

Sample Quality

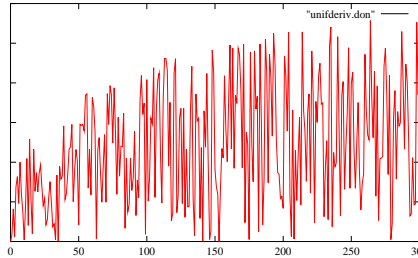
descriptive analysis of data

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS
Porto Alegre, Brazil – October 30th, 2017



CONTROL OF EXPERIMENTS (1)



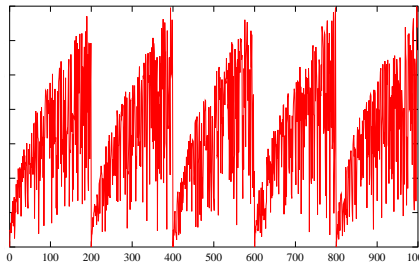
Tendency analysis

non homogeneous experiment

⇒ model the evolution of experiment
estimate and compensate tendency

explain why

CONTROL OF EXPERIMENTS (2)



Periodicity analysis

periodic evolution of the experimental environment ?

⇒ model the evolution of experiment

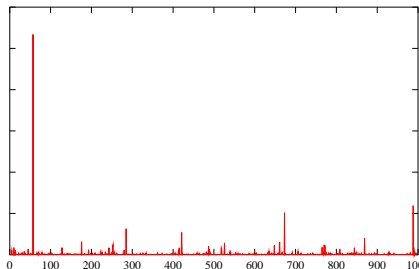
Fourier analysis of the sample

Integration on time (sliding window analysis) Danger : size of the window

Wavelet analysis

explain why

CONTROL OF EXPERIMENTS (3)



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

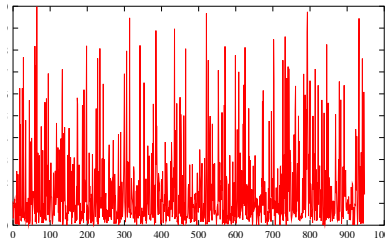
⇒ threshold by quantile

Danger : choice of the percentage : indicate the rejection value

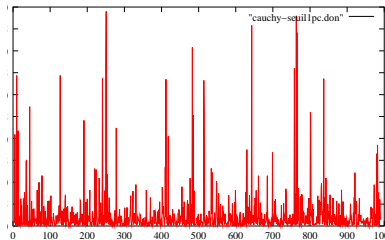
explain why

CONTROL OF EXPERIMENTS (4)

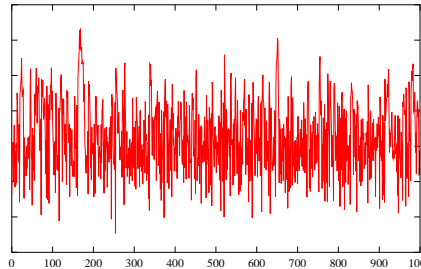
Threshold value : 10



Threshold percentage : 1%



CONTROL OF EXPERIMENTS (5)



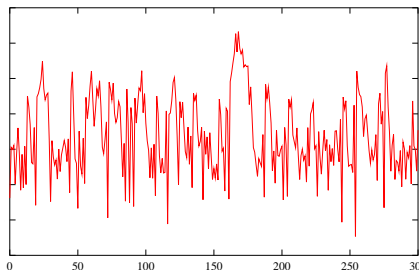
looks like correct experiments

Statistically independent

Statistically homogeneous

CONTROL OF EXPERIMENTS (5BIS)

Zooming



Autocorrelation

Danger time correlation among samples

experiments impact on experiments

⇒ stationarity analysis

autocorrelation estimation (ARMA)

EXPERIMENTAL RESULTS

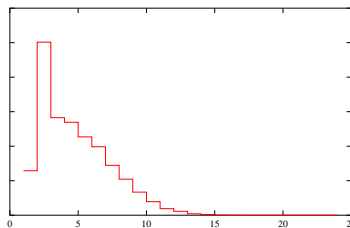
- ▶ Deterministic (controlled error non significant (white noise))
- ▶ Statistic (the system is non deterministic)

Sample analysis

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

DISTRIBUTION ANALYSIS

Summarize data in a **histogram**



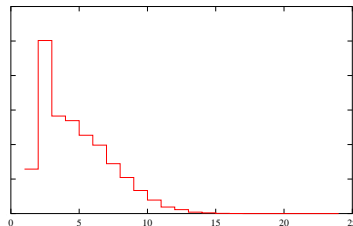
Shape analysis

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

⇒ **Central tendency analysis**

⇒ **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

⇒ **Use it carefully**

⇒ **Predictor function**

MEDIAN VALUE

Median

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

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

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

⇒ **Randomized algorithms**

MEAN VALUE

Mean

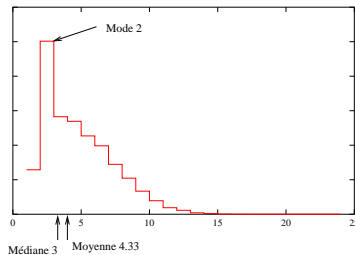
- ▶ **Vector space**
- ▶ Average of values

$$\text{Mean} = \frac{1}{\text{Sample_Size}} \sum x_i = \sum_x x \cdot f_x.$$

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

\Rightarrow **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

- ▶ $H(f) \geq 0$
- ▶ $H(f) = 0$ 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 :

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

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

For deciles

$$d_i = \operatorname{argmax}_i \left\{ \sum_{j \leq i} f_j \leq \frac{i}{10} \right\}.$$

Utilization as quantile/quantile plots to compare distributions

VARIABILITY (3)

Vectorial data

Quadratic error for the mean

$$\text{Var}(X) = \frac{1}{n} \sum_1^n (x_i - \bar{x}_n)^2.$$

Properties :

$$\text{Var}(X) \geq 0;$$

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

$$\text{Var}(X + \text{cste}) = \text{Var}(X);$$

$$\text{Var}(\lambda X) = \lambda^2 \text{Var}(X).$$