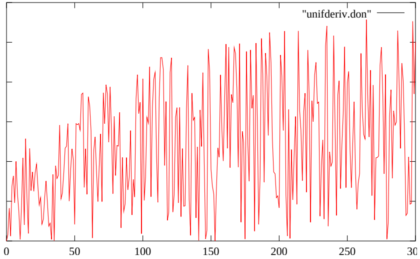
Sample Quality

descriptive analysis of data

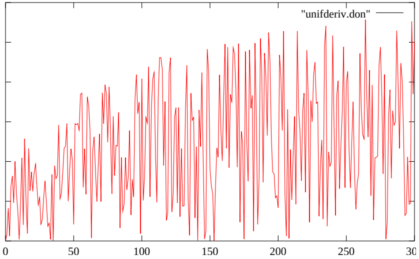
Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS
Porto Alegre, Brazil – October 2018

Control of experiments (1)



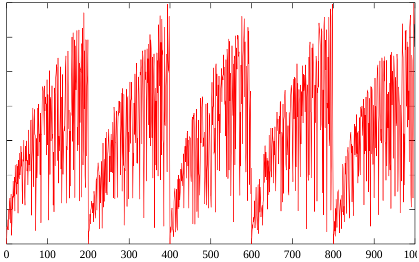
Control of experiments (1)



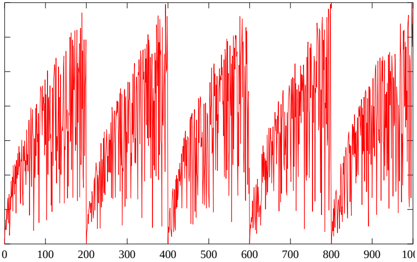
Tendency analysis

non homogeneous experiment
⇒ model the evolution of experiment
estimate and compensate tendency
explain why

Control of experiments (2)



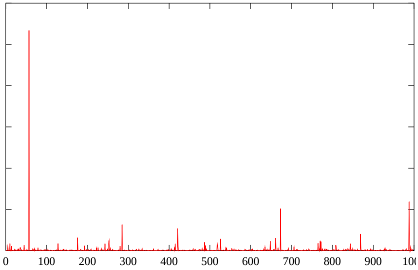
Control of experiments (2)



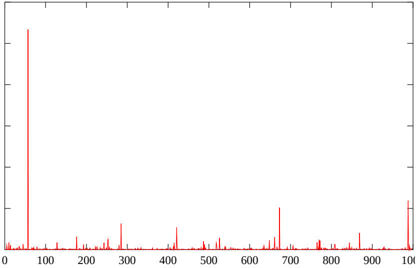
Periodicity analysis

periodic evolution of the experimental environment ?
⇒ 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)

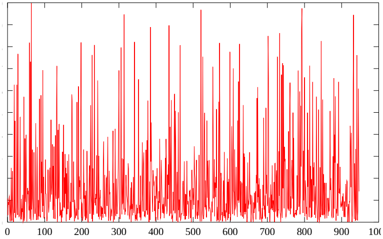


Non significant values

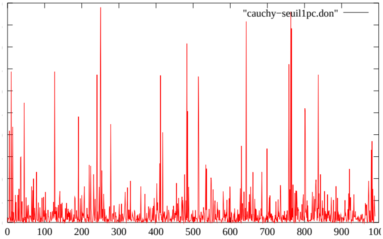
extraordinary behaviour of experimental environment
rare events with different orders of magnitude
⇒ threshold by value
Danger : choice of the threshold : indicate the rejection rate
⇒ threshold by quantile
Danger : choice of the percentage : indicate the rejection value
explain why

Control of experiments (4)

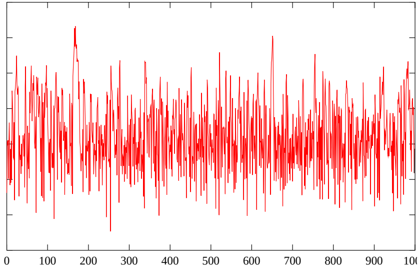
Threshold value : 10



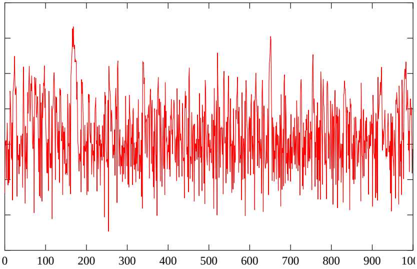
Threshold percentage : 1%



Control of experiments (5)



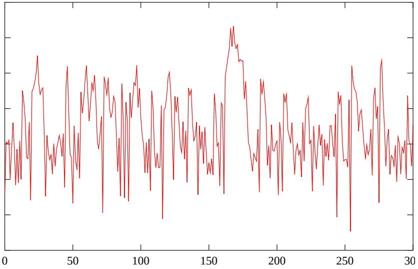
Control of experiments (5)



looks like correct experiments
Statistically independent
Statistically homogeneous

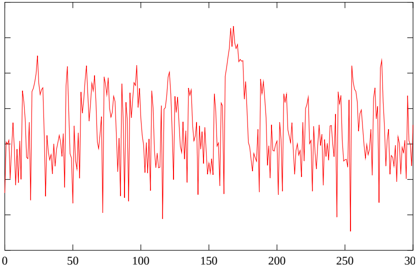
Control of experiments (5bis)

Zooming



Control of experiments (5bis)

Zooming



Autocorrelation

Danger time correlation among samples
experiments impact on experiments
⇒ stationarity analysis
autocorrelation estimation (ARMA)

Experimental results

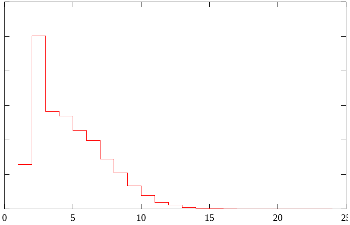
- Deterministic (controlled error non significant (white noise))
- Statistic (the system is non deterministic)

Sample analysis

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

Distribution analysis

Summarize data in a **histogram**

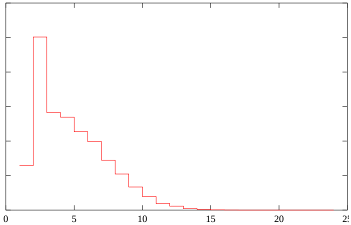


Shape analysis

- 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

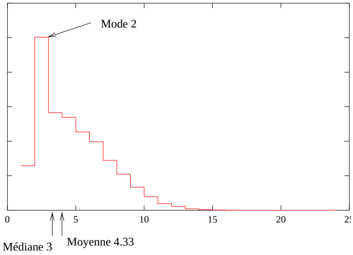
- Vector space
- Average of values

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

- 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

⟹ **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 :

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