# Data Statistics Introduction

## on the nature of observations

Arnaud Legrand, Lucas Mello Schnorr, Jean-Marc Vincent

Master MOSIG
Grenoble


# DATA STATISTICS INTRODUCTION



UGA IM²AG UFR Université Grenoble Alpes

# DATA PRODUCTION

## First question : Why this dataset has been produced ? (purpose)

- Who organized the study ?
- What was the question to be answered by the statistical analysis ?
- Who will be the target of the analysis ?

# DATA PRODUCTION

## First question : Why this dataset has been produced ? (purpose)

- Who organized the study ?
- What was the question to be answered by the statistical analysis ?
- Who will be the target of the analysis ?

## Second question : Which approach has been used ? (method)

- Exhaustive collected information
- Designed survey on a population
- Designed Experiments

# DATA PRODUCTION

## First question : Why this dataset has been produced ? (purpose)

- Who organized the study ?
- What was the question to be answered by the statistical analysis ?
- Who will be the target of the analysis ?

## Second question : Which approach has been used ? (method)

- Exhaustive collected information
- Designed survey on a population
- Designed Experiments

## Third question : How this dataset has been practically produced ? (observations)

- Nature of the items in the Data set
- Characterization of data
- Semantic of Data

**Take time to analyse the production process**

# ANALYSIS OF THE SET OF VARIABLES

## Identification of the variables types

- Type of the variables (numbers, identifiers, ...)
- Set of values taken by the variables (bounds, sets,...)
- Properties of the variables (positive,...)

# ANALYSIS OF THE SET OF VARIABLES

## Identification of the variables types

- Type of the variables (numbers, identifiers, ...)
- Set of values taken by the variables (bounds, sets,...)
- Properties of the variables (positive,...)

## Identification of the variables role

- When these variables has been collected ?
- Why these variables have been chosen ?

# ANALYSIS OF THE SET OF VARIABLES

## Identification of the variables types

- Type of the variables (numbers, identifiers, ...)
- Set of values taken by the variables (bounds, sets,...)
- Properties of the variables (positive,...)

## Identification of the variables role

- When these variables has been collected ?
- Why these variables have been chosen ?

## Identification of the variables semantic

- What is the interpretation of the variables values ? (size, weight, ...)
- What are the relations between variables (structure) ?

**Take time to build a serious metadata document**

UGA UFR IM²AG Université Grenoble Alpes

# ANALYSIS OF THE TYPE OF VARIABLES

## Nominal Variables : classification, membership (qualitative)

- Values in an unstructured set
- Examples : color, gender, ...
- Methods : grouping
- Operators : $=$, $\neq$

# ANALYSIS OF THE TYPE OF VARIABLES

## Nominal Variables : classification, membership (qualitative)

- Values in an unstructured set
- Examples : color, gender, ...
- Methods : grouping
- Operators : $=$, $\neq$

## Ordinal Variables : Comparison, Level (qualitative)

- Values in an ordered set
- Examples : ranking, opinion, ...
- Methods : sorting
- Operators : $\leqslant$, $\geqslant$

# ANALYSIS OF THE TYPE OF VARIABLES

## Nominal Variables : classification, membership (qualitative)

- Values in an unstructured set
- Examples : color, gender, ...
- Methods : grouping
- Operators : $=$, $\neq$

## Ordinal Variables : Comparison, Level (qualitative)

- Values in an ordered set
- Examples : ranking, opinion, ...
- Methods : sorting
- Operators : $\leqslant$, $\geqslant$

## Quantitative Variables : Quantities

- Real values (ratio is significant)
- Examples : amount, duration, cost ...
- Methods : sum, difference
- Operators : $+$, $-$, ($\times$, / )

**Take time to define precisely the variables properties**

# USAGE OF VARIABLES

## Response Variables

- Quantity asked by the question
- Examples : response time, iteration duration, ...

# USAGE OF VARIABLES

## Response Variables

- Quantity asked by the question
- Examples : response time, iteration duration, ...

## Explanatory Variables

- Variables that could affect the response variable
- Examples : size, load, ...

# USAGE OF VARIABLES

## Response Variables

- Quantity asked by the question
- Examples : response time, iteration duration, ...

## Explanatory Variables

- Variables that could affect the response variable
- Examples : size, load, ...

## Univariate or Multivariate

- Univariate : one variable is involved
- Multivariate : several variables are involved

**Take time to identify the response/explanatory variables**

# DATA STATISTICS INTRODUCTION

# DATA PRODUCTION PROCESS

## Global Process

$$\text{Question} \Longrightarrow \text{Experiment, Survey} \Longrightarrow \text{Decision}$$

**Decision = Risk**

# DATA PRODUCTION PROCESS

## Global Process

$$\text{Question} \Longrightarrow \text{Experiment, Survey} \Longrightarrow \text{Decision}$$

$$\text{Decision} = \text{Risk}$$

## Quality of Data

Specification of the Data

- Error model for the values
- Experimental/Survey bias
- Analysis limitations

**Evaluate the Quality of the Decision**

# CRITERIA FOR THE QUALITY OF DATA (FROM EUROSTAT)

## Relevance

- degree to which statistics meet current and potential needs
- could extend to varying needs

# CRITERIA FOR THE QUALITY OF DATA (FROM EUROSTAT)

## Relevance

- degree to which statistics meet current and potential needs
- could extend to varying needs

## Accuracy

- Closeness of computations or estimates to the (unknown) exact or true values
- Variability (random error) and bias (systematic error)
- Sources of errors (experimental, coverage sampling...)

# Criteria for the Quality of Data (from EuroStat)

## Relevance

- degree to which statistics meet current and potential needs
- could extend to varying needs

## Accuracy

- Closeness of computations or estimates to the (unknown) exact or true values
- Variability (random error) and bias (systematic error)
- Sources of errors (experimental, coverage sampling...)

## Timeliness

- delay between the reference point and the availability date
- trade-off against accuracy,

# CRITERIA FOR THE QUALITY OF DATA (FROM EUROSTAT)

## Comparability

- measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas, non-geographical domains, or over time

# CRITERIA FOR THE QUALITY OF DATA (FROM EUROSTAT)

## Comparability

- measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas, non-geographical domains, or over time

## Coherence

- adequacy to be reliably combined in different ways
- compatibility of measures

# CRITERIA FOR THE QUALITY OF DATA (FROM EUROSTAT)

## Comparability

- measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas, non-geographical domains, or over time

## Coherence

- adequacy to be reliably combined in different ways
- compatibility of measures

## Accessibility

- Accessibility refers to the physical conditions under which users can obtain data
- Clarity refers to the data's information environment

Extracted from *Handbook on Data Quality Assessment Methods and Tools* EuroStat Report (2013)

# Other Criteria for the Quality of Data (from Berti-Equille (2007))

## Interpretability

- availability of the supplementary information and metadata
- covers the underlying concepts

## Unicity

- one physical observation is represented by a unique object in the Dataset
- no duplicates

## Conformity to Norm

- use the standardized encoding (reals, strings, statistical variables)

## Consistency

- duplicated informations have the same value

# PRE-PROCESSING OF DATA

**Before any analysis : check the Data**

## Question on the Quality

- Are there missing values ? almost yes
- How many sampling are missing ?
- Is there a bias for missing data or randomly spread ?
- Is the bias in the dataset sufficiently important to modify the analysis (estimators, tests,...) ?

**Give potential explanations**

# PRE-PROCESSING OF DATA

**Before any analysis : check the Data**

## Question on the Quality

- Are there missing values ? almost yes
- How many sampling are missing ?
- Is there a bias for missing data or randomly spread ?
- Is the bias in the dataset sufficiently important to modify the analysis (estimators, tests,...) ?

**Give potential explanations**

## Identification of Data Problems

Model of the Dataset (types, semantic,...)

- Missing Data (none or partial value)
- Non relevant
- Duplicated

**Give potential explanations**

# PRE-PROCESSING OF DATA (2)

## Distributions of Data Problems

Analyse the position of missing values in the Dataset

- MCAR, Missing Completely at random (unpredictable missing)
- MAR, Missing at random (predictable values : model)
- MNAR, Non missing at random

# PRE-PROCESSING OF DATA (2)

## Distributions of Data Problems

Analyse the position of missing values in the Dataset

- MCAR, Missing Completely at random (unpredictable missing)
- MAR, Missing at random (predictable values : model)
- MNAR, Non missing at random

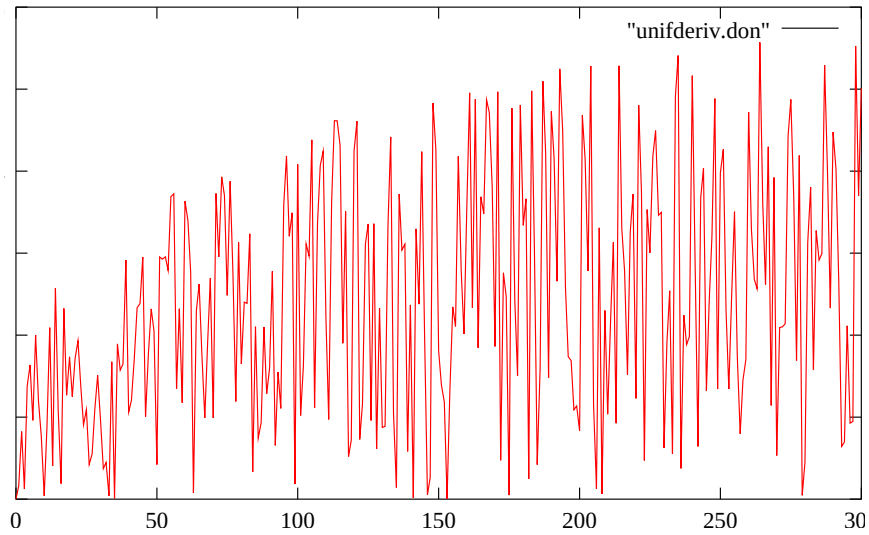
## Processing Missing Data

- Do nothing
- Remove samples with missing values
- Weighted analysis
- Value imputation (central tendency, EM, regression, random hot deck, neighbouring,...)

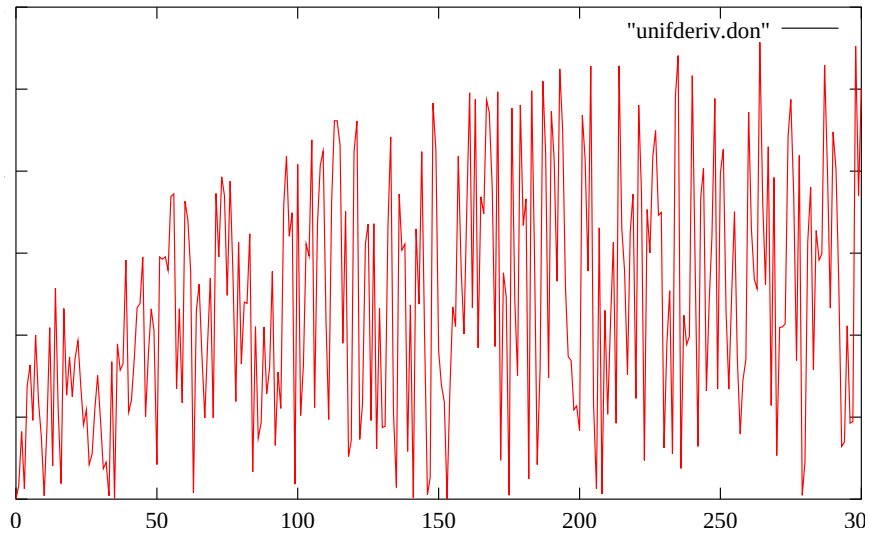
**Report the method that has been used**

# DATA STATISTICS INTRODUCTION

# CONTROL OF EXPERIMENTS (1)



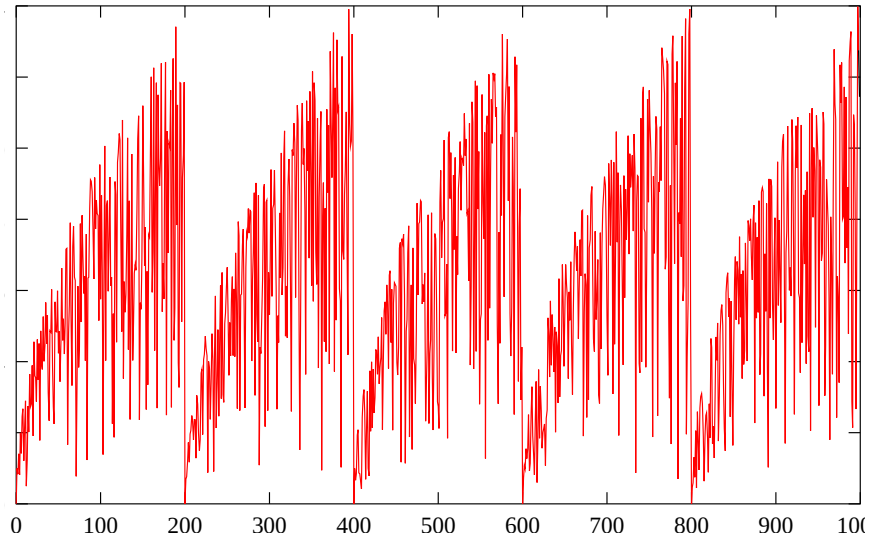
# CONTROL OF EXPERIMENTS (1)



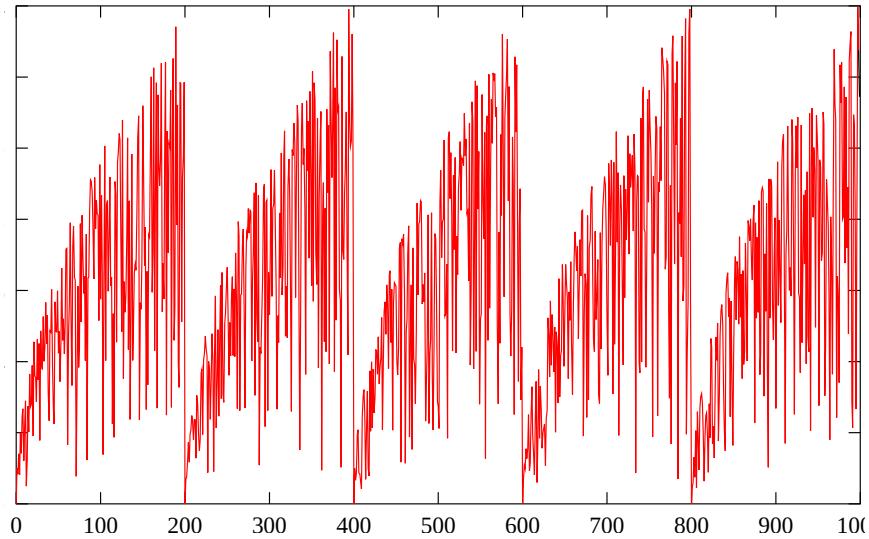

**Tendency analysis**

**non homogeneous experiment**
$\Rightarrow$ model the evolution of experiment
estimate and compensate tendency
**explain why**

# CONTROL OF EXPERIMENTS (2)



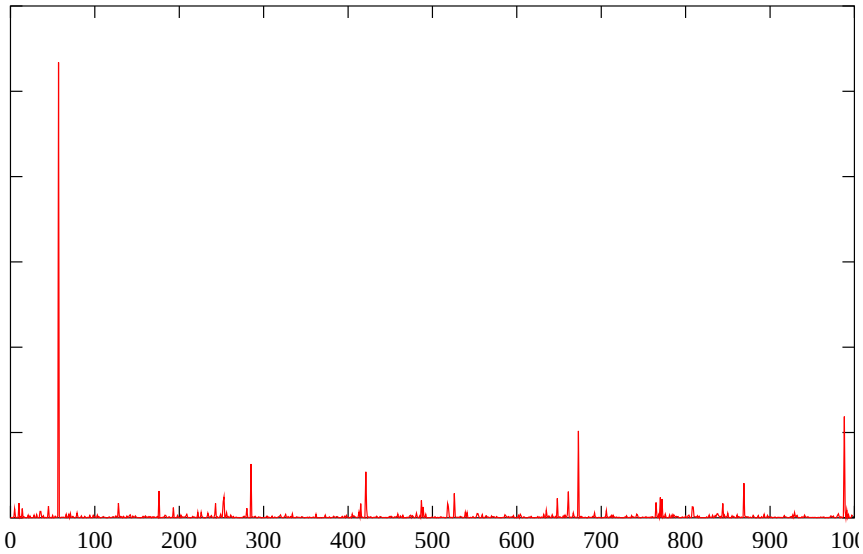
# CONTROL OF EXPERIMENTS (2)



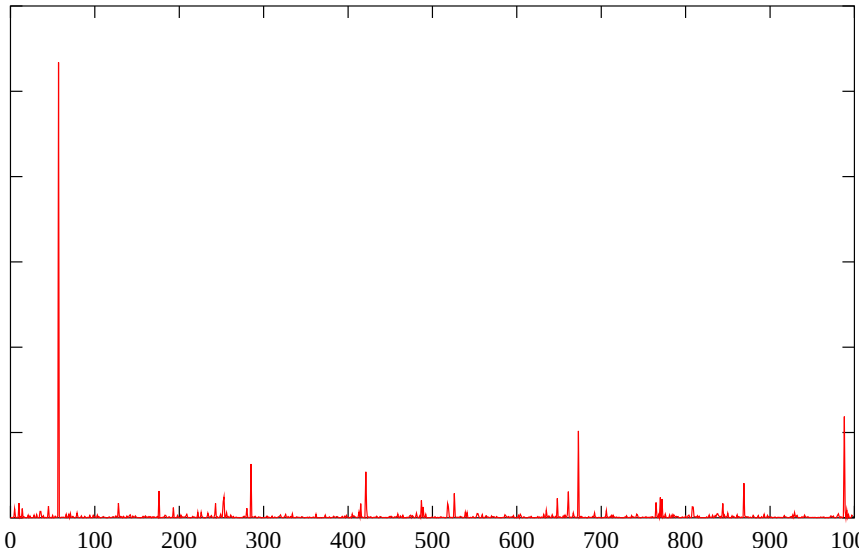

**Periodicity analysis**

**periodic evolution of the experimental environment ?**
$\Rightarrow$ model the evolution of experiment
Fourier analysis of the sample
Integration on time (sliding window analysis) Danger : size of the window
Wavelet analysis
**explain why**

# CONTROL OF EXPERIMENTS (3)



# CONTROL OF EXPERIMENTS (3)

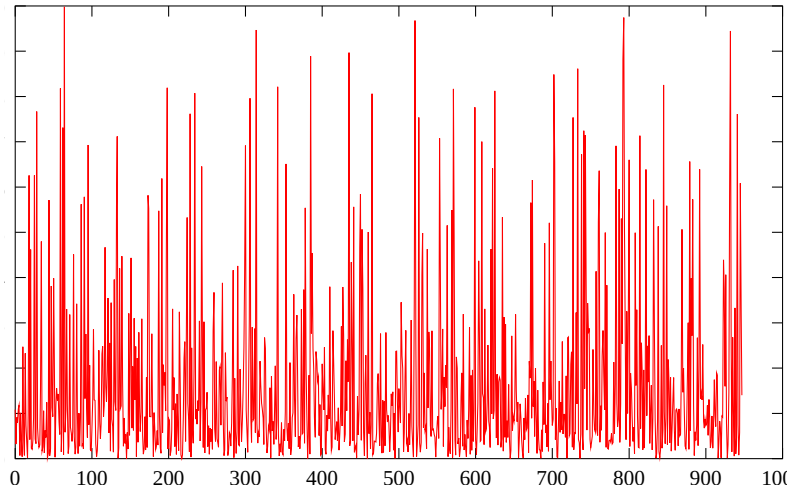



## Non significant values

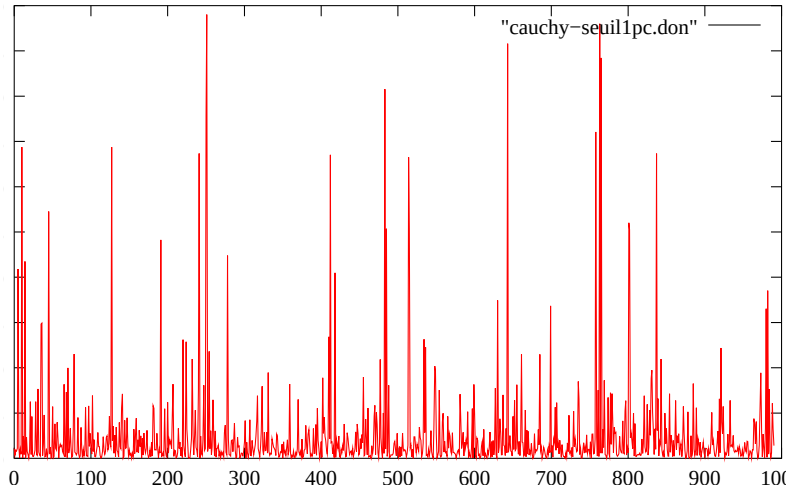
**extraordinary behaviour of experimental environment**
rare events with different orders of magnitude
$\Rightarrow$ threshold by value
Danger : choice of the threshold : indicate the rejection rate
$\Rightarrow$ threshold by quantile
Danger : choice of the percentage : indicate the rejection value
**explain why**

# CONTROL OF EXPERIMENTS (4)

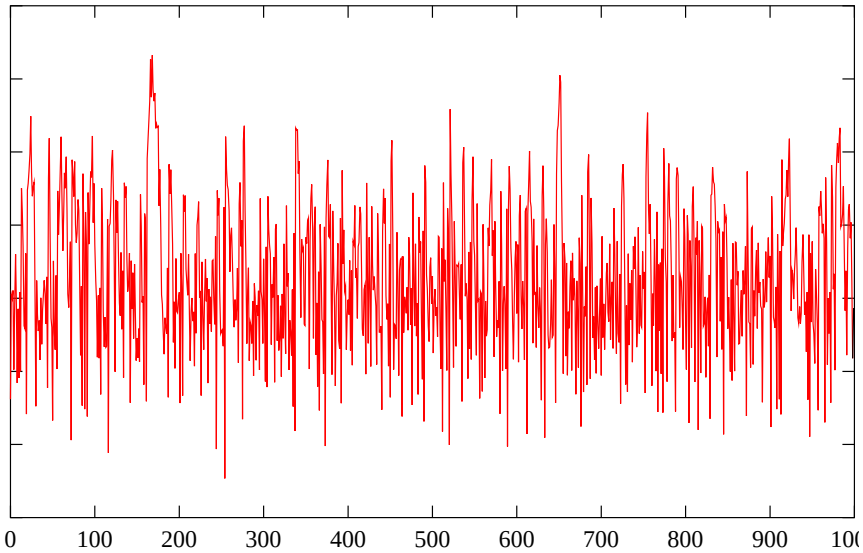
Threshold value : 10



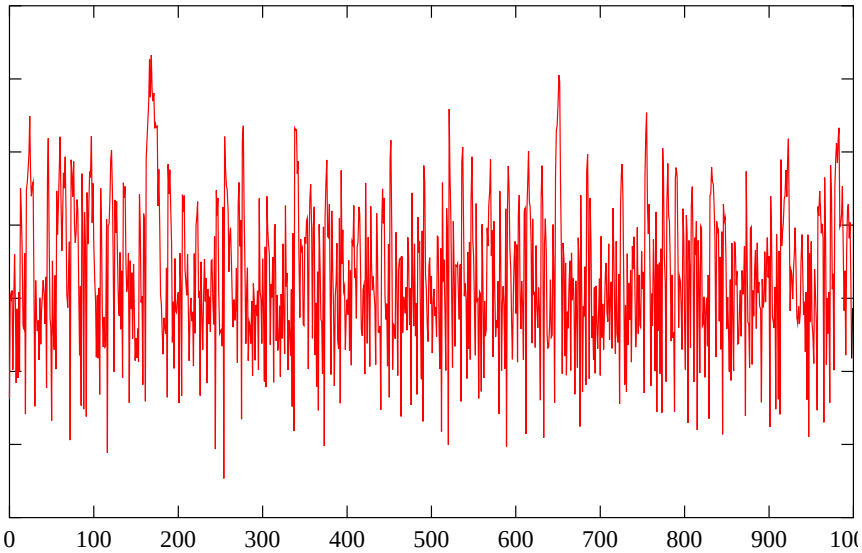

Threshold percentage : 1%



# CONTROL OF EXPERIMENTS (5)



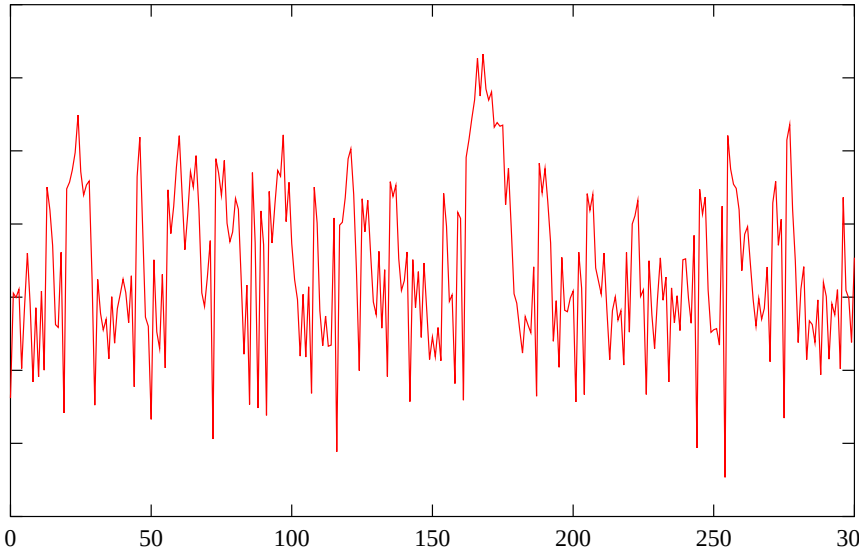
# CONTROL OF EXPERIMENTS (5)




**looks like correct experiments**
Statistically independent
Statistically homogeneous

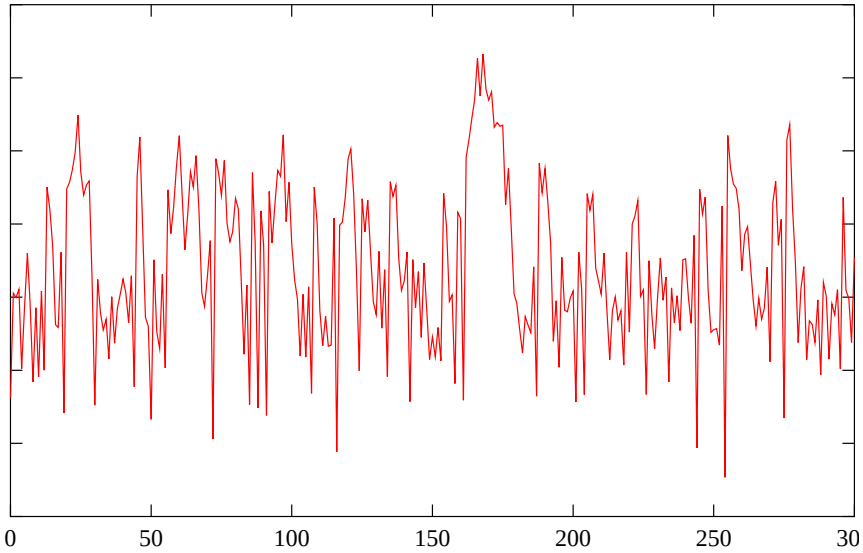
# CONTROL OF EXPERIMENTS (5BIS)

Zooming



# CONTROL OF EXPERIMENTS (5BIS)

Zooming




**Autocorrelation**

Danger time correlation among samples
**experiments impact on experiments**
$\Rightarrow$ stationarity analysis
autocorrelation estimation (ARMA)

# EXPERIMENTAL RESULTS

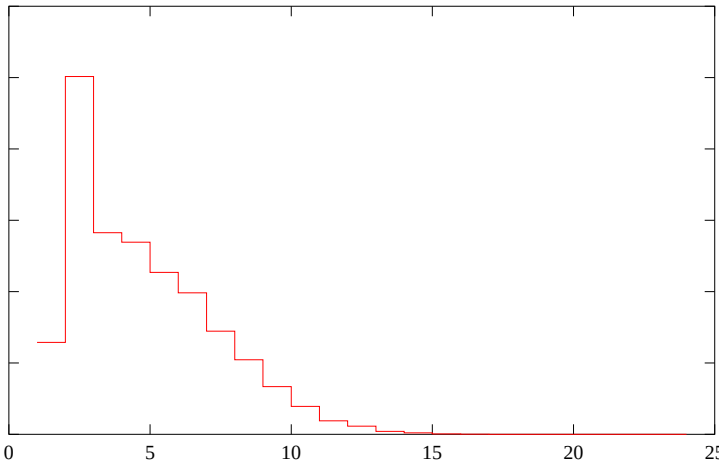
- Deterministic (controlled error non significant (white noise))
- Statistic (the system is non deterministic)

## Sample analysis

- Identification of the response set
- Structure of the response set (measure)

# DISTRIBUTION ANALYSIS

Summarize data in a **histogram**

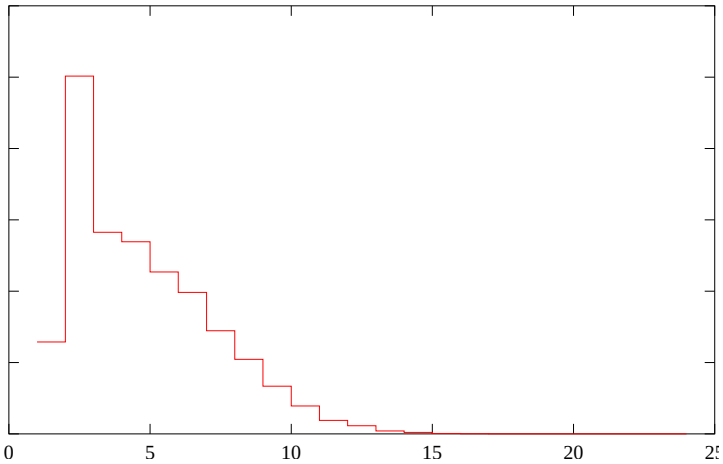



## Shape analysis

- unimodal / multimodal
- variability
- symmetric / dissymmetric (skewness)
- flatness (kurtosis)

$\Longrightarrow$ **Central tendency analysis**
$\Longrightarrow$ **Variability analysis around the central tendency**

# MODE VALUE




## Mode

- **Categorical data**
- Most frequent value
- highly unstable value
- for continuous value distribution depends on the histogram step
- interpretation depends on the flatness of the histogram

$\Longrightarrow$ **Use it carefully**
$\Longrightarrow$ **Predictor function**

# MEDIAN VALUE

## Median

- **Ordered data**
- Split the sample in two equal parts

$$\sum_{i \leqslant Median} f_i \leqslant \frac{1}{2} \leqslant \sum_{i \leqslant Median+1} f_i .$$

- more stable value
- does not depends on the histogram step
- difficult to combine (two samples)

$\Longrightarrow$ **Randomized algorithms**

# MEAN VALUE

## Mean

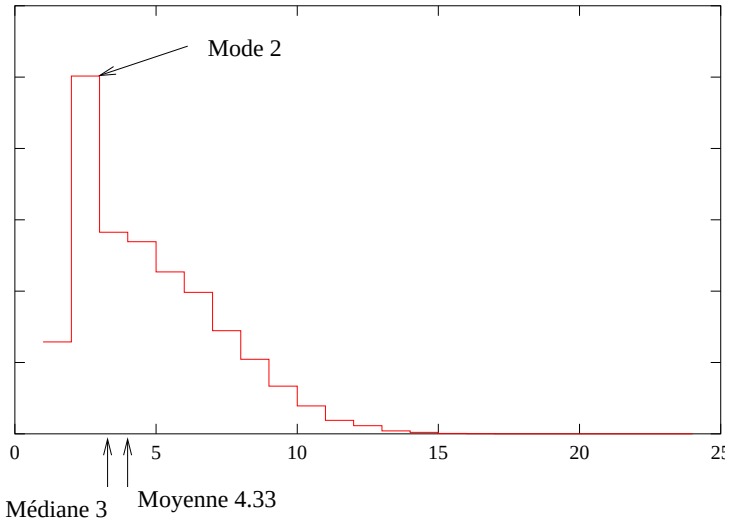
- **Vector space**
- Average of values

$$Mean = \frac{1}{Sample\_Size} \sum x_i = \sum_x x.f_x.$$

- stable value
- does not depends on the histogram step
- easy to combine (two samples $\Rightarrow$ weighted mean)

$\Longrightarrow$ **Additive problems (cost, durations, length,...)**

# CENTRAL TENDENCY




## Complementarity

- Valid if the sample is "Well-formed"
- **Semantic of the observation**
- Goal of analysis

$\Longrightarrow$ **Additive problems (cost, durations, length,...)**

# CENTRAL TENDENCY (2)

## Summary of Means

- Avoid means if possible
  Loses information
- Arithmetic mean
  When sum of raw values has physical meaning
  Use for summarizing times (not rates)
- Harmonic mean
  Use for summarizing rates (not times)
- Geometric mean
  Not useful when time is best measure of perf
  Useful when multiplicative effects are in play

# VARIABILITY

## Categorical data (finite set)

$f_i$ : empirical frequency of element $i$
Empirical entropy

$$H(f) = \sum_i f_i \log f_i.$$

Measure the empirical distance with the uniform distribution

- $H(f) \geqslant 0$
- $H(f) = 0$ iff the observations are reduced to a unique value
- $H(f)$ is maximal for the uniform distribution

# VARIABILITY (2)

**Ordered data**

Quantiles : quartiles, deciles, etc
Sort the sample :

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

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

For deciles

$$d_i = argmax_i\{\sum_{j \leqslant i} f_j \leqslant \frac{i}{10}\}.$$

Utilization as quantile/quantile plots to compare distributions

# VARIABILITY (3)

## Vectorial data

Quadratic error for the mean

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

## Properties :

$$\begin{aligned} Var(X) &\geqslant 0; \\ Var(X) &= \overline{x^2} - (\bar{x})^2, \text{ où } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \\ Var(X + cste) &= Var(X); \\ Var(\lambda X) &= \lambda^2 Var(X). \end{aligned}$$