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 20, 2011

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 $P : \mathcal{F} \to [0, 1]$ is a function returning an event's probability.
- Since many computer science experiments are based on time measurements, we focus on continuous variables.

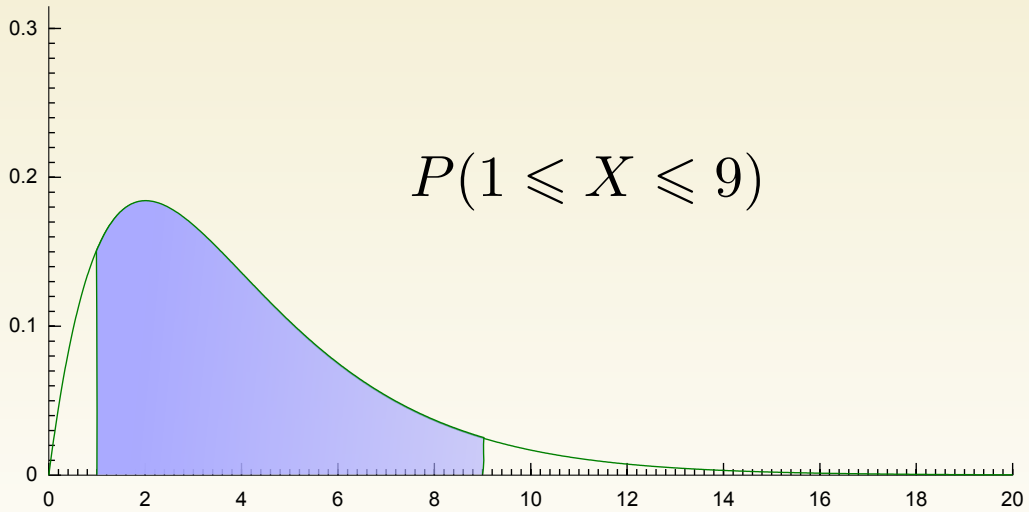
$$X : \Omega \to \mathbb{R}$$

Probability Distribution

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.

$$\begin{aligned}\mathrm{E}[X] &= x_1 p_1 + x_2 p_2 + \ldots + x_k p_k \\ &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\end{aligned}$$

  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

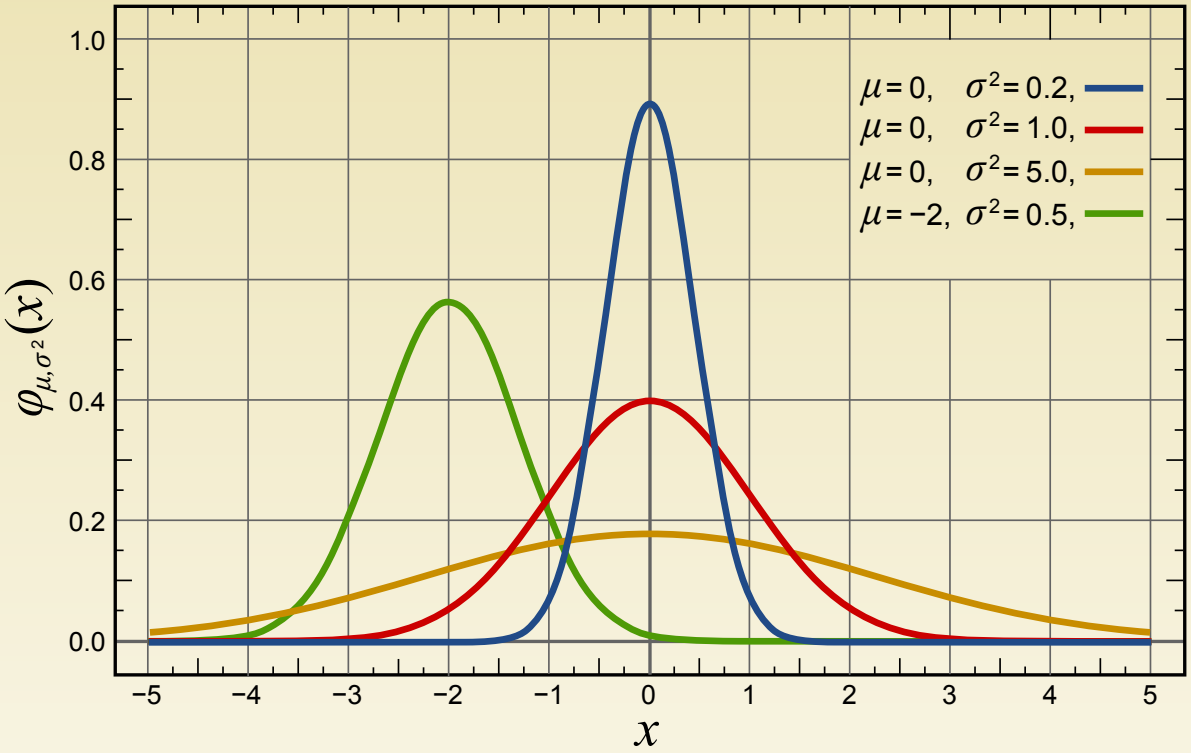
- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n (i.e., a sequence of **independent** and **identically distributed** random variables with expected values μ and variances σ^2).
- The sample average of these random variables is:

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

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

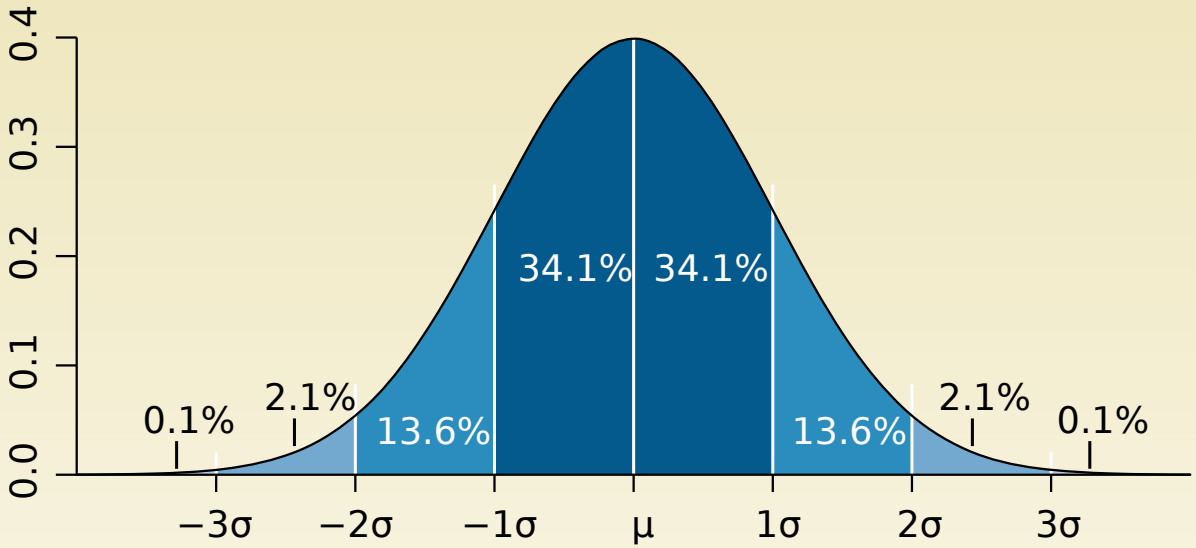
$$S_n \xrightarrow[n\to\infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The Normal Distribution



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

The Normal Distribution

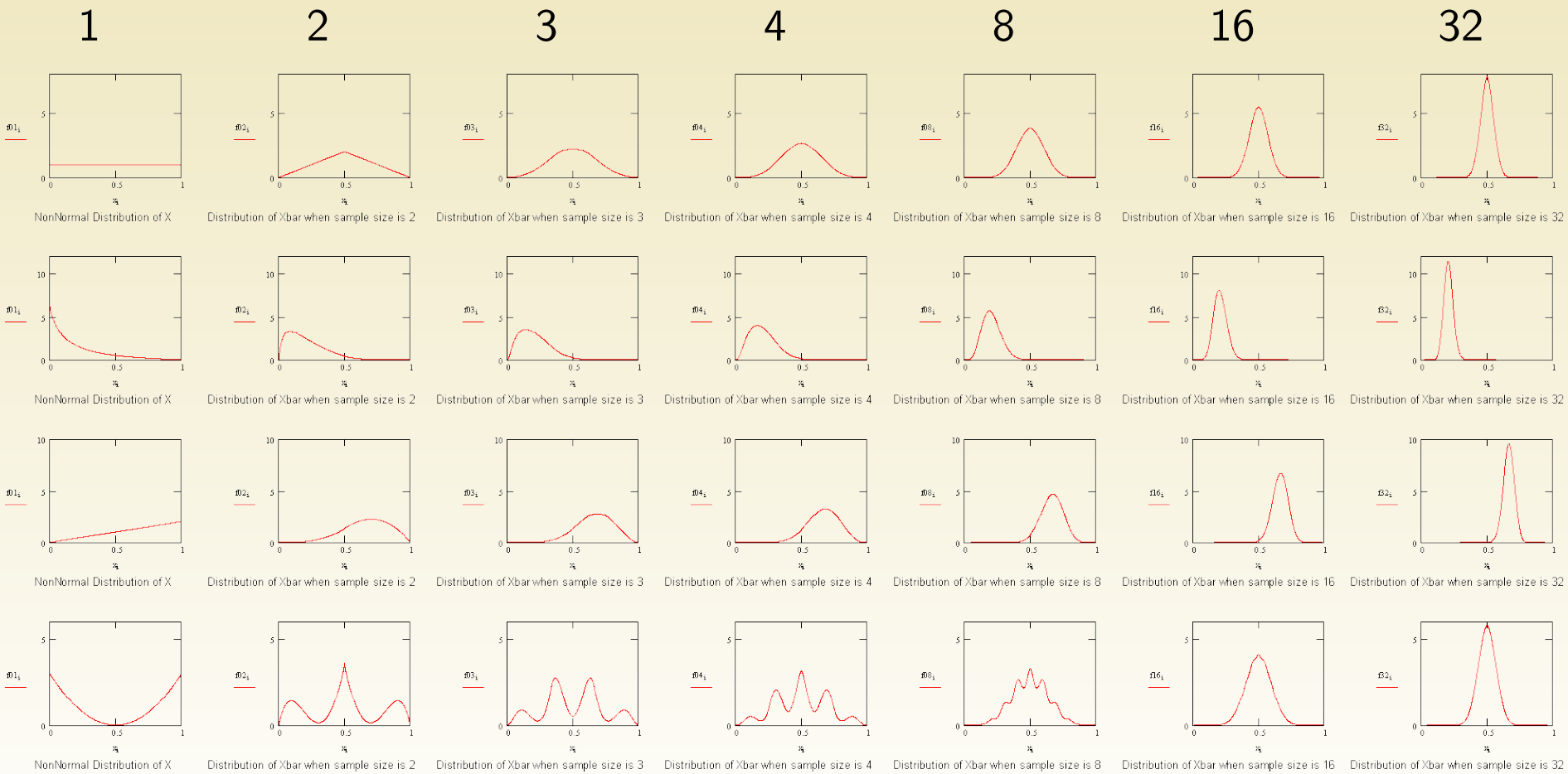


The smaller the variance the more "spiky" the distribution.

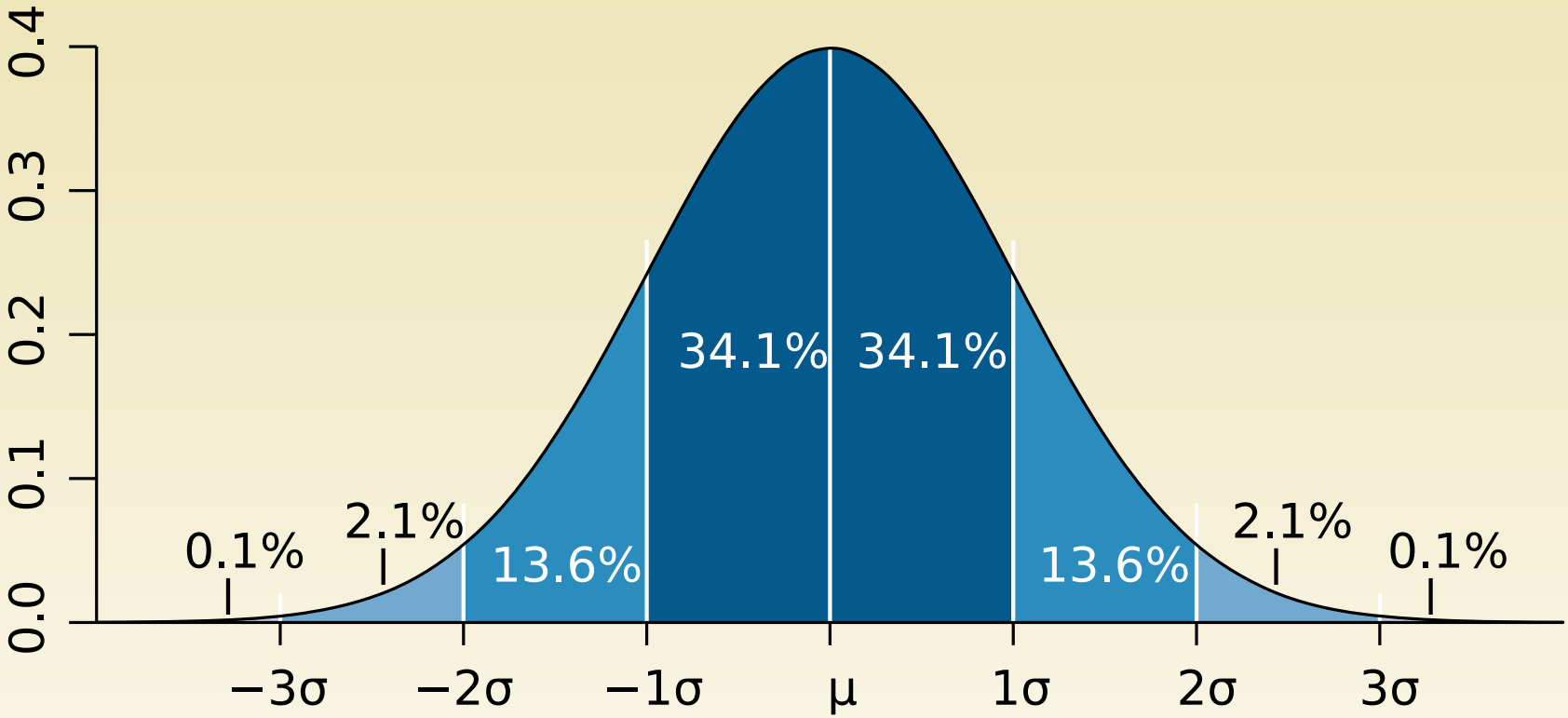
- Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set.
- Two standard deviations from the mean (medium and dark blue) account for about 95%
- Three standard deviations (light, medium, and dark blue) account for about 99.7%

CLT Illustration

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



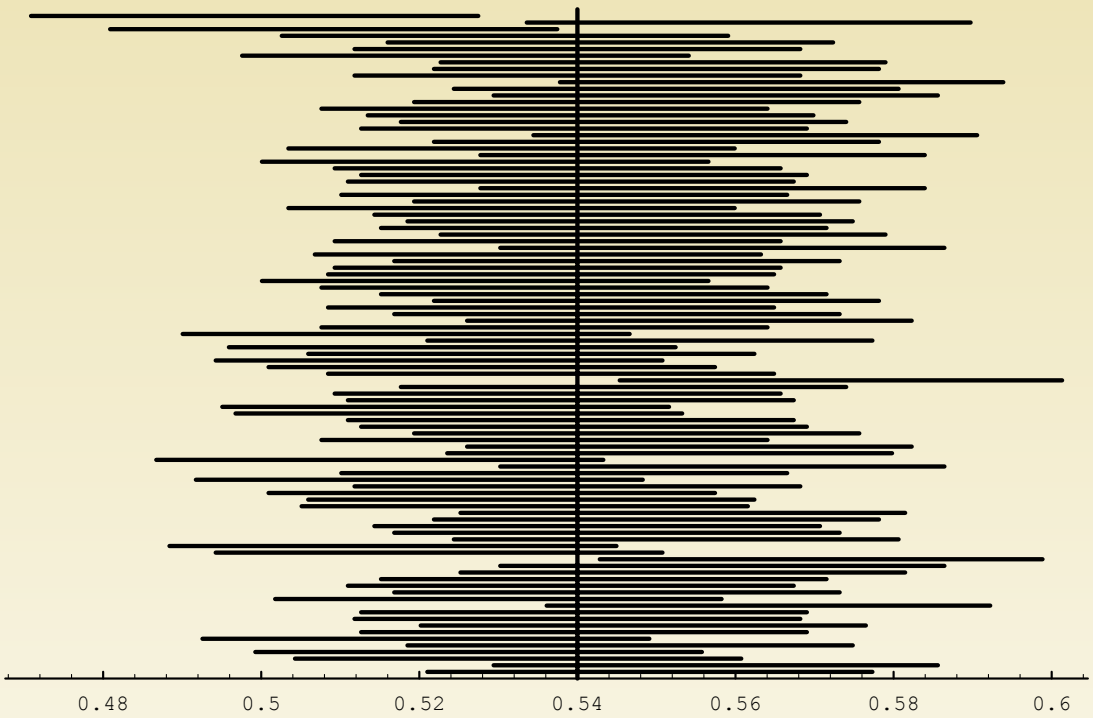
CLT consequence: confidence interval



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



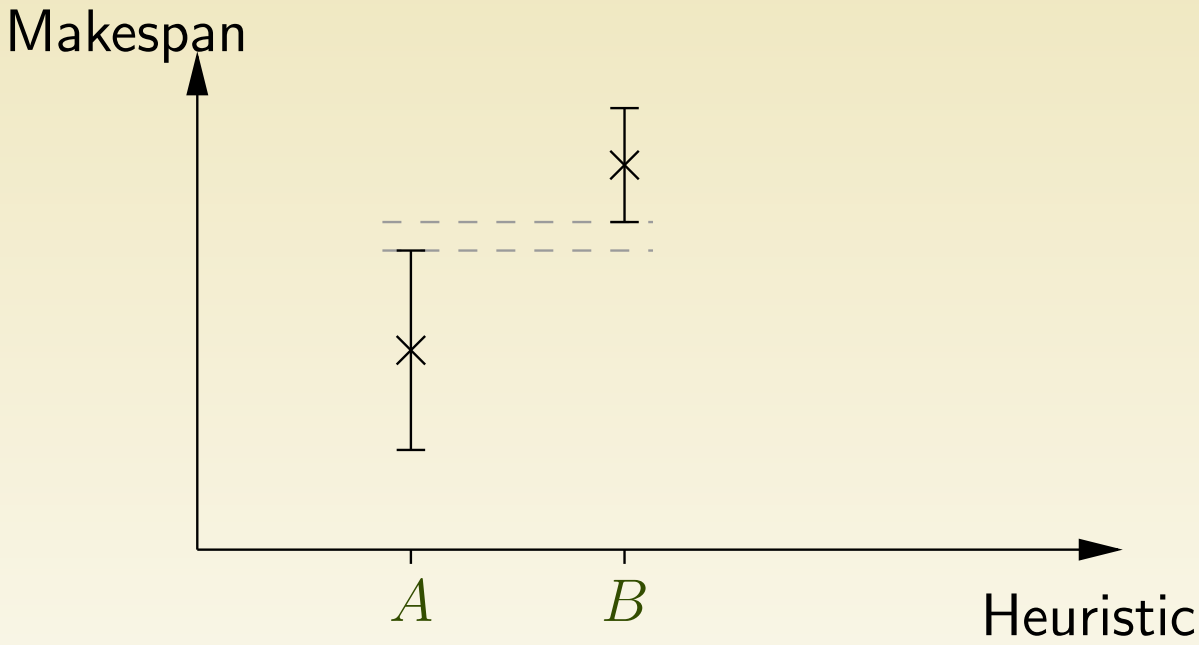
When n is large:

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

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

Comparing Two Alternatives

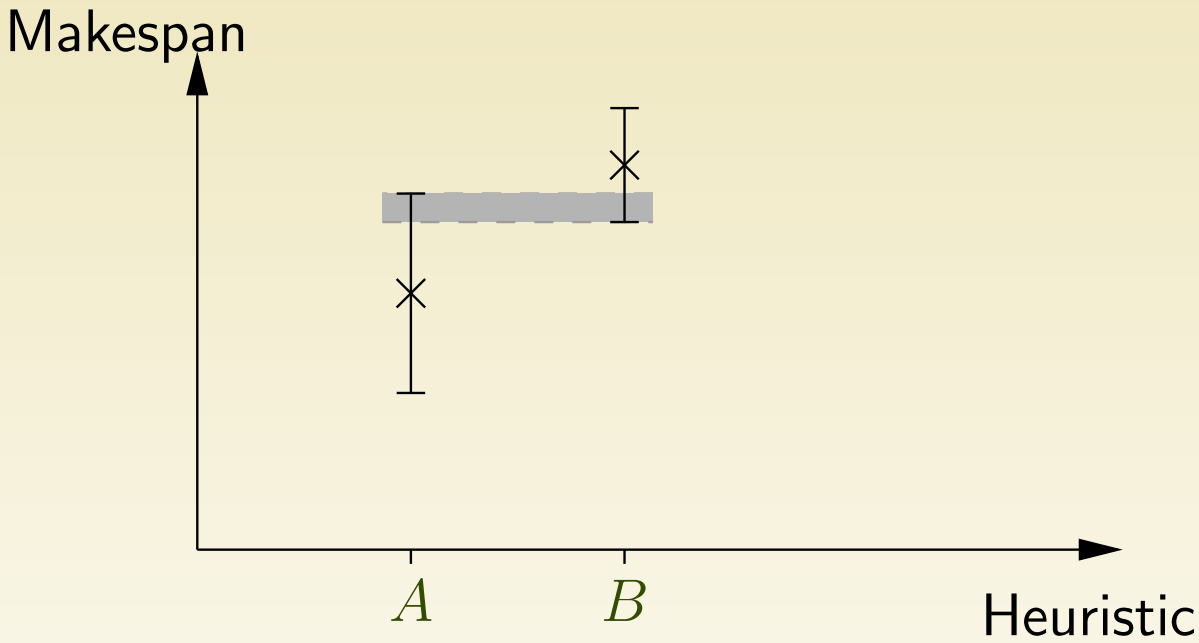
Assume, you have evaluated two scheduling heuristics A and B on n different DAGs.



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

Comparing Two Alternatives

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

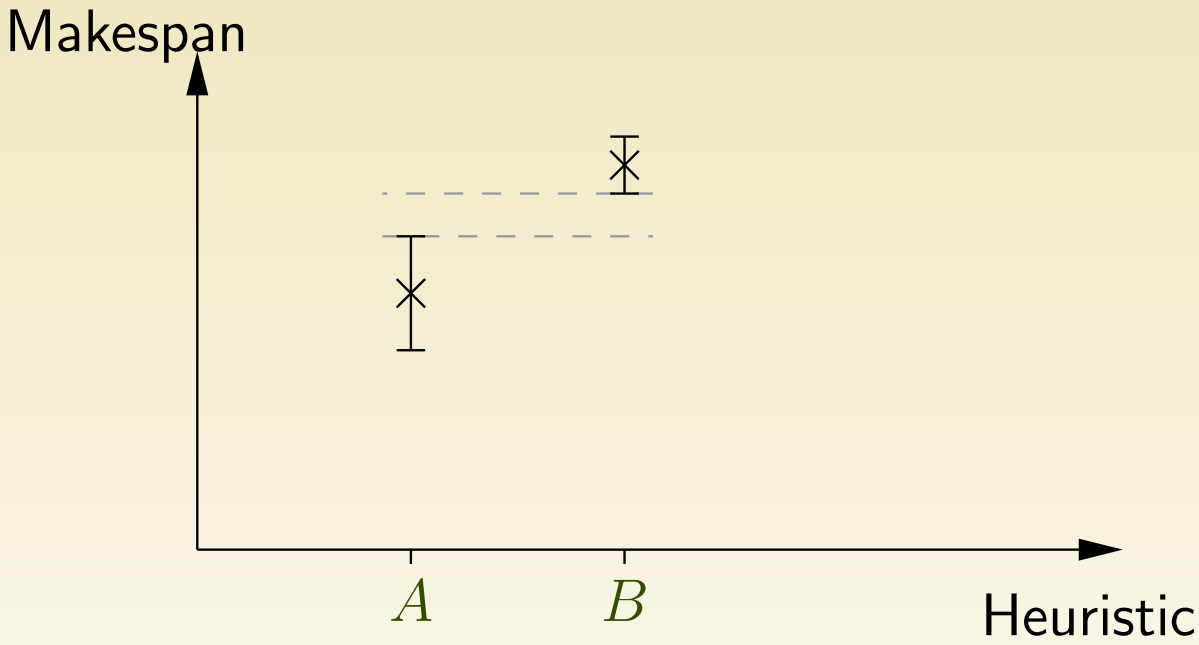


The two 95% confidence intervals do overlap $\rightsquigarrow$??.

Reduce C.I ?

Comparing Two Alternatives

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

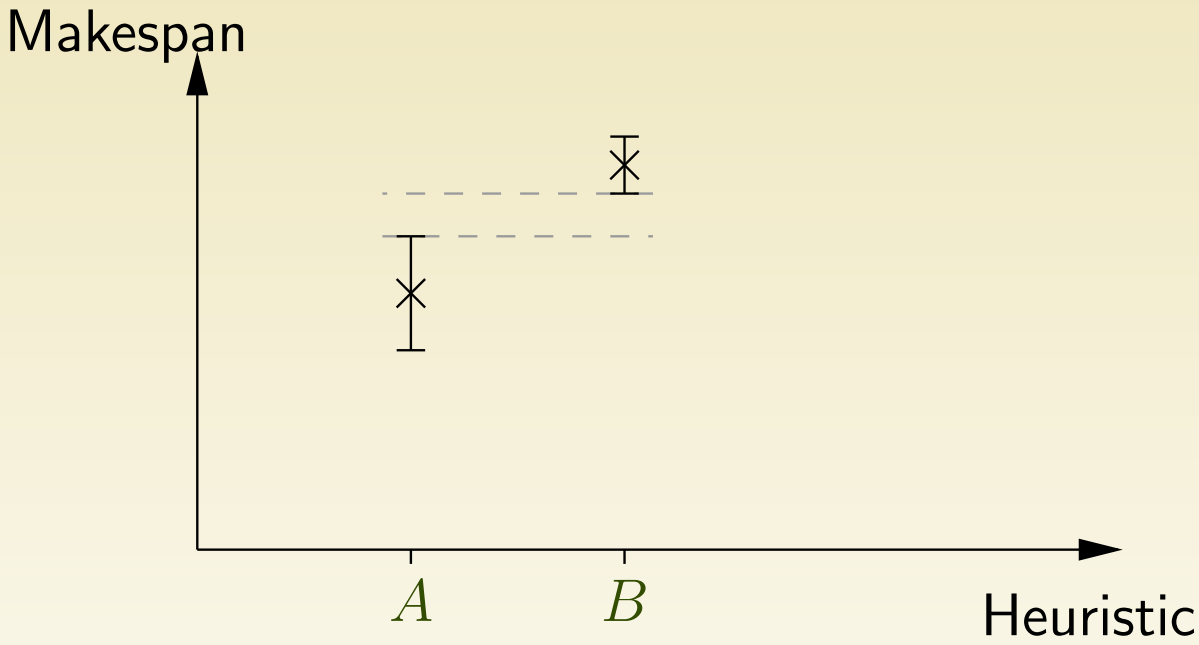


The two 70% confidence intervals do not overlap $\rightsquigarrow \mathrm{P}(\mu_A < \mu_B) > 49\%$.

Let's do more experiments instead.

Comparing Two Alternatives

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



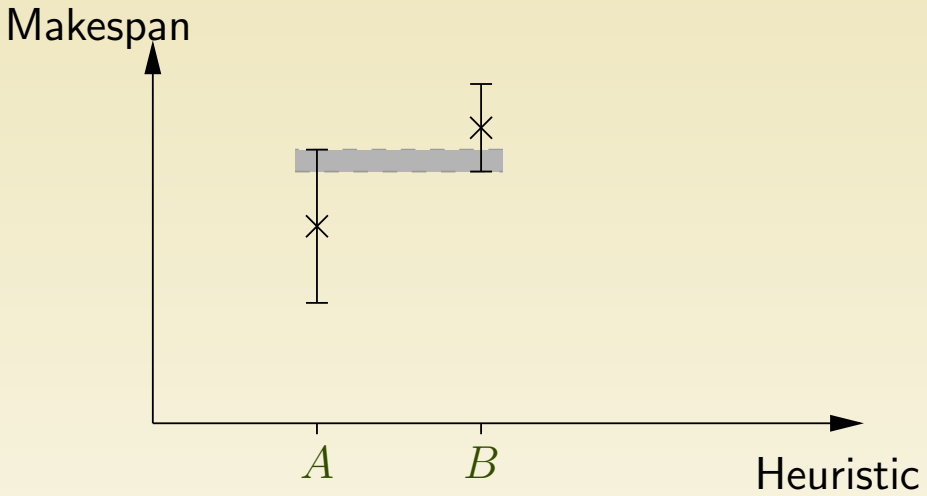
The width of the confidence interval is proportionnal to $\frac{\sigma}{\sqrt{n}}$.

Halving C.I. requires 4 times more experiments!

Try to reduce variance if you can...

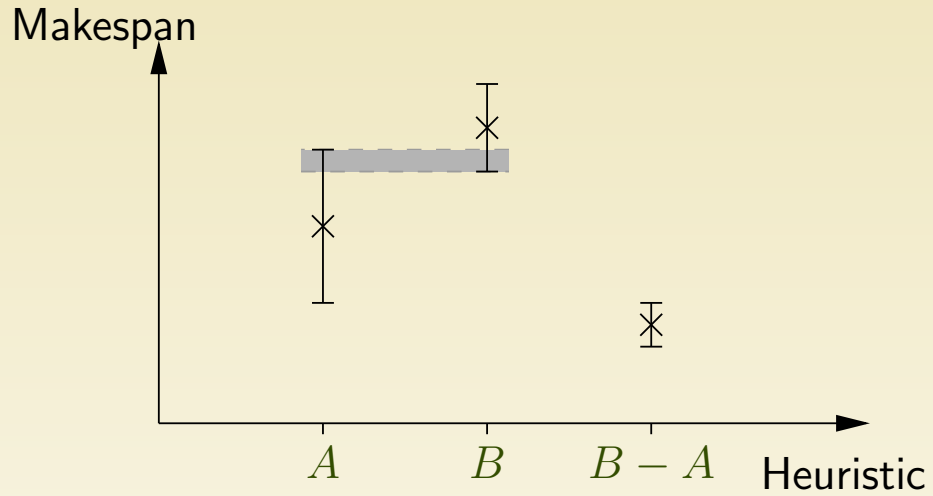
Comparing Two Alternatives with Blocking

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



Comparing Two Alternatives with Blocking

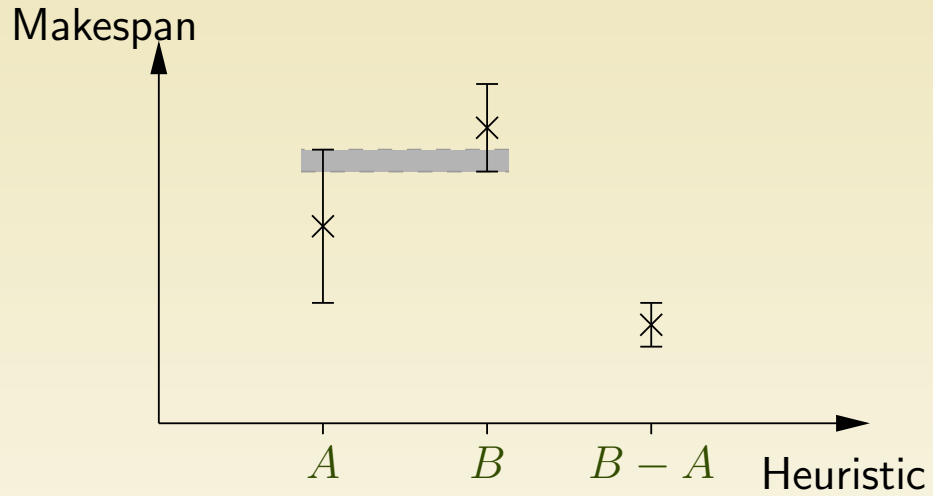
- C.I.s overlap because variance is large. Some DAGS have an intrinsically longer makespan than others, hence a large $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$



- 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.

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 $\mathrm{Var}(B)$



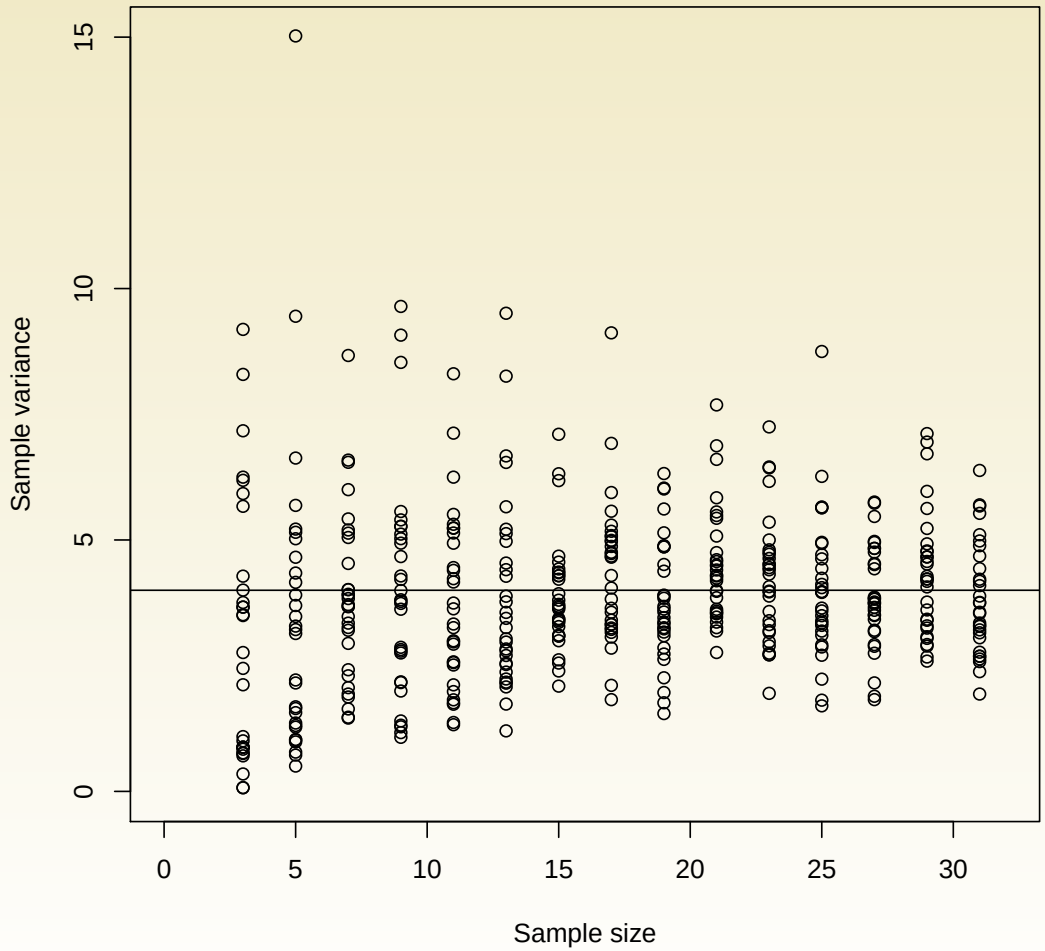
- The previous test estimates μ_A and μ_B independently.
  $\mathrm{E}[A] < \mathrm{E}[B] \Leftrightarrow \mathrm{E}[B - A] < 0$.
  In the previous evaluation, the same DAG is used for measuring A_i and B_i, hence we can focus on $B - A$.
  Since $\mathrm{Var}(B - A)$ is much smaller than $\mathrm{Var}(A)$ 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 $\sigma = \sqrt{\mathrm{Var}(X)}$ 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.



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.**

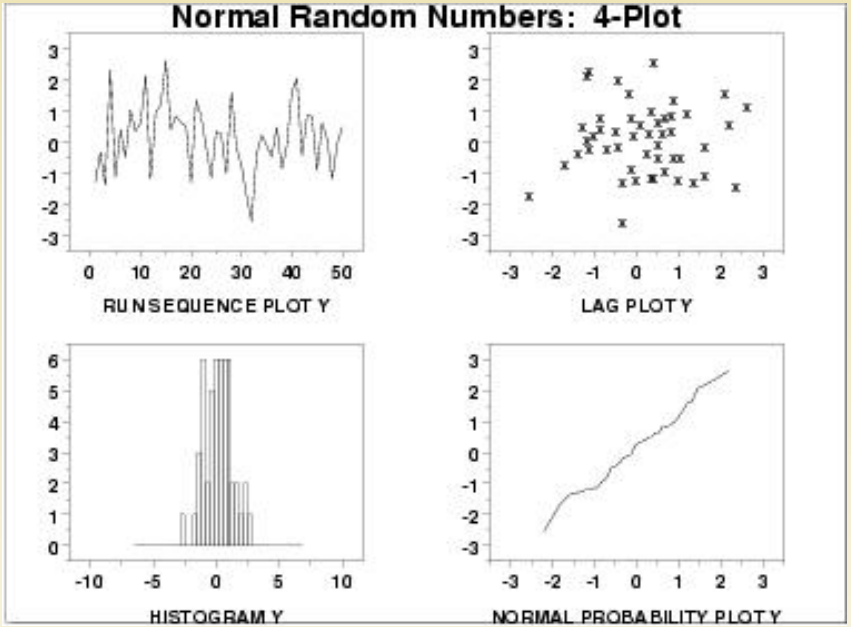
Key Hypothesis

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.

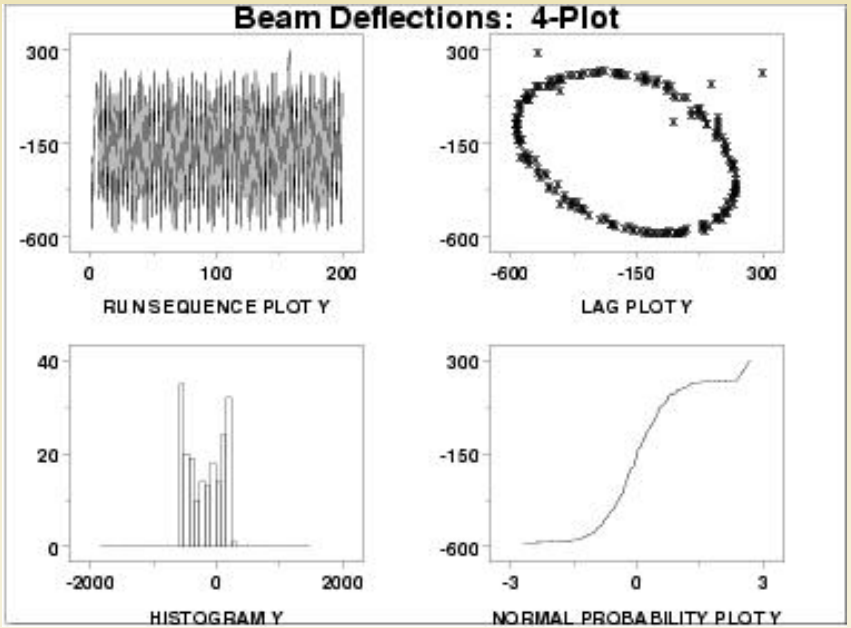
Fixed Variation: If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.

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.

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.

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

Other important issues:

- Parsimony
- Pseudo-replication
- Experimental vs. observational data

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.

Other important issues:

- Parsimony
- Pseudo-replication
- Experimental vs. observational data

  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.

Parsimony

The principle of parsimony is attributed to the 14th century English philosopher William of Occam:

> *"Given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation"*

Parsimony

The principle of parsimony is attributed to the 14th century English philosopher William of Occam:

> *"Given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation"*

- Models should have as few parameters as possible
- Linear models should be preferred to non-linear models
- Models should be pared down until they are *minimal adequate*

Parsimony

The principle of parsimony is attributed to the 14th century English philosopher William of Occam:

> *"Given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation"*

- Models should have as few parameters as possible
- Linear models should be preferred to non-linear models
- Models should be pared down until they are *minimal adequate*

This means, a variable should be retained in the model only if it causes a significant increase in deviance when removed from the current model.

> *A model should be as simple as possible. But no simpler.*
>
> *– A. Einstein*

Replication vs. Pseudo-replication

Measuring the same configuration several times is not replication. It's pseudo-replication and may be biased.
Instead, test other configurations (with a good randomization).

In case of pseudo-replication, here is what you can do:

- average away the pseudo-replication and carry out your statistical analysis on the means
- carry out separate analysis for each time period
- use proper time series analysis

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments.

Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
This may be OK in the early stages but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments.

Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
This may be OK in the early stages but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Strong inference Essential steps:

1. Formulate a clear hypothesis
2. devise an acceptable test

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments.

Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
This may be OK in the early stages but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Strong inference Essential steps:

1. Formulate a clear hypothesis
2. devise an acceptable test

Weak inference It would be silly to disregard all observational data that do not come from designed experiments. Often, they are the only we have (e.g. the trace of a system).
But we need to keep the limitations of such data in mind. It is possible to use it to derive hypothesis but not to test hypothesis.

Computer scientists tend to either:

- vary one parameter at a time and use a very fine sampling of the parameter range,
- or run thousands of experiments for a week varying a lot of parameters and then try to get something of it. Most of the time, they (1) don't know how to analyze the results (2) realize something went wrong and everything need to be done again.

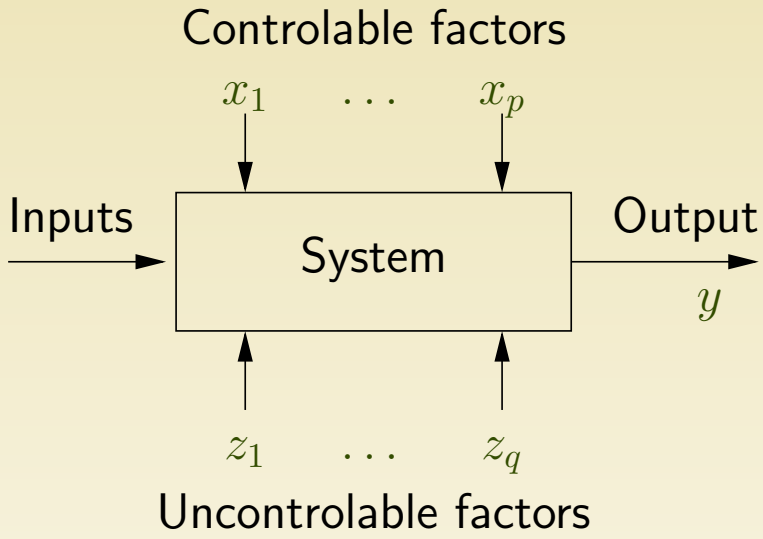
These two flaws come from poor training and from the fact that C.S. experiments are almost free and very fast to conduct.
Most strategies of experimentation have been designed to

- provide sound answers despite all the randomness and uncontrollable factors;
- maximize the amount of information provided by a given set of experiments;
- reduce as much as possible the number of experiments to perform to answer a given question under a given level of confidence.

- Clearly define the kind of system to study, the kind of phenomenon to observe (state or evolution of state through time), the kind of study to conduct (descriptive, exploratory, prediction, hypothesis testing, ...).
- For example, the set of experiments to perform when studying the stabilization of a peer-to-peer algorithm under a high churn is completely different from the ones to perform when trying to assess the superiority of a scheduling algorithm compared to another over a wide variety of platforms.
- It would be also completely different of the experiments to perform when trying to model the response time of a Web server under a workload close to the server saturation.

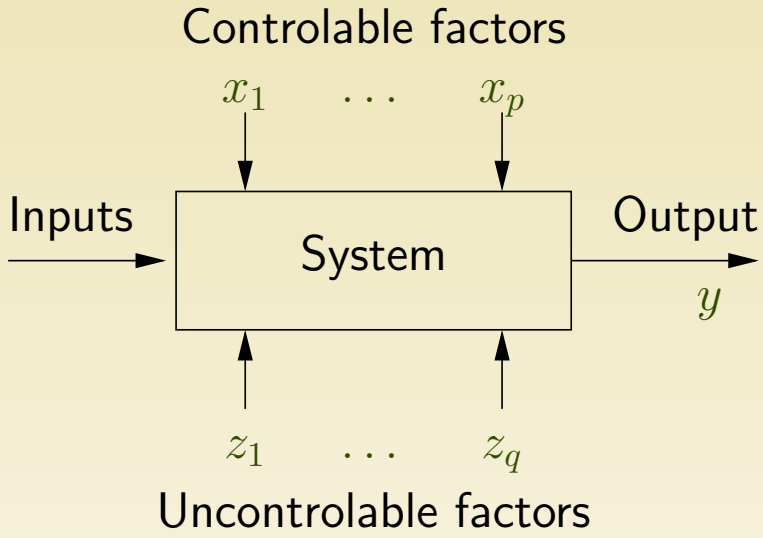
This first step enables to decide on which kind of design should be used.

The system under study is generally modeled though a black-box model:



- In our case, the response could be the makespan of a scheduling algorithm, the amount of messages exchanged in a peer-to-peer system, the convergence time of distributed algorithm, the average length of a random walk, . . .
- Some of these metrics are simple while others are the result of complex aggregation of measurements. Many such responses should thus generally be recorded so as to check their correctness.

Some of the variables (x_1,. . . ,x_p) are controllable whereas some others (z_1, . . . , z_q) are uncontrollable.



- In our case typical controllable variables could be the heuristic used (e.g., FIFO, HEFT, . . .) or one of their parameter (e.g., an allowed replication factor, the time-to-live of peer-to-peer requests, . . .), the size of the platform or their degree of heterogeneity, ...
- In the case of computer simulations, randomness should be controlled and it should thus be possible to completely remove uncontrollable factors. Yet, it may be relevant to consider some factors to be uncontrollable and to feed them with an external source of randomness.

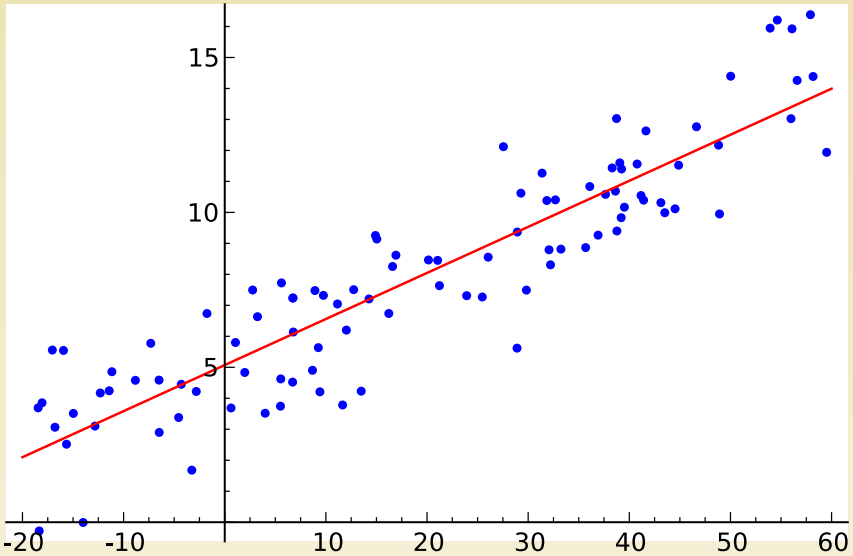
The typical case studies defined in the first step could include:

- determining which variables are most influential on the response y (*factorial designs*, *screening designs*). This allows to distinguish between *primary factors* whose influence on the response should be modeled and *secondary factors* whose impact should be averaged. This also allows to determine whether some factors interact in the response;
- deriving an analytical model of the response y as a function of the primary factors x. This model can then be used to determine where to set the primary factors x so that response y is always close to a desired value or is minimized/maximized (*analysis of variance*, *regression model*, *response surface methodology*, ...);
- determining where to set the primary factors x so that variability in response y is small;
- determining where to set the primary factors x so that the effect of uncontrollable variables $z_1, \ldots, z_q$ is minimized (*robust designs*, *Taguchi designs*).

Linear Regression

$$Y = a + bX + \varepsilon$$

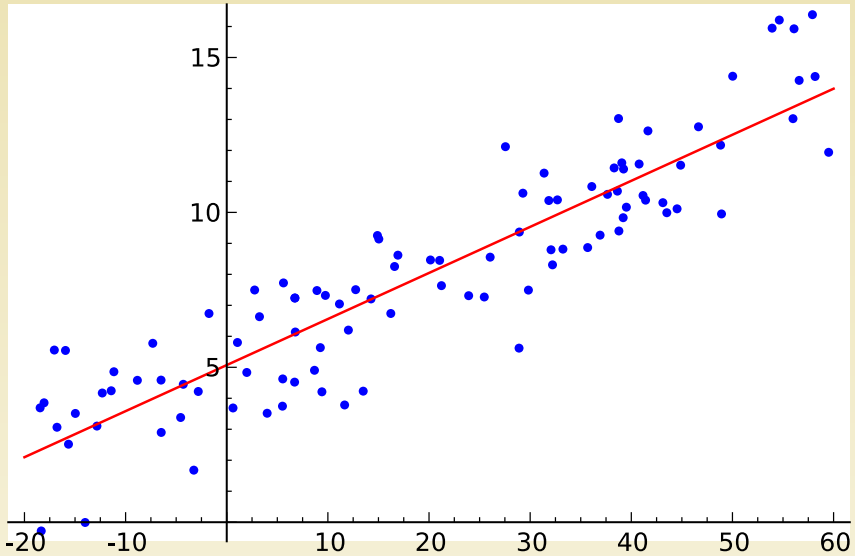
- Y is the response variable
- X is a continuous explanatory variable
- a is the intercept
- b is the slope
- ε is some noise



Linear Regression

$$Y = a + bX + \varepsilon$$

- Y is the response variable
- X is a continuous explanatory variable
- a is the intercept
- b is the slope
- ε is some noise



When there are 2 explanatory variables:

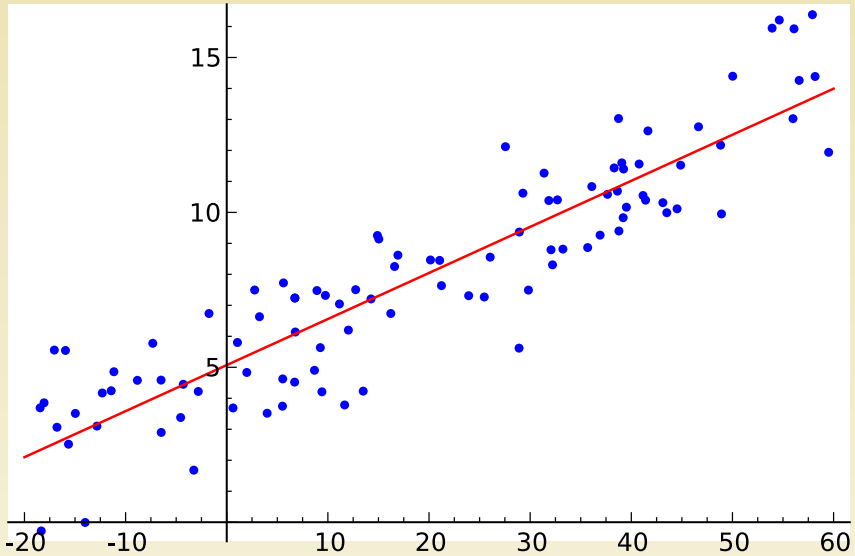
$$Y = a + b^{(1)}X^{(1)} + b^{(2)}X^{(2)} + b^{(1,2)}X^{(1)}X^{(2)} + \varepsilon$$

ε is generally assumed to be independent of $X^{(k)}$, hence it needs to be checked once the regression is done.

Linear Regression

$$Y = a + bX + \varepsilon$$

- Y is the response variable
- X is a continuous explanatory variable
- a is the intercept
- b is the slope
- ε is some noise



When there are 2 explanatory variables:

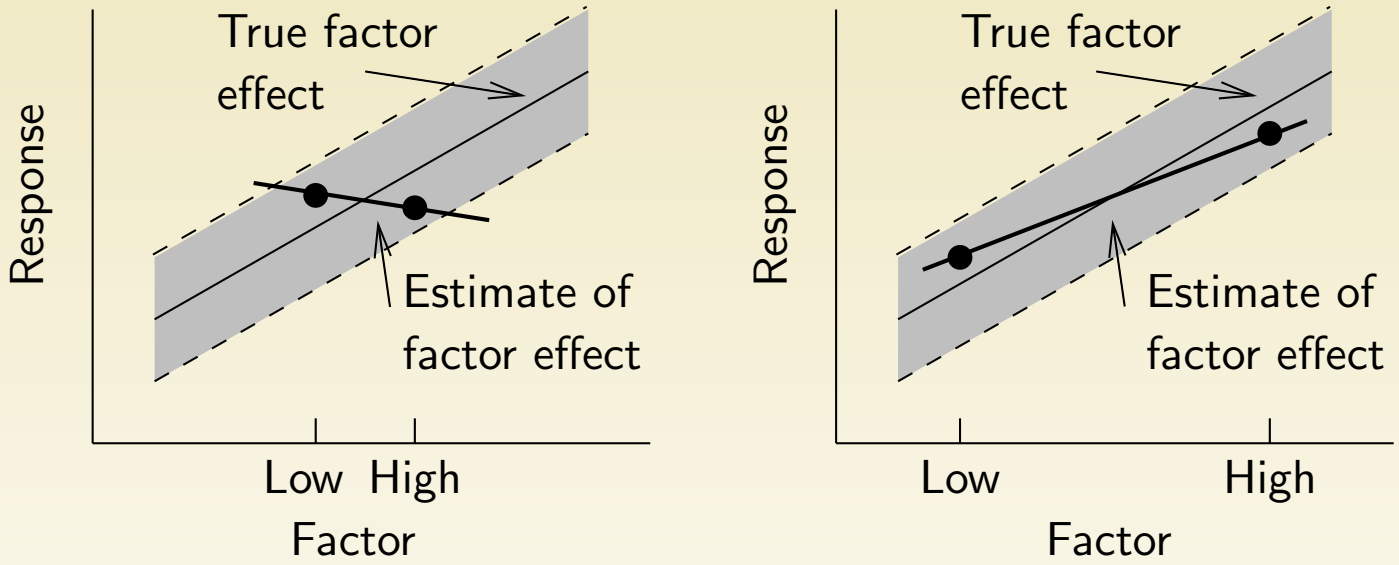
$$Y = a + b^{(1)}X^{(1)} + b^{(2)}X^{(2)} + b^{(1,2)}X^{(1)}X^{(2)} + \varepsilon$$

ε is generally assumed to be independent of $X^{(k)}$, hence it needs to be checked once the regression is done.

- Although your phenomenon is not linear, the linear model helps for initial investigations (as a first crude approximation).
- You should always wonder whether there is a way of looking at your problem where it is linear.

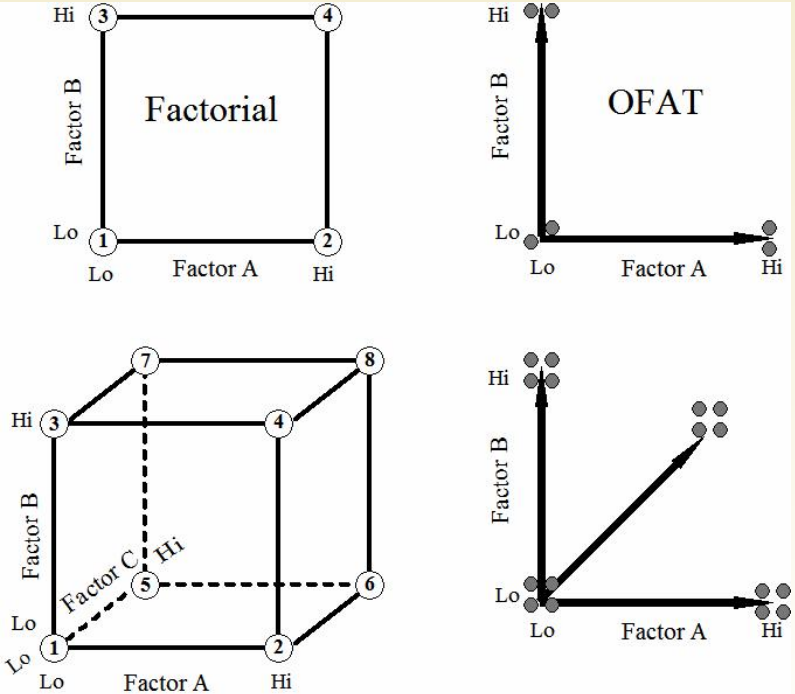
2-level factorial design

- Decide a low and a high value for every factor



2-level factorial design

- Decide a low and a high value for every factor
- Test every (2^p) combination of high and low values, possibly replicating for each combination.
- By varying everything, we can detect interactions right away.



2-level factorial design

- Decide a low and a high value for every factor
- Test every (2^p) combination of high and low values, possibly replicating for each combination.
- By varying everything, we can detect interactions right away.
- Standard way of analyzing this: ANOVA (ANalysis Of VAriance) enable to discriminate real effects from noise.
  - $\rightsquigarrow$ enable to prove that some parameters have little influence and can be randomized over (possibly with a more elaborate model)
  - $\rightsquigarrow$ enable to easily know how to change factor range when performing steepest ascent method.