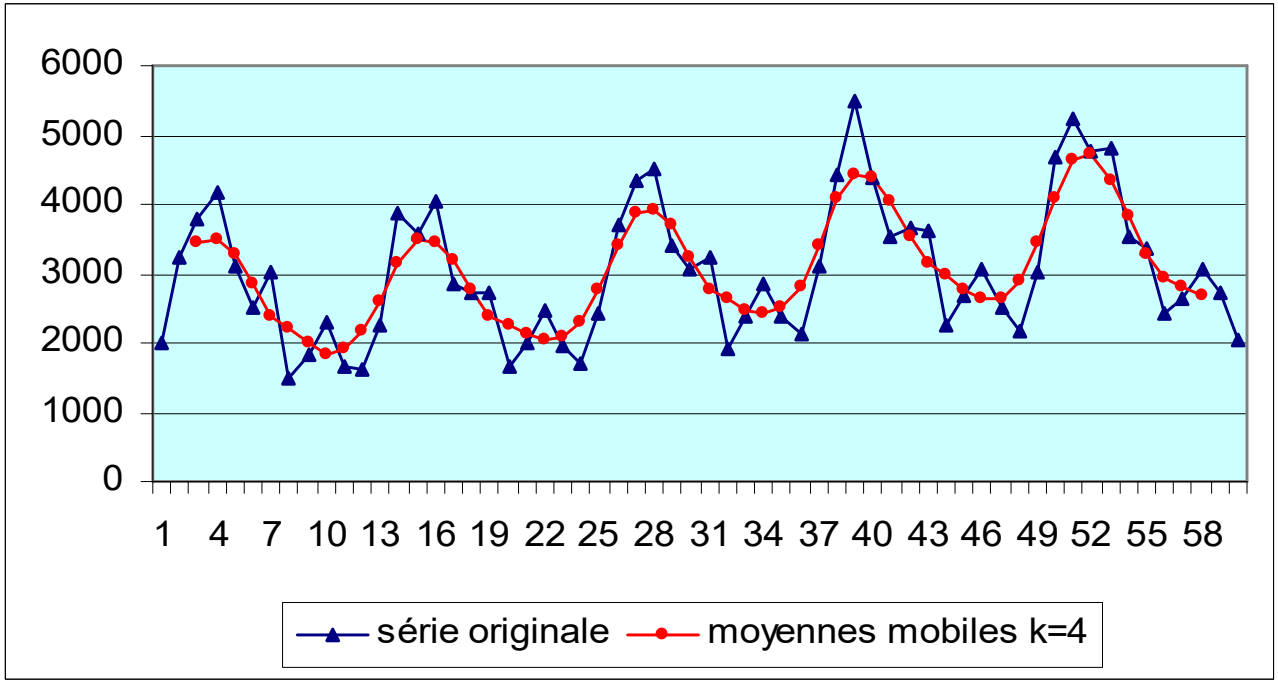
Il faut lisser une seconde fois en utilisant k = 2 (à partir des valeurs lissées une première fois

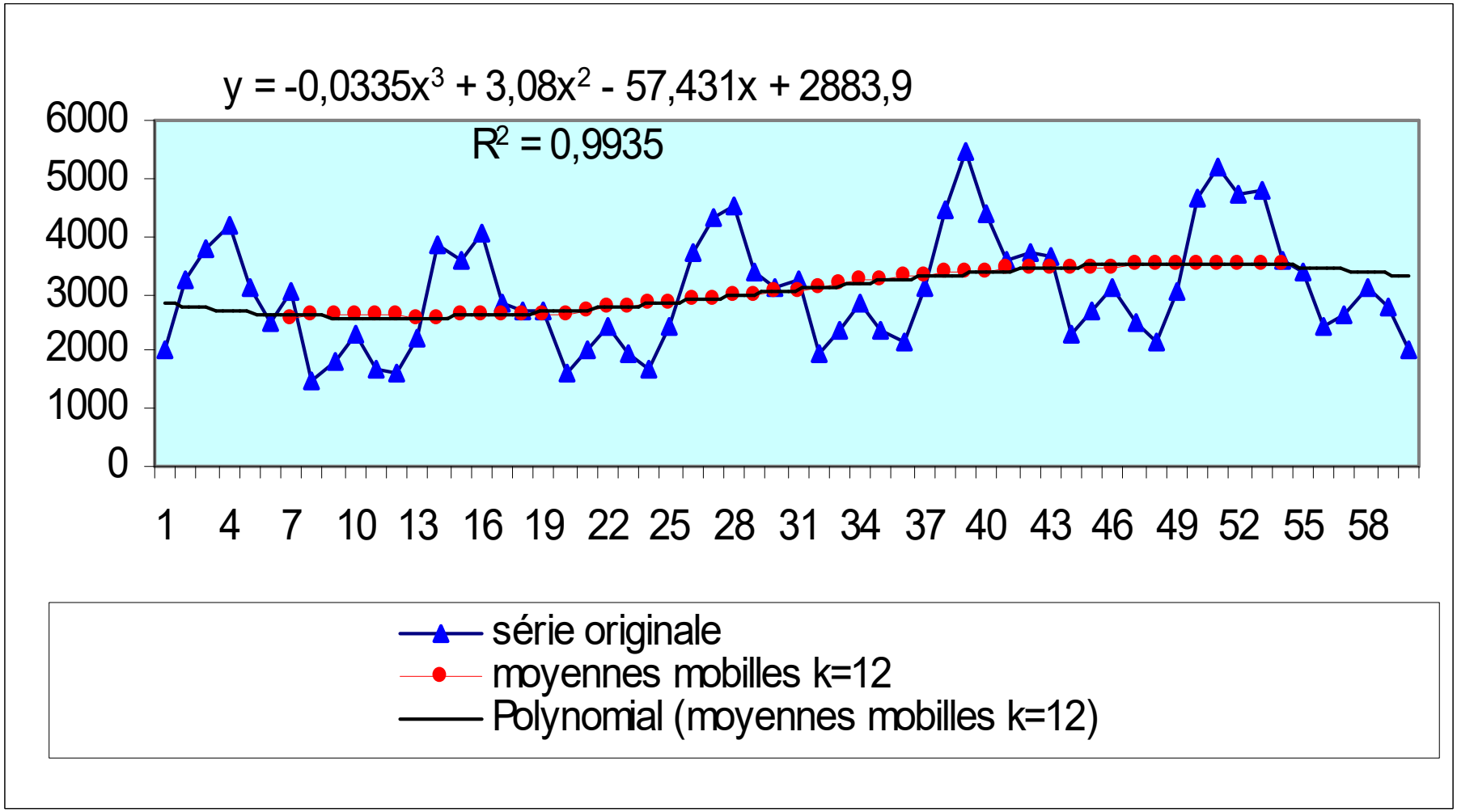
$$\frac{y_1/2 + y_2 + y_3 + y_4 + y_5/2}{4} = M_4(3) = \frac{1}{2}\left[M_4(2.5) + M_4(3.5)\right]$$

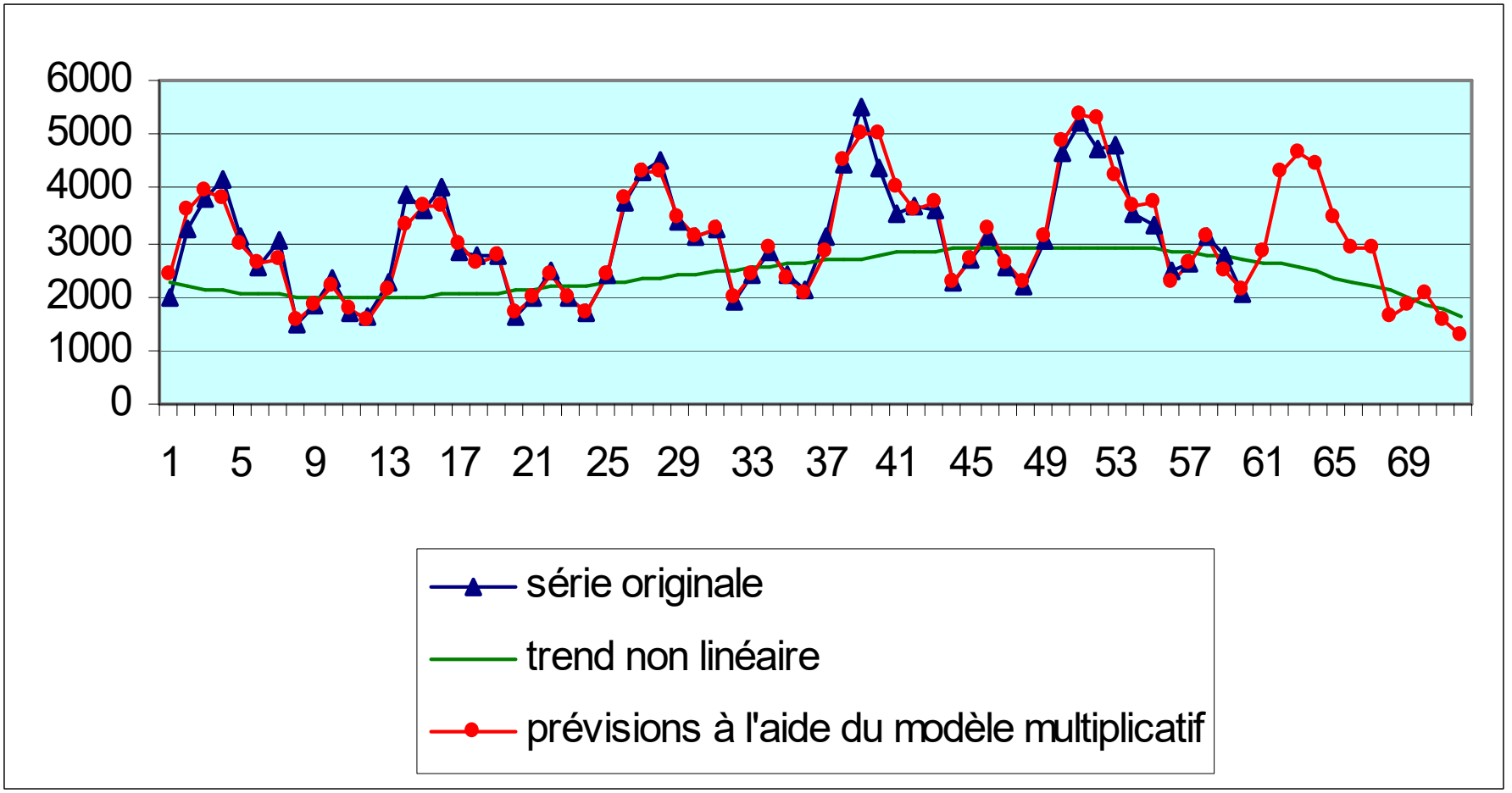
$$\frac{y_2/2 + y_3 + y_4 + y_5 + y_6/2}{4} = M_4(4) = \frac{1}{2}\left[M_4(3.5) + M_4(4.5)\right]$$

Remarque On ne peut pas déterminer la moyenne mobile pour $2 \cdot m$ valeurs extrêmes du temps t, c'est à dire *m* de chaque côté (dans notre exemple pour t =1, t = 2, t = 5 et t = 6)


6000
5000
4000
3000
2000
1000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58
série originale
moyennes mobiles k=3


6000
5000
4000
3000
2000
1000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58
série originale
moyennes mobiles k=4


y = -0,0335x3 + 3,08x2 - 57,431x + 2883,9
R2 = 0,9935
6000
5000
4000
3000
2000
1000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58
série originale
moyennes mobilles k=12
Polynomial (moyennes mobilles k=12)


6000
5000
4000
3000
2000
1000
0
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
série originale
trend non linéaire
prévisions à l'aide du modèle multiplicatif