Introduction to Design of Experiments

Jean-Marc Vincent and Arnaud Legrand

Laboratory ID-IMAG
MESCAL Project
Universities of Grenoble
{Jean-Marc.Vincent,Arnaud.Legrand}@imag.fr

November 10, 2014

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

Continuous random variable

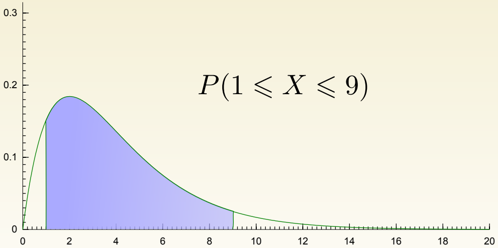
- A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement.
  Such a variable enables to model uncertainty that may result of *incomplete information* or *imprecise measurements*.
  Formally $(\Omega, \mathcal{F}, P)$ is a probability space where:
  - Ω, the sample space, is the set of all possible outcomes (e.g., $\{1, 2, 3, 4, 5, 6\}$)
  - $\mathcal{F}$ if the set of events where an event is a set containing zero or more outcomes (e.g., the event of having an odd number $\{1, 3, 5\}$)
  - The probability measure $P : \mathcal{F} \rightarrow [0, 1]$ is a function returning an event's probability.
- Since many computer science experiments are based on time measurements, we focus on continuous variables.

$$X : \Omega \rightarrow \mathbb{R}$$

A probability distribution (a.k.a. probability density function or p.d.f.) is used to describe the probabilities of different values occurring.

A random variable X has density f, where f is a non-negative and integrable function, if:

$$P[a \leqslant X \leqslant b] = \int_a^b f(x)\,\mathrm{d}x$$



Expected value

- When one speaks of the "expected price", "expected height", etc. one means the expected value of a random variable that is a price, a height, etc.

$$\begin{aligned}\mathrm{E}[X] &= x_1p_1 + x_2p_2 + \ldots + x_kp_k \\ &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\end{aligned}$$

  The expected value of X is the "average value" of X.

  It is **not** the most probable value. The mean is one aspect of the distribution of X. The median or the mode are other interesting aspects.

- The variance is a measure of how far the values of a random variable are spread out from each other.
  If a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

How to estimate Expected value ?

To empirically estimate the expected value of a random variable, one repeatedly measures observations of the variable and computes the arithmetic mean of the results.

Unfortunately, if you repeat the estimation, you may get a different value since X is a random variable ...

Central Limit Theorem

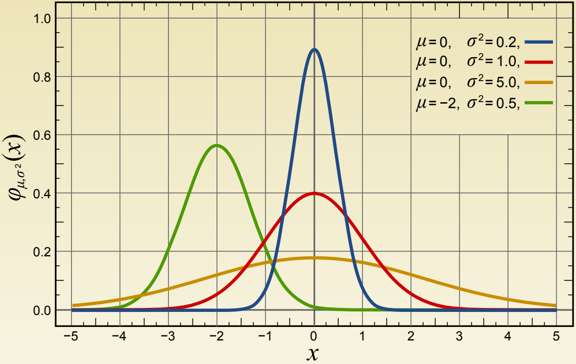
- Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n (i.e., a sequence of **independent** and **identically distributed** random variables with expected values μ and variances σ^2).
- The sample average of these random variables is:

$$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

  S_n is a random variable too.
- For large n's, the distribution of S_n is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

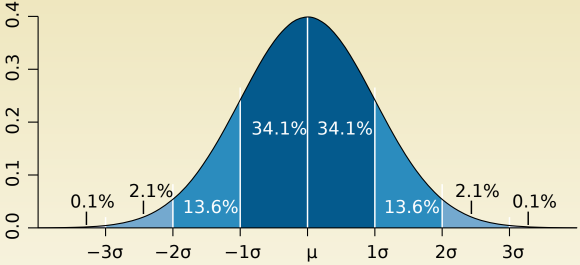
$$S_n \xrightarrow[n\to\infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The Normal Distribution



The smaller the variance the more “spiky” the distribution.

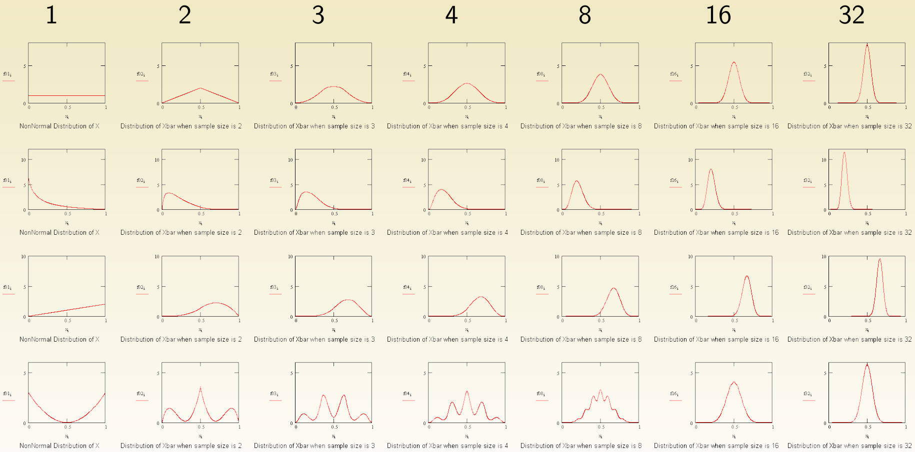
The Normal Distribution



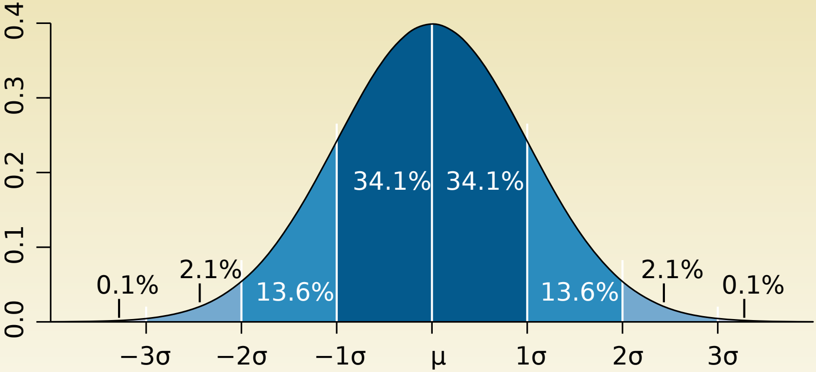
The smaller the variance the more "spiky" the distribution.

- Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set.
- Two standard deviations from the mean (medium and dark blue) account for about 95%
- Three standard deviations (light, medium, and dark blue) account for about 99.7%

Start with an arbitrary distribution and compute the distribution of S_n for increasing values of n.



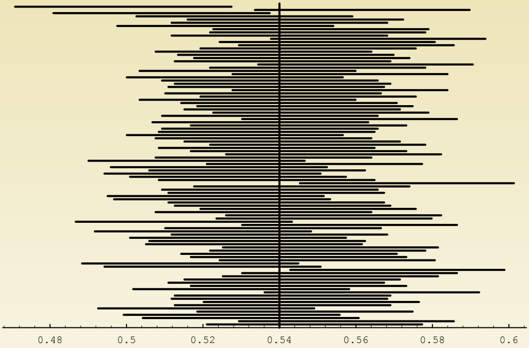
CLT consequence: confidence interval



When n is large:

$$\mathrm{P}\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = \mathrm{P}\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

CLT consequence: confidence interval



When n is large:

$$\mathrm{P}\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = \mathrm{P}\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

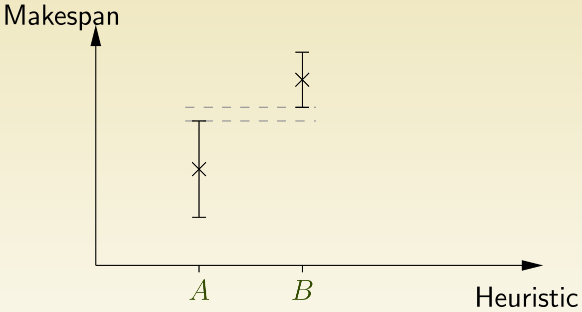
There is 95% of chance that the true mean lies within $2\frac{\sigma}{\sqrt{n}}$ of the sample mean.

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

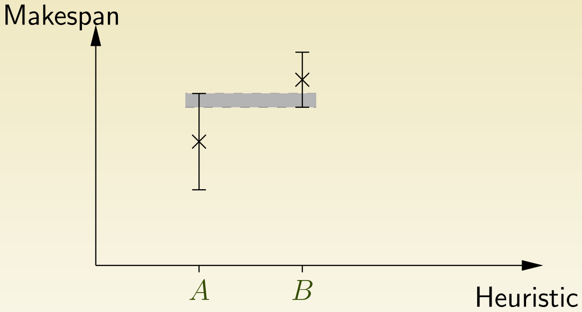
Comparing Two Alternatives

Assume, you have evaluated two scheduling heuristics A and B on n different DAGs.



The two 95% confidence intervals do not overlap $\leadsto \mathrm{P}(\mu_A < \mu_B) > 90\%$.

Assume, you have evaluated two scheduling heuristics A and B on n different DAGs.

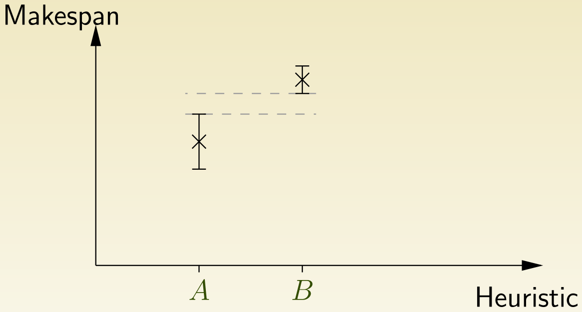


The two 95% confidence intervals do overlap $\rightsquigarrow$??.

Reduce C.I ?

Comparing Two Alternatives

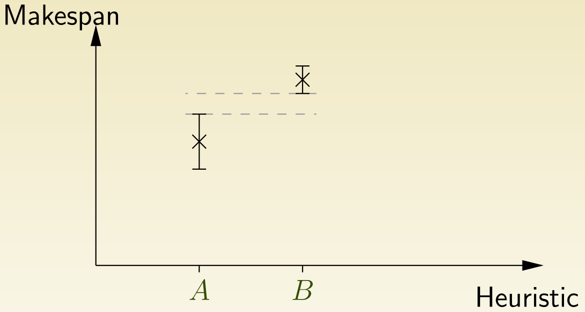
Assume, you have evaluated two scheduling heuristics A and B on n different DAGs.



The two 70% confidence intervals do not overlap $\rightsquigarrow \mathrm{P}(\mu_A < \mu_B) > 49\%$.

Let's do more experiments instead.

Assume, you have evaluated two scheduling heuristics A and B on n different DAGs.



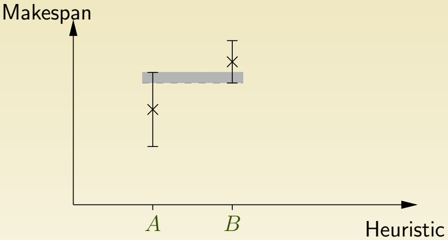
The width of the confidence interval is proportionnal to $\frac{\sigma}{\sqrt{n}}$.

Halving C.I. requires 4 times more experiments!

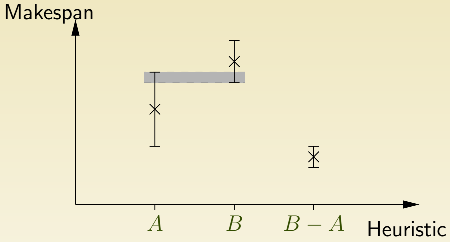
Try to reduce variance if you can...

Comparing Two Alternatives with Blocking

- C.I.s overlap because variance is large. Some DAGS have an intrinsically longer makespan than others, hence a large $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$



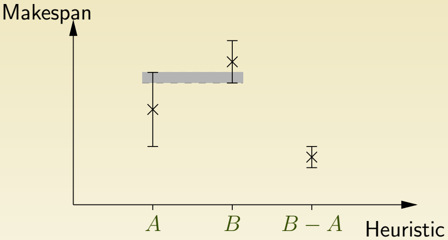
- C.I.s overlap because variance is large. Some DAGS have an intrinsically longer makespan than others, hence a large $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$



- The previous test estimates μ_A and μ_B independently.
  $\mathrm{E}[A] < \mathrm{E}[B] \Leftrightarrow \mathrm{E}[B-A] < 0$.
  In the previous evaluation, the same DAG is used for measuring A_i and B_i, hence we can focus on $B-A$.
  Since $\mathrm{Var}(B-A)$ is much smaller than $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$, we can conclude that $\mu_A < \mu_B$ with 95% of confidence.

Comparing Two Alternatives with Blocking

- C.I.s overlap because variance is large. Some DAGS have an intrinsically longer makespan than others, hence a large $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$



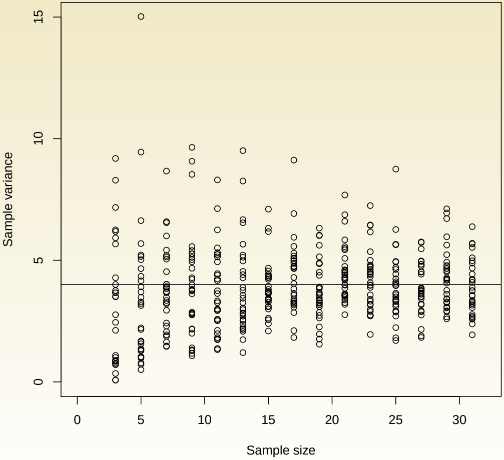
- The previous test estimates μ_A and μ_B independently.
  $\mathrm{E}[A] < \mathrm{E}[B] \Leftrightarrow \mathrm{E}[B - A] < 0$.
  In the previous evaluation, the same DAG is used for measuring A_i and B_i, hence we can focus on $B - A$.
  Since $\mathrm{Var}(B - A)$ is much smaller than $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$, we can conclude that $\mu_A < \mu_B$ with 95% of confidence.
- Relying on such common points is called blocking and enable to reduce variance.

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\text{Var}(X)}$ but we only have the sample variance, not the true variance.

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.



How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

  A1: How many can you afford ?

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

  A1: How many can you afford ?
  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work).

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

  A1: How many can you afford ?
  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work).
- With less than 30, you need to make the $C.I.$ wider using e.g. the Student law.

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

  A1: How many can you afford ?
  A2: 30. . .

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work).
- With less than 30, you need to make the $C.I.$ wider using e.g. the Student law.
- Once you have a first C.I. with 30 samples, you can estimate how many samples will be required to answer your question. If it is too large, then either try to reduce variance (or the scope of your experiments) or simply explain that the two alternatives are hardly distinguishable...

How Many Replicates ?

- The CLT says that "when n goes large", the sample mean is normally distributed.
  The CLT uses $\sigma = \sqrt{\mathrm{Var}(X)}$ but we only have the sample variance, not the true variance.

  Q: How Many Replicates ?

  A1: How many can you afford ?
  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work).
- With less than 30, you need to make the $C.I.$ wider using e.g. the Student law.
- Once you have a first C.I. with 30 samples, you can estimate how many samples will be required to answer your question. If it is too large, then either try to reduce variance (or the scope of your experiments) or simply explain that the two alternatives are hardly distinguishable...
- **Running the right number of experiments enables to get to conclusions more quickly and hence to test other hypothesis.**

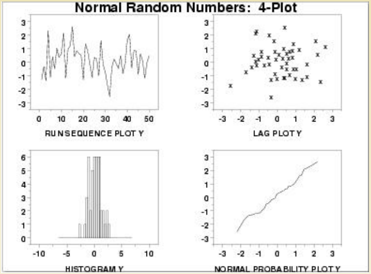
Key Hypothesis

The hypothesis of CLT are very weak. Yet, to qualify as replicates, the repeated measurements:

- must be independent (take care of warm-up)
- must not be part of a time series (the system behavior may temporary change)
- must not come from the same place (the machine may have a problem)
- must be of appropriate spatial scale

Perform graphical checks

Simple Graphical Check



Fixed Location: If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.

Fixed Variation: If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.

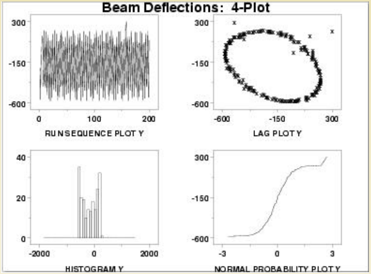
Independence: If the randomness assumption holds, then the lag plot will be structureless and random.

Fixed Distribution : If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then

- the histogram will be bell-shaped, and
- the normal probability plot will be linear.

If you see several modes, you may want to investigate further is there is not another hidden parameter you should take into account.

Simple Graphical Check



Fixed Location: If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.

Fixed Variation: If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.

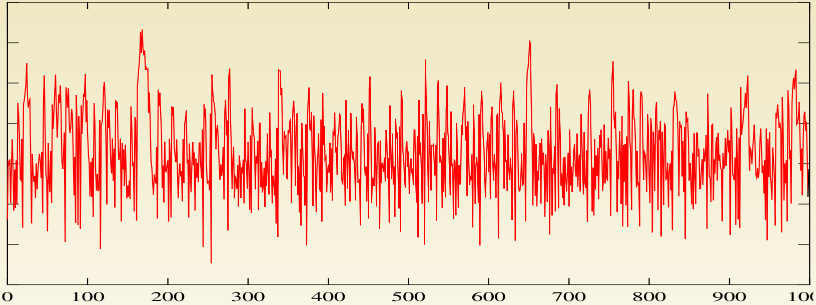
Independence: If the randomness assumption holds, then the lag plot will be structureless and random.

Fixed Distribution : If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then

- the histogram will be bell-shaped, and
- the normal probability plot will be linear.

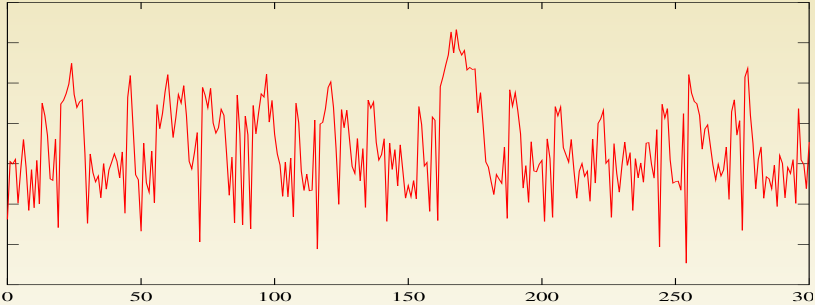
If you see several modes, you may want to investigate further is there is not another hidden parameter you should take into account.

Temporal Dependancy



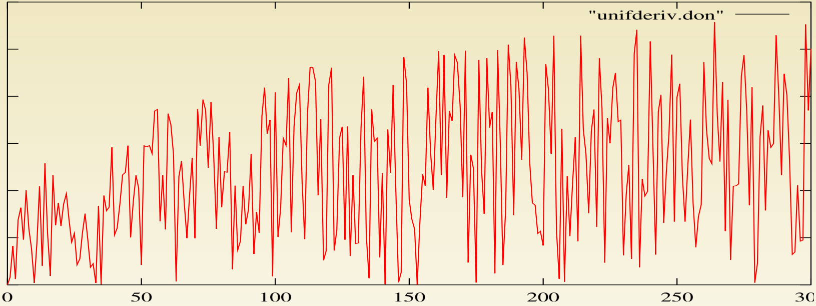
- Looks independant and statistically identical

Temporal Dependancy



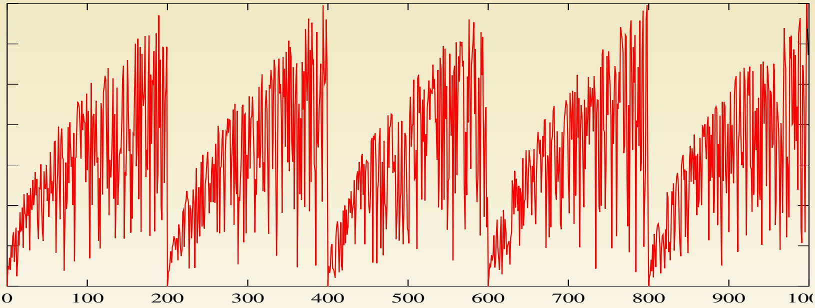
- Looks independant and statistically identical
- Danger: temporal correlation $\rightsquigarrow$ study stationnarity.

Detect Trends



- Model the trend: here increase then saturates
- Possibly remove the trend by compensating it (multiplicative factor here)

Detect Periodicity



May depend on sampling frequency or on horloge resolution.

- Study the period (Fourier)
- Use time series

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea.

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea.
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \ldots$.

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea.
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \ldots$.
- Even better, randomize your run order. You should flip a coin for each configuration and start with A on head and with B on tail...

  $$A, B, \quad B, A, \quad B, A, \quad A, B, \ldots.$$

  With such design, you will even be able to check whether being the first alternative to run changes something or not.

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea.
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \dots$.
- Even better, randomize your run order. You should flip a coin for each configuration and start with A on head and with B on tail...

  $$A, B, \quad B, A, \quad B, A, \quad A, B, \dots.$$

  With such design, you will even be able to check whether being the first alternative to run changes something or not.
- Each configuration you test should be run on different machines. You should record as much information as you can on how the experiments was performed (http://expo.gforge.inria.fr/).

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

**If you replicate thoroughly and randomize properly,
you will not go far wrong.**

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

If you replicate thoroughly and randomize properly,
you will not go far wrong.

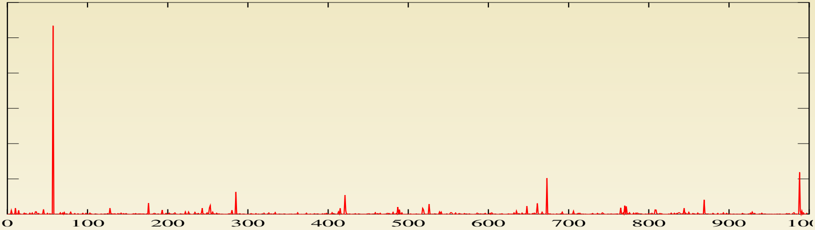
> It doesn't matter if you cannot do your own advanced statistical analysis. If you designed your experiments properly, you may be able to find somebody to help you with the statistics.
>
> If your experiments is not properly designed, then no matter how good you are at statistics, you experimental effort will have been wasted.

No amount of high-powered statistical analysis can turn a bad experiment into a good one.

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

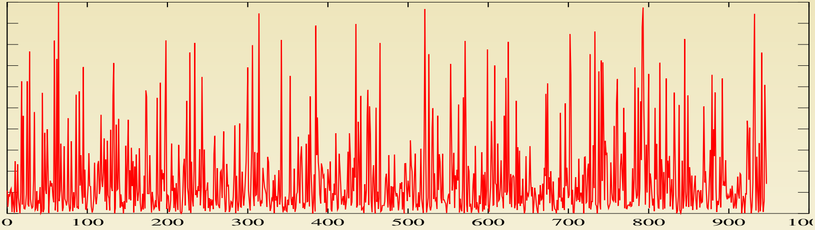
Abnormal measurements



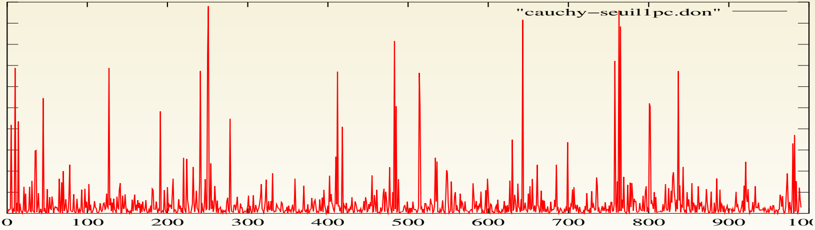
- Rare events: interpretation
- Get rid of it using:
  - a threshold value: what is the right threshold ?
  - quantiles: what is the good rejection rate ?

Thresholds:

Reject values larger than 10 $\rightsquigarrow$ 5% of rejection



Reject values larger than 50 $\rightsquigarrow$ 1% of rejection

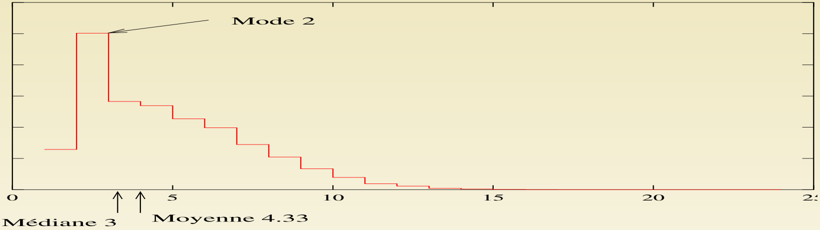


Actually, here, the samples are generated using the Cauchy distribution, which is pathological for most ideas you may come up with. :)

Outline

1. Confidence Intervals
2. Using Confidence Intervals
3. Design of Experiments: Early Intuition
4. Getting rid of Outliers
5. Issues when studying something else than the mean

Summarizing the distribution

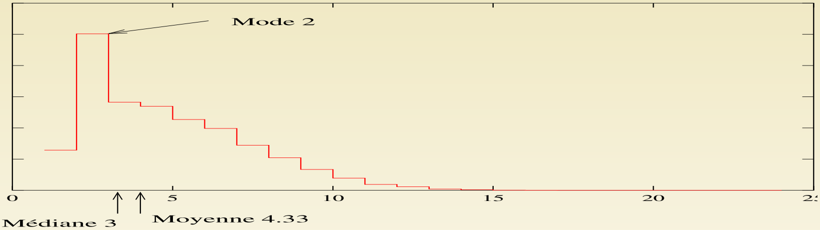


What is the shape of the histogram:

- uni/multi-modal
- symmetrical or not ($\leadsto$ skewness)
- Flat of not ($\leadsto$ kurtosis)

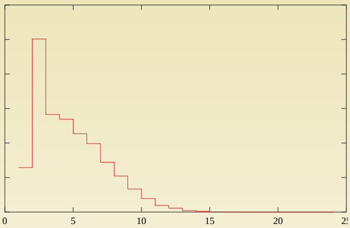
Summarize with central tendancy

Summarizing the distribution



- Mode: the most probable value (higly depends on the bin size)
- Median: splits the samples in half (rather unstable)
- Mean: average "cost" (can simply estimate confidence intervals)

Mode value
histogram



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

- **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**

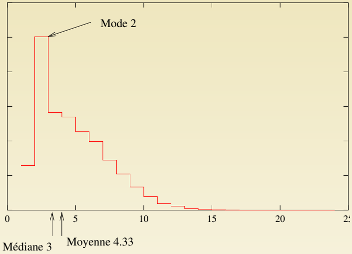
Mean

- **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,...)**



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

Computational aspects

- Mode : computation of the histogram steps, then computation of max $O(n)$ “off-line”
- Median : sort the sample $O(nlog(n))$ or $O(n)$ (subtile algorithm) “off-line”
- Mean : sum values $O(n)$ “on-line” computation

Computational aspects

- Mode : computation of the histogram steps, then computation of max $O(n)$ “off-line”
- Median : sort the sample $O(nlog(n))$ or $O(n)$ (subtile algorithm) “off-line”
- Mean : sum values $O(n)$ “on-line” computation

Is the central tendency significant ?
$\Rightarrow$ Explain variability.

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:

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

Roadmap for a good data analysis

1. Plot the sample (various representations)
2. Describe the results (data analysis)
3. Preliminary processing : remove or flag outliers, estimate or flag missing values
4. Propose a stochastic model : establish the hypothesis : independence (time correlation, auto-correlation), stationarity, same probability law
5. Summarize data by a histogram
6. Comment the shape (modal/skewness/flatness/...)
7. Estimate the central tendency of the sample : choose the central index
8. Estimate the accuracy of the result (confidence intervals)
9. Propose a visualization

References

Raj Jain.
Art of Computer Systems Performance Analysis.
John Wiley and Sons, 1991.
Chapter 12 Summarizing Measured Data and Chapter 13 Comparing Systems Using Sample Data.

David J. Lilja.
Measuring Computer Performance: A Practitioner's Guide.
Cambridge University Press, 2009.
Chapter 3 Average Performance and Variability and Chapter 4 Errors in Experimental Measurements.

Douglas C. Montgomery.
Design and Analysis of Experiments.
John Wiley and Sons, 2009.
Read the introduction for the positionning problem.

Sheldon M. Ross.
Introductory Statistics, Third Edition.
Academic Press, 3 edition, 2010.
Chapter 3 for descriptive statistics, Chapter 8 for estimation and confidence intervals.