**Université Paris Ouest**

**M1 – Modélisation Appliquée**

# Distribution d'une variable quantitative continue

Laurent Ferrara

Février 2013

# Objectif

- On s'intéresse à une variable aléatoire quantitative continue X de loi inconnue $P_\theta$ à densité $f_\theta$ .
- On recueille $(x_1, \ldots, x_n)$, une observation du n-échantillon issue de cette variable d'intérêt.
- On cherche à ajuster un modèle statistique à ces données.

→ **Inférence** : Comment utiliser cet échantillon pour estimer la loi de distribution empirique et en tirer des conclusions sur la loi de distribution théorique de X : $P_\theta$ . ?

# Plan de la présentation

**1. Descriptions graphiques d'une distribution**

**2. Descriptions numériques d'une distribution**

**3. Rappels des distributions continues usuelles**

**4. Outils de comparaison de distributions**

**5. Tests de comparaison de distributions**

# 1. Descriptions graphiques

**1) Histogramme**

**Objectif : estimer la densité de distribution empirique**

- On range les données par ordre croissant $x_{(1)}, \ldots, x_{(n)}$.
- On regroupe les données en J classes égales de largeur h
- Une classe est un intervalle semi-ouvert : $B_j = ]b_{j-1}, b_j]$
- Le milieu de chaque classe j est : $m_j = (b_{j-1} + b_j)/2$
- La largeur de chaque classe j est : $h = b_j - b_{j-1}$

donc $B_j = ](j-1)h,\ jh]$

- L'histogramme est la fonction $f_H$ définie par, pour tout x appartenant au support,

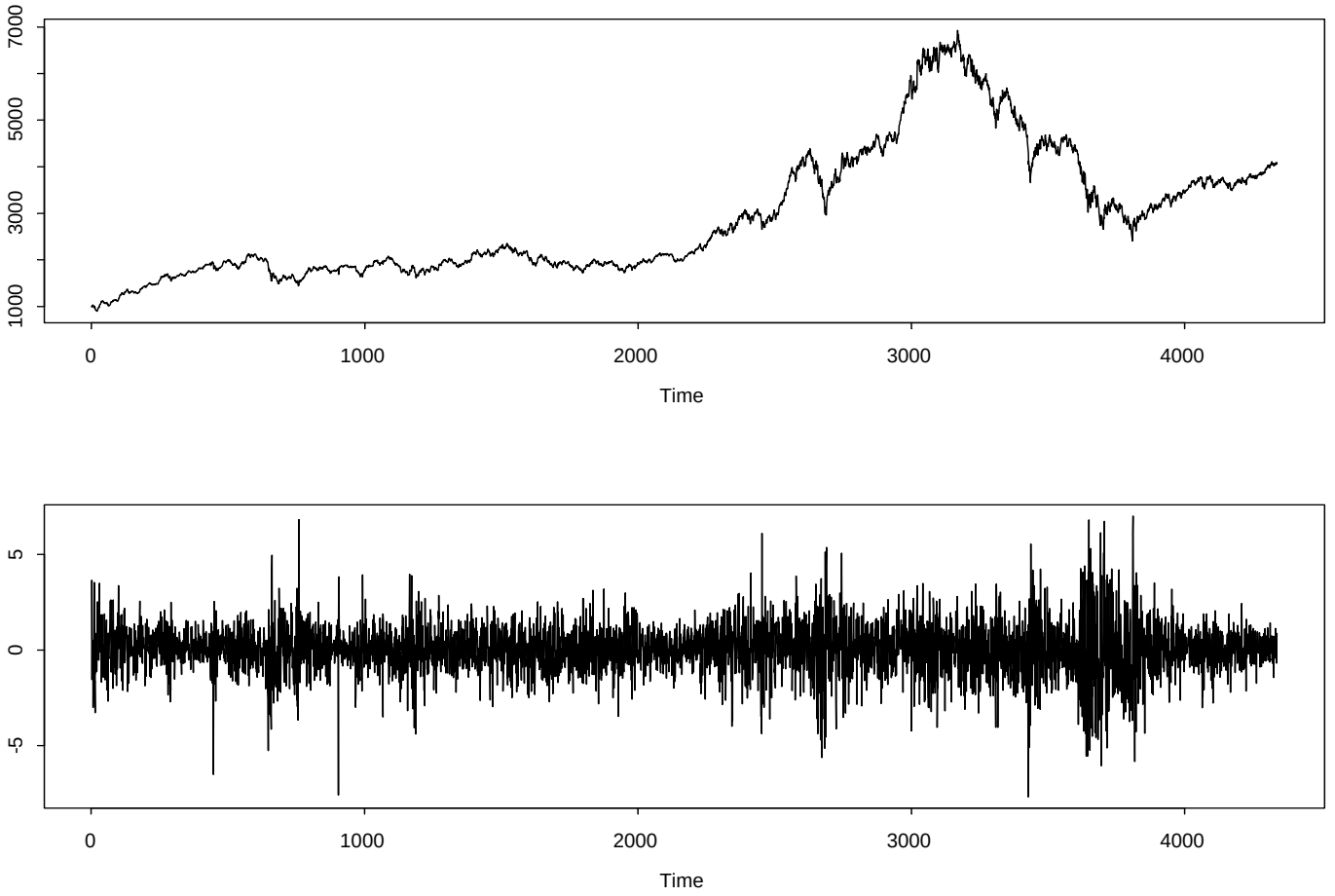
$$f_H(x) = \frac{1}{nh} \times \left\{ nbre\ de\ x_i\ dans\ la\ classe\ B_j\ contenant\ x \right\}$$

$$f_H(x) = \frac{1}{nh} \sum_{i=1}^{n} 1_{\{X_i = x_i \in B_j\}} = \frac{1}{nh} \sum_{i=1}^{n} 1_{\{x_i \in [m_j \pm h/2]\}}$$

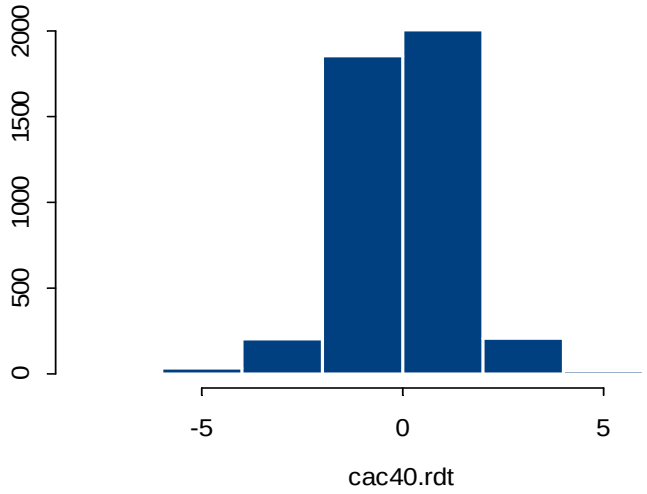
- <u>Pb pratique :</u>
  - Choix de b0 ?
  - Choix de h ? = Choix du nbre de classes J ?

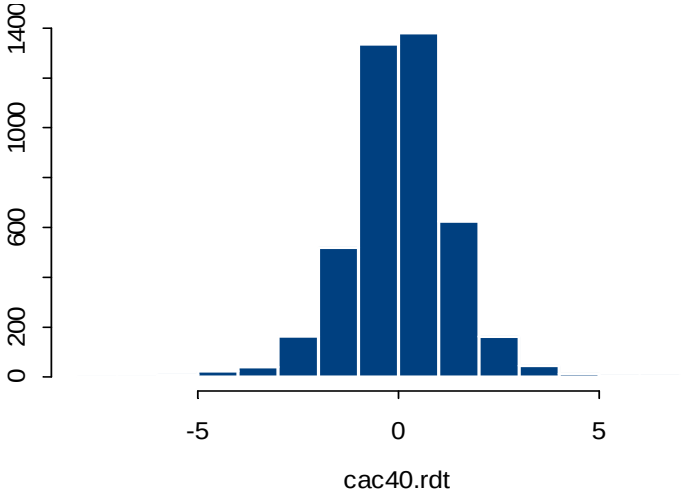
- Attention : $b_0 \leq x_{(1)} \quad et \quad x_{(n)} \leq b_J$

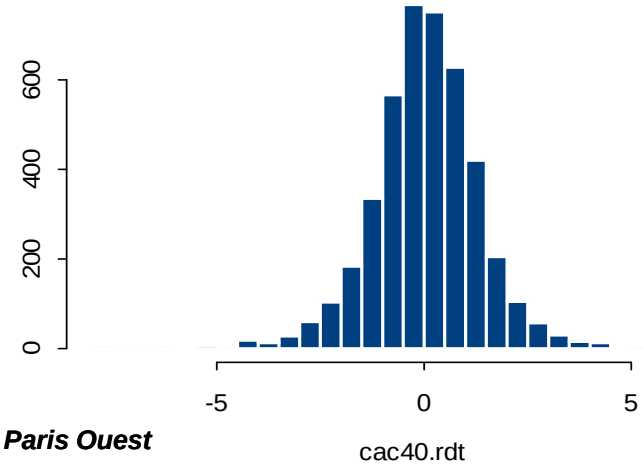
# Exemple : Rendements journaliers du CAC 40 de 1987 à 2004 (n=4337)

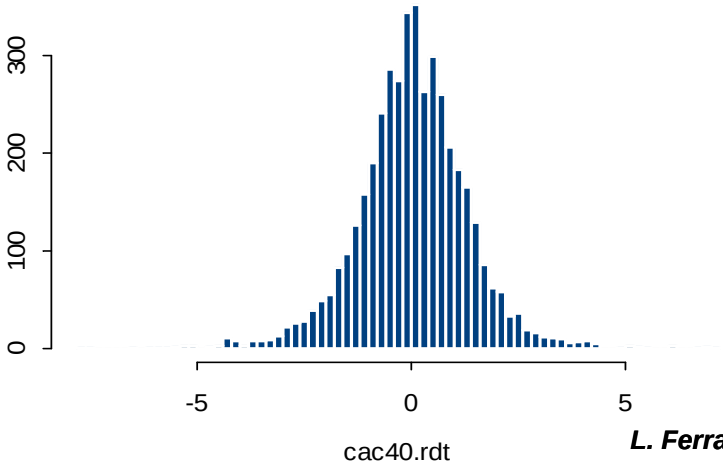


# Exemple : Rendements journaliers du CAC 40 de 1987 à 2004 (n=4337)



## 2) Estimation non paramétrique par la méthode des noyaux

→ *Même objectif que l'histogramme*

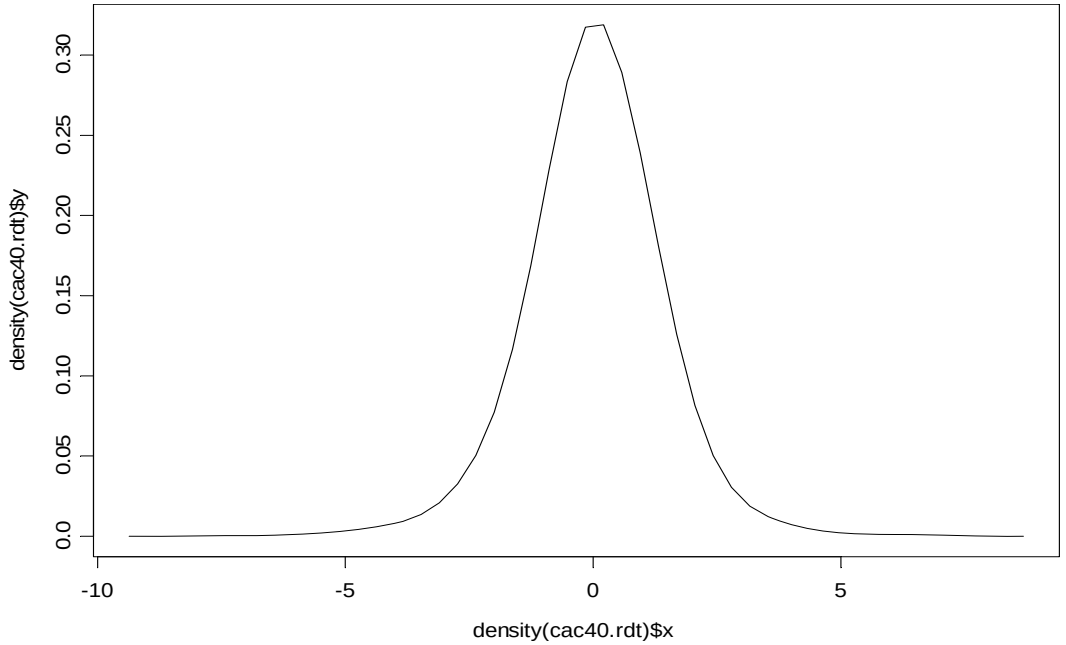
- 2 problèmes pratiques avec l 'histogramme
  - Choix de la fenêtre h ? (= choix du nombre de classes J)
  - Choix de l'origine $b_0$?

- 2 problèmes majeurs avec l'histogramme:
  - Perte d'information en identifiant tous les points de la classe au point central de cette classe.
  - la densité de distribution est supposée être lisse alors que l 'histogramme ne l'est pas. (alternative: interpoler linéairement les centres des classes)

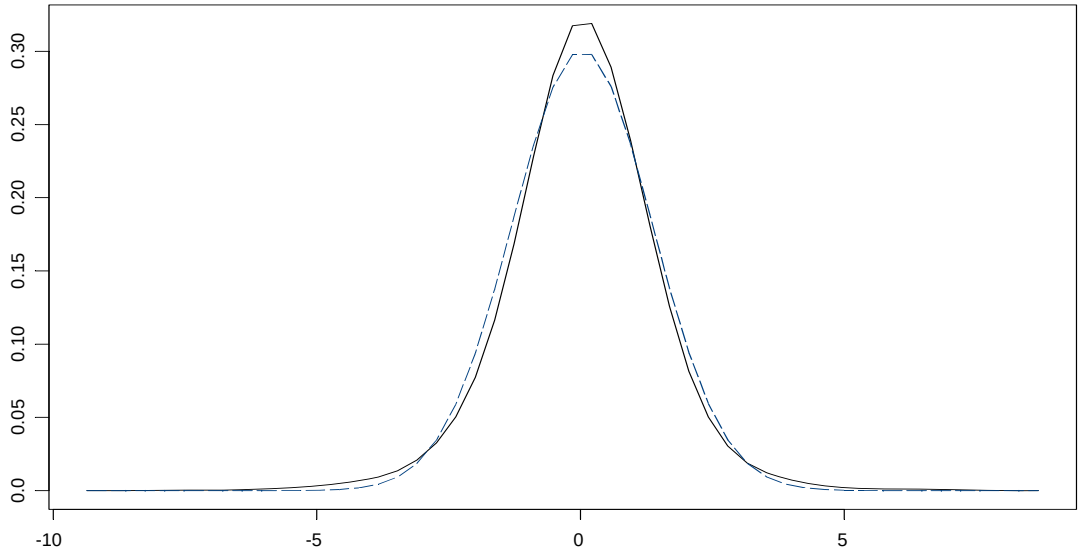
# Commandes S-Plus pour obtenir les graphiques des noyaux

```
> x<-seq(-1,1,0.01)
# Uniforme
> yunif<-ifelse(-0.5<=x&x<=0.5,1,0)
> plot(x,yunif,type="l")
# Triangle
> ytriangle<-(1-abs(x))
> plot(x,ytriangle,type="l")
# Epanechnikov
> yepanech<-0.75*(1-x**2)
> plot(x,yepanech,type="l")
#Quartic
> yquartic<-15/16*(1-x**2)**2
> plot(x,yquartic,type="l")
# Gaussien
> x<-seq(-4,4,0.01)
> ygauss<-dnorm(x)
> plot(x,ygauss,type="l")
```

Densité empirique NP
du CAC
avec un noyau
gaussien




Comparaison densité
empirique NP
et densité théorique
Gaussienne



## 3) Fonction de distribution cumulative (cdf) ou fonction de répartition

- Soit X une v.a. de densité $f_\theta$ . La fonction cdf

$$F(x) = \int_{-\infty}^{x} f_\theta(u)du$$

- Fonction croissante comprise entre 0 et 1

- La fonction cdf est estimée par cdf empirique

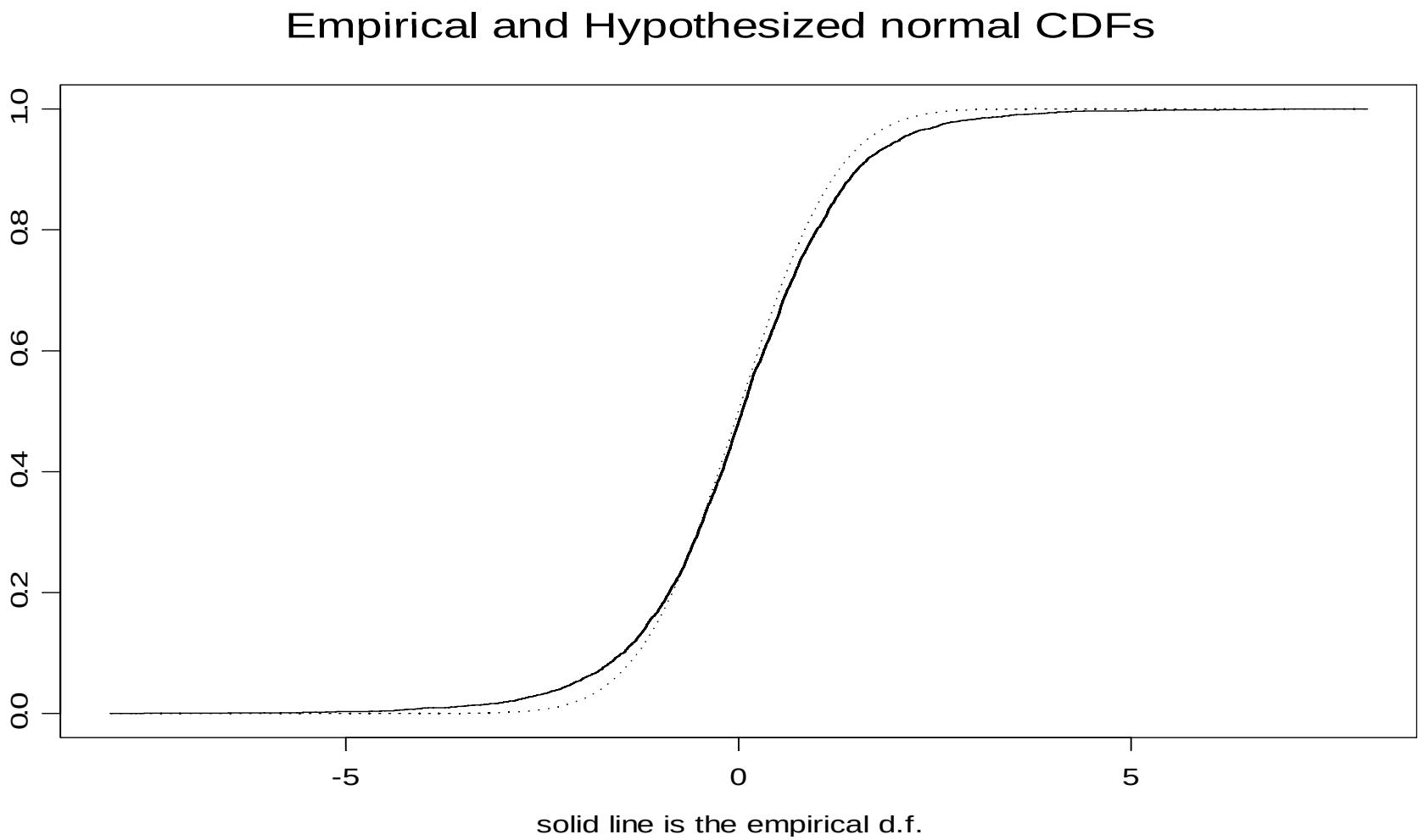
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$$

## 3) Fonction de distribution cumulative (cdf)

- Moins facile à interpréter que la densité de distribution

- Fonction utile pour certains calculs tq : les percentiles ou quantiles de la distribution

# Exemple Cac 40 : cdf empirique et cdf théorique Gaussienne



# 2. Descriptions numériques

- Soit ($x_1$ , …, $x_n$ ) une observation d'un n-échantillon issu de la v.a. X de loi inconnue. Les caractéristiques principales de la loi de distribution de X sont

**2.1 position**

**2.2 dispersion**

**2.3 symétrie**

**2.4 épaisseur des queues**

**2.5 multimodalités**

## 2.1 Mesures de position

- La moyenne: $\overline{X} = \sum_{i=1}^{n} \omega_i x_i \quad où \quad \sum_{i=1}^{n} \omega_i = 1$

Moyenne arithmétique $\rightarrow$ poids uniformes : $\omega_i = 1/n, \quad \forall i$

Moyenne arithmétique : estimateur de E(X) (moment d'ordre 1)

- La médiane : *med(X)*

  med(X) est la valeur qui partage l'échantillon en 2 parties, les valeurs de la première partie étant plus petites que med(X) ; les valeurs de la seconde étant plus grandes.

  C 'est une statistique de rang.

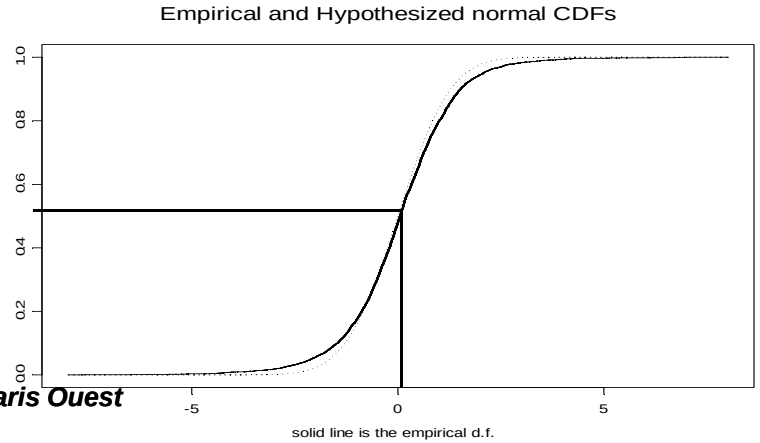
# 2.1 Mesures de position

Soit $x_{(1)}$, …, $x_{(n)}$, l'échantillon rangé par ordre croissant.

Si n est impair : $med(X) = x_{((n+1)/2)}$,

Si n est pair : $med(X) = (x_{(n/2)} + x_{(n/2+1)}) /2$.

La médiane satisfait : $F_n(med(X)) = 0.5$



## 2.1 Mesures de position

- Les quantiles

  Généralisation de la médiane en permettant de découper l'échantillon en un nombre fini de sous-parties:

  4 parties $\rightarrow$ quartiles

  10 parties $\rightarrow$ déciles

  Pour $0 < \alpha < 1$, le quantile d'ordre $\alpha$ est défini par :

$$q_\alpha = F_n^{-1}(\alpha)$$

## 2.1 Mesures de position

- Le mode

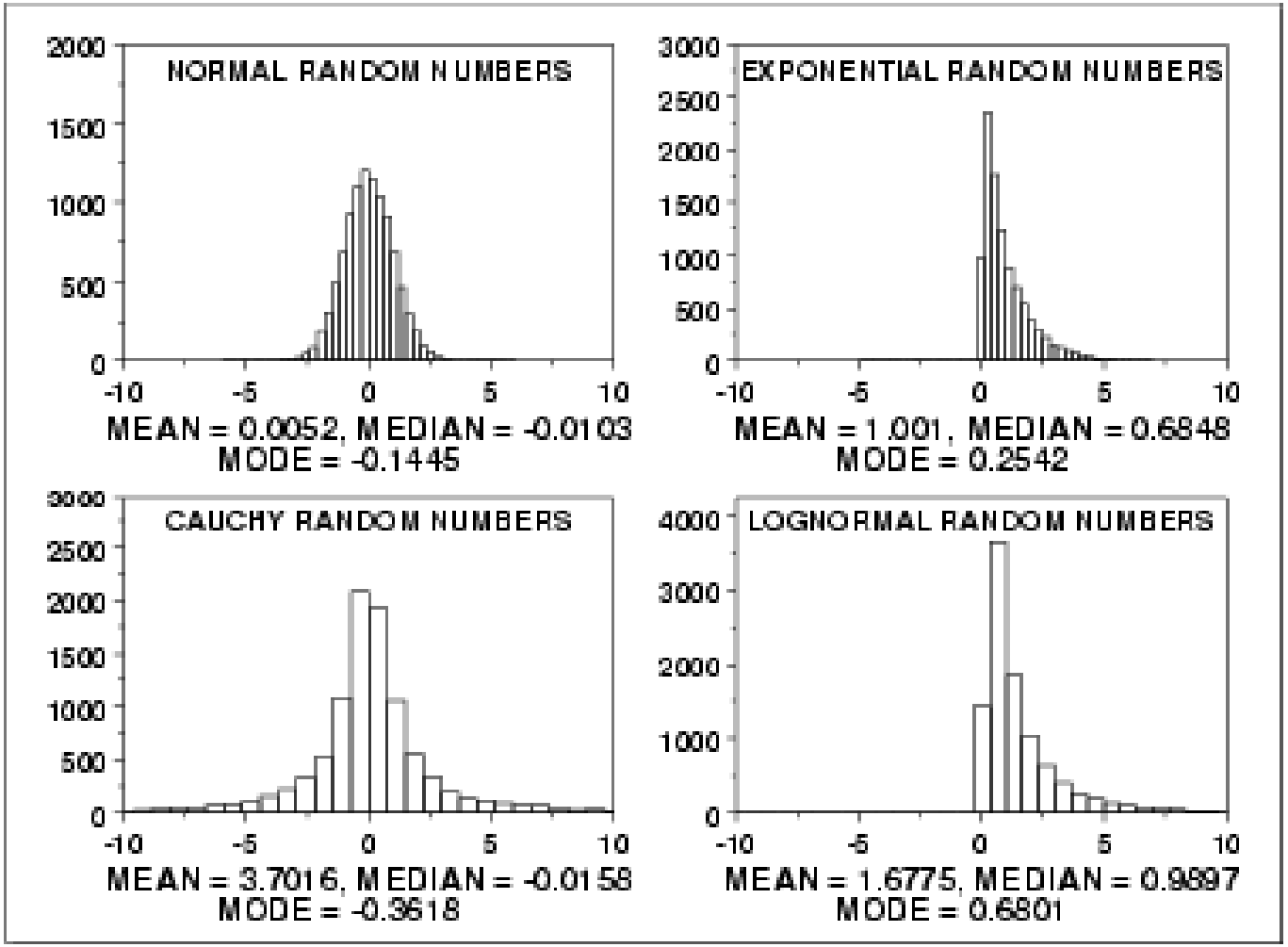
  C'est la valeur de l'échantillon qui apparaît avec la plus grande fréquence.

  Valeur utilisée principalement pour des données qualitatives.

  Pour une variable continue, on choisit la valeur pour laquelle la densité de distribution empirique est maximum.

  Il peut y en avoir des maximums locaux, impliquant plusieurs modes : distribution multimodale.

# Pourquoi différentes mesures de position ?




→ Trouver un estimateur raisonable en cas de non-Gaussianité

# Pourquoi différentes mesures de position ?

Médiane ou moyenne?

- En cas de distribution symétrique: médiane = moyenne
- Dissymétrie à droite : moyenne > médiane
- Dissymétrie à gauche : moyenne < médiane
- Médiane plus robuste aux valeurs aberrantes (« outliers ») ou événements rares

- Mesures alternatives:
  - Mid-mean : moyenne pour les données entre les quantiles 0.25 et 0.75
  - Trimmed-mean : moyenne de l ’échantillon tronqué

## 2.2 Mesures de dispersion

Soit $x_1$ , …, $x_n$ .

2 questions se posent :

1) Comment sont dispersées les valeurs près du centre de la distribution ?

2) Comment sont dispersées les valeurs dans les queues de la distribution ?

Les différentes mesures ci-après donnent plus ou moins de poids à chacune des ces 2 composantes.

## 2.2 Mesures de dispersion

- *Variance empirique* définie par :
  (Moment centré d'ordre 2)

$$s^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

- *Ecart-type* empirique défini par :

$$s(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

- *Range* défini par :

$$x_{(n)} - x_{(1)}$$

- *Average Absolute Deviation* :

$$AAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{X}\right|$$

## 2.2 Mesures de dispersion

- *Median Absolute Deviation* : $MAD = med(|x_i - med(X)|)$

- *Ecart Inter-Quartile* : $IQ = q_{0.75} - q_{0.25}$

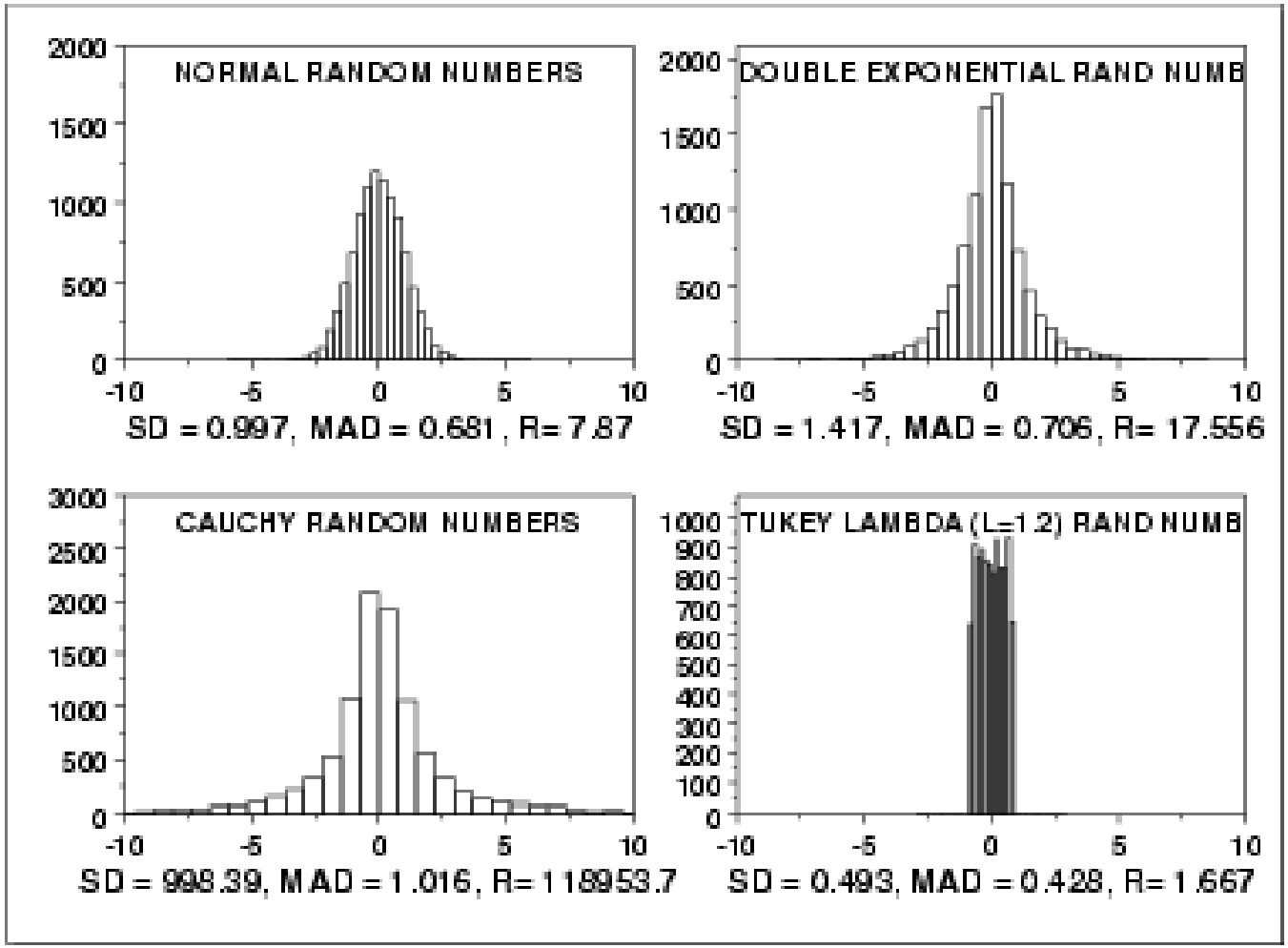
Variance, Ecart-type, AAD et MAD mesurent simultanément les deux aspects de la variabilité.

AAD et MAD ne sur-pondèrent pas les comportement dans les queues.

Range ne mesure que la variabilité des queues

IQ ne mesure que la variabilité centrale

# Pourquoi différentes mesures de dispersion ?




→ Trouver un estimateur raisonable en cas de non-Gaussianité

## 2.3 Mesures de symétrie

Le skewness mesure l'asymétrie, basé sur le moment centré d'ordre 3, $m^3(X)$, tq :

$$m_3(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3$$

et :

$$Sk(X) = \frac{m_3(X)}{s^3(X)}$$

*Sk(X)* est nul si la distribution est symétrique et élevé sinon.

Asymétrie positive = $Sk > 0$ = la queue droite est plus épaisse

Asymétrie négative = $Sk < 0$ = la queue gauche est plus épaisse

## 2.4 Mesure d'épaisseur des queues

La kurtosis mesure l'épaisseur des queues de distribution, basée sur le moment centré d'ordre 4, $m^4(X)$, tq :

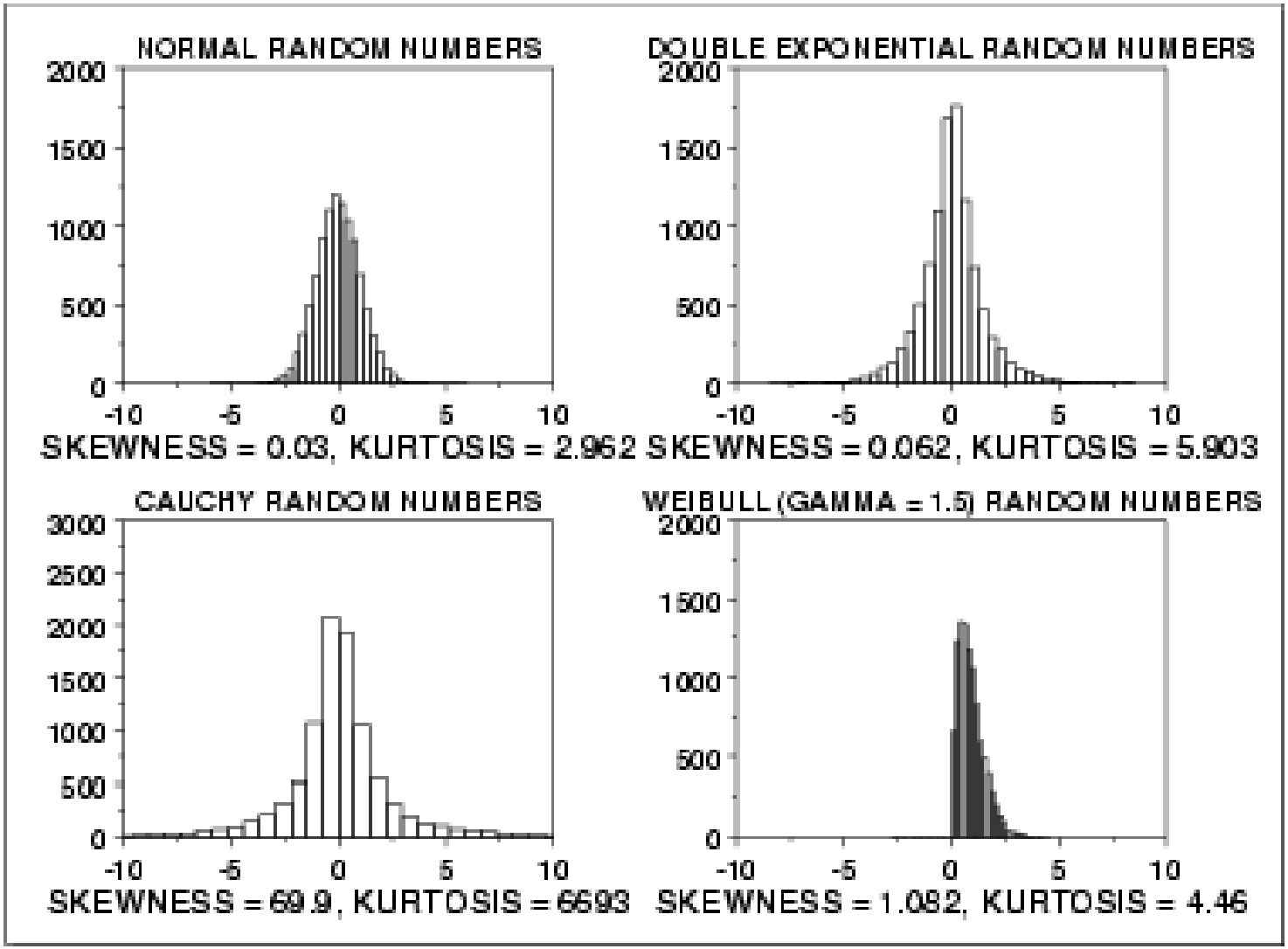
$$m_4(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^4$$

et :

$$K(X) = \frac{m_4(X)}{s^4(X)}$$

*K(X)* est égal à 3 si la distribution est Gaussienne. Une mesure utilisée est l 'Excess Kurtosis définie par : *K(X) - 3*

Un EK(X) > 0 indique des queues de distribution plus épaisses que celles de la loi Normale, et inversement.


NORMAL RANDOM NUMBERS
2000
1500
1000
500
0
-10
-5
0
5
10
SKEWNESS = 0.03, KURTOSIS = 2.962
DOUBLE EXPONENTIAL RANDOM NUMBERS
2000
1500
1000
500
0
-10
-5
0
5
10
SKEWNESS = 0.062, KURTOSIS = 5.903
CAUCHY RANDOM NUMBERS
3000
2500
2000
1500
1000
500
0
-10
-5
0
5
10
SKEWNESS = 69.9, KURTOSIS = 6593
WEIBULL (GAMMA = 1.5) RANDOM NUMBERS
2000
1500
1000
500
0
-10
-5
0
5
10
SKEWNESS = 1.082, KURTOSIS = 4.46

## 2.5 Multimodalités

- Distribution à plusieurs modes :

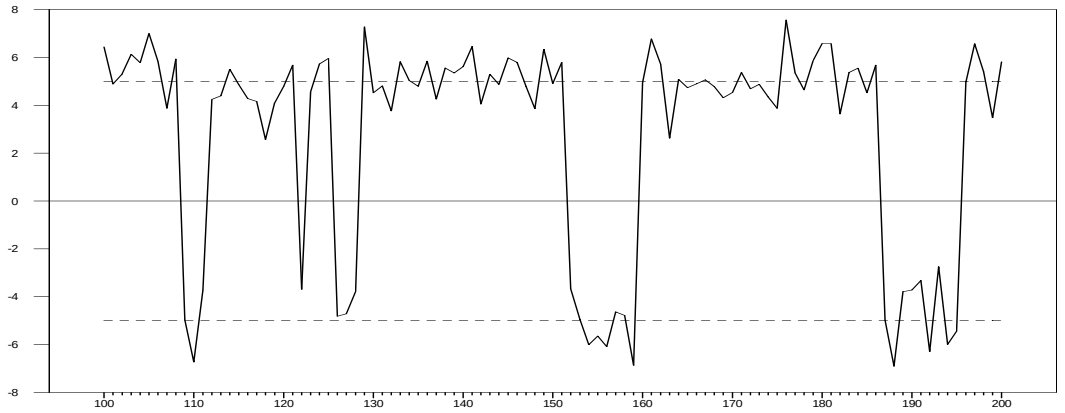
approximation par une mélange de lois unimodales

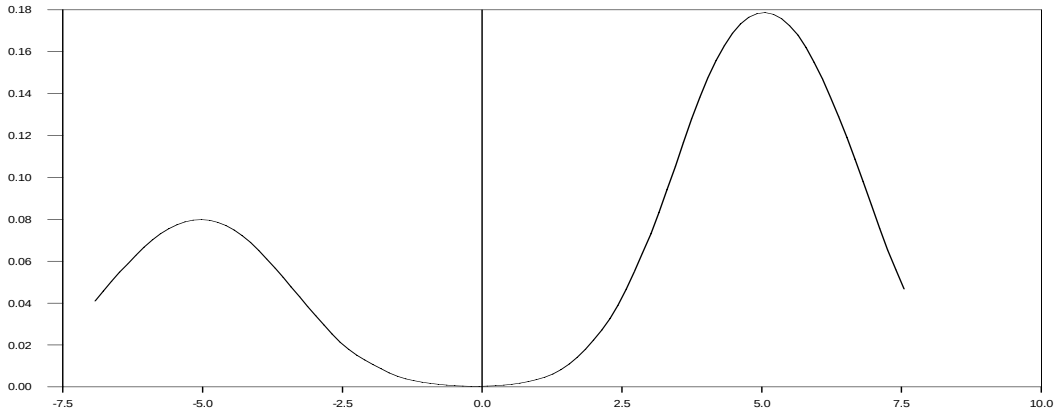
$\rightarrow$ en général, mixture de lois Normales

Indicateur de présence de non-linéarité dans les données

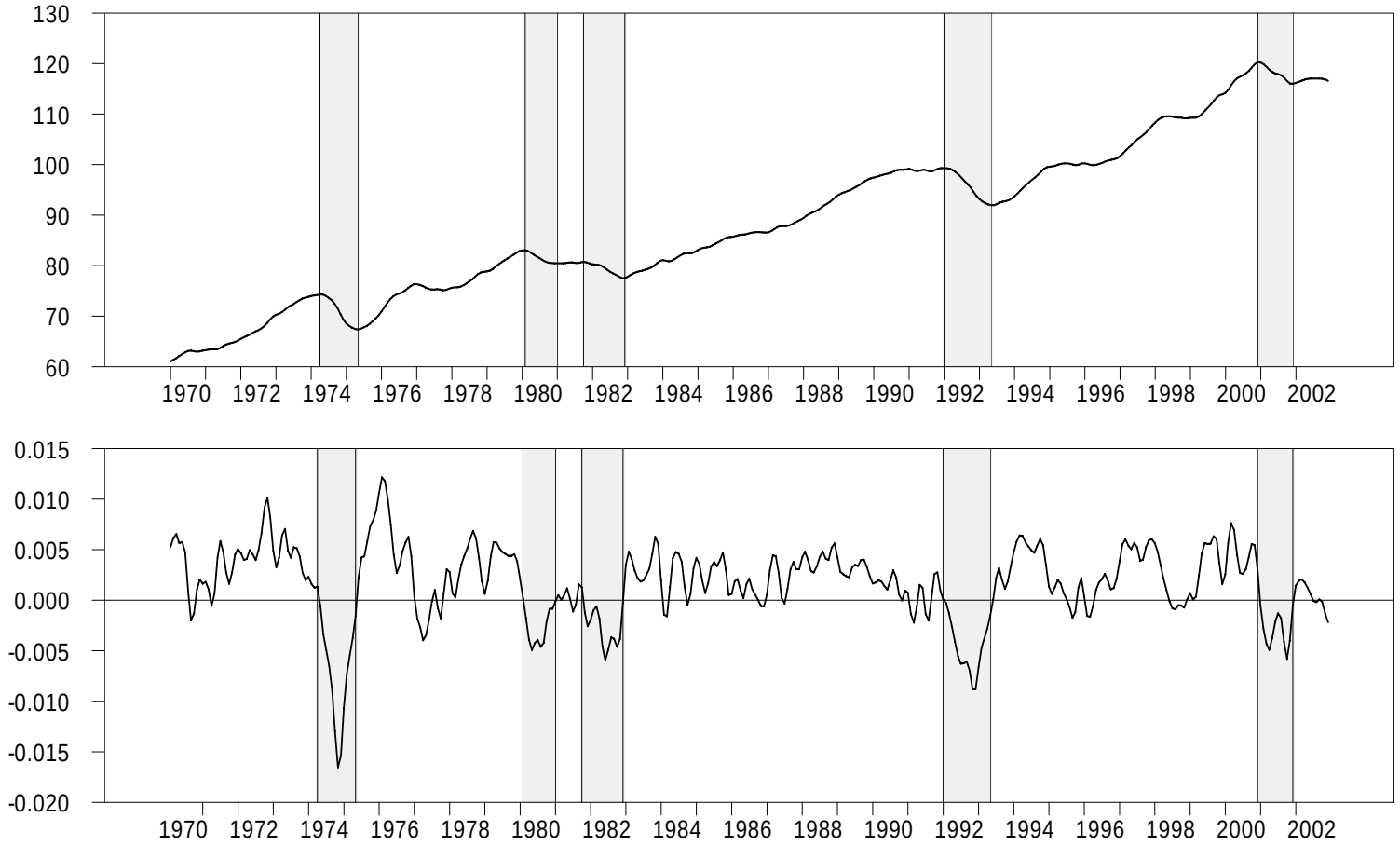
$\rightarrow$ utilisation de modèles non-linéaires ou linéaire par morceaux

# Exemple de distribution bimodale

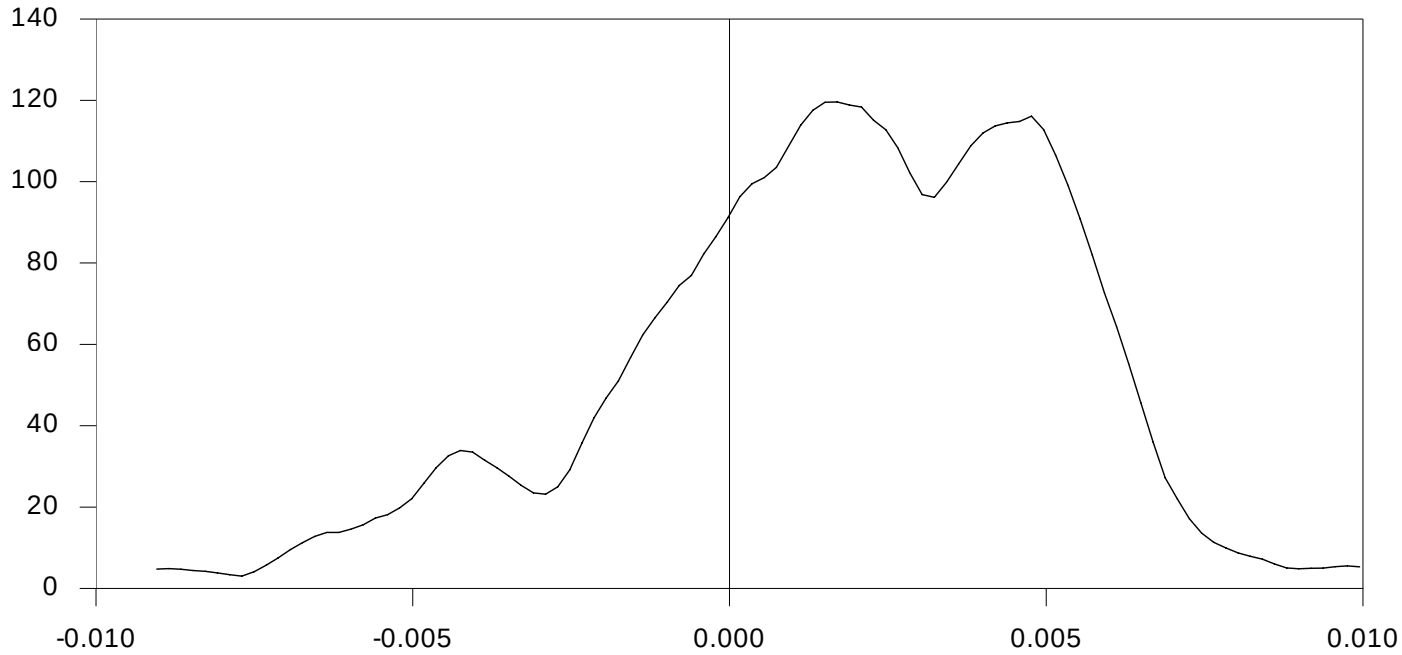


# Exemple de distribution trimodale :
# Indice de la production industrielle de zone euro



# Exemple de distribution trimodale



# Exemple : Rendements journaliers du CAC 40 de 1987 à 2004 (n=4337)

```
> summary(cac40.rdt)

Regular Time Series:
Observations: 4337

     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 -0.07678 -0.00681  0.00041  0.00032  0.00784  0.07002

Time Parameters :
 start deltat frequency
 2     1      1
> skewness(cac40.rdt)
[1] -0.1245464
> kurtosis(cac40.rdt)
[1] 2.833693
```

# 3. Distributions continues usuelles

A partir d'une observation d'un n-échantillon, on cherche à identifier la loi de X à une loi connue, à partir de ses caractéristiques empiriques observées précédemment.

- Gaussienne
- Student
- Uniforme
- Chi-2
- Fischer
- Log-Normale
- Exponentielle

- La loi de Gauss (ou Normale)

La va X suivant une loi Normale N (m,$\sigma^2$) a la densité suivante:

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(u-m)^2}{2\sigma^2})$$

Loi Normale standard pour m = 0 et $\sigma$ =1

cdf: $$\Phi(x) = P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{(u-m)^2}{2\sigma^2}) du$$

- La loi de Gauss (ou Normale)

Propriétés (Rappel) :

$$\Phi(-x) = 1 - \Phi(x)$$

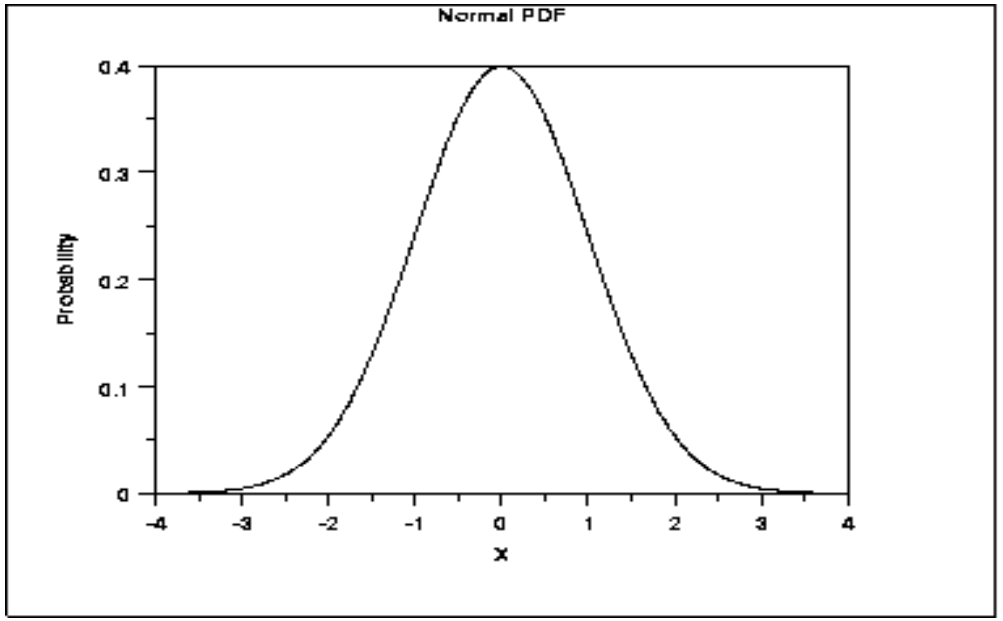
$$P(-x \leq X \leq x) = 2\Phi(x) - 1$$

$$X \approx N(m, \sigma^2) \Leftrightarrow \frac{X - m}{\sigma} \approx N(0,1)$$

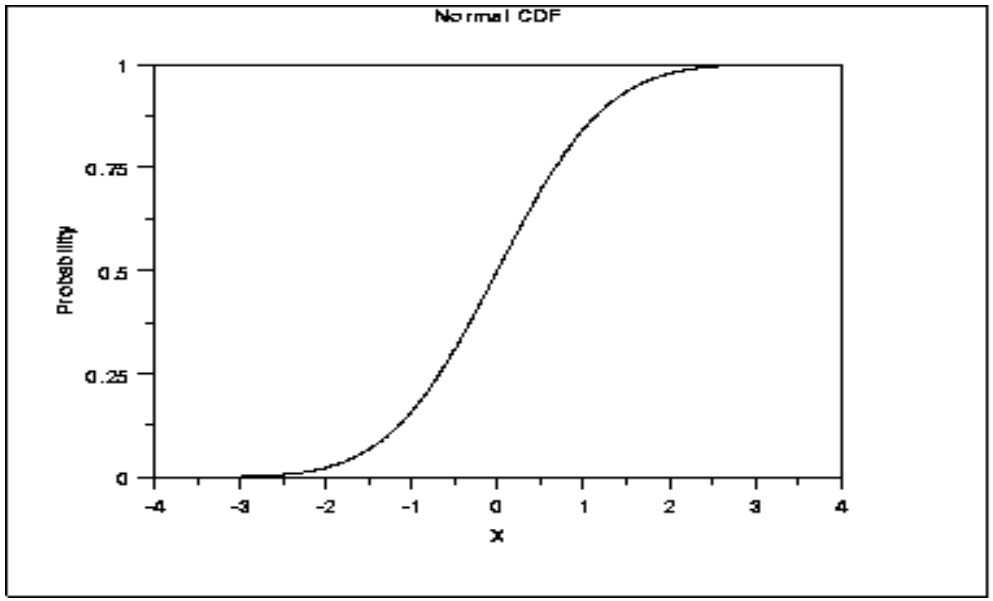
$$P(m - \sigma \leq X \leq m + \sigma) \approx 68\%$$

$$P(m - 2\sigma \leq X \leq m + 2\sigma) \approx 95\%$$

Distributions de Gauss standard




cdf de Gauss standard



- Caractéristiques de la loi de Gauss

Moyenne = Mediane = Mode = m

Ecart-type = $\sigma$

Skewness = 0

Kurtosis = 3

- La loi Uniforme

La v.a. X suivant une loi Uniforme dans l ’intervalle [a,b] a la densité suivante:

$$f(u) = \frac{1}{b-a} 1_{[a,b]}$$

a est le paramètre de position

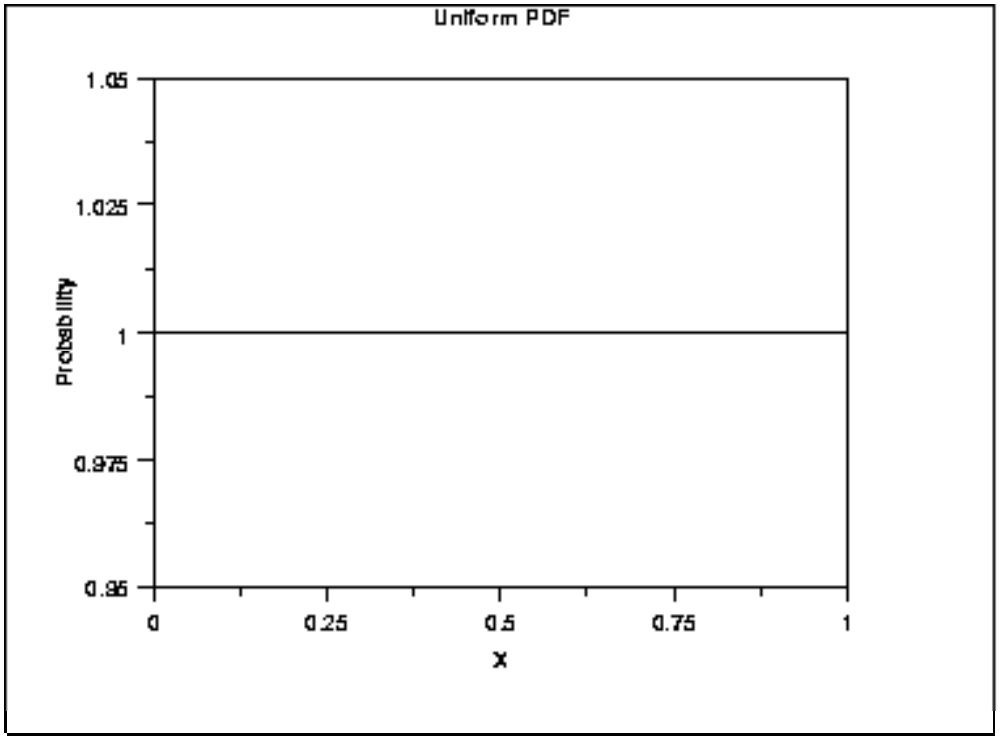
b-a est le paramètre de dispersion

Distribution standard uniforme : a=0, b=1

cdf:

$$F(x) = P(X \leq x) = x$$

# Distribution Uniforme



- Loi Uniforme

Moyenne = Mediane = (a+b)/2

Variance = $\frac{(b-a)^2}{12}$

Skewness = 0

Kurtosis = 9/5

- La loi de Student

Soit $X_0$ , $X_1$ , …, $X_n$ , n+1 v.a. iid selon la loi Normale standard.
Alors la v.a. T tq:

$$T = \frac{X_0}{\sqrt{1/n \sum X_i^2}}$$

suit une loi de Student à n degrés de liberté

Densité:

$$f(u) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} (1+u^2/n)^{-(n+1)/2}$$

- La loi de Student

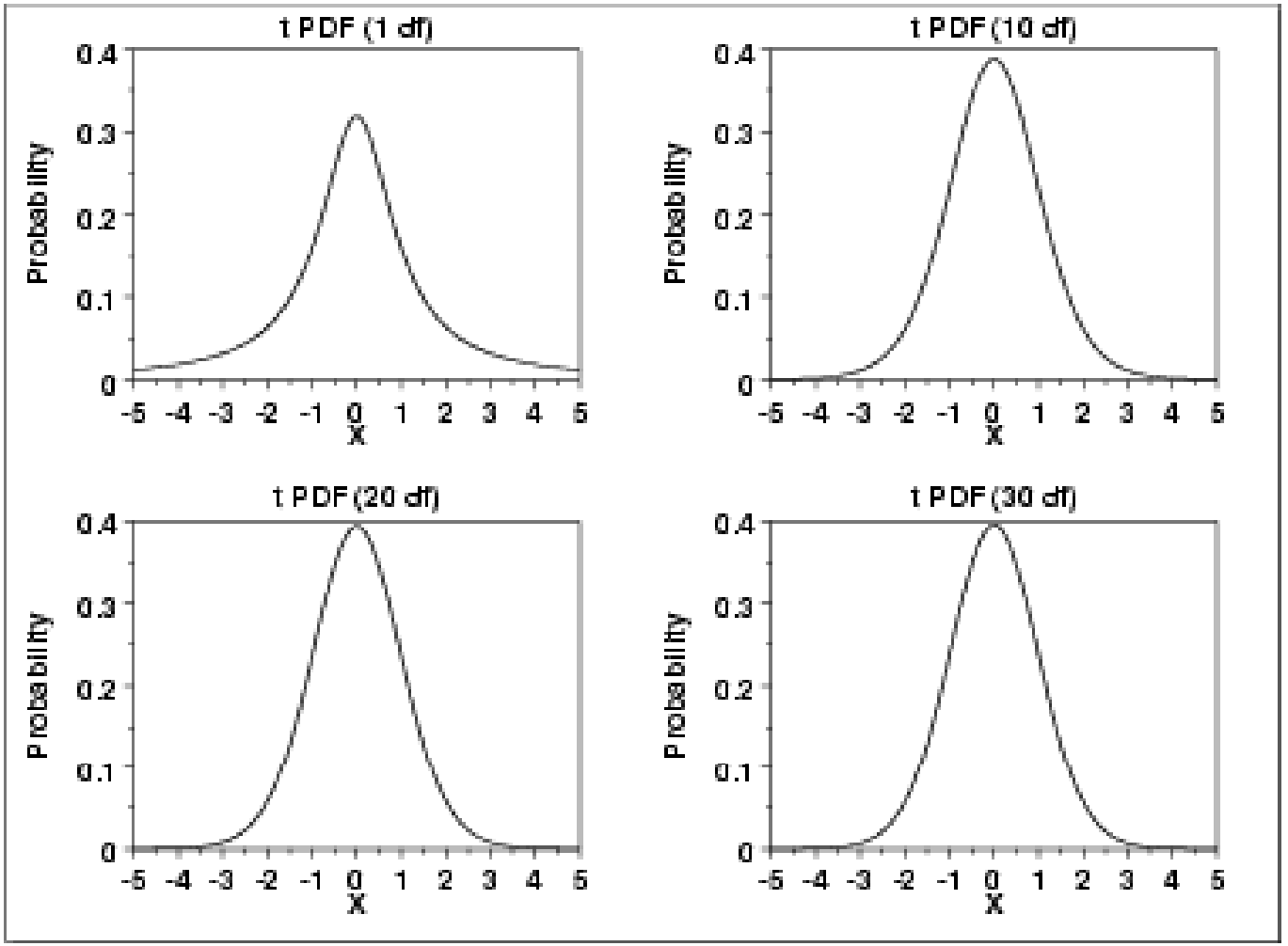
avec la fonction Gamma, pour $a > 0$, tq:

$$\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx$$

On rappelle que :

$$\Gamma(a) = (a-1)\Gamma(a-1)$$
$$\Gamma(1/2) = \sqrt{\pi}$$
$$\Gamma(n) = (n-1)!, n \in N$$

# Distributions de Student



- La loi de Student

Moyenne = Mediane = 0

Variance = $\frac{n}{n-2}$ n > 2

Skewness = 0

Kurtosis = $\frac{3(n-2)}{n-4}$ n > 4

- Loi du Chi-2

Soit $X_1$ , …, $X_n$ , *n* v.a. iid selon la loi Normale standard.
Alors la v.a. Z tq:

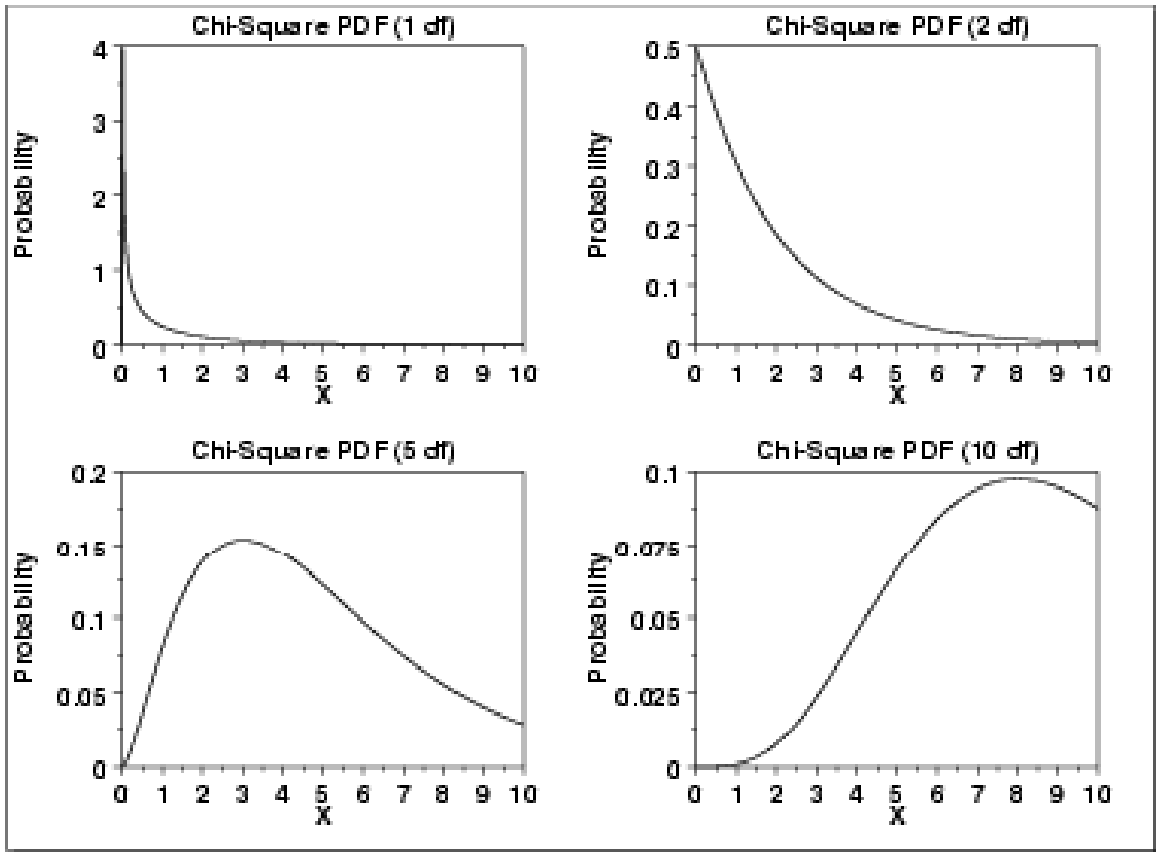
$$Z = \sum_{i=1}^{n} X_i^2$$

suit une loi du Chi-2 à n degrés de liberté

Densité pour u ≥ 0:

$$f(u) = \frac{1}{2^{n/2}\Gamma(n/2)} \exp(-u/2) u^{(n/2-1)}$$

# Distributions du Chi-2



- La loi du Chi-2

Moyenne = n

Mediane = n - 2/3 , lorsque n grand

Mode = n - 2 , pour n > 2

Variance = $2n$

Skewness = $\frac{2^{1/5}}{\sqrt{n}}$

Kurtosis = $3+\frac{12}{n}$

- Loi de Fischer

Soit $X_1$, …, $X_n$, $X_{n+1}$, …, $X_{n+m}$, $n+m$ v.a. iid selon la loi N(0,1)
Alors la v.a. F tq:

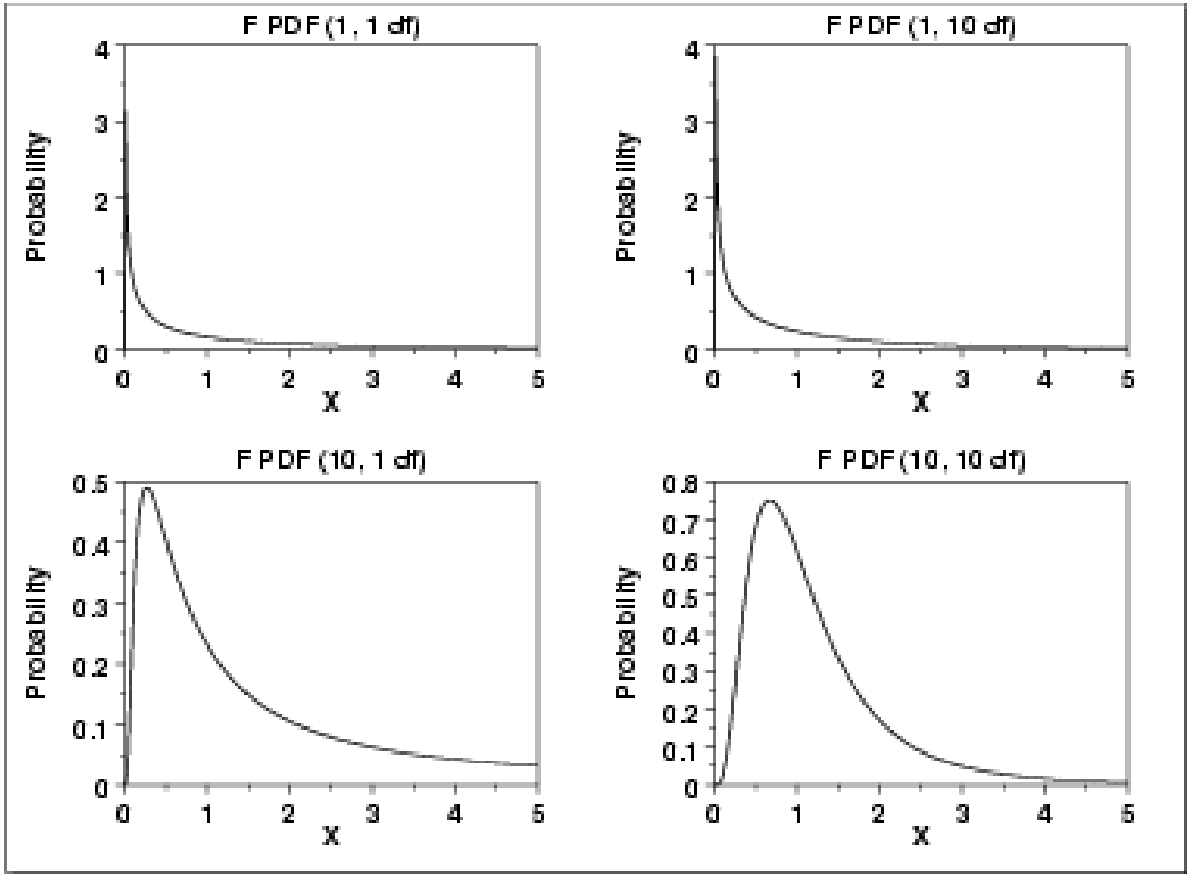
$$F = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\frac{1}{m}\sum_{i=n+1}^{m} X_i^2}$$

suit une loi de Fischer à (n,m) degrés de liberté

Densité pour $u \geq 0$:

$$f(u) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} n^{n/2} m^{m/2} u^{n/2-1} (m+nu)^{-(n+m)/2}$$

# Distributions de Fischer



- Loi de Fischer

Moyenne = $\frac{m}{m-2}$ , m > 2

Mode = $\frac{m(n-2)}{n(m+2)}$ , n > 2

Ecart-type = $\sqrt{\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}}$ $\nu_2 > 4$

- Loi exponentielle

La va X suivant une loi exponentielle a la densité suivante:

$$f(u) = \frac{1}{\beta} \exp(-\frac{(u-\mu)}{\beta})$$

pour $u \geq \mu$ et $\beta > 0$.

$\mu$ : paramètre de position et $\beta$ : paramètre de dispersion

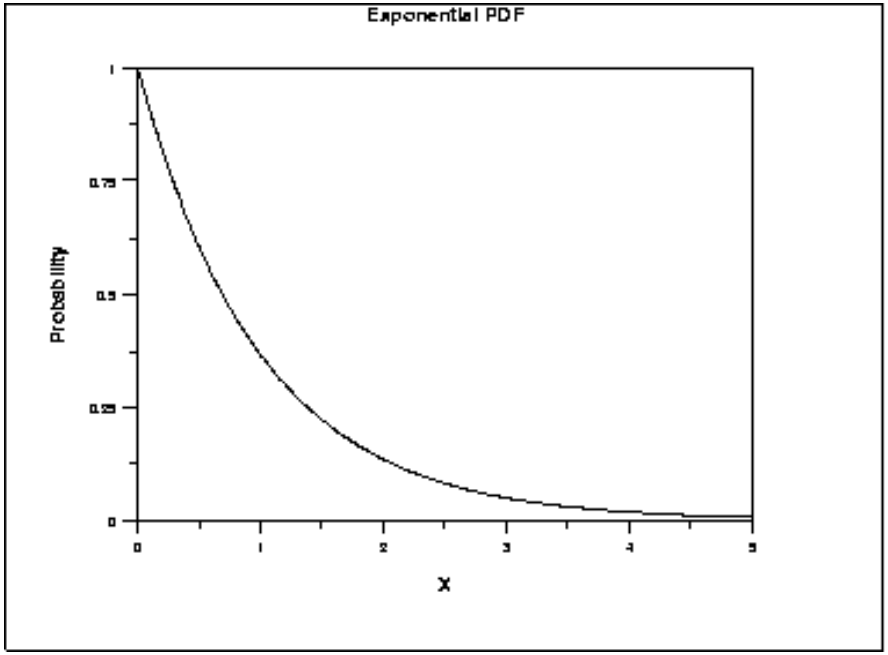
Loi exponentielle standard pour $\mu = 0$ et $\beta = 1$

cdf: $x \geq 0$ et $\beta > 0$

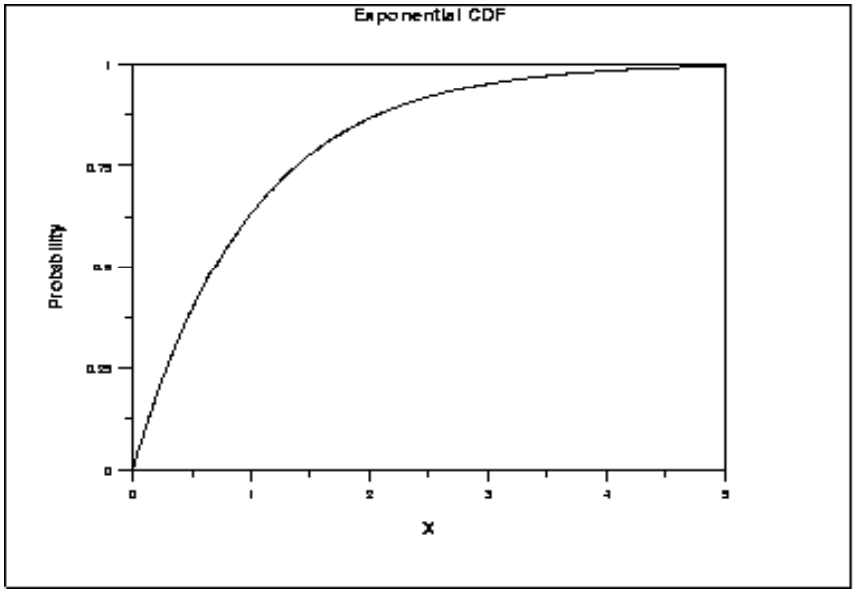
$$F(x) = P(X \leq x) = 1 - \exp(-x/\beta)$$

# Distribution exponentielle standard

Densité




cdf



- La loi exponentielle

Moyenne = $\beta$

Mediane = $\beta \ln(2)$

Mode = 0

Variance = $\beta^2$

Skewness = 2

Kurtosis = 9

- Loi log-Normale

La va X suit une loi log-normale si log(X) suit une loi Normale. Sa densité est la suivante :

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{u}\exp(-\frac{(\log(u)-m)^2}{2\sigma^2})$$

u ≥ 0 et σ > 0 (m = position, σ = dispersion)

cdf: x ≥ 0 et σ > 0

$$F(x) = \Phi(\frac{\ln(x)}{\sigma}) \qquad x \geq 0; \sigma > 0$$

- Loi Log-normale standard

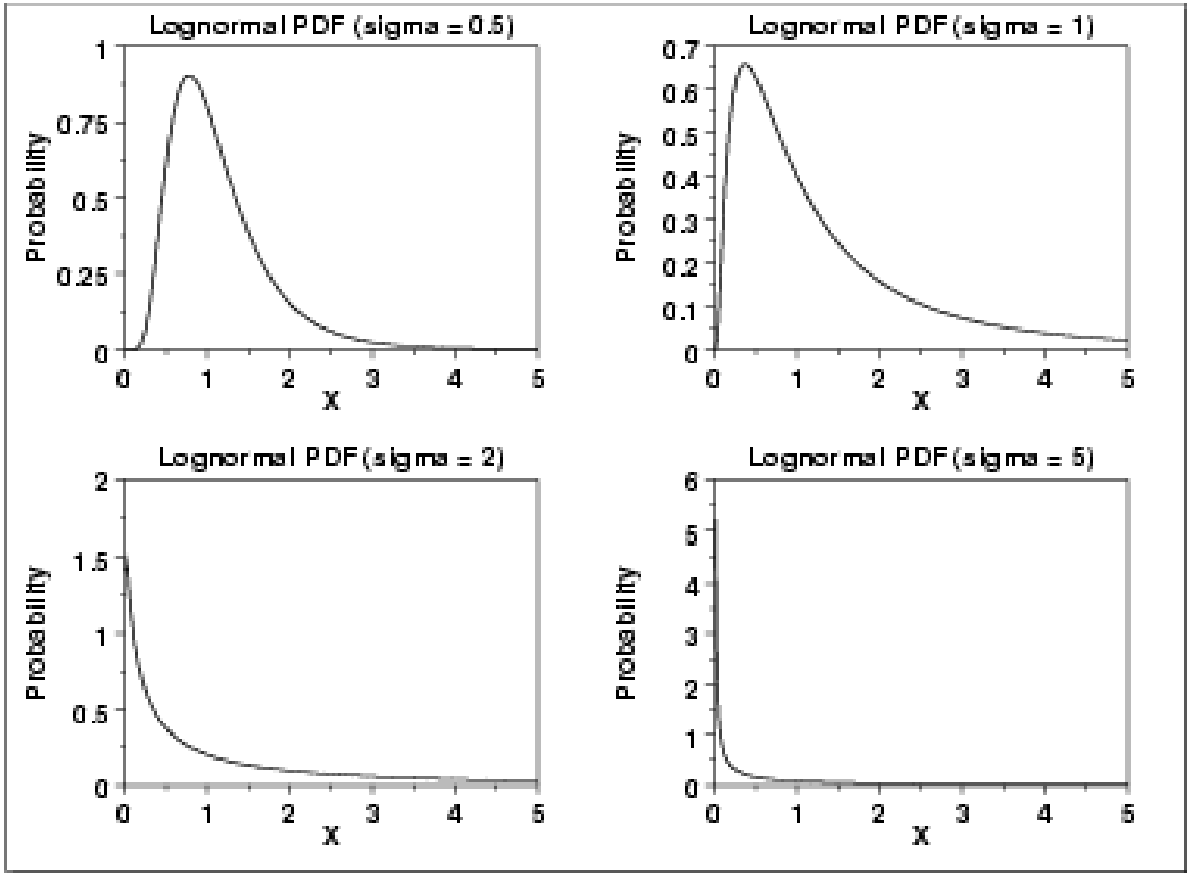
Moyenne = $\exp(\frac{\sigma^2}{2})$

Variance = $\exp(\sigma^2)(\exp(\sigma^2)-1)$

Skewness = $\exp(\sigma^2+2)\sqrt{(\exp(\sigma^2)-1)}$

Kurtosis = $\exp(\sigma^2)^4+2\exp(\sigma^2)^3+3\exp(\sigma^2)^2-3$

# Distributions log-Normale



- Loi Skewed-Normal

Sa densité est la suivante : $f(x) = 2\phi(x)\ \Phi(\alpha x)$

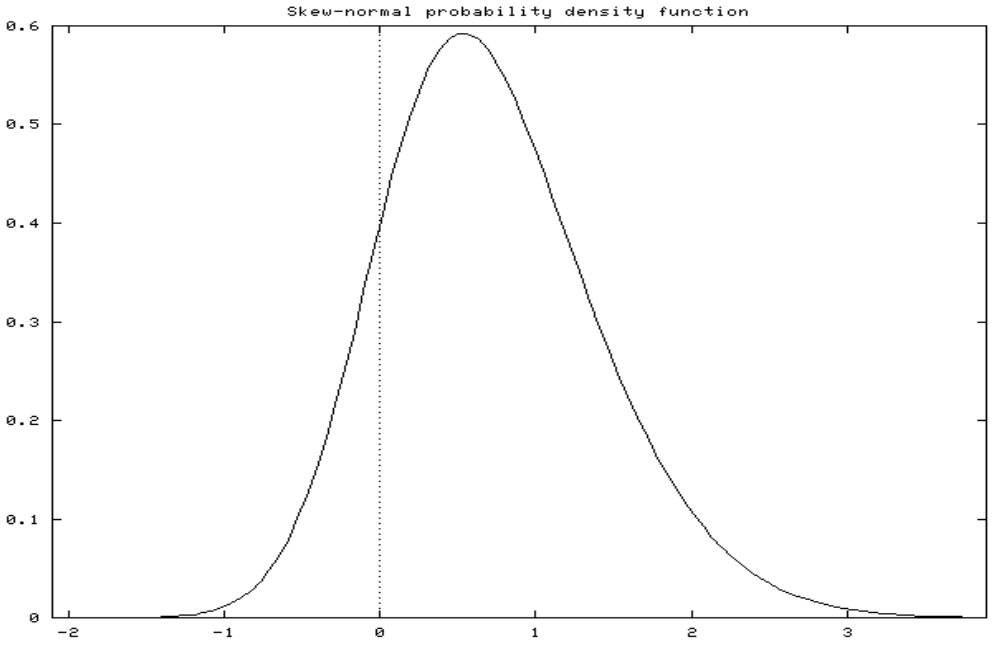
avec : $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}, \qquad \Phi(\alpha x) = \int_{-\infty}^{\alpha x} \phi(t)\, dt$

$\rightarrow$ Si $\alpha = 0$, on retrouve la Normale

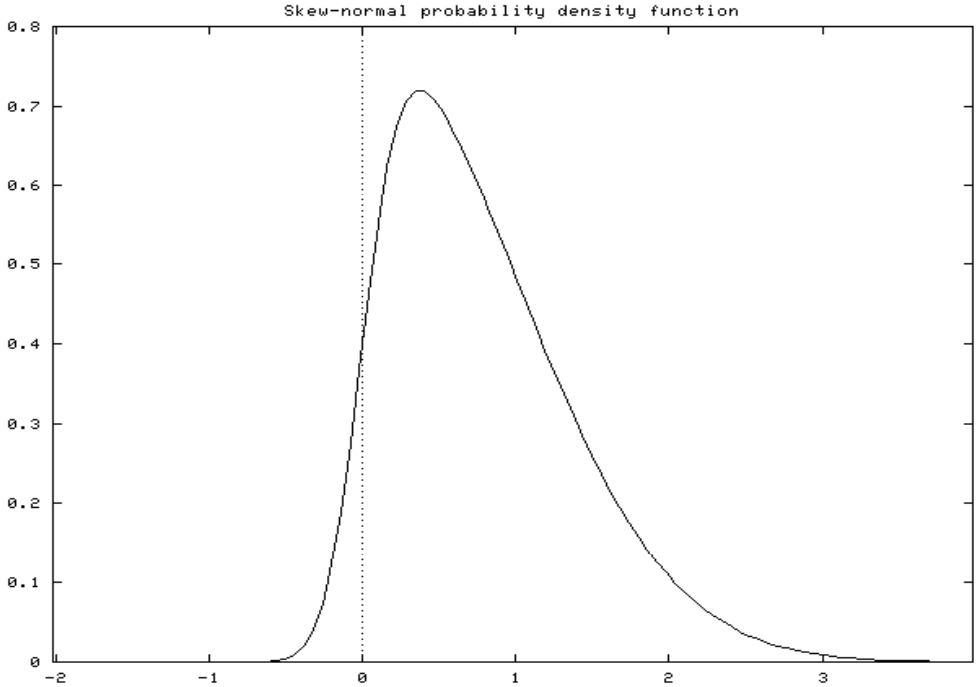
$\rightarrow$ Quand $\alpha$ augmente, le skewness augmente aussi

$\rightarrow$ Quand $\alpha$ change de signe, la densité prend la forme opposée

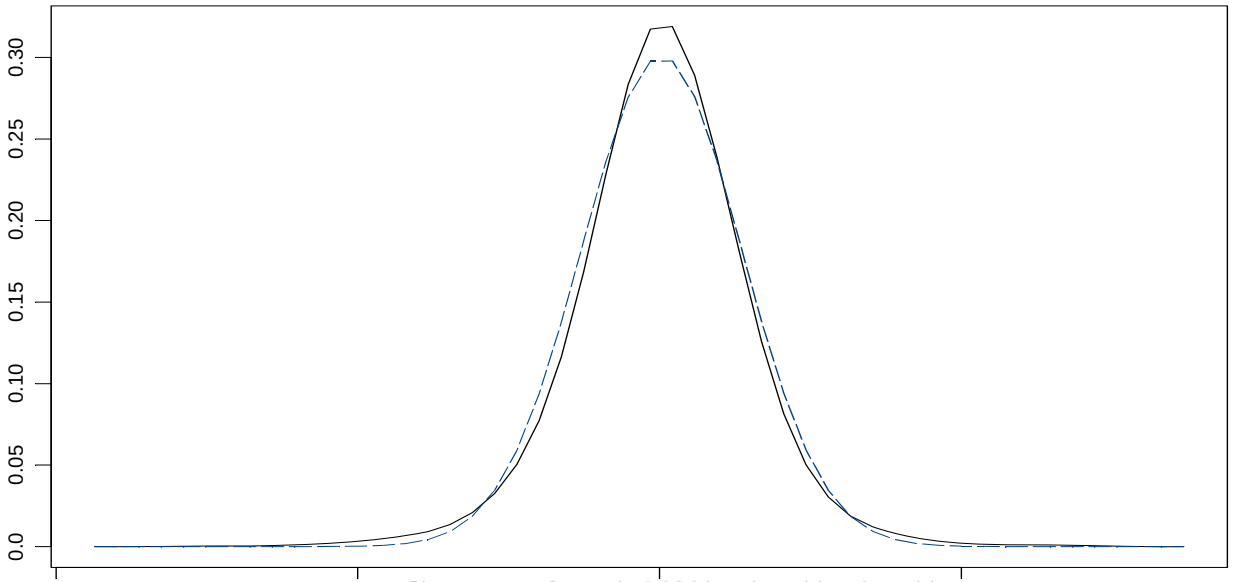
$\alpha = 2$



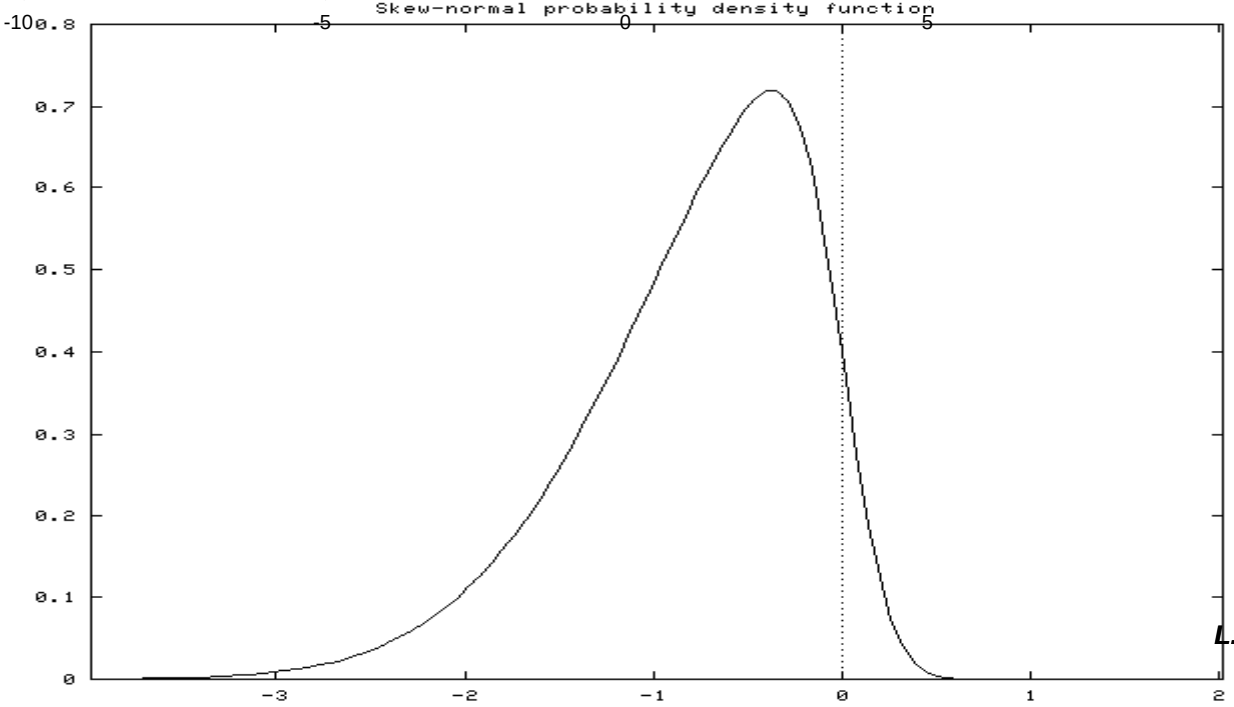

$\alpha = 5$



$\alpha = 0$




$\alpha = -5$



# 4. Comparaison de distributions

Outils graphiques de comparaisons de distributions :

1) Box-Plot
2) QQ- Plot

Outils plus formels : Tests d ’hypothèses

## 4.1 le Box-Plot

- Permet une représentation graphique de la distribution basée sur les résumés numériques de position et dispersion

- Utile pour comparer les distribution de 2 populations

- Mise en évidence de valeurs aberrantes

**Principe :**

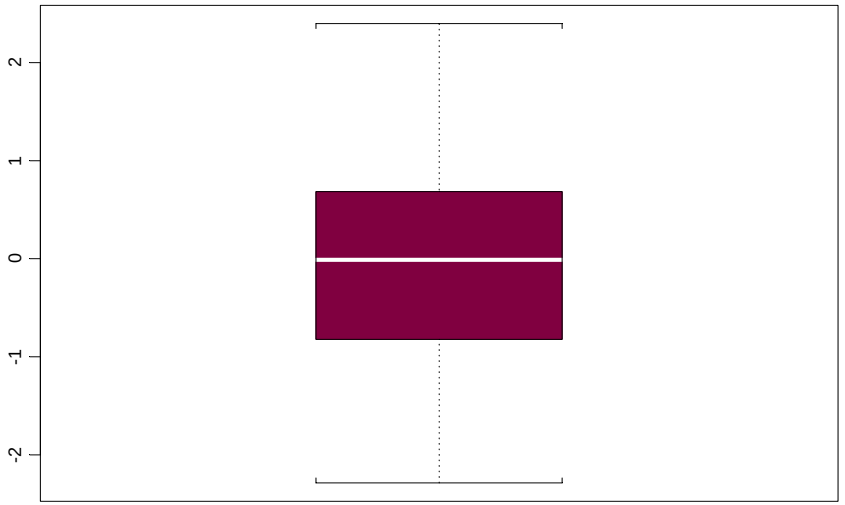
Les quartiles encadrent la médiane.

Est outlier :

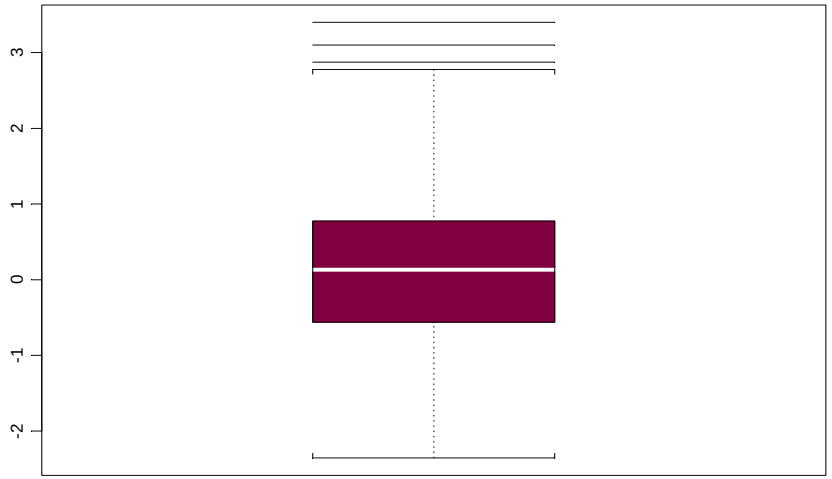
toute valeur supérieure à q(0.75)+1.5*IQ

toute valeur inférieure à q(0.25)-1.5*IQ

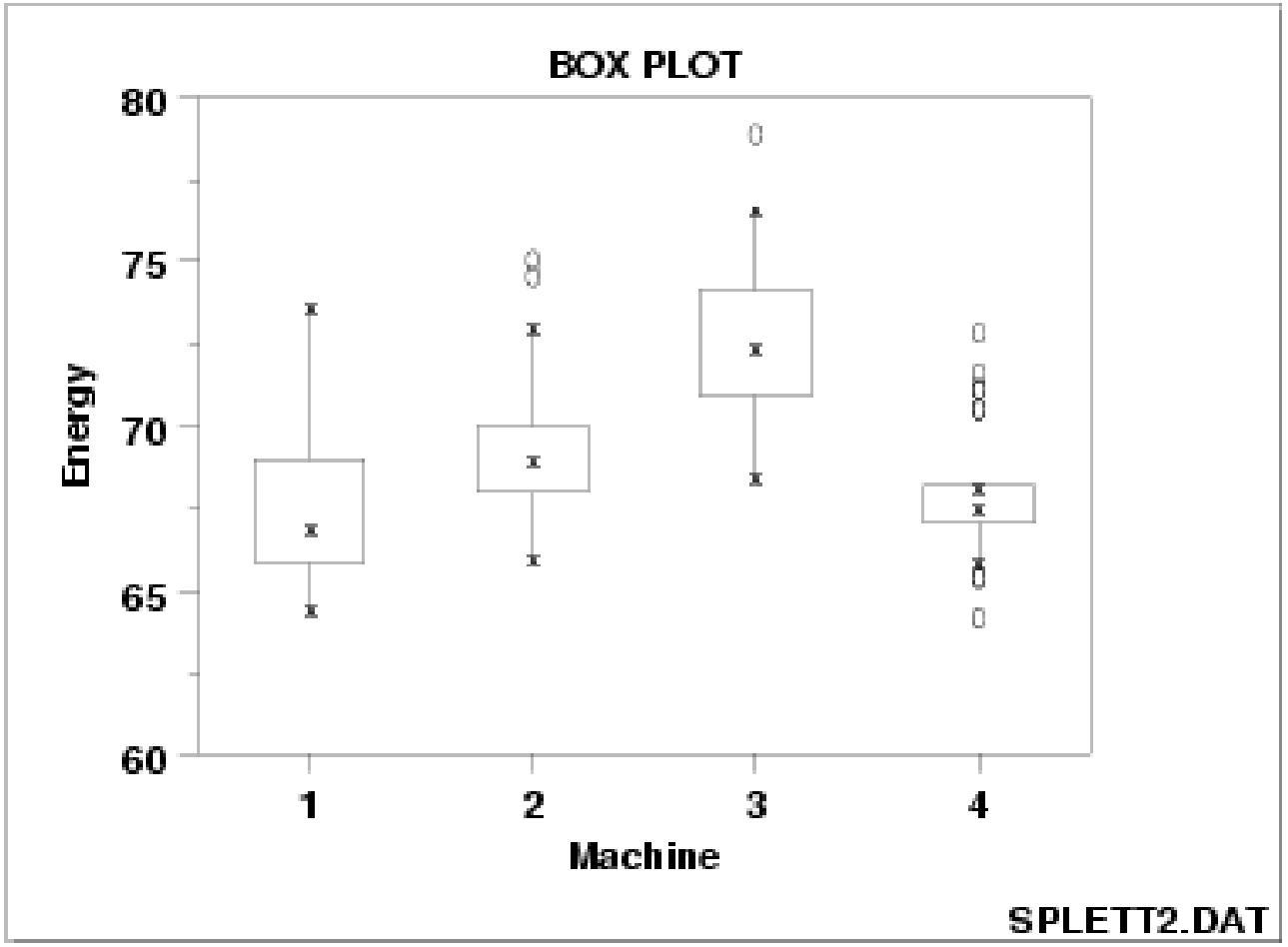
N(0,1)




t(20)



# Exemple : 4 types d'appareils à produire de l'énergie



# Exemple : Rendements du CAC 40

## 4.2 le QQ-Plot

- Permet de comparer la distribution d'une variable avec celle d'une autre variable ou d'une loi théorique à partir des quantiles

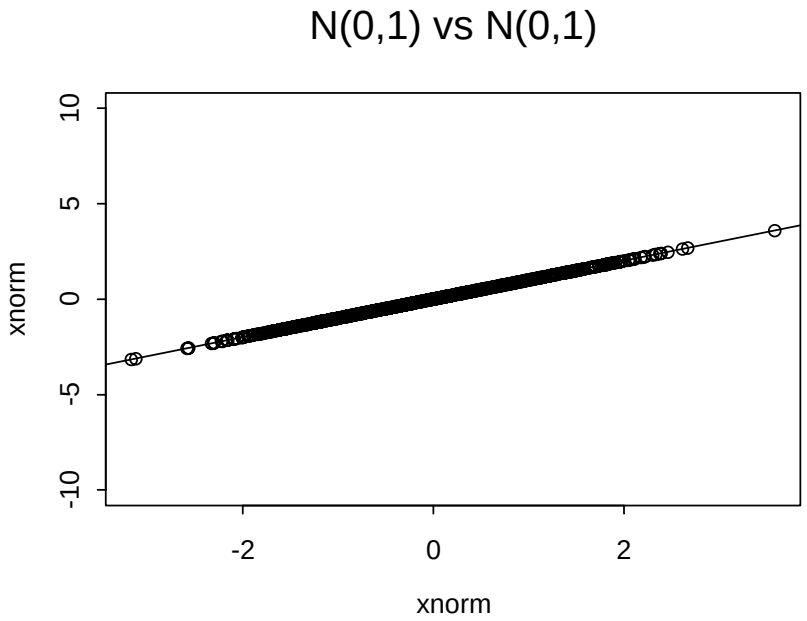
- Diagramme (« scatter-plot ») des quantiles de X1 contre X2
  - Quantiles empiriques de X1 contre quantiles empiriques X2
  - Quantiles empiriques de X1 contre quantiles théoriques d'une loi donnée

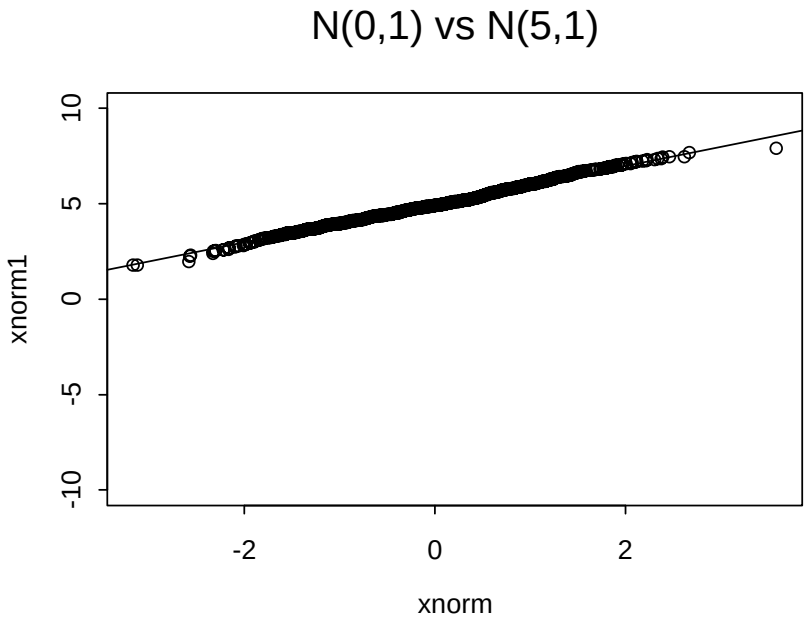
- Si les 2 lois sont proches, le diagramme est proche d'une ligne droite de référence

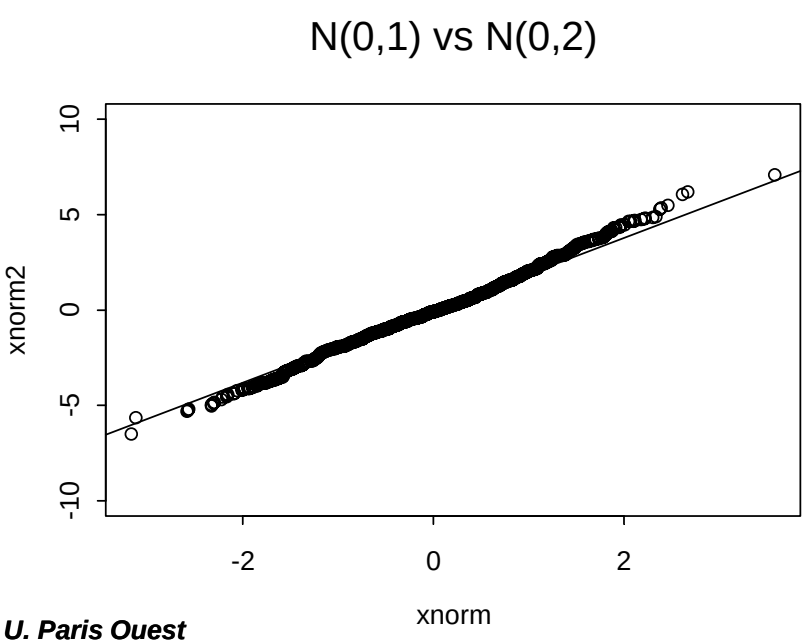
<u>Avantage:</u>

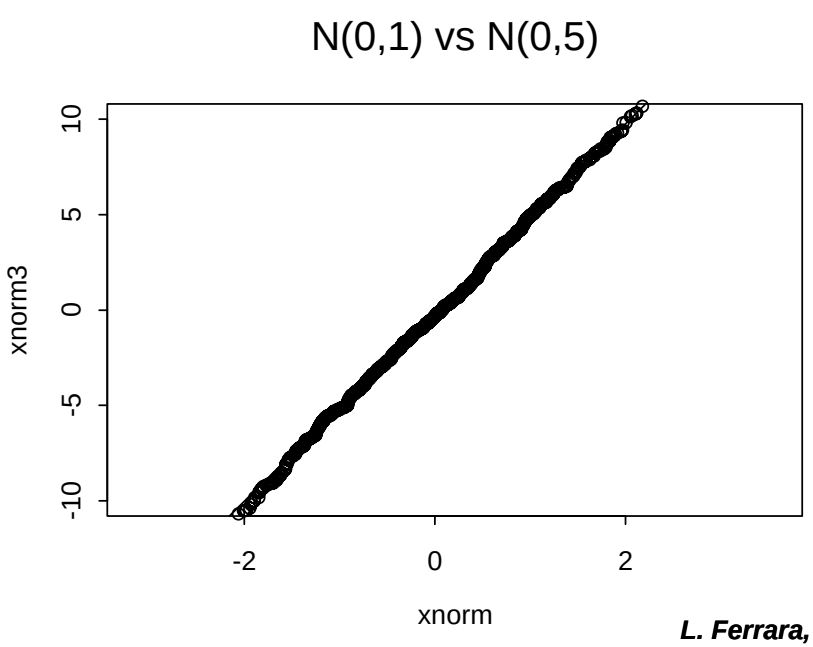
les 2 jeux de données n'ont pas besoin d'être de même taille

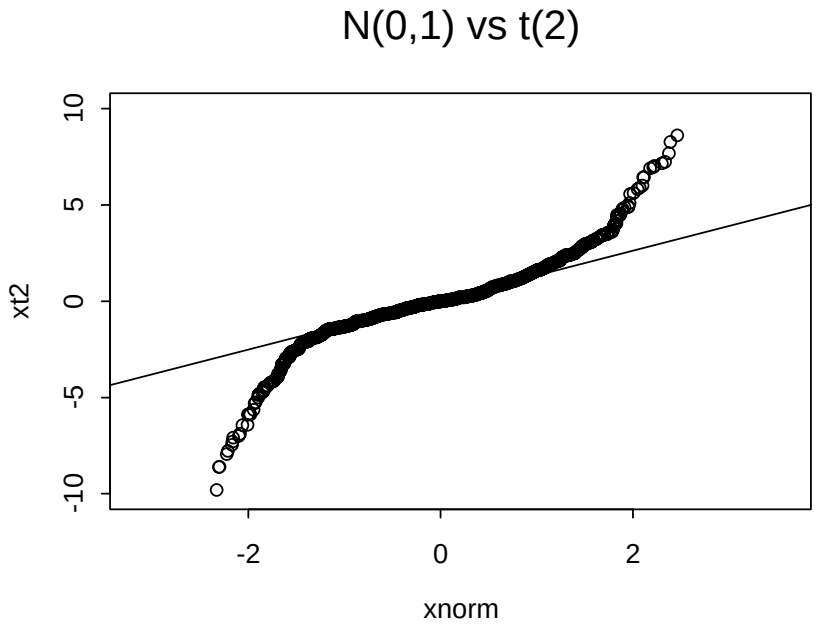
# Exemples

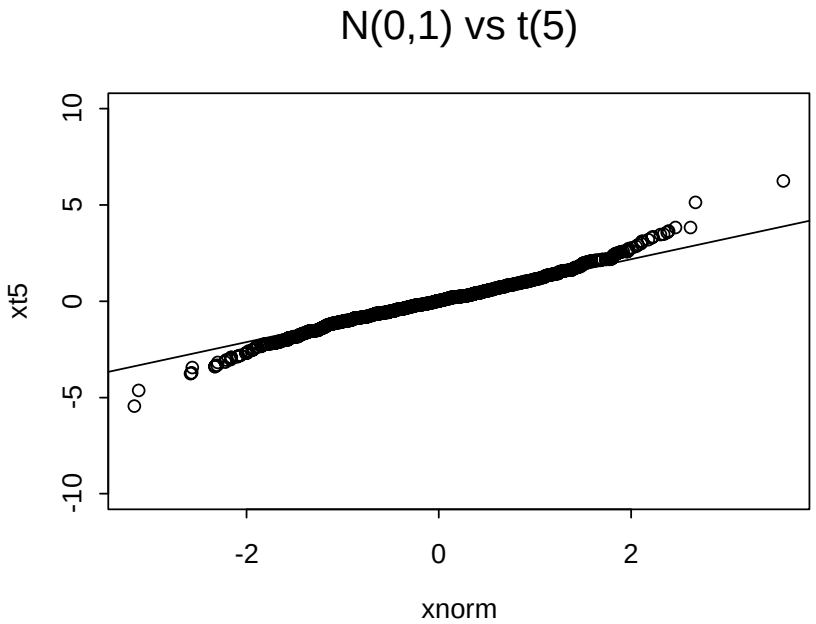

N(0,1) vs N(0,1)
xnorm
xnorm
-10
-5
0
5
10
-2
0
2

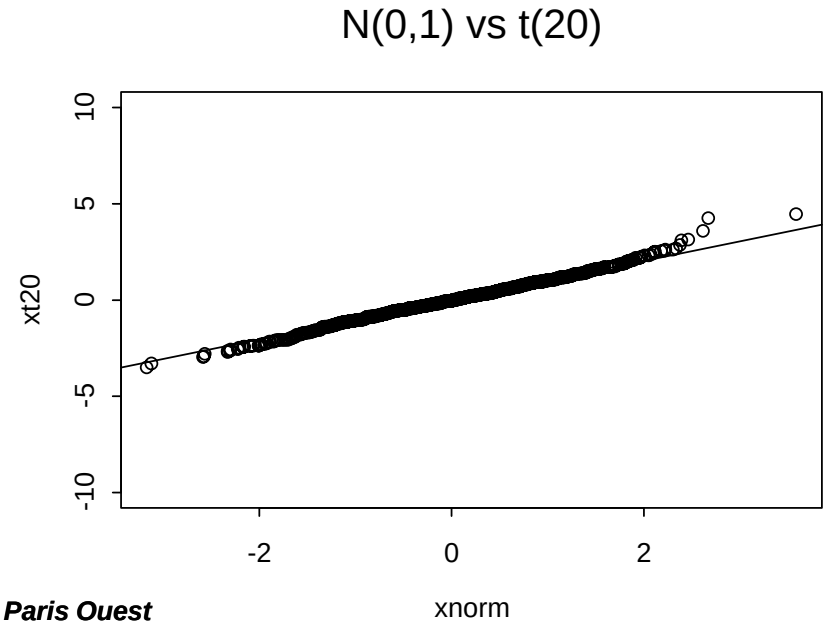

N(0,1) vs N(5,1)
xnorm1
xnorm
-10
-5
0
5
10
-2
0
2

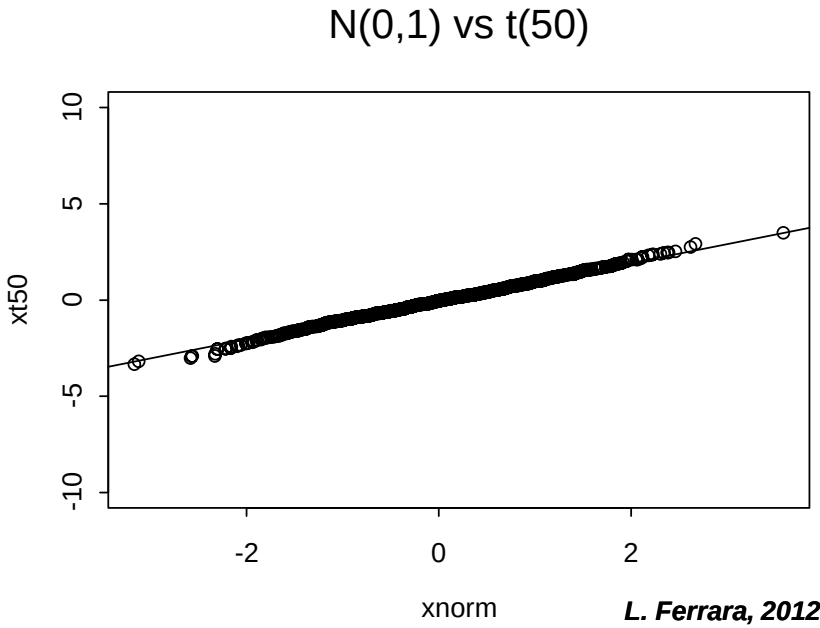

N(0,1) vs N(0,2)
xnorm2
xnorm
-10
-5
0
5
10
-2
0
2


N(0,1) vs N(0,5)
xnorm3
xnorm
-10
-5
0
5
10
-2
0
2

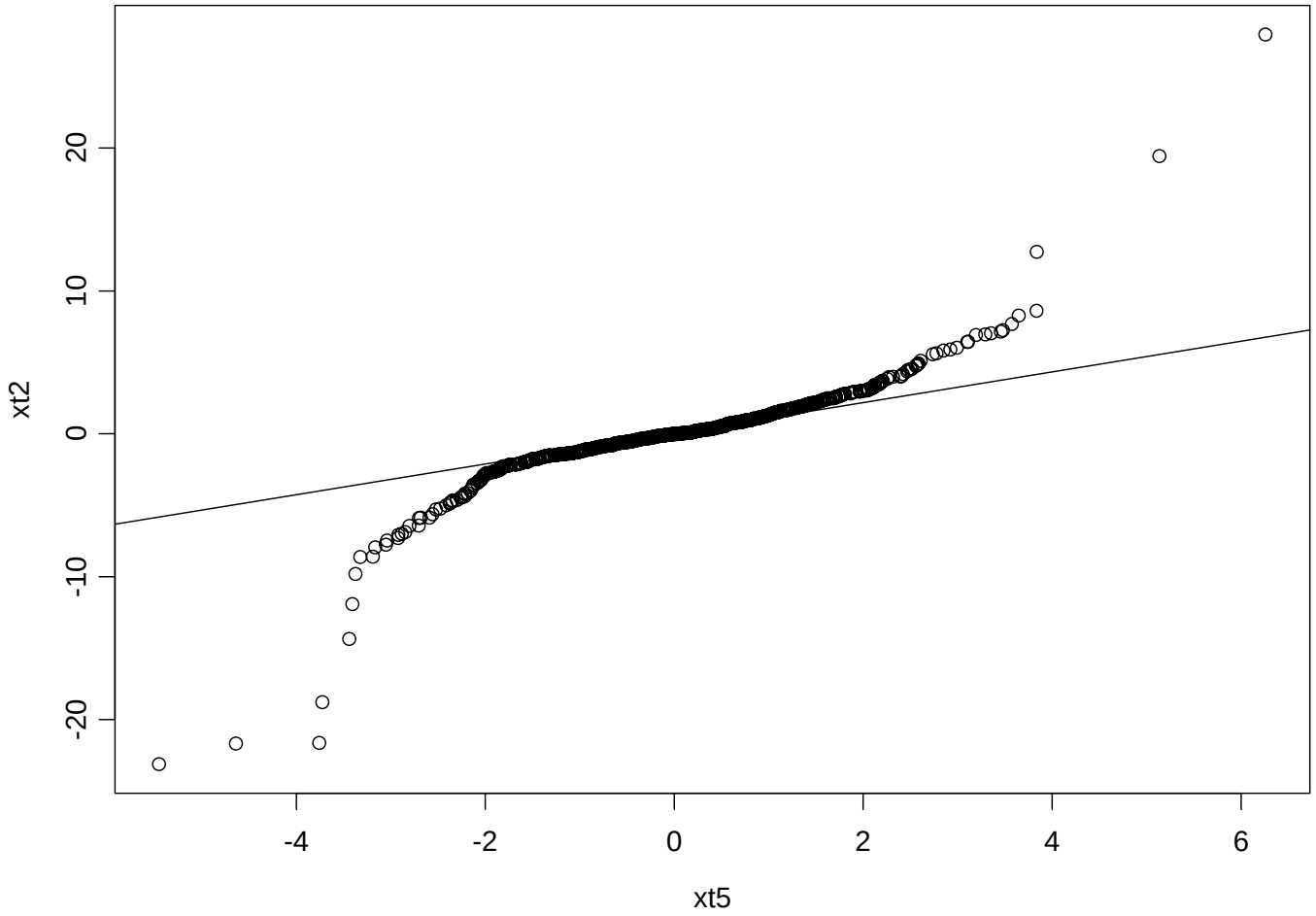

N(0,1) vs t(2)
xt2
xnorm
10
5
0
-5
-10
-2
0
2


N(0,1) vs t(5)
xt5
xnorm
10
5
0
-5
-10
-2
0
2

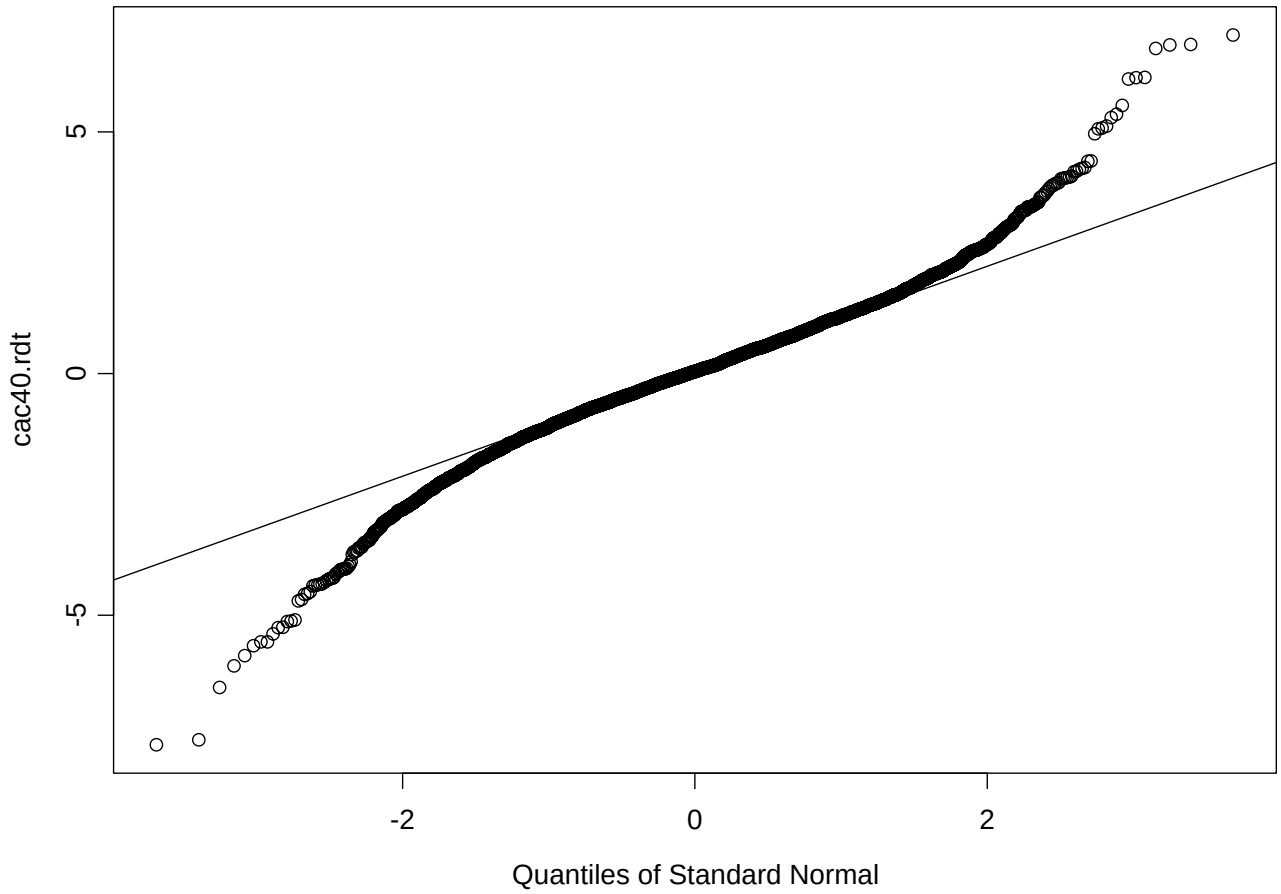

N(0,1) vs t(20)
xt20
xnorm
10
5
0
-5
-10
-2
0
2


N(0,1) vs t(50)
xt50
xnorm
10
5
0
-5
-10
-2
0
2

# Exemple : t(5) vs t(2)



# Exemple : Rendements du CAC 40



# Exemple : Rendements du CAC 40