Statistical Modeling part 2

# What is the design of experiments?

The design of experiments is a way allowing us to model and simulate some entities in order to collect data and so achieve some purposes. Moreover we can make comparisons between different configurations and study the side effects over the variables.

## Principles

* Randomization: we assign random values to each factor that is non controlled by the experiment;
* Replication: we repeated the experiment many times in order to reduce the variability between answers;
* Statistical homogeneity: we need it to compare different alternatives from the results;

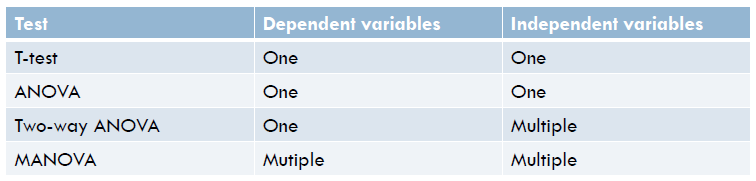
## Purposes

* Isolate the effects of each variable in input;
* determine the effects using the interactions;
* determine the magnitude of the experimental error;
* Get the max information for a given effort.

## Terminology

* The output is called response variable and it is a measured value;
* The input variables we can change are the factors;
* The levels are factors with a specific value, they can be continuos or discrete.
* A replication is a re-run of an experiment using the same inputs, it is useful to compute the impact of the measurement error;
* a interaction is the effect of one factor, it depends on the lavel of another factor;
* the scenario is one of the possible combinations of factors, our purpose is to analyze it.

## Tests



With ANOVA, we separate the variation in:

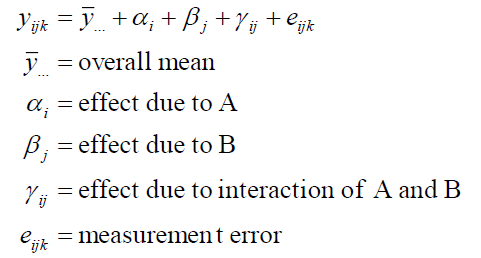
* the variation in one system: it is caused by the random measurement errors;
* the variation between systems: it is caused by the random error and the real differences

ANOVA is a one factor experimental design, what if we have more factors?

Considering the two-factor experiments, we have two factors A and B, so we can interpret the output in this way:

* the effects due to A;
* the effects due to B;
* the effects of the interactions of these factors (AB);
* the experimental error.

Each measurement is the composition of the overall mean, the effects, the errors and the interactions:







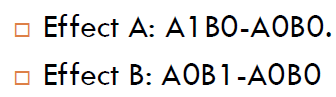
In case of no replication (n=1), the SSE is equals to 0, so we haven’t informations about the errors in the measurements.

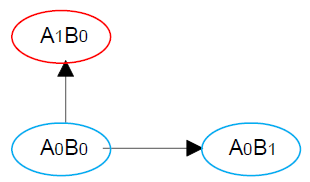
As a consequence, we cannot separate the effects of interactions from the measurement noise, this is a reason why the replication is important!

In fact, in case of no replication, SSAB is the difference of SST, SSA and SSB, so SSE=SST-SSA-SSB-SSAB is equals to zero.

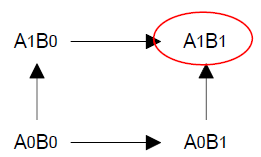
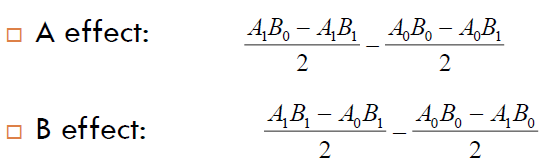
## Factorial designs

Now we consider a situation where there is no a factorial design: given two factors A and B, each factor can modify the level, but we consider it only for one factor per time:





Considering the factorial design, we take in consideration the interactions, so we can reach more states:



So, what is a factorial design? It is a type of design where we control k factors on different levels. Each level is an experiment and in total they are k.

The easiest factorial design is 2^k where li=2.

These is a problem in this type of design: considering k factors, v levels and n interactions, totally we will do n\*v^k executions, it is a huge number!

So, what can we do? We use the fractional factorial design, a type of design where the number of levels is fixed on 2. The number of executions can be huge, but in is case we can manage it.

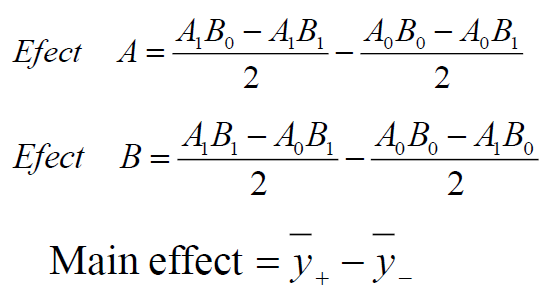
In poor words, we consider the factors as binary variables.

# Yates algorithm

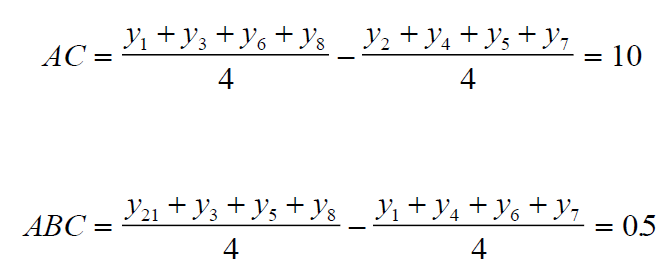
The Yates algorithm is a way to simplify the computing of interactions on a fractional factorial design, it has the following advantages:

* it determine the tendency with experiments economy, also called smoothness;
* it can evolve in a composide design (so, local explorations);
* it is the base for fractional factorial design, so there is a fast vision of multiple factors;
* Trivial analysis and interpretation.

## Computing of the effects



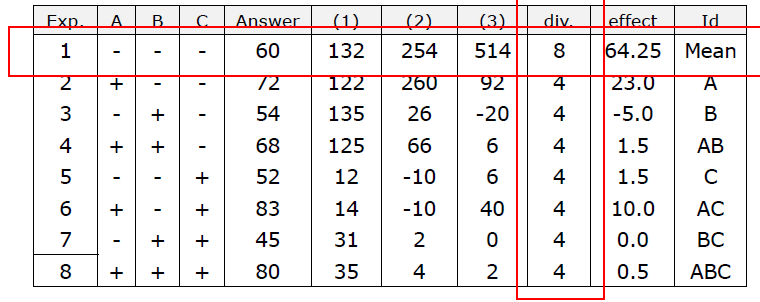
## Interaction effects



## How does it works?

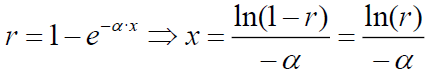
We use a table to make the computing systematic, so:

* in the ith column we add the answer of the matrix of the experimental design in the standard form;
* we add an auxiliary column as factor exists;
* we add a new column dividing the first value of the last column by the number of scenarios E and the others by half E.
* In the last column the first value is the mean of the answers, the last value are the effects;
* the correspondence value-effect is represented by a + in the rows of the matrix. if there is only one + in a column B, it means that it is the principal effect on B. More + corresponds to interactions of factors.



## How to perform a design of experiment

* Set the purposes;
* Select the process variables;
* Define the experiment design;
* Execute it;
* Check the consistence of data considering the assumptions;
* Analyze and interpret the results, detect effect of main factors and interactions;



### Example: clean industry

A clean industry clean some kind of pieces, each cleaning take a time represented by an exponential distribution with a parameter u depending on the technology used in the process.

Each factor add time to the process in this way:

* the cleaning time for a piece takes from 10 to 50 seconds;
* the amount of machines we use for glueing the pieces can be from 1 to 5, each one reduce the time by 1 seconds if it increase the number over two;
* the amount of workers finishing pieces, one worker take 1 second to finish a piece, 2 of theme take 0.5 seconds.

Considering the previous rules, we can say the following:

* the purpose is to detect the effects and the interactions of the three main factors;
* the process variables are the cleaner, the machines and the workers;
* we define 2\*k experimental design;
* We use excel or Python to emulate the behaviour of each model;
* We use the same software to detect the effects and the interactions.

## How can we compute the effects?

We can compute the effects using the answers that owns variability, in fact it goes to the effects:

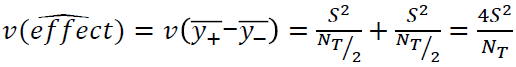


Then we calculate the variance:



The variance of the effect is the sum of the means of scenarios + and -:





# Kinds of simulations

The terminate simulation is a type of simulation that goes on until a condition won’t be satisfied, usually the time.

The non terminate simulations is a simulation without conditions.

## Transient elimination

Given k as the number of deleted observations, m as the amount of observation in a single run and n is the number of runs, we can compute:

* the average of the j-th observation over the others:



* the overall average:



* Considering k, we compute the average of the j-th observation with the first k ones:



* and the the relative change:



## Batch means

We can divide the execution in batch, each one of size N/m. In this way we work with mean values of the observations, each one is considered independent