Analyse de données expérimentales

Jean-Marc.Vincent@imag.fr

ID-IMAG Laboratory

http://www-id.imag.fr/ jvincent

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

INP Grenoble

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

GRENOBLE 1 UNIVERSITÉ JOSEPH FOURIER SCIENCES,TECHNOLOGIE,MÉDECINE

Analyse de données expérimentales

1. Visualisation de l'échantillon
2. Analyse de la tendance centrale
3. Analyse de la variabilité
4. Estimation

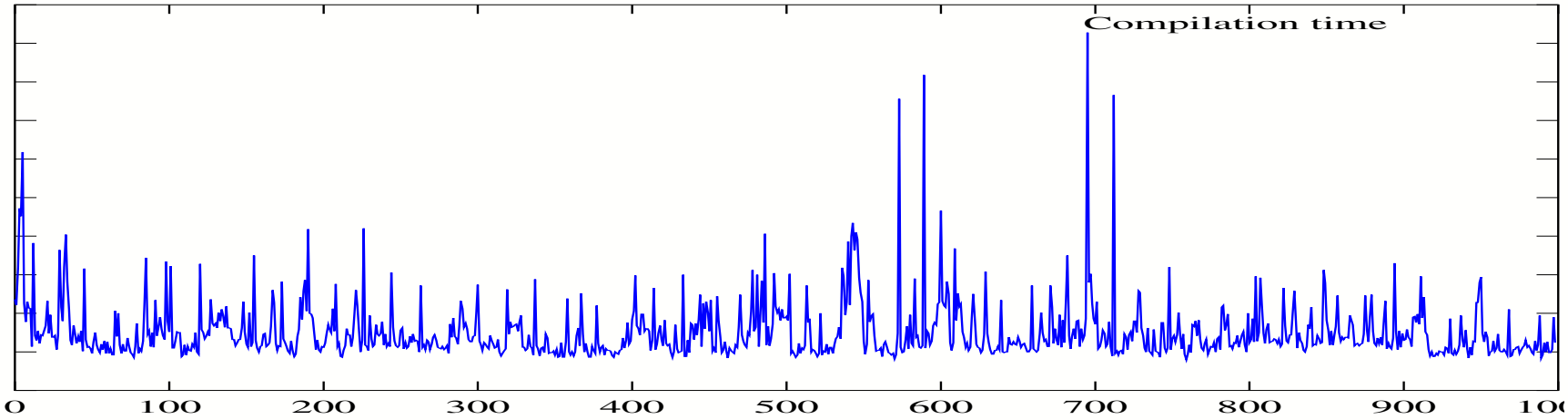
Référence:

The Art of Computer Systems Performance Analysis : Techniques for Experimental Design, Measurment, Simulation and Modeling.

Raj Jain *Wiley 1991 (nouvelles versions)*

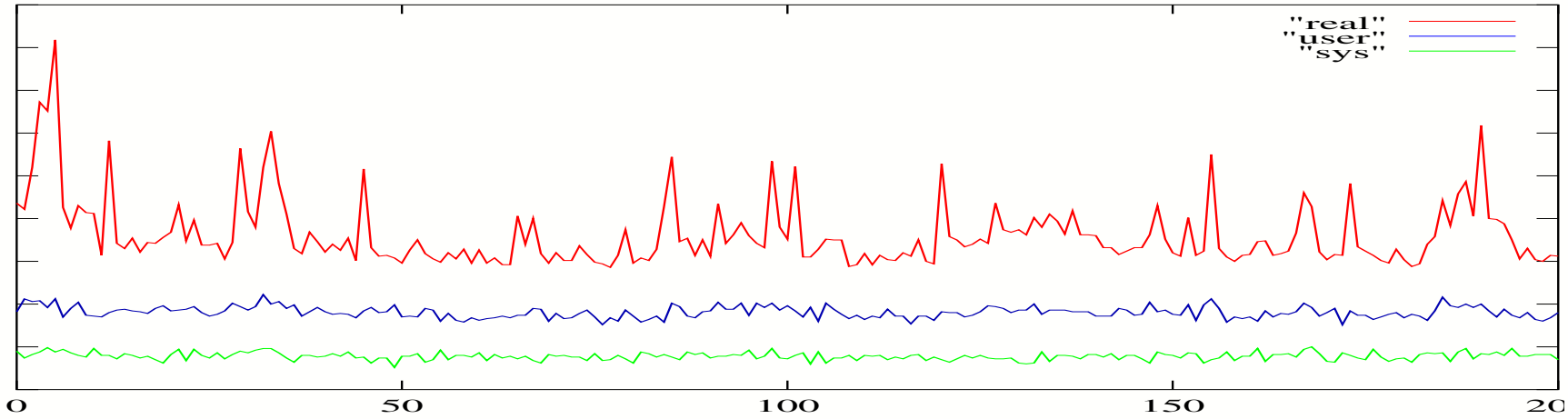
Ex : mesure de temps de compilation

Fichier de données brutes (obtenues par `time`)
Taille de l'échantillon : 200
Structure : `real sys user`



Problème: Analyser la variabilité du temps d'exécution de la tâche en vue de son exportation.

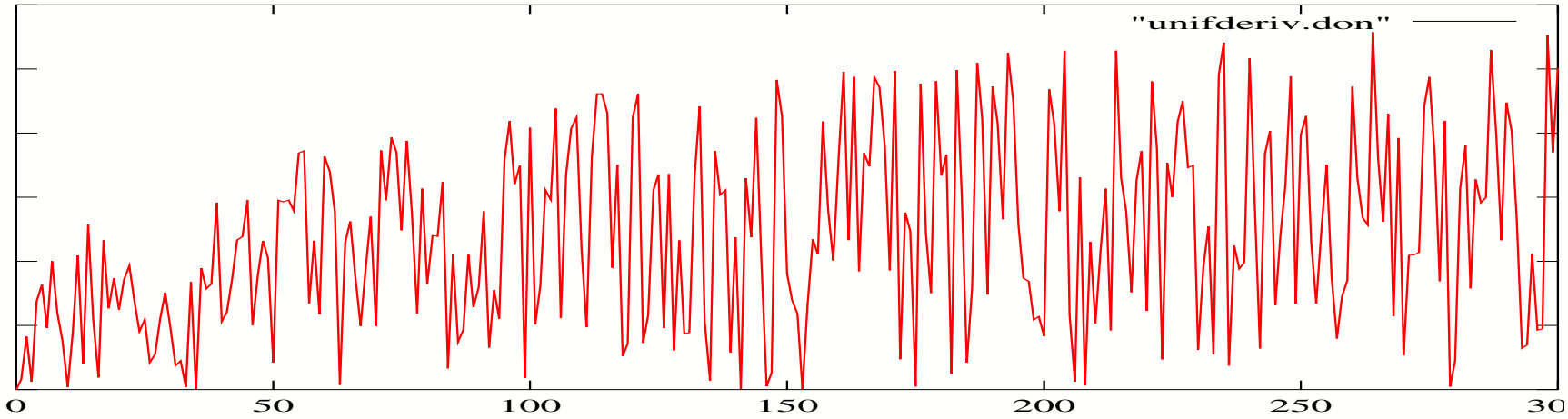
Ex : mesure de temps de compilation (2)



Interprétation :

- expérimentation semble stable
- perturbation de l'environnement (facteur 5)
- exportation possible si tcom«0.5

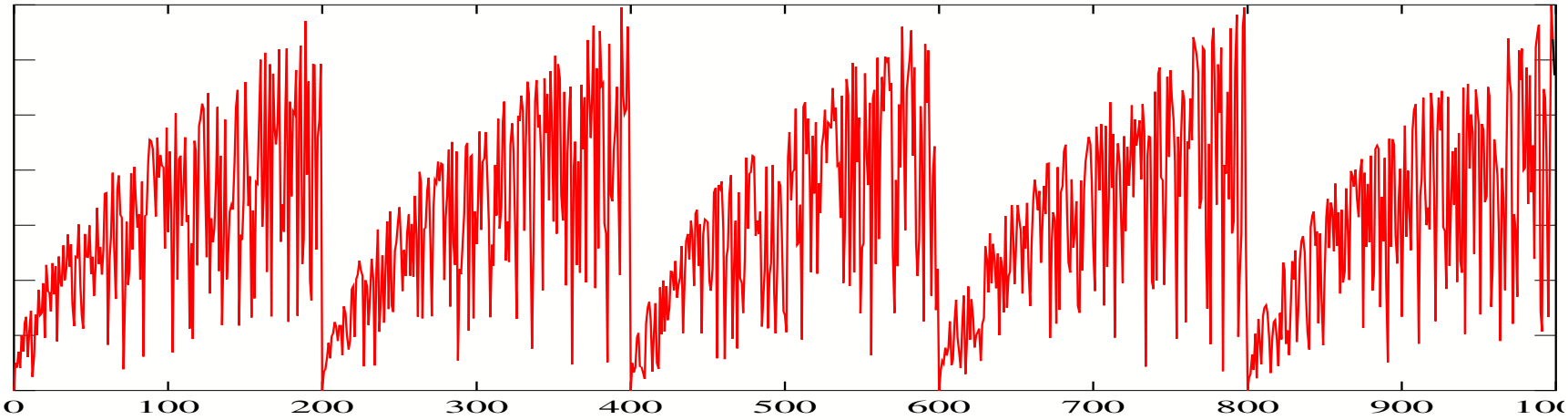
Détection de tendance



$\Rightarrow$ Modéliser la tendance : (ici croissance et saturation)

$\Rightarrow$ éliminer la tendance par compensation (ici facteur multiplicatif)

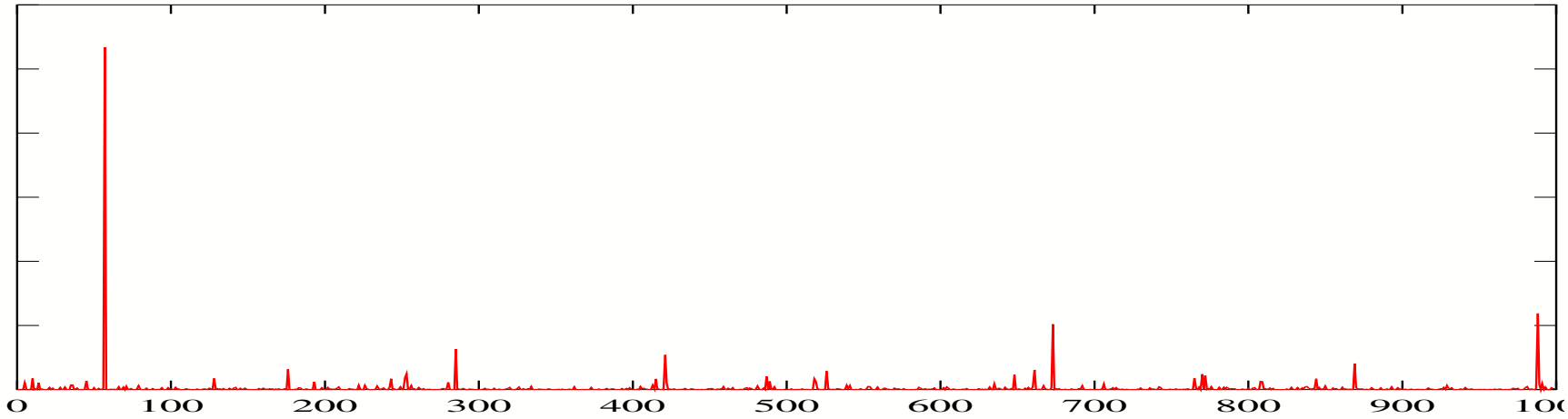
Détection de périodicité



Danger : dépendance de l'échantillonage ou de l'horloge
$\Rightarrow$ analyse de la période (Fourier),
$\Rightarrow$ filtrage temporel

Danger: largeur de fenêtre (analyse en ondelette)

Mesures aberrantes

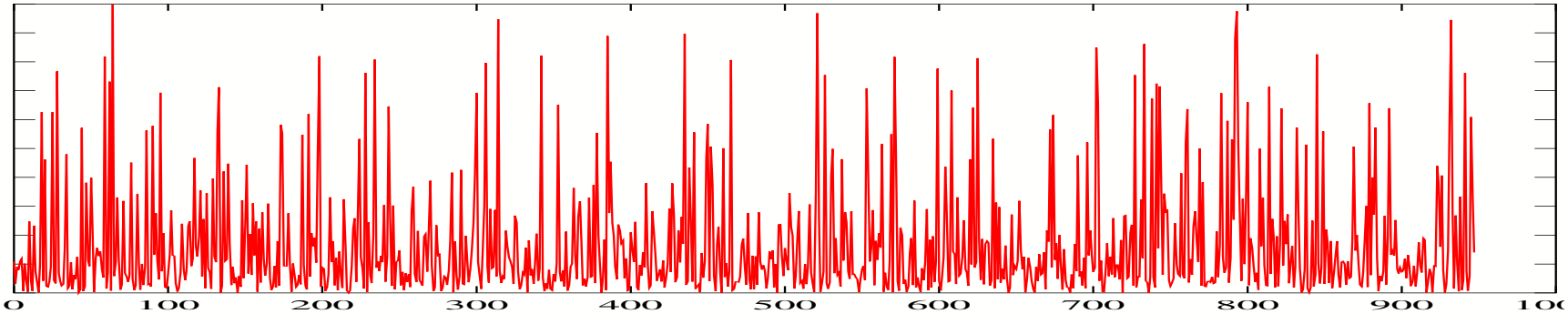


événements rares : interprétation
$\Rightarrow$ seuillage par valeur
Danger : valeur de seuil : fournir le taux de rejet des valeurs
$\Rightarrow$ seuillage par quantile

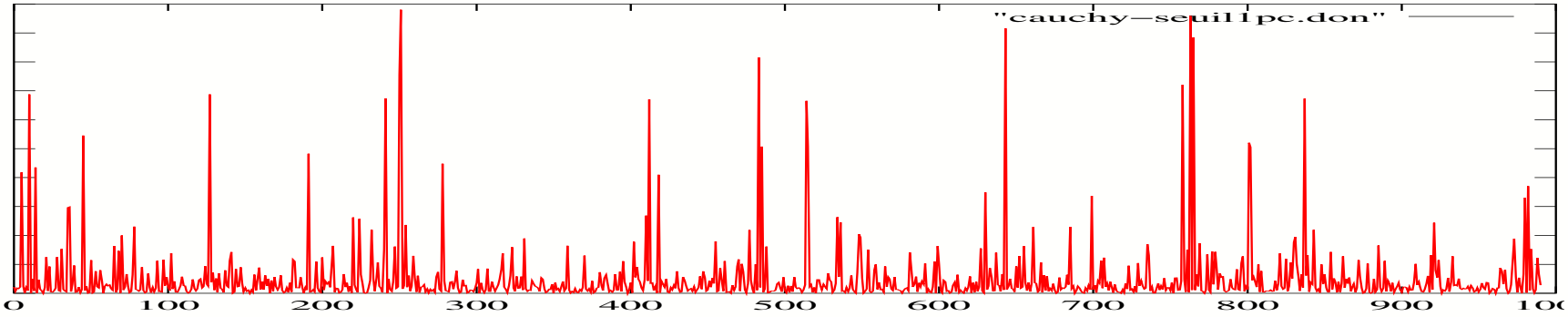
Danger : fixer le pourcentage maximum d'élimination : fournir la valeur minimale de rejet

Seuillage

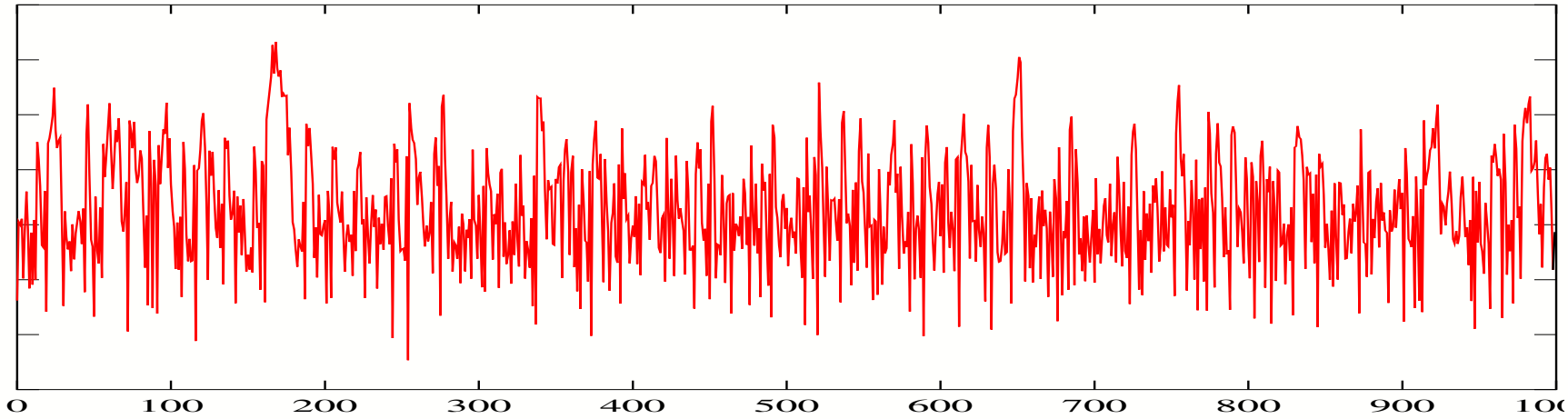
Seuillage à la valeur 10 : rejet 5%



Seuillage à 1% : valeur 50



Détection de dépendance temporelle

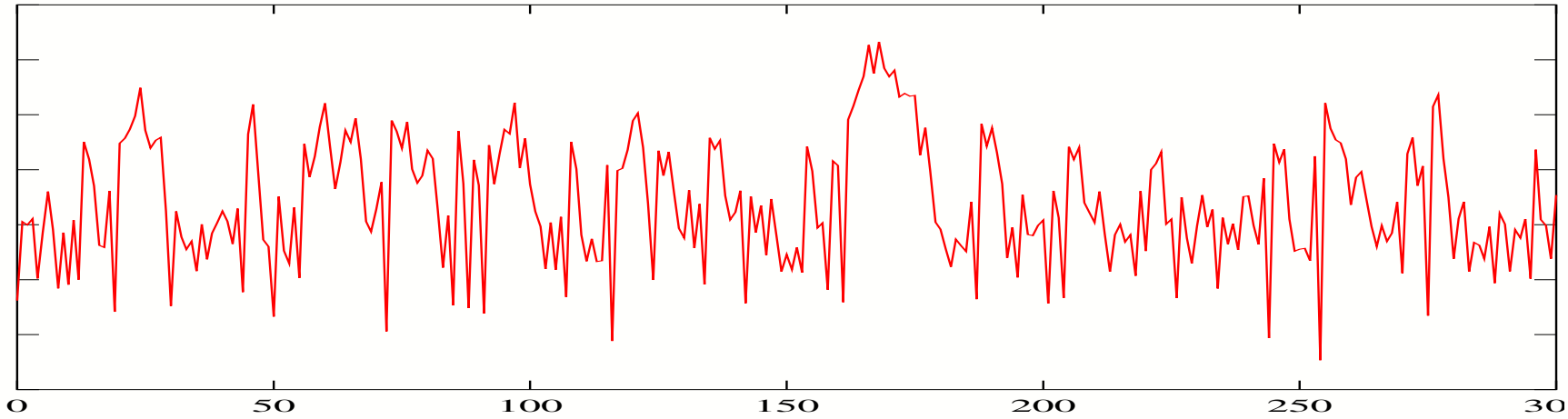


Semble correct, expériences indépendantes

expériences indépendantes;

expériences statistiquement identiques

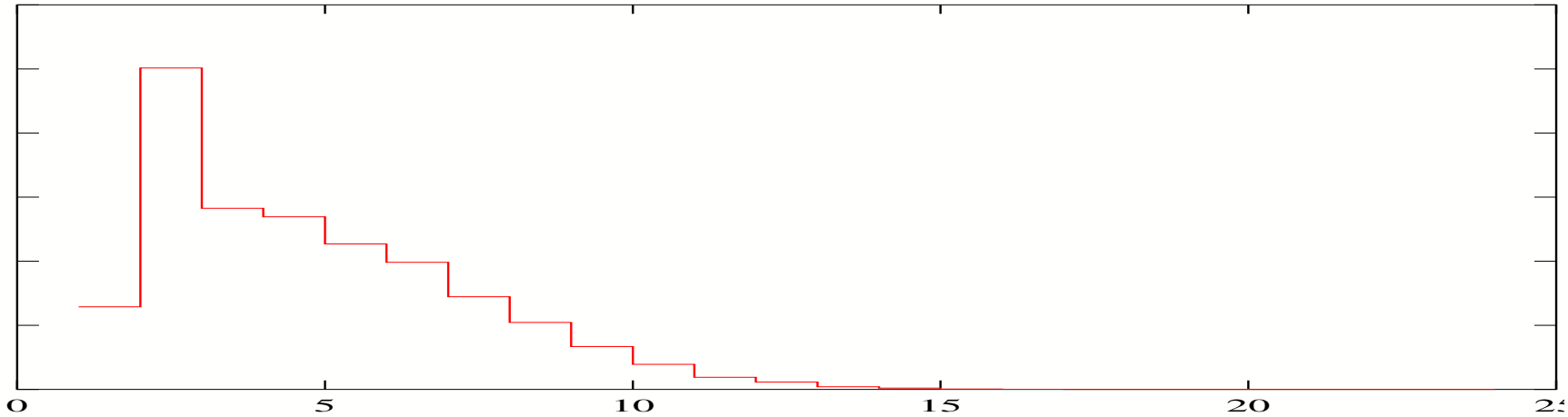
Détection de tendance



Danger: Phénomène de corrélation temporelle

$\Rightarrow$ étudier la “stationnarité”

Représentation de la distribution



Forme de l'histogramme :

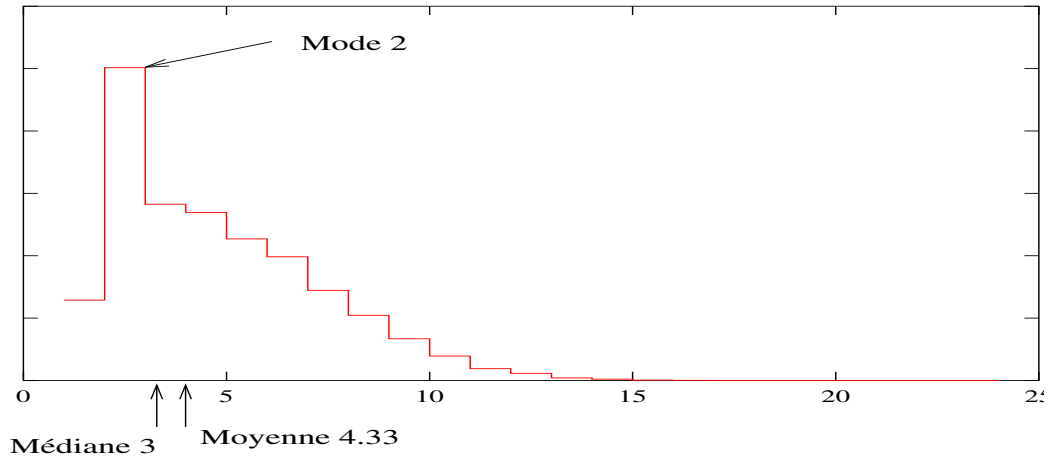
$\Rightarrow$ uni/multi-modal

$\Rightarrow$ symmétrique ou non $\rightarrow$ skewness

$\Rightarrow$ aplati ou non $\rightarrow$ kurtosis

$\Longrightarrow$ représentation par la **TENDANCE CENTRALE**

Tendance centrale : mode



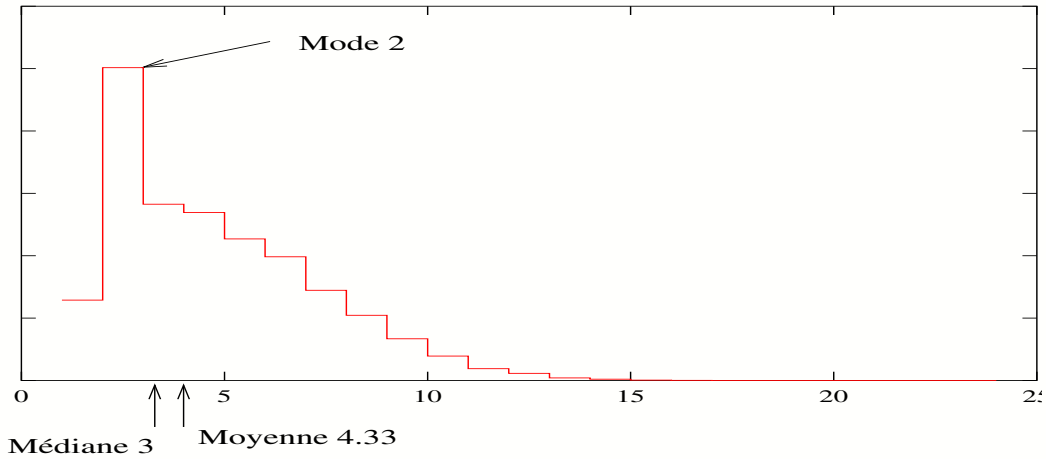
$\longrightarrow$ valeur la plus fréquente $\Rightarrow$ variables qualitatives (dépend du pas de l'histogramme)

$\longrightarrow$ maximise la probabilité de succès

$\longrightarrow$ très instable et pas de sens pour les lois aplaties

Ressource critique, goulot d'étranglement

Tendance centrale : médiane

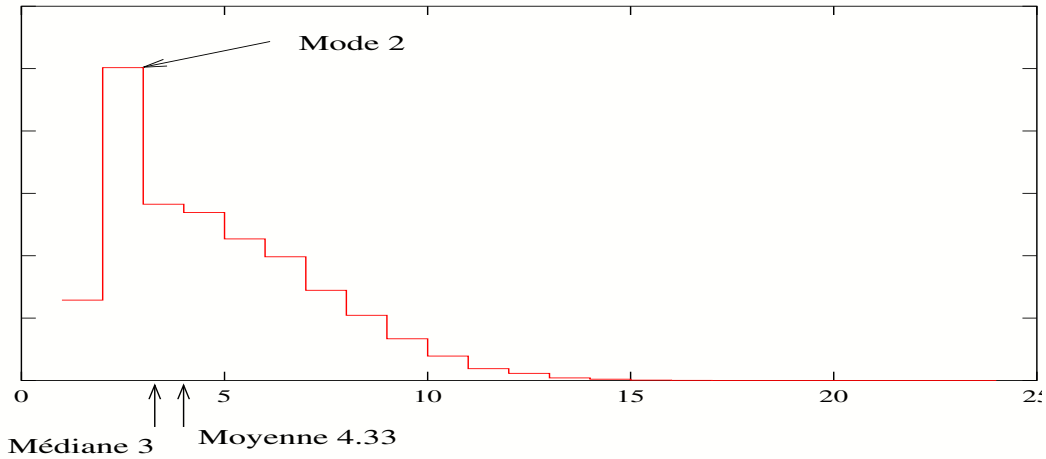


$\longrightarrow$ sépare l'échantillon en 2 $\Rightarrow$ variables ordonnées

$\longrightarrow$ instable mais se combine avec les quantiles

temps d'attente, charge

Tendance centrale : moyenne



$\longrightarrow$ notion de coût moyen $\Rightarrow$ variables dans un espace vectoriel

$\longrightarrow$ la somme a du **sens**

ex : gain moyen $\longrightarrow$ moyenne géométrique

temps d'interarrivée, coût

Tendance centrale : calcul

Mode : choix du pas de l'histogramme + maximum $O(n)$
"off-line"
Médiane : tri des valeurs $O(nlog(n))$
Moyenne : somme des valeurs $O(n)$ "on-line"

$\longrightarrow$ EXPLIQUER LA VARIABILITE AUTOUR DE LA TENDANCE CENTRALE

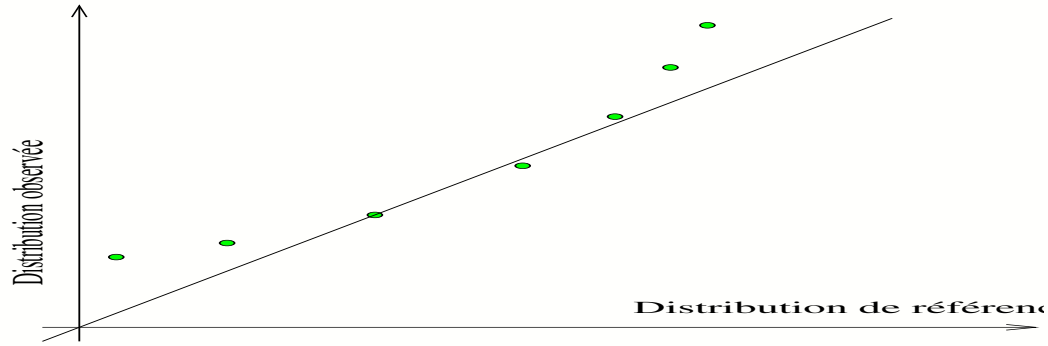
Description de la variabilité

Quantiles : quartiles, déciles, etc $\longrightarrow$ Ordonner l'échantillon :

$$(x_1, x_2, \cdots, x_n) \longrightarrow (x_{(1)}, x_{(2)}, \cdots, x_{(n)});$$

$$Q_1 = x_{(n/4)};\ Q_2 = x_{(n/2)} = médiane;\ Q_3 = x_{(3n/4)}.$$

Comparaison avec des distributions de références : diagramme quantile/quantile



Première idée de test

Variance

Ecart quadratrique à la moyenne

$$Var(X) = \frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})^2.$$

Propriétés:

$$\begin{aligned} Var(X) &= \overline{x^2} - (\bar{x})^2, \text{ où } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \\ Var(X + cste) &= Var(X); \\ Var(\lambda X) &= \lambda^2 Var(X). \end{aligned}$$

Synthèse : analyse de données expérimentales

1. Spécification des quantités à observer (nature), fixer les paramètres expérimentaux
2. Collecter les données
3. Contrôler "de visu" l'échantillon collecté : détection de corrélations, tendances, éléments aberrants
4. Déterminer les tendances centrales pertinentes
5. Analyser la dispersion autour des tendances centrales
6. Tenter d'expliquer le résultat observé