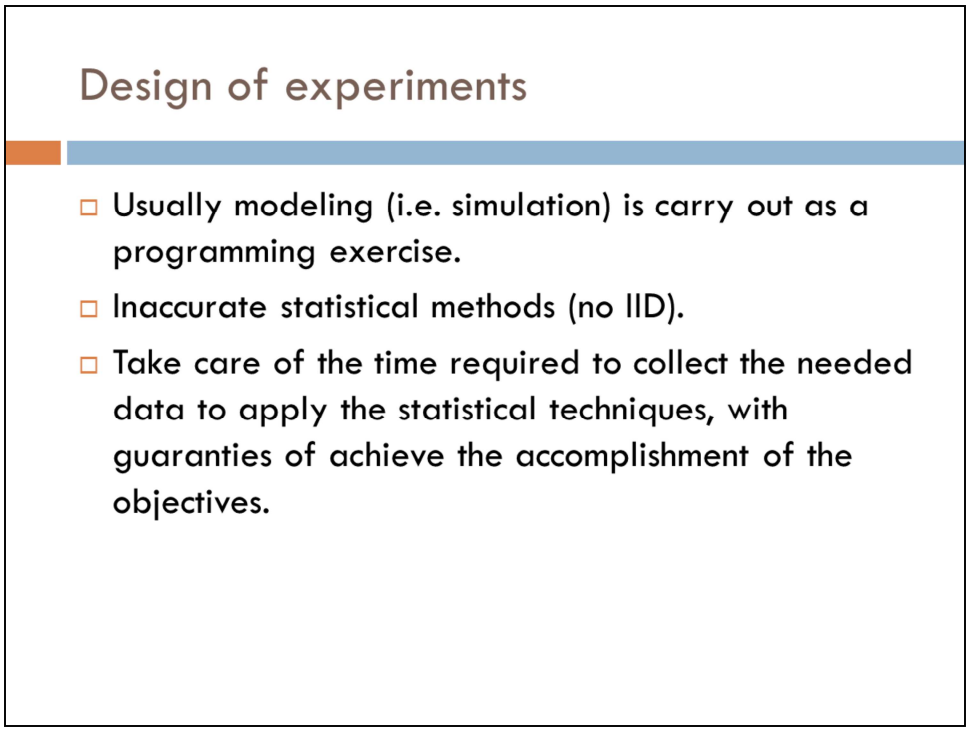
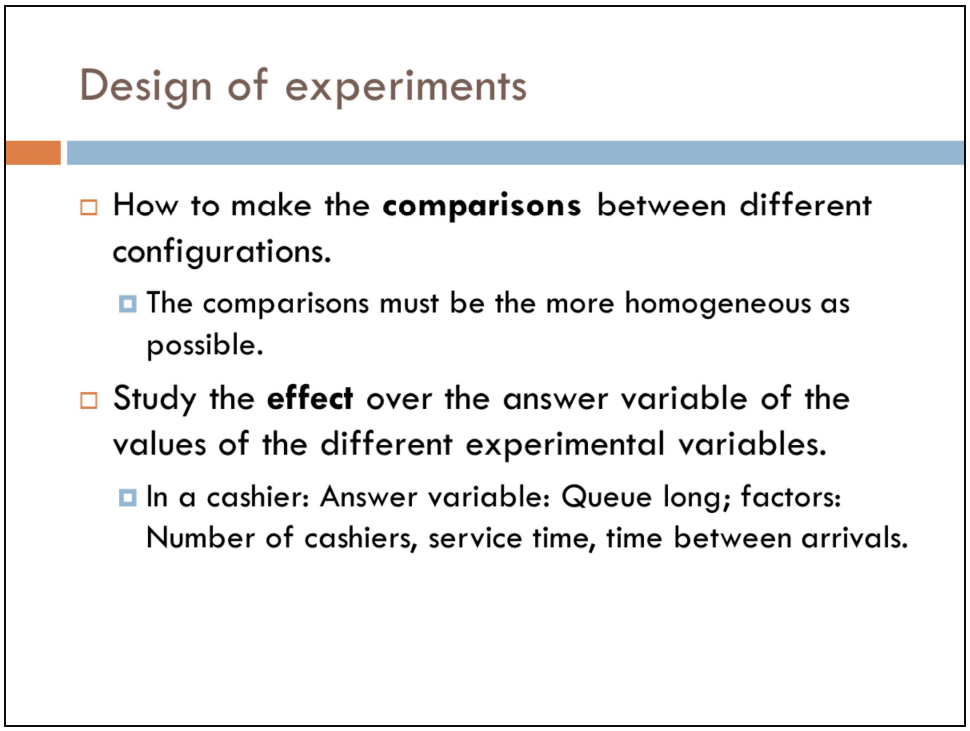
Dear Students, Today we are going to review the Design of Experiments parts, take this video as a support material that can be used to complete your understanding on this part of the subject.



One of the biggest problems of modelling, is consider that it is a programming exercise. It is not, the model is a product that will be useless until we do experiments with it, to test the assumptions and to understand if our models are working well.

Hence, it is needed to define an experimental design that fits on our needs.

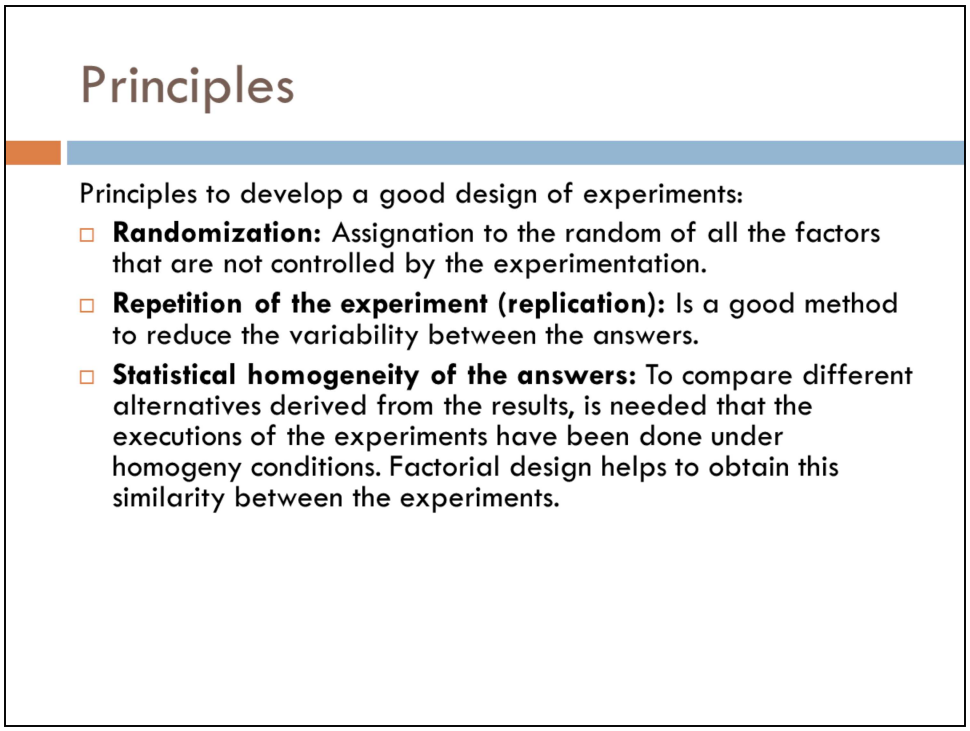
Also, there are problems regarding this, because, sometimes, we use inaccurate statistical methods (no independence between the observations is a typical concern), or we didn't take care of the time needed to do the experimentation, or the time needed to collect the needed data to apply the statistical techniques, with guaranties of achieve the objectives we define at the beginning of the modeling project.



The Design of experiments will help us to understand how to make the **comparisons** between different configurations. To compare different configurations we must assure that they are as much homogeneous as possible.

Also, the goal of the design of experiments is to study the effect over the answer variable of the values of the different experimental variables.

Consider this example, a cashier, where one want to analyze the longitude of the queue in front this cashier. There are several variables that can affect this longitude, like the number of cashiers, the service time, the time between arrivals, among others. All this variables will affect the answer variable we want to analyze, hence are possible factors for our analysis.



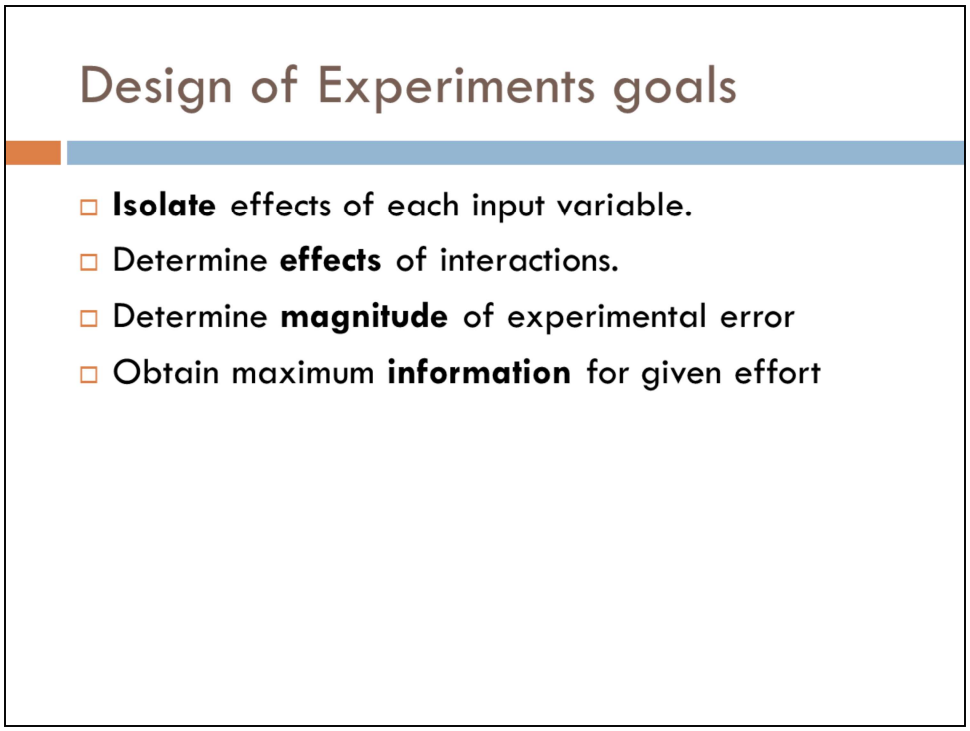
To assure that we are doing a good experimental design, it is needed to follow some basic principles, the Randomization, the replication and assure the Statistical homogeneity of the answers.

The randomization is the assignation to the random of all the factors that are not controlled by the experimentation. It means that if we have some knowledge that is not fully understood by the specialists that are working on the modeling process, we need to assign this to probability distributions that will help us to understand this behavior.

Since in the model we are using this random distribution, it will be needed to repeat the experiment several times, to assure that we are obtaining a result that makes sense. The repetition of the experiment (named replication) is the method we will use to reduce the variability between the answers, and to understand the real nature of the gaussian answer we will obtain form the answer we are analyzing.

Finally it will be needed to assure that we are comparing the right elements, the statistical homogeneity of the answers. We must assure that the experiments have been done under homogeny conditions.

As we will see later, factorial design helps to obtain this similarity between the experiments.



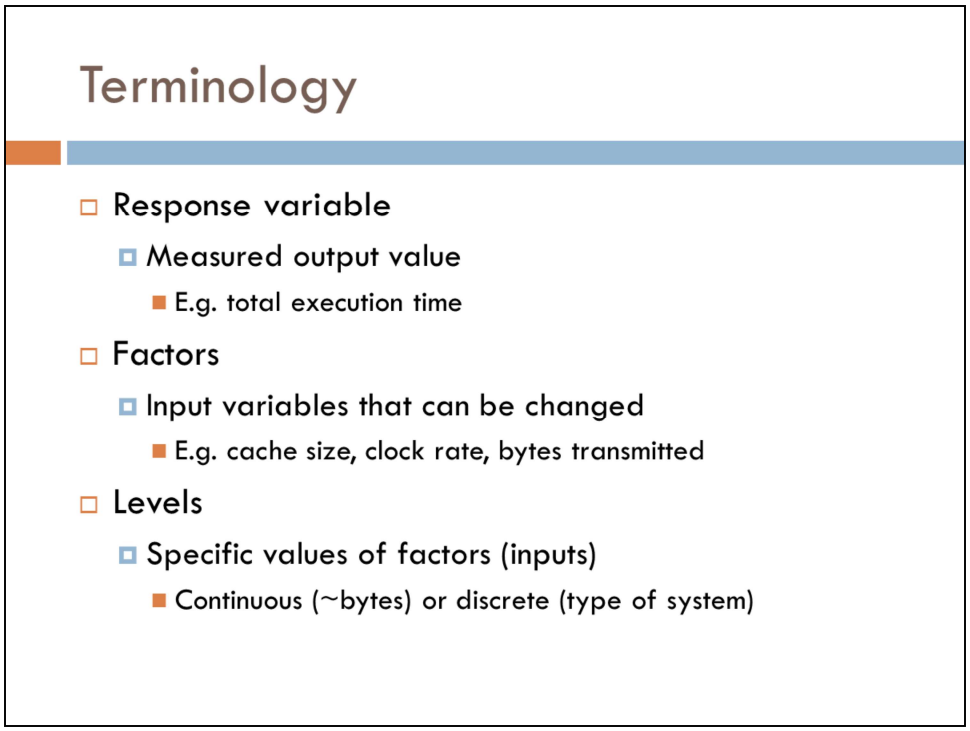
The goals we will want to achieve in a design of experiments are:

Isolate effects of each input variable.

Determine **effects** of interactions.

Determine **magnitude** of experimental error

Obtain maximum **information** for given effort

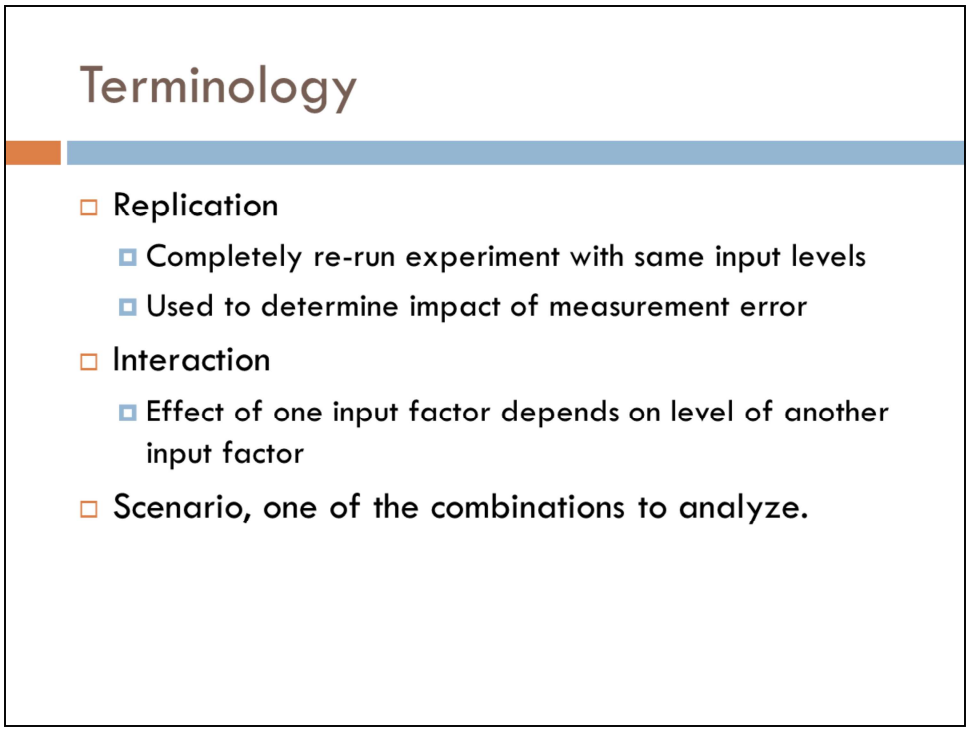


The terminology we are following is presented on this and the next slides.

The answer variable (or response variable) are the subject of our analysis. Notice that if there are different elements we want to analyze, then we must do this analysis, for each one of the answer variables we want. Be aware of this and select accurately the answer variables that are going to provide more information for your understanding of the system.

The factors are those variables that you control, and that can be changed in order to improve the behavior of the answer variable. Notice that improve can be maximize or minimize its value. Hence the factors are the subset of the variables that you are going to change to improve the answer.

The levels are the possible values of the factors. Is clear that if you have “n” factors and “l” levels, the combinations are going to be “n” multiplied by “l”, implying that usually the definition of a good experimental design is a tradeoff between the time you have to do the experiments nd the amount of alternatives you are going to analyze.



Replication is a main concern in modelling, since we often use probability distributions, that implies that every time we execute a new model, we will obtain results that will be slightly different, following a gaussian curve. Since we must assure that the value, we obtain from the model makes sense, it will be needed to assure that the number of observation of the normal curve are enough. This will be the leiv motive of the analysis of replications we will do. First calculate the number of times we must repeat the execution of the experiment, with the same parameters, but with different random number streams, and to determine how we are going to do this replications.

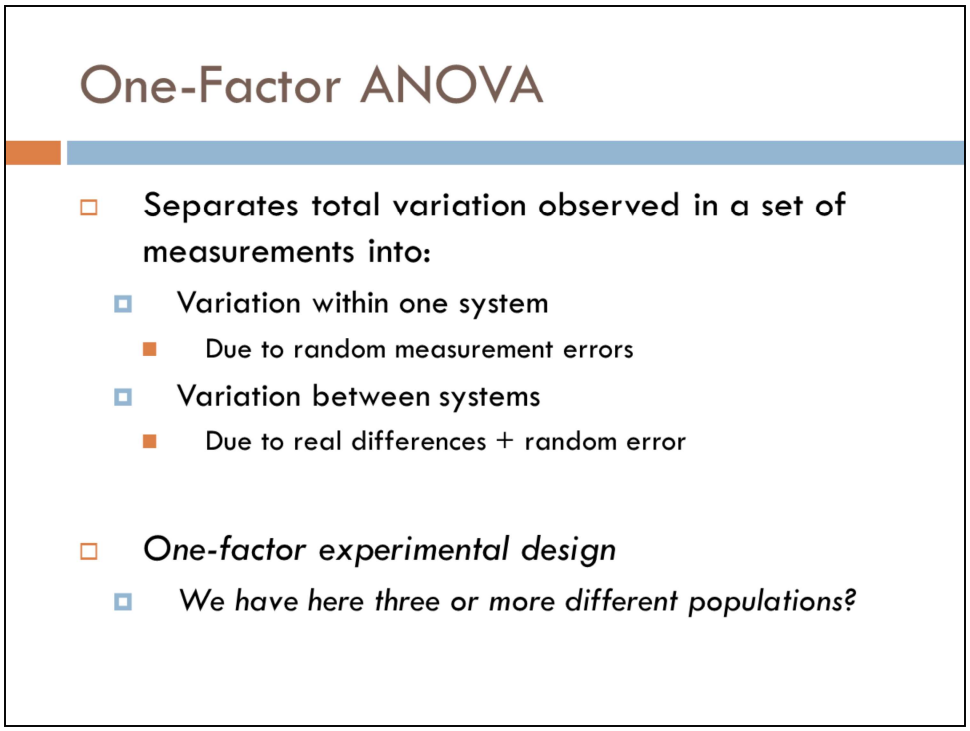
Finally, it will be needed to see the interaction between the different factors, this implies to be able to detect the effects that exists in the answer variable, due to the interaction that maybe exist in each one of the factors, (with all the possible combinations).

The scenario is just a combination to be analyzed.

Tests

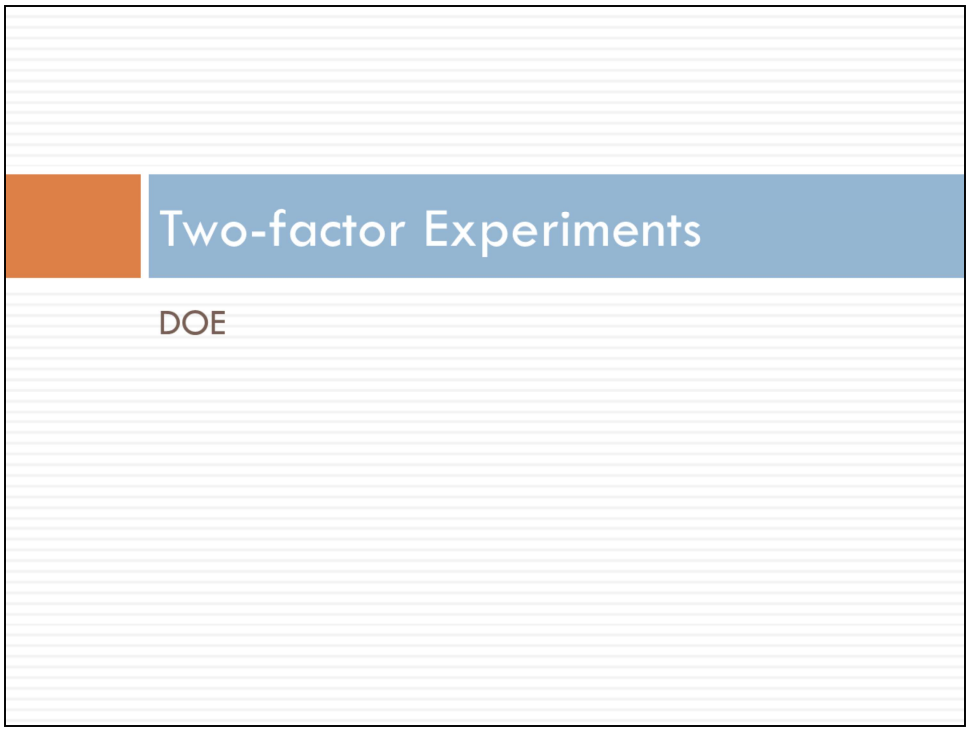
Test	Dependent variables	Independent variables
T-test	One	One
ANOVA	One	One
Two-way ANOVA	One	Multiple
MANOVA	Mutiple	Multiple

To connect with other parts of the course were we discuss regarding the t-test, ANOVA and MANOVA, see this table, one can detect that in each one of the different alternatives we increase the number of independent variables, in order to make the analysis more accurate, detecting interactions between the different factors.



In a one factor ANOVA, we analyze, for each one of the different observations of the answer, that are representing the replication in our modelling process, the variation due to the random measurement error, within one of the classes we propose, and the variation between the different classes, that represents the real differences we expect to find.

The question we want to answer in this kind of analysis is, *We have here three or more different populations?*



Let's go to review fast ANOVA and compare it with the techniques we will use in Design of Experiments.

Two-factor Experiments

- Two factors (inputs)
  - A, B
- Separate total variation in output values into:
  - Effect due to A
  - Effect due to B
  - Effect due to interaction of A and B (AB)
  - Experimental error

In a two factor ANOVA we will see the effects of two factors, independent variables, on the dependent variable, the answer we want to analyze, and determine is these factors generates different groups.

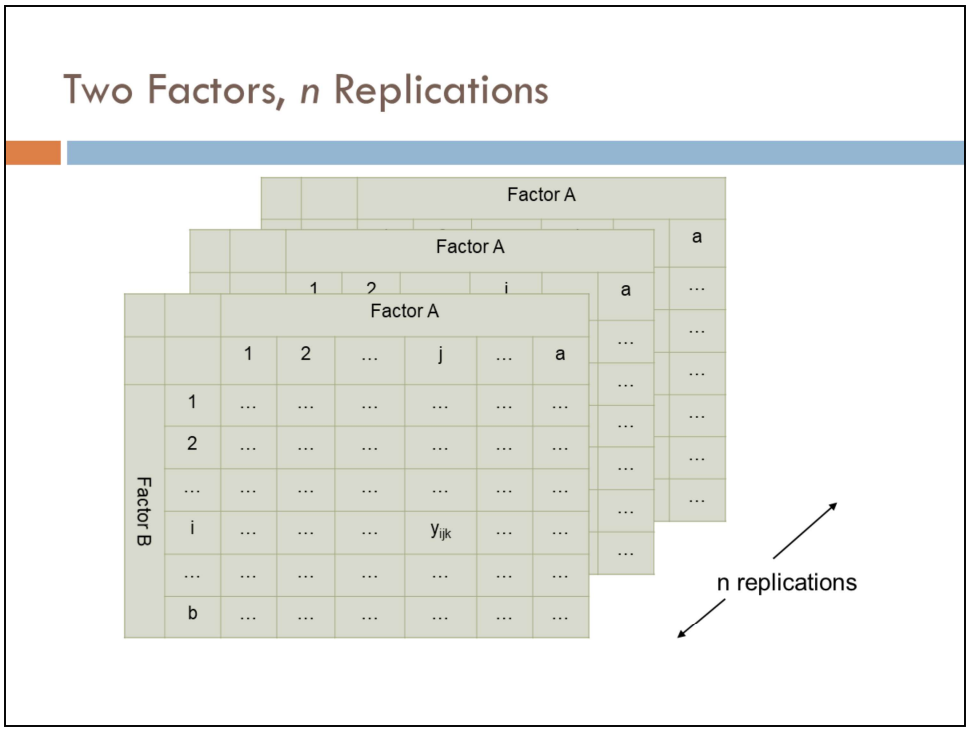
In the slide we can see, that the method can be used to calculate the effect due to A, the effect due to B and the effect due to the interaction between A and B. Also we will be able to calculate the experimental error.

Example – User Response Time

- A = degree of multiprogramming
- B = memory size
- AB = interaction of memory size and degree of multiprogramming

	B (Mbytes)		
A	32	64	128
1	1,125	1,105	1,075
2	1,26	1,225	1,18
3	1,405	1,33	1,25
4	1,75	1,725	1,35

Let's go to see an example. Here we have the results obtained by executing a program in different scenarios. Theses scenarios are defined by the combination of the levels of several factors, specifically two, the A, that is the degree of multiprogramming, and B, that is the memory size. On the table we can see the time needed to execute this program depending on the level for factor A, and factor B.



Notice that if one of the rules that define the behavior of the factors A and B, or any of other variable used in the model, is defined by a probability distribution, we must do replications. These replications are going to generate more tables like the one presented on the previous slide, and conceptually can be seen as the image we present on this slide. On the analysis we will use each one of this tables, as a new observation, for our analysis. Without the replications, obtaining different observation we cannot proceed further in the analysis.

One-factor ANOVA

- Each individual measurement is composition of
  - Overall mean
  - Effect of alternatives
  - Measurement errors

$$y_{ij} = \bar{y}_{..} + \alpha_i + e_{ij}$$

$\bar{y}_{..}$ = overall mean

α_i = effect due to A

e_{ij} = measurement error

In our well-known one factor ANOVA this will be the different parameters to consider in our calculus, while in the two-factor ANOVA.

Two-factor ANOVA

- Factor A – *a* input levels
- Factor B – *b* input levels
- *n* measurements for each input combination
- *abn* total measurements

We will expand the analysis with a new factor.

Two-factor ANOVA

- Each individual measurement is composition of
  - Overall mean
  - Effects
  - Measurement errors
  - Interactions

$$y_{ijk} = \bar{y}_{...} + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

$\bar{y}_{...}$ = overall mean

α_i = effect due to A

β_j = effect due to B

γ_{ij} = effect due to interaction of A and B

e_{ijk} = measurement error

That lead to a be able to calculate the main effect of each one of the two factors and its interactions.

Sum-of-Squares

- As before, use sum-of-squares identity

$$SST = SSA + SSB + SSAB + SSE$$

- Degrees of freedom
  - $df(SSA) = a - 1$
  - $df(SSB) = b - 1$
  - $df(SSAB) = (a - 1)(b - 1)$
  - $df(SSE) = ab(n - 1)$
  - $df(SST) = abn - 1$

And we will be able to calculate this using our well-known formulas.

Sum-of-Squares

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{...}\right)^2}_{SS_{Total}} = \underbrace{r\cdot b\cdot\sum_{i=1}^{a}\left(\bar{Y}_{i..}-\bar{Y}_{...}\right)^2}_{SS_A} + \underbrace{r\cdot a\cdot\sum_{j=1}^{3}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2}_{SS_B}$$
$$+\underbrace{r\times\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2}_{SS_{A\times B}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2}_{SS_{within}}$$

$$MS_{within} = SS_{within}\Big/df_{within}$$

Two-Factor ANOVA table

Source	Degrees of Freedom	SS	MS	F
A	a-1	SS_A	MS_A	MS_A / MS_{within}
B	b-1	SS_B	MS_B	MS_B / MS_{within}
$A \times B$	(a-1)(b-1)	$SS_{A\times B}$	$MS_{A\times B}$	$MS_{A\times B} / MS_{within}$
Within	ab(r-1)	SS_{within}	MS_{within}	
Total	abr-1	SS_{Total}		

From this we can build the table, that we will use to calculate the main elements, and that we will obtain in the well known statistical software like R.

Replications are needed

- If no replications, n=1
  - SSE = 0, no information regarding the errors in the measurements.
- Cannot separate effect due to interactions from measurement noise
- Must *replicate* each experiment at least twice

Notice that if we don't have replications, numerically makes no sense the calculus.

Replications are needed

- If n=1 (only one measurement of each configuration)
  - SSE = 0
- Since
  - SSE = SST – SSA – SSB – SSAB
- and
  - SSAB = SST – SSA – SSB

This is clear on the expressions.

Example

- Output = user response time (seconds)
- Want to separate effects due to
  - A = degree of multiprogramming
  - B = memory size
  - AB = interaction
  - Error
- Need replications to separate error

	B (Mbytes)		
A	32	64	128
1	1,125	1,105	1,075
2	1,26	1,225	1,18
3	1,405	1,33	1,25
4	1,75	1,725	1,35

On the example we propose, we need more replications, hence we repeat the experiment, with all the parameters equal.

Example

A	B (Mbytes) 32	64	128
1	1,125	1,105	1,075
	1,14	1,095	1,055
2	1,26	1,225	1,18
	1,24	1,245	1,15
3	1,405	1,33	1,25
	1,38	1,295	1,305
4	1,75	1,725	1,35
	1,125	1,105	1,075

Consider that we repeat the experiment two times, hence for each configuration we will obtain two values. On this table you can see the solution for this experiment that uses two replication for each scenario (configuration of the parameters).

We here do not consider if this two replications are enough for our analysis, it is clear that if the variability of the answers is huge, two replications will be not enough. We will discuss this later.

Two-Factor ANOVA table

	A	B	AB	Error
Sum of squares	SSA	SSB	$SSAB$	SSE
Deg freedom	1	1	1	$2^m(n-1)$
Mean square	$s_a^2 = SSA/1$	$s_b^2 = SSB/1$	$s_{ab}^2 = SSAB/1$	$s_e^2 = SSE/[2^m(n-1)]$
Computed F	$F_a = s_a^2/s_e^2$	$F_b = s_b^2/s_e^2$	$F_{ab} = s_{ab}^2/s_e^2$	
Tabulated F	$F_{[1-\alpha;1,2^m(n-1)]}$	$F_{[1-\alpha;1,2^m(n-1)]}$	$F_{[1-\alpha;1,2^m(n-1)]}$	

With the table for the two-factor ANOVA we can build the solution.

Example

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	0,842854	3	0,280951	460,2617	1,2E-12	3,490295
Columns	0,128802	2	0,064401	105,5034	2,43E-08	3,885294
Interaction	0,107915	6	0,017986	29,46473	1,65E-06	2,99612
Within	0,007325	12	0,00061			
Total	1,086896	23				

The solution for the problem is presented on this slide. One can see the p.value for the factor A, B and the interaction of A and B. Since the p.value is really small (lower than 0.05) we can conclude that here we haver different groups depending on A, B and the interaction of A and B. Analogous to say that A, B and their interaction affects the answer.

$n2^k$ Contrasts

- Effects of A, B and interactions.

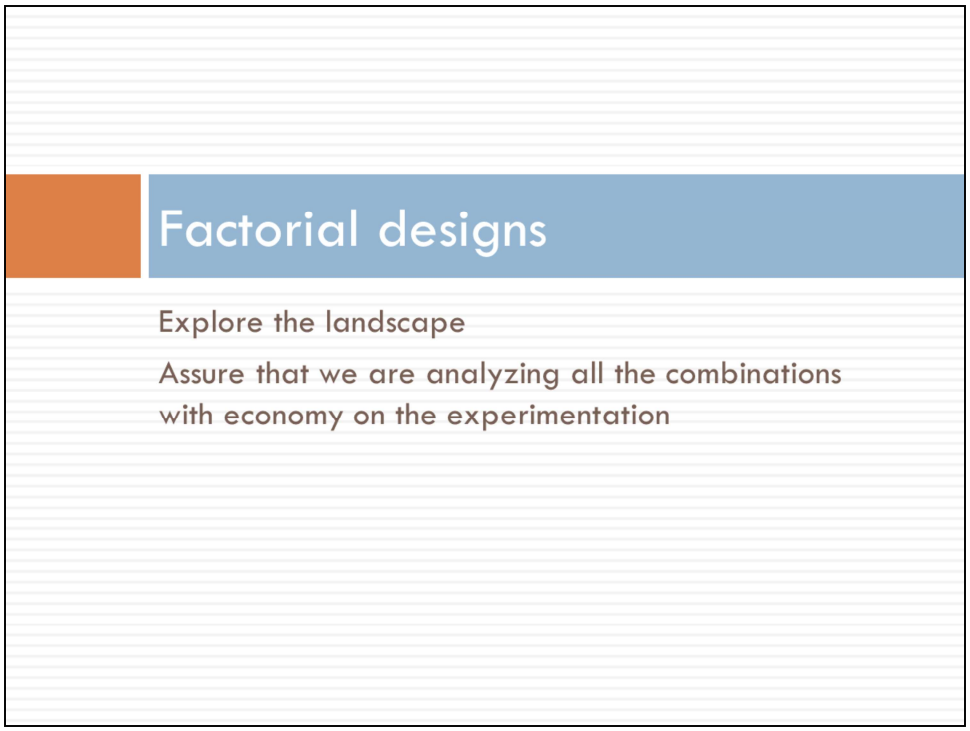
$$w_A = y_{AB} + y_{Ab} - y_{aB} - y_{ab}$$

$$w_B = y_{AB} - y_{Ab} + y_{aB} - y_{ab}$$

$$w_{AB} = y_{AB} - y_{Ab} - y_{aB} + y_{ab}$$

Notice that the main effects and the interactions can be calculated as is presented in this slide. If we want to calculate the main effect for the factor A, degree of multiprogramming, we can see the results we obtain in all the different combination. We keep the level for A fixed, and we calculate the mean of the two results, and we subtract the mean of the same but now considering the second level for "A". As we can see this is going to provide information regarding A, because in each part of the expression we use all the levels of the other factors, in that case just B.

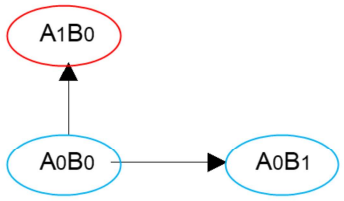
For the other expressions we will use the same schema. It is worth to mention that for the calculus of the interactions the expression calculates, in each part of the expression the different levels of each factor, analyzing precisely how this interaction happens.



On this section we will start with the first approach for the design of experiments. We expect to have no constrains on the number of factors to analyze, also, we expect to reduce the number of scenarios to compare and execute.

No factorial designs

- To fix two factors and modify all the levels of a third until find a good solution. Fixing this level, start the exploration for the other factors.
- Effect A: A1B0-A0B0.
- Effect B: A0B1-A0B0

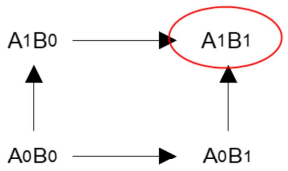


We can start with this approach. On this experiment we will analyze what happens on a system where we have two factors, a factor A with several levels {0,…} and a factor B, that also owns several different levels from {0,..]. The first scenario to consider, is a scenario where the factor A uses as a level 0, the same for the factor B. On this scenario we will obtain a answer. This answer will be good or bad, depending on the results we will obtain in other scenarios, hence we change the levels for the factors. We can start changing the level for the factor A, in the example to level 1. We are going to obtain a new answer. Then we can change the level for the factor B, obtaining a new answer. Once we have these three answer, one may ask what will be the next steep, and, considering that the second scenario (the red one) have the better answer, we can continue the exploration from these branch. Modifying the levels for A, but keeping B fixed. Also think about what happens if other factor exists, in the red scenario we will fix A and B levels and we will start working with C.

The problem arises because we will do not analyze in this approach, the interaction between A and B (or C). Hence we will lose the possibility to find a better answer, because the interaction is really good.

Factorial designs

- Take in consideration the interactions.
- A effect: $\frac{A_1B_0 - A_1B_1}{2} - \frac{A_0B_0 - A_0B_1}{2}$
- B effect: $\frac{A_1B_1 - A_0B_1}{2} - \frac{A_0B_0 - A_1B_0}{2}$



To solve this problem we will use factorial designs, where we will explore all the space in order to find the best answer. This approach is nice, however the number of scenarios to consider will growth exponentially.

Notice that with this approach we can calculate the effects and interactions of the factors we analyze. Notice also that in this approach we calculate the scenario with the interaction that we miss in the previous approach, in the red circle.

Factorial designs

- Controlling "k" factors.
- "l" levels for each factor ("li" levels for the l factor).
- $l_1 \cdot l_2 \cdot \ldots \cdot l_k$ experiments
- The easiest factorial design is the 2^k with li = 2 $\forall$i = 1,..,k.

On factorial designs we are going to control "k" factors, the factors that we want to analyze. To reduce the number of scenarios to analyze, we limit the number of possible levels of the factors to only "l", usually this l is equal to 2, defining the well-known experimental design schema 2k factorial designs.

A Problem

- *Full factorial design with replication*
  - Measure system response with all possible input combinations
  - Replicate each measurement n times to determine effect of measurement error
- k factors, v levels, n replications

  $\rightarrow n\, v^k$ experiments
- *Example:*
  - $k = 5$ input factors, $v = 4$ levels, $n = 3$
  - $\rightarrow 3(4^5) = 3{,}072$ executions!

Think in this problem, we want to analyze a system with 5 factors, that can be modified in four different ways, levels, each one of them. In this case, consider that we only do 3 replications.

With this configuration, the number of experiments is going to be 3072 different executions. A value that can be affordable or not, depending on the time and the resources needed for each one of the executions to do.

Fractional Factorial Designs: $n2^k$ Experiments

- Special case of generalized *m*-factor experiments
- Restrict each factor to two possible values
  - High, low
  - On, off
- Find factors that have largest impact
- Full factorial design with only those factors

Because we want to reduce the number of executions to perform, and since the number of factors, and replications cannot be changed, although we will see that the number of replications can be reduced a little later, the only alternative is to fix the levels we analyze for each factor.

It is common to fix the levels to two, because one can be a big value, the other a small value, allowing an exploration of the response surface, and allowing to understand the trend of the modification of the answers depending on the levels.

Finding Sum of Squares Terms

Sum of n measurements with (A,B) = (High, Low)	Factor A	Factor B
yAB	High	High
yAb	High	Low
yaB	Low	High
yab	Low	Low

To define this type of experiment we will use a table like the one presented on this slide. On the table, each row will represent a scenario, while the columns will represent the different factors that we will use.

In this case we have two factors, A and B and, because we constrain the levels to two, only 4 different scenarios.

On each cell we can see the level that we will use for each scenario.

On the selected scenario, on red, we can see that the level used for the Factor A is “Hight”, and for the Factor B is also “Hight”.

Contrasts for $n2^k$ with $k = 2$ factors

Measurements	Contrast		
	W_a	W_b	W_{ab}
y_{AB}	+	+	+
y_{Ab}	+	-	-
y_{aB}	-	+	-
y_{ab}	-	-	+

$$w_a = y_{AB} + y_{Ab} - y_{aB} - y_{ab}$$
$$w_b = y_{AB} - y_{Ab} + y_{aB} - y_{ab}$$
$$w_{ab} = y_{AB} - y_{Ab} - y_{aB} + y_{ab}$$

Usually in the books, and literature, the levels for the factors, are going to be represented as a plus sign, or with a minus sign. Plus usually represents the value that is proposed to improve the answer, despite we are maximizing or minimizing the value; the minus sign represents the current value, level, for the factor. Usually the current state of the factor if we have a system that exists.

2^k Matrix

Experiment	Factor 1	Factor 2	....	Factor k	Answer
1	-	-		-	R1
2	+	-		-	R2
3	-	-		-	R3
4	+	-		-	R4
5	-	+		-	R5
6	+	+		-	R6
2^k	+	+		+	$R2^k$

If we have more factors, we will add more columns and, the number of scenarios to consider will increase at a rate of 2 to the number of factors we analyze.

Important Points

- Experimental design is used to
  - Isolate the effects of each input variable.
  - Determine the effects of interactions.
  - Determine the magnitude of the error
  - Obtain maximum information for given effort
- Expand 1-factor ANOVA to *k* factors
- Use $n2^k$ design to reduce the number of experiments needed
  - But loses some information
  - Useful to underline the tendency with economy of experiments.

Here you have some important point to consider at this point. Notice that the number of factor to be analyzed on a 2k factorial design is not a problem from the point of view of the technique, but from the point of view of the tie needed to obtain the answers.

Exercise 1

- We have on a factory three different Machines:
  - A, with speed from 2 to 10
  - B, with speed of 2 and 3
  - C, with speed of 2 but that can be changed for other machine with a speed of 3 for 1000€.
- Define a table for an 2^k experimental design that allows to analyze this.

Try this exercise. The machines in this scenario are the factors we will analyze, and the levels are proposed on the slide. Notice that we will be use a 2k factorial design, hence only two levels for factor will be used. What factors will be selected.

Please, stop the video and try it by yourself. Drawn the experimental table and define the levels you will use.

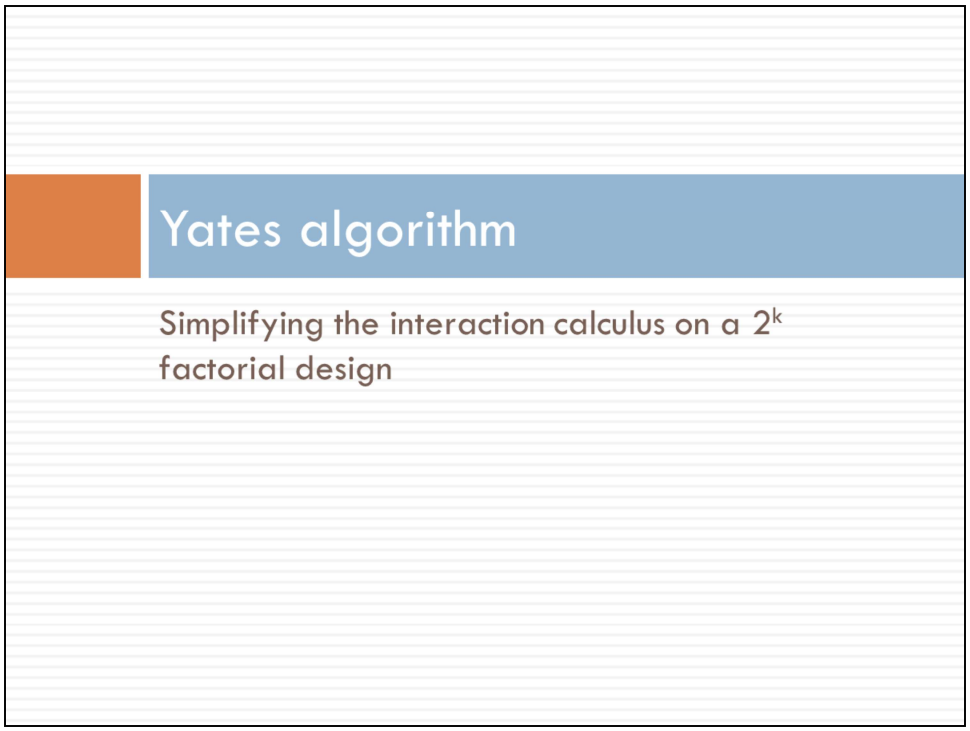
Solution

	A	B	C	Answer
1	- (means 2)	-(means 2)	- (means 2)	
2	-	-	+ (means 3, 1000€)	
3	-	+(means 3)	-	
4	-	+	+	
5	+ (means 10)	-	-	
6	+	-	+	
7	+	+	-	
8	+	+	+	

This is the solution for the proposed exercise. Here we see that we have three different factors, hence three different columns. For each column we see the combinations of the levels for the factors we have.

Notice that in the table we add the meaning of the sign, minus and plus, for each factor.

Did you do the definition of the two factorial design experiment correctly?



Notice that in the approach we follow at this point the calculus of the main effects and interactions rely on the expressions we present previously. If the number of factors we will use are huge, then the possibility to add an error increases. On this section we will present a method, the Yates algorithm that will simplify this interactions calculus.

2^k factorial designs

Advantages

- Determination of the tendency with experiments economy (smoothness).
- Possibility to evolve to composite designs (local exploration).
- Basis for factorial fractional designs (rapid vision of multiple factors).
- Easy analysis and interpretation.

We will start reviewing the main advantages of the 2k factorial designs.

First is the determination of the tendency of the answer with experiments economy (smoothness), using only two levels reduce the number of scenarios to consider.

Next, we have the possibility to evolve to composite designs, using local exploration for some factors and levels.

This will represent a basis for factorial fractional designs that allows a rapid vision of multiple factors.

And finally, and interesting, allows an easy analysis and interpretation.

2^k Matrix example

Experiment	A	B	C	Answer
1	-	-	-	60
2	+	-	-	72
3	-	+	-	54
4	+	+	-	68
5	-	-	+	52
6	+	-	+	83
7	-	+	+	45
8	+	+	+	80

Here you can see table for a 2k factorial design with three factors, on the last column we see the result we obtain. Consider that this result is obtained not from a single execution of the scenario, but for a set of replications although the values are integer values. Later we will see how to assure that this number of replications are correctly executed and are enough.

Effects calculus

$$Efect \quad A = \frac{A_1B_0 - A_1B_1}{2} - \frac{A_0B_0 - A_0B_1}{2}$$

$$Efect \quad B = \frac{A_1B_1 - A_0B_1}{2} - \frac{A_0B_0 - A_1B_0}{2}$$

$$\text{Main effect} = \bar{y}_+ - \bar{y}_-$$

Remember that the expressions we will use in this analysis will be, for the main effect of factor A and B, the expressions presented on this slide. In this case is for a 2 to 2 factorial design.

Effects calculus example

$$Main\ effect = \bar{y}_{+} - \bar{y}_{-}$$

$$A = \frac{72+68+83+80}{4} - \frac{60+54+52+45}{4} = 23$$

$$B = \frac{54+68+45+80}{4} - \frac{60+72+52+83}{4} = -5$$

$$C = \frac{52+83+45+80}{4} - \frac{60+72+54+68}{4} = 1.5$$

Adding the values we know and using the expressions for a 3 to 2 factorial design, notice that the divisor now is 4, we obtain this expressions that express the main effect for factor A, B and C.

Interactions for 2 and 3 factors

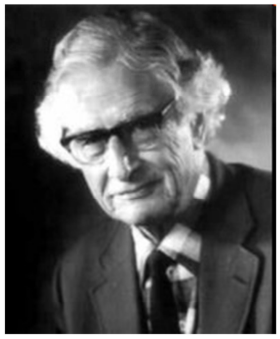
$$AC = \frac{y_1 + y_3 + y_6 + y_8}{4} - \frac{y_2 + y_4 + y_5 + y_7}{4} = 10$$

$$ABC = \frac{y_{21} + y_3 + y_5 + y_8}{4} - \frac{y_1 + y_4 + y_6 + y_7}{4} = 0.5$$

In this slide we can see the interactions between the factor A and C, and the interactions between A, B and C.

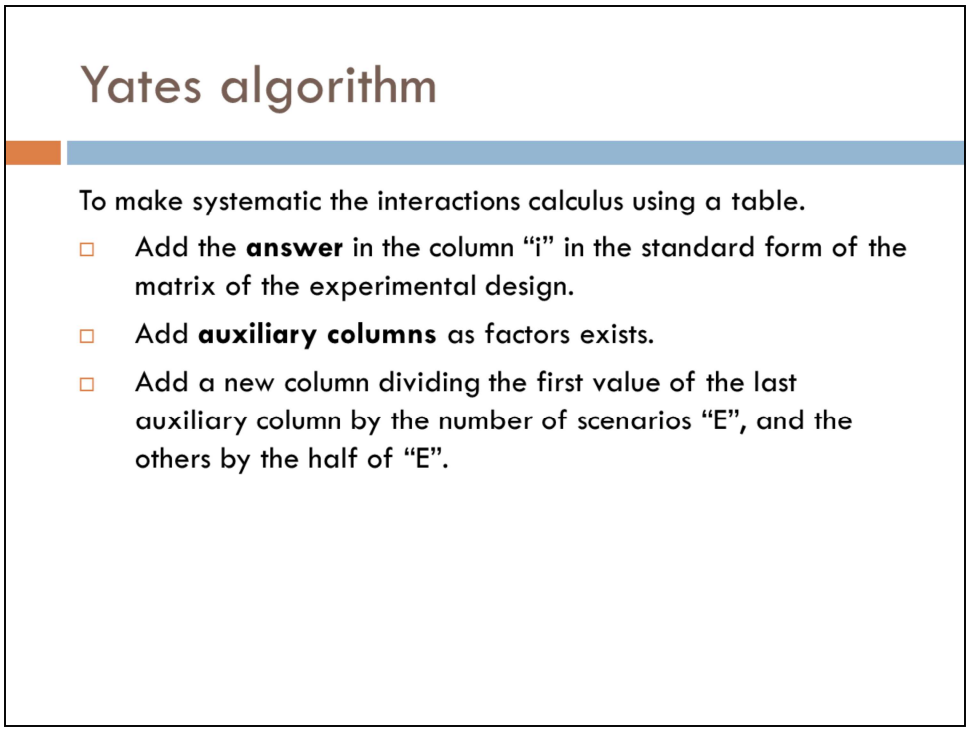
One can notice that this expressions are not difficult but, increases the complexity with the number of factors to analyze. Hence it will be desirable to use a method that automatically calculates this main effects and interactions.

Frank Yates



- A pioneer of the Operation research of the s.XX.

Hopefully, Frank Yates proposed a method to calculate in a systematic way the main effects and interactions. This method is the method that we will use if we are analyzing a 2k factorial design.



The Yates algorithm works as we present on this slide. First, we add the answer column to the matrix of the experimental design, as we see on the previous slides. Remember, we must assure that the number of replications of the answers is enough. We will add, next one auxiliary column for each factor we have. As an example, if we have 3 factors, we will add 3 new auxiliary columns.

Then we will add a new column that will be populated with the value of the last auxiliary column divided by a specific value. The first cell divided by the number of the scenarios, the other divided by half of the number of scenarios. As an example, if we have 8 different scenarios, the first cell will be divided by 8, while the others by 4. As you can imagine, this is going to reproduce systematically the expressions we use to calculate the main factors and interactions. The first cell will represent the mean of all the scenarios we execute, and a value, we will use as a reference for the analysis of main effects and interactions.

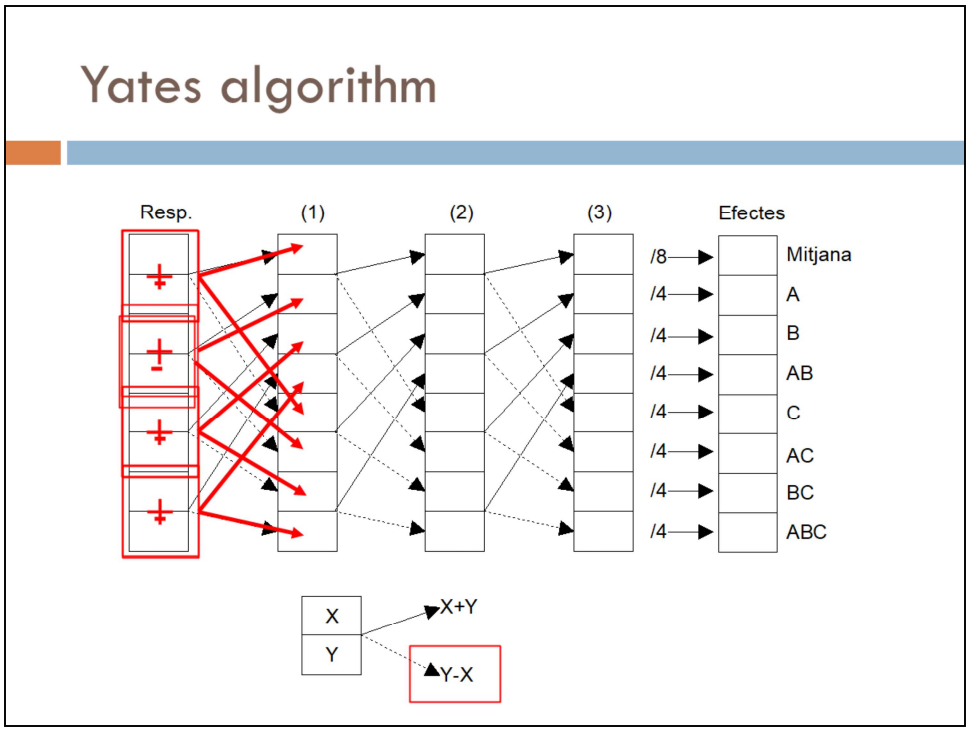
Yates algorithm

- In the last column the first value is the mean of the answers, the last values are the effects.
- The correspondence between the values and effects is done through localize the + values in the corresponding rows of the matrix. A value with a single + in the B column is representing the principal effect of B. A row wit two + on A and C corresponds to the interaction of AC, etc.

Since in the last column, the first cell is going to represent the mean, and this is going to be use as a reference value, it is strongly recommended to write in the first row an scenario that represents the base case,, that means, the case that currently exist if we are modeling something that exist, or the worse scenario, in a more general system, a system that maybe does not exist.

This represents that the first row contains all the levels to "-".

The values that we will obtain in the next rows are showing information regarding main effects or interactions depending on where we will find a "+" sign. If we analyze a row, where the "+" sign only appears on the column of the first factor, the value we will see on the last position of the Yates table is the main effect for this factor. If we see a row with two ore more "+" signs, this implies that we are analyzing the interactions of these factors.



To calculate the values of the intermediate columns, the auxiliary columns we add, the Yates algorithm follows this rule. The first half of the column is going to be filled by adding the first two cells of the prior column. The second half of the column is going to be filled subtracting the second value minus the first value, for the first position, the fourth position minus the third position, and so on.

In the slide you can see an example for a 2 to 3 factorial design.

Notice that when we do the subtraction in the second part of the column, we change the order of the values, we subtract the first cell from the second one.

We repeat this schema for all the auxiliary columns we have.

Yates algorithm example

Exp.	A	B	C	Answer	(1)	(2)	(3)	div.	effect	Id
1	-	-	-	60	132	254	514	8	64.25	Mean
2	+	-	-	72	122	260	92	4	23.0	A
3	-	+	-	54	135	26	-20	4	-5.0	B
4	+	+	-	68	125	66	6	4	1.5	AB
5	-	-	+	52	12	-10	6	4	1.5	C
6	+	-	+	83	14	-10	40	4	10.0	AC
7	-	+	+	45	31	2	0	4	0.0	BC
8	+	+	+	80	35	4	2	4	0.5	ABC

This is an example of the Yates algorithm. Here you can see that for this 2 to 3 factorial design we have the answers for each scenario, and from this answers, we calculate the values for the auxiliary columns. Finally we divide the value for the appropriate number, that we can see on the div column, obtaining the effect.

Let's go to review the results we obtain in the last column.

As one can see, the first row is representing the base case, all the levels are set to "-". The value that we obtain in this first row is 64.25, that is the mean of the values we obtain in all the scenarios. This is a value that can be considered a reference. Obtaining a value that is from the same magnitude of this implies that we are obtaining a considerable effect for the factor or the interactions. As an example, on the second row the value we obtain is 23, and we can see that the only "+" sign corresponds to the column A, hence here we are analyzing the main effect for A.

On the last row, the "+" sign is on the tree factors, hence on this last row we are analyzing the interaction, and remember only the interaction, of the three factors.

Consider in this example that we want to maximize the value, hence the effect of A, that is 23, and we can compare it with 64.25, is important, C is also important, but not because alone is going to improve the answer, but because

with the interaction with A, obtaining a value of 10, is also relevant to increase the value.

In the other side, B is not good to improve the value of the answer.

Wooden industry example

- Wooden industry that allows to reduce the cost.
- 4 variables to consider
  - Change the light to natural light (open the ceiling).
  - Increase the speed of the machines.
  - Increase the lubricant use.
  - Increase the working space.

Now we are going to do an exercise. Try this example, in it we are going to see what happens in a wooden factory. We want to MINIMIZE the time needed to work with the wooden pieces. We know that we have four factors to consider.

We can change the light to natural, opening the windows in the ceiling, we can also increase the speed of the machines, the saws, we can increase the lubricant use to avoid that the saws become stopped due to the high temperature, and finally, we can buy a new space in order to increase the space we have to work.

Wooden industry example

Comb.	1	2	3	4	Description	obs.
(1)	-	-	-	-		71
a	+	-	-	-	Natural light	61
b	-	+	-	-	Increase the speed of the machines	90
ab	+	+	-	-		82
c	-	-	+	-	Increase the use of lubricant	68
ac	+	-	+	-		61
bc	-	+	+	-		87
abc	+	+	+	-		80
d	-	-	-	+	Increase the working space.	61
ad	+	-	-	+		50
bd	-	+	-	+		89
abd	+	+	-	+		83
cd	-	-	+	+		59
acd	+	-	+	+		51
bcd	-	+	+	+		85
abcd	+	+	+	+		78

This table shows the results from the different scenarios we have, a 2 to 4 factorial design. The results are on the last column. Apply the Yates algorithm to calculate the main effects and interactions. What is the more important factors to consider to improve the behavior of the system?.

Stop thee video and do the exercise, the solution, on the next slides.

Wooden industry example

Comb.	obs.	1	2	3	4	Effect	Description
(1)	71						
a	61						
b	90						
ab	82						
c	68						
ac	61						
bc	87						
abc	80						
d	61						
ad	50						
bd	89						
abd	83						
cd	59						
acd	51						
bcd	85						
abcd	78						

The first think to do is to write this table, adding an auxiliary column for each one of the factors we have, hence 4 auxiliary columns. At this point we are able to calculate the Yates algorithm.

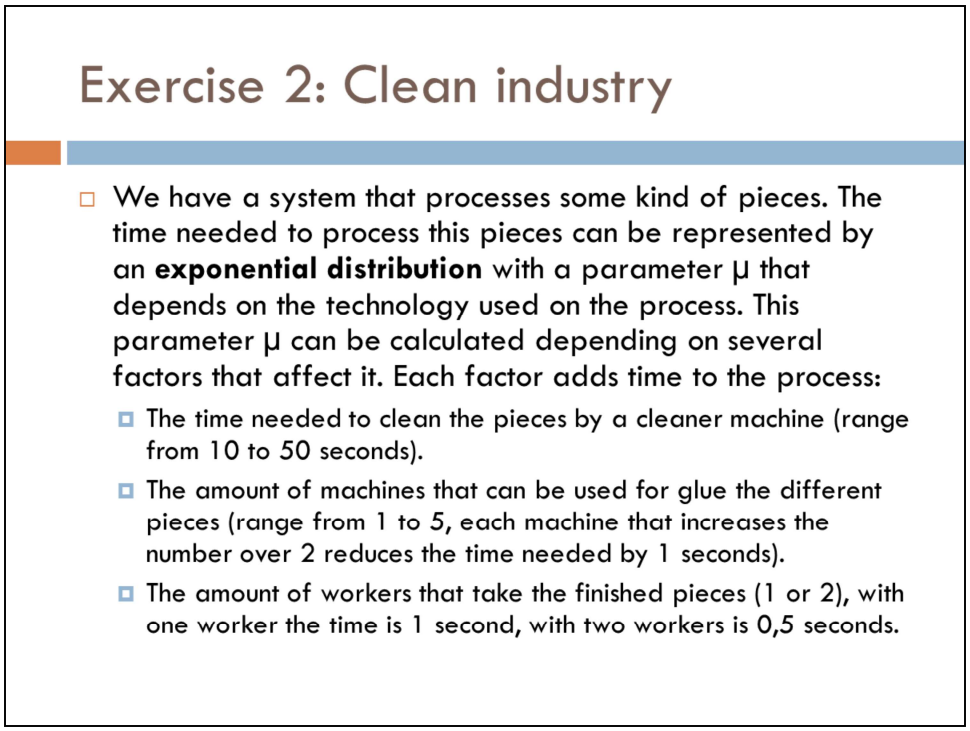
Wooden industry example

Comb.	obs.	1	2	3	4	Effect	Description
(1)	71	132	304	600	1156	72,25	Mean
a	61	172	296	556	-64	-8	A
b	90	129	283	-32	192	24	B
ab	82	167	273	-32	8	1	AB
c	68	111	-18	78	-18	-2,25	C
ac	61	172	-14	114	6	0,75	AC
bc	87	110	-17	2	-10	-1,25	BC
abc	80	163	-15	6	-6	-0,75	ABC
d	61	-10	40	-8	-44	-5,5	D
ad	50	-8	38	-10	0	0	AD
bd	89	-7	61	4	36	4,5	BD
abd	83	-7	53	2	4	0,5	ABD
cd	59	-11	2	-2	-2	-0,25	CD
acd	51	-6	0	-8	-2	-0,25	ACD
bcd	85	-8	5	-2	-6	-0,75	BCD
abcd	78	-7	1	-4	-2	-0,25	ABCD

This is the table with the Yates algorithm calculated. As you can see, the last column describes if we are analyzing a main effect or an interaction.

It is worth not mention that A, C and D decreases the time, but B increases the answer, remember that we want to minimize it, hence it seems that modify B is not a good idea. Remember that B represents increase the speed of the machines. In this case, that is a simplification of a real case, re workers become afraid of a saw working at double speed and takes extra prevention to work with them, using more time.

Did you notice that the interactions almost have no effect?.



This is a second example. Here we are going to work with Excel to define the model, the answer and the table that represents the Yates algorithm. This Excel is provided as a supplementary material, but don't review it yet, try to do it by yourself.

The model, that represents the factory is a huge simplification, since is only an exponential distribution. The parameter of this exponential distribution, the parameter mu is going to be calculated, depending on the values for the factors we use.

Three factors can be considered: The time needed to clean the pieces by a cleaner machine, that is a range from 10 to 50 seconds. The amount of machines that can be used for glue the different pieces, again, a range from 1 to 5. In this case every machine over 2, as an example using 2, 3 or more machines), reduces the time needed by 1 seconds. Finally, the last factor are the workers we have, we can use one or two. The parameter of the exponential distribution is going to be increased with one worker by one second, while if we are using two workers, we are working faster, by 0,5 seconds.

Perform a DOE for the proposed system

- Set the objectives.
- Select the process variables.
- Define an experimental design.
- Execute the design.
- Check that the data are consistent with the experimental assumptions.
- Analyze and interpret the results, detect effects of main factors and interactions.
- Remember:

$$r = 1 - e^{-\alpha \cdot x} \Rightarrow x = \frac{\ln(1-r)}{-\alpha} = \frac{\ln(r)}{-\alpha}$$

To work in this example we can review wat will be the different steeps to do. This are presented on this slide.

Set the objectives.

Select the process variables.

Define an experimental design.

Execute the design.

Check that the data are consistent with the experimental assumptions.

Analyze and interpret the results, detect effects of main factors and interactions.

Remember also tht the expression that we will use to generate an exponential distribution is presented at the end of the slide. The parameter R is a random number, RANDOM in Excel, from 0 to 1, while the alpha is the parameter of the exponential distribution we calculate depending on the selection of the levels on the factors, differs in each scenario.

The result that we will obtain from this expression, is the time needed to do the operation, small value is better.

Perform a DOE for the proposed system

- Set the objectives.
  - Detect the effects and the interactions of the three main factors
- Select the process variables.
  - Cleaner, Machines, workers.
- Define an experimental design.
  - We define a 2k experimental design.
- Execute the design.
  - Using Excel we "simulate" the behavior for each proposed model.
- Check that the data are consistent with the experimental assumptions.
  - Independence, homoscedasticity, normality, etc?.
- Analyze and interpret the results, detect effects of main factors and interactions.
  - Done on the Excel

Applying this to our problem we obtain this. Read this slide and notice that exponential distribution acts as a complete model of our factory.

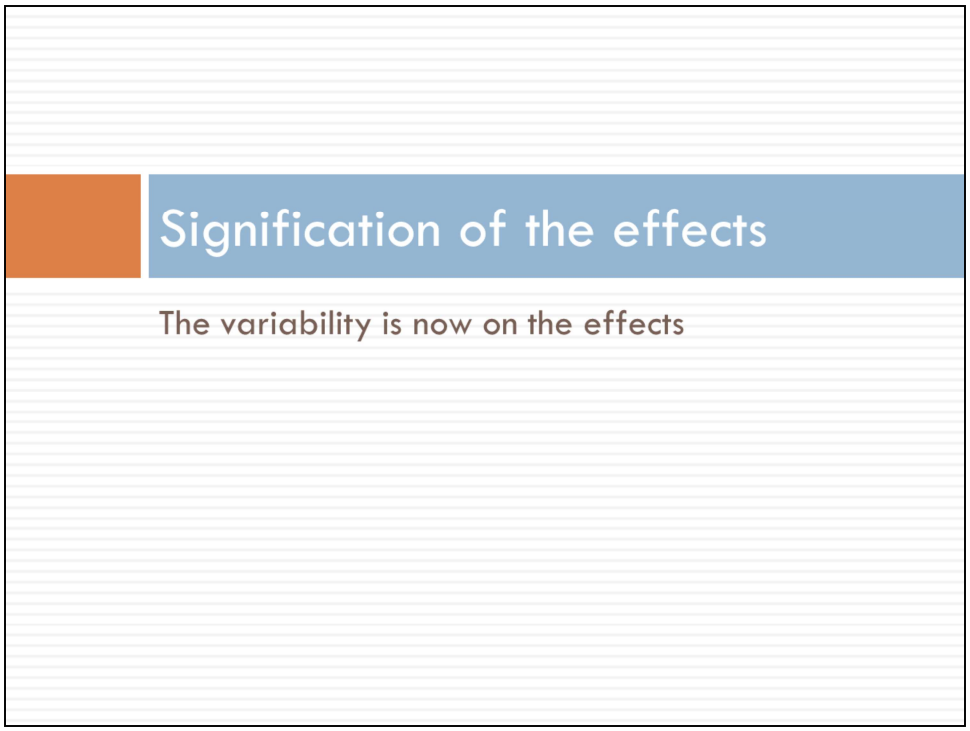
Answer

											Yates				
Cleaner	Machines	Workers	VALUES			μ	1/μ	x1	x2	mean					
-	-	-	50	0	1	51	0,019607843	20,8852	20,2611	20,57315	55,67501	206,082	246,6539	**30,83174**	Mean
-	-	+	50	0	0,5	50,5	0,01980198	33,83603	36,36768	35,10185	150,407	40,57198	-8,42686	**-2,10671**	Workers
-	+	-	50	-4	1	47	0,021276596	9,909855	142,0044	75,95712	25,41431	13,02142	84,47532	**21,11883**	Machines
-	+	+	50	-4	0,5	46,5	0,021505376	17,76593	131,1337	74,44984	15,15767	-21,4483	-2,71458	**-0,67864**	Machines*Workers
+	-	-	10	0	1	11	0,090909091	42,759	0,040152	21,39958	14,5287	94,73195	-165,51	**-41,3775**	Cleaner
+	-	+	10	0	0,5	10,5	0,095238095	5,702481	2,326982	4,014731	-1,50728	-10,2566	-34,4697	**-8,61743**	Cleaner*Workers
+	+	-	10	-4	1	7	0,142857143	10,48071	8,740406	9,610556	-17,3848	-16,036	-104,989	**-26,2471**	Cleaner*Machines
+	+	+	10	-4	0,5	6,5	0,153846154	5,977532	5,116704	5,547118	-4,06344	13,32141	29,35739	**7,339348**	Cleaner*Machines*Workers

$$r = 1 - e^{-\alpha \cdot x} \Rightarrow x = \frac{\ln(1-r)}{-\alpha} = \frac{\ln(r)}{-\alpha}$$

Finally the Excel must look like this. Notice that we do two replications of the experiment. Remember that if you are using the RANDOM instruction in Excel, every time that you modify something in the Excel a new random number stream is generated. If all is correct, the conclusions are not going to be dependent on the streams, but is this the case?, are two replications enough?. We will discuss this on the next part of the course.

This first part of the Experimental Design section finish here. Try to do it by yourself and analyze what happens with the conclusions.



The effects depends on the number of replications we do, not for the conclusions if this number is enough, but to reduce the confidence interval they have. On this section we will learn how to calculate the confidence interval for the effects we obtain in our analysis.

Have the effects significations?

- The effects have been calculated using the answers that owns variability.
  - The variability goes to the effects.
    - $\hat{A} = \bar{y}(+) - \bar{y}(-)$

Notice that the effects have been calculated through the answers, and the answers owns a variability. Consider a simple linear regression model, we are obtaining a confidence interval as an answer, the same happens in a simulation model. This implies that the effects also owns a variability that can be calculated and analyzed.

We have replications

- Calculation of the variance
  - assuming that the amount of replications in each scenario is equal.
    - $S_R^2 = \frac{s_1^2+s_2^2+..+s_N^2}{N}$
  - assuming that the amount of replications in each scenario is different.
    - $S_R^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+..+(n_N-1)s_N^2}{n_1+n_2+..+n_N-N}$,
  - Where $N=2^k$, and $Nt=v*2^k$.

We can calculate the variance for each experiment, using one expression if the number of replications in each scenario is the same, or assuming that the number of replications in each scenario is no the same. This second case is more general, since changing the levels, hence the scenarios, can cause an increase on the variability of the scenario.

We have replications

- Variance for an effect
  - $\widehat{effect} = \overline{y_+}\text{-}\overline{y_-}$
  - Where $\overline{y_+}$ is the mean of the Nt/2 scenarios (+ sign), and $\overline{y_-}$ is the mean of the Nt/2 scenarios (- sign).
- The calculus is:
  - $v(\widehat{effect}) = v(\overline{y_+} - \overline{y_-}) = \frac{S^2}{N_T/_2} + \frac{S^2}{N_T/_2} = \frac{4S^2}{N_T}$

Then we can calculate the variance of the effects, calculating the mean of the scenarios with the “plus” sign, minus the scenarios with the “minus” sign. The calculus is summarized on the last expression of this slide.

Example

x1	x2	Mean	S2			
60	64	62	8			
72	74	73	2			
54	55	54,5	0,5			
68	70	69	2		v(efect)	12,1942
52	54	53	2			3,492019
83	87	85	8			
45	50	47,5	12,5			
80	85	82,5	12,5			
		S2=24,38839				

This is an example of the calculus. Notice that we have here two replications, on the first two columns of the table, and the mean of this two replications. With this we can calculate the variance, on the last column, and the variability of the effect, on the cell v(effect) that defines a confidence interval from 12.1 to 3.4

Signification test

- Hypothesis testing
  - H_0: effect=0
  - H_1: effect≠0
- Statistic of the test
  - $t = \frac{\widehat{effect} - 0}{S_{effect}}$
  - t-Student with $n_1+n_2+..+n_n$-N degrees of freedom ($N=2^k$)
- Calculus of the I.C.
  - $\widehat{effect} \pm x * S_{effect}$

With this we can do a signification test, to detect if zero belongs to the interval.