Replications

Calculate the number of replications needed for the Exercise 2

											Yates					
Cleaner	Machines	Workers	VALUES			μ	1/μ	x1	x2	mean						
-	-	-	50	0	1	51	0,019607843	8,50041	70,78097	39,64069	91,98358	155,1265	183,0445	**22,88056**	Mean	
-	-	+	50	0	0,5	50,5	0,01980198	78,80559	25,88018	52,34289	63,14288	27,918	-41,2648	**-10,3162**	Workers	
-	+	-	50	-4	1	47	0,021276596	108,3006	4,530386	56,41547	12,03321	-36,9859	-24,9891	**-6,24728**	Machines	
-	+	+	50	-4	0,5	46,5	0,021505376	4,652516	8,802295	6,727405	15,88479	-4,27889	-72,239	**-18,0597**	Machines*Workers	
+	-	-	10	0	1	11	0,090909091	1,966561	7,281746	4,624153	12,70219	-28,8407	-127,208	**-31,8021**	Cleaner	
+	-	+	10	0	0,5	10,5	0,095238095	7,493775	7,324339	7,409057	-49,6881	3,851583	32,70698	**8,176745**	Cleaner*Workers	
+	+	-	10	-4	1	7	0,142857143	15,95828	6,990312	11,47429	2,784904	-62,3903	32,69229	**8,173071**	Cleaner*Machines	
+	+	+	10	-4	0,5	6,5	0,153846154	0,557606	8,263392	4,410499	-7,06379	-9,8487	52,54156	**13,13539**	Cleaner*Machines*Workers	

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

Cleaner	Machines	Workers	x1	x2	Mean	Confidence interval (alfa=0.05)	n* 5%	n* 6%
-	-	-	20,885	20,261	20,57	3,964	29,71	20,628
-	-	+	33,836	36,368	35,10	16,08	168,01	116,67
-	+	-	9,9099	142	75,95	839,18	97.654,01	67815,28
-	+	+	17,766	131,13	74,45	720,21	74.869,67	51992,82
+	-	-	42,759	0,0402	21,40	271,39	128.673,31	89356,46
+	-	+	5,7025	2,327	4,01	21,44	22.825,53	15851,06
+	+	-	10,481	8,7404	9,61	11,05	1.059,13	735,50
+	+	+	5,9775	5,1167	5,55	5,46	777,56	539,97

Bank model

A bank is planning the requirements for its ATMs, in the frame of a new expansion. There are places for up to six ATMs, not all these places have to be used. You can buy three types of ATM: ATM generals (to give cash, balances, mini statements and change PIN), cash machines for the payment of money and ATMs that provide the full statements of the account. The Bank has a policy that customers would not have to wait more than 5 minutes in *most* cases (usually performed with the 99% of cases).

Answer

For the problem of the Bank, we select the following variables as factors, determining the next factorial design. We are interested in analyzing the service time of the ATM.

	GENERAL ATM. (A)	INCOME ATM. (B)	MANAGEMENT ATM. (C)	ANSWER
E1	-	-	-	8.7
E2	-	-	+	8.7
E3	-	+	-	8.7
E4	-	+	+	8.7
E5	+	-	-	1.4
E6	+	-	+	1.4
E7	+	+	-	1.4
E8	+	+	+	1.4

Where "+" means 2 servers while "-" one.

Calculate the effects of (A), (B) and (C).

Bank model main effects and interactions calculus

Calculate the main effects of the factors (A), (B) and (C) for the Banc model with and without using the Yates algorithm.

Answer

To calculate the main effects of (A):

$$Main\_Effect_A = \frac{(E_5 - E_1) + (E_6 - E_2) + (E_7 - E_3) + (E_8 - E_4)}{4}$$
$$= \frac{(1.4 - 8.7) + (1.4 - 8.7) + (1.4 - 8.7) + (1.4 - 8.7)}{4} = \frac{-29.2}{4} = -7.3$$

To calculate the main effects of (B):

$$Main\_Effect_B = \frac{(E_3 - E_1) + (E_4 - E_2) + (E_7 - E_5) + (E_8 - E_6)}{4}$$
$$= \frac{(8.7 - 8.7) + (8.7 - 8.7) + (1.4 - 1.4) + (1.4 - 1.4)}{4} = \frac{0}{4} = 0$$

To calculate the main effects of (C):

$$Main\_Effect_C = \frac{(E_2 - E_1) + (E_4 - E_3) + (E_6 - E_5) + (E_8 - E_7)}{4}$$
$$= \frac{(8.7 - 8.7) + (8.7 - 8.7) + (1.4 - 1.4) + (1.4 - 1.4)}{4} = \frac{0}{4} = 0$$

In view of the results it can be concluded that the factor (A) GENERAL ATM, is the only determining factor to be taken into account to improve the behavior of the system.

We can calculate the same using the Yates algorithm.

	General ATM. (A)	Income ATM. (B)	Management ATM. (C)	Answer	Auxiliary columns			Effects	
1	-	-	-	8.7	17,4	34,8	40,4	5,05	Mean
2	-	-	+	8.7	17,4	5,6	0	0	(C)
3	-	+	-	8.7	2,8	0	0	0	(B)
4	-	+	+	8.7	2,8	0	0	0	(BC)
5	+	-	-	1.4	0	0	-29,2	-7,3	(A)
6	+	-	+	1.4	0	0	0	0	(AC)
7	+	+	-	1.4	0	0	0	0	(AB)
8	+	+	+	1.4	0	0	0	0	(ABC)

The effect is due to A (General ATM), that is to say, the number of ATM's that we will use. Therefore it is recommend to increase the number of ATM's.

9 factors and 20K€

We have a limited budget to analyze the 9 different factors that can be considered to improve the behavior of the system. Each individual experiment costs 100€ and we have a total budget of 20.000€ to be destined to experimentation. Define an experiment design, with this constraint, considering that we need at least 3 replications for each experiment.

Answer

If we can spend all the budget in an initial experimentation without considering the possible inconvenient that can appear due to possible high variability of certain variables (that lead us to increase the number of needed replications that is initial considered by 3), the maximum amount of experiment is 64:

$$64\ experiment * 3\ replications * 100€ = 19200€$$

This design doesn't allow a complete experimentation considering all the factors we have on our model, we have 9 factors that implies a 2^9 experiments meaning 512 experiments with 3 replications each one of them.

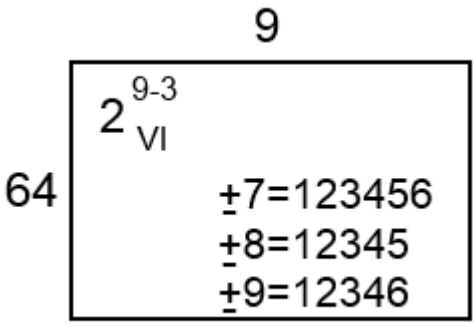
This implies that we need to reduce the number of variables to be considered using a fractional factorial design.

To do this it is needed to define a confounding pattern and select those factors that are going to be confounded.

The design will be defined as: 2^{9-3}

Hence 3 confounding patterns must be defined.

We select in our case the last four variables to be confounded using the first 6 variables.



Other possible resolution of the problem can be based on the PB design.

SEO Part one

We are going to face with a SEO problem; we try to understand what are the main factors that determines the position of a website on a search result. The main concern is the huge volume of data we need to consider. We are facing a table with more than 300.000 observations (that represents the results for a specific search), and more than 200 variables (factors) in each individual (observation).

The structure of the data is similar to the presented on this table.

Keyword	Pos	url	Doc_length (words)	AdSense	chars	(..)
anniversary gift	1	http://url1.com	1451	0	42786	..
anniversary gift	2	http://url2.com	1887	1	46354	..
anniversary gift	3	http://url3.com	1202	3	18632	..
anniversary gift	4	http://url4.com	1503	0	59325	..
anniversary gift	5	http://url5.com	1832	2	65323	..
..	..	..	..	..	..	..
Halloween costumes	1	http://url11.com	1451	0	42786	..
Halloween costumes	2	http://url12.com	1887	1	46354	..
Halloween costumes	3	http://url13.com	1202	3	18632	..
Halloween costumes	4	http://url14.com	1503	0	59325	..
Halloween costumes	5	http://url15.com	1832	2	65323	..
..	..	..	..	..	..	..

We want to analyze it to understand if is possible to detect what are the main factors (and the values needed) in order to improve our position in a search engine regarding a specific keyword.

What are the techniques you are going to use here? **Detail clearly every step to be done**.

Answer

Here we want to predict the position of a website due to a search. In this case a regression model could be useful if we can determine the factors that imply this position. Since the dimensionality is very high it is imperative to use some dimensionality reduction techniques, like PCA or clustering in order to make this analysis feasible.

The specific steps to be done are:

1. **Goals of the analysis:** To identify the main factors that determine the position of a website regarding a keyword.
2. **Design of the analysis**. It is needed to reduce the dimensionality of the problem. We can start with a PCA and clustering to define different subsets of the data to work. We can use first PCA to discard all variables that are not important in our analysis. Later we can define a regression or a simulation model to predict the position.
3. **Hypotheses of the Analysis**. Since we use regression analysis, normality, homoscedasticity and independence on the data must be assured. Here we must decide with what to do with missing data.
4. **Analytical procedure.** We estimate the model and we evaluate the fit to the data. In this step may appear unusual observations (outliers) or influential whose influence on the estimates and the goodness of fit must be analyzed.
5. **Interpretation of the results.** Such interpretations can lead to additional specifications or model variables with which you can return back to steps 3) and 4)
6. **Analysis Validation.** Is to establish the validity of the results obtained by analyzing whether the results, obtained with the sample, is generalized to the population from which it comes. This sample can be divided into several parts in which the model is re-estimated and the results are compared. Other techniques that can be used here are resampling techniques (jackknife and bootstrap)

SEO Part two

With the previous information, you will be able to define a model but, a detailed analysis makes clear that is needed to define a complex simulation model to obtain accurate results. This model needs more than one hour to complete a single replication.

With this model, we can parametrize our websites to try to find the best combinations to improve the position of the website in the search engines. Although the number of factors in this simulation model is just 10 (not the 200 previously analyzed) the amount of experiments to be conducted will be huge.

Following these two assumptions:

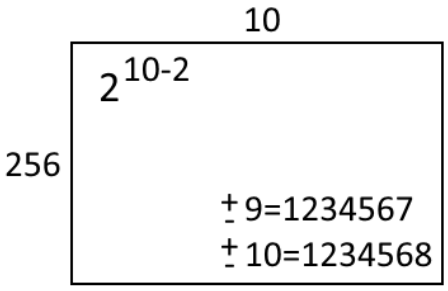
1. For the distinct factors two levels are enough.
2. 3 replications are enough for a preliminary analysis.

What is the experimental design we need define to obtain results in less than 5 weeks? We have just a single computer.

Discuss the proposed solution.

Answer

In this case we have 10 factors, 2 levels by factor, 3 replications for each experiment, and we know that one simulation needs at least one hour. Been optimistic, the number of hours needed to complete the experiment in a full factorial design is 3*2^10=3073 hours, but we want answers in 5 weeks, meaning that we have only 840 hours. It is needed to apply a fractional factorial design. With this proposed design:



The amount of experiments to be conducted is reduced to 256*3=768, which fits on our requirements.

10 factors model

We have a limited budget to analyze the different 10 factors to consider on our model. Each individual experiment costs 1000€ and we have a total budget of 56.000€ to be destined to experimentation. Define an experiment design, with this constraint, considering that we need at least 2 replications for each experiment. We do not want to consider nothing but the main effects.

Answer

Since we do not want to consider nothing but the main effect a Plackett and Burman design will be enough.

In that case, the table that defines our experiment will be similar to this.

	Variable										
Test	A	X1	B	C	D	X2	E	F	G	X3	X4
1	+	+	-	+	+	+	-	-	-	+	-
2	+	-	+	+	+	-	-	-	+	-	+
3	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	-	-	-	+	-	+	+	-
5	+	+	-	-	-	+	-	+	+	-	+
6	+	-	-	-	+	-	+	+	-	+	+
7	-	-	-	+	-	+	+	-	+	+	+
8	-	-	+	-	+	+	-	+	+	+	-
9	-	+	-	+	+	-	+	+	+	-	-
10	+	-	+	+	-	+	+	+	-	-	-
11	-	+	+	-	+	+	+	-	-	-	+
12	-	-	-	-	-	-	-	-	-	-	-

The amount of experiments will be constrained to 12 and the replications to 2, hence we have a total of 24 experiments that are less than the maximum of experiments we can execute due to the budget:
$mV^k*1000<=56000$ hence $k<=4$

The ceramic strength

We want to determine the effect of machining factors on ceramic strength, our response variable is the ceramic strength.

We have 3 factors and, for each factor, different values.

- Factor 1, the table speed is going from .025 m/s to .125 m/s, a real value.
- Factor 2, the down feed rate is going from .05 mm to .125 mm, a real value.
- Factor 3, the direction has two levels, longitudinal and transverse.

Define a DOE to determine what is the best scenario regarding our response variable. How do you deal with the randomness of the experiment? What are you going to apply to determine the interactions and main effects? Justify your answers.

Answer

In that case, we need to define a design that limits the possible infinite amount of experiments we can perform, since Factor 1 and Factor 2 are real values.

We propose to define a 2^k factorial design with the next levels for the 3 factors we have.

Factor	**Positive**	**Negative**
1	.025 m/s	.125 m/s
2	.05 mm	.125 mm
3	longitudinal	transverse

With this the table, we have is composed by $2^3 = 8$ experiments as is shown in the next table.

Experiment	**Factor 1**	**Factor 2**	**Factor 3**	**Answer**
1	-	-	-	?
2	-	-	+	?
3	-	+	-	?

4	-	+	+	?
5	+	-	-	?
6	+	-	+	?
7	+	+	-	?
8	+	+	+	?

Since the answer depends on an experiment that deals with random, it is needed to replicate the scenario. In this case, we are in a FINITE scenario, and we want to analyze the loading process, hence INDEPENDENT REPETITIONS will be the best technique to deal with randomness.

		Finite	**No finite**
Loading period needed		Independent repetitions	Independent repetitions
Loading period unneeded		Independent repetitions erasing the loading period/ Batch means	Batch means

To determine the number of replications needed in each experiment (row) it is needed to calculate the half range for each experiment, and the desired half range. We can apply the next expression to determine if the number of replications is enough.

$$n* = n(\frac{h}{h*})^2$$

where:

n = initial number of replications.

*n** = total replications needed.

h = half-range of the confidence interval for the initial number of replications.

*h** = half-range of the confidence interval for all the replications (the desired half-range).

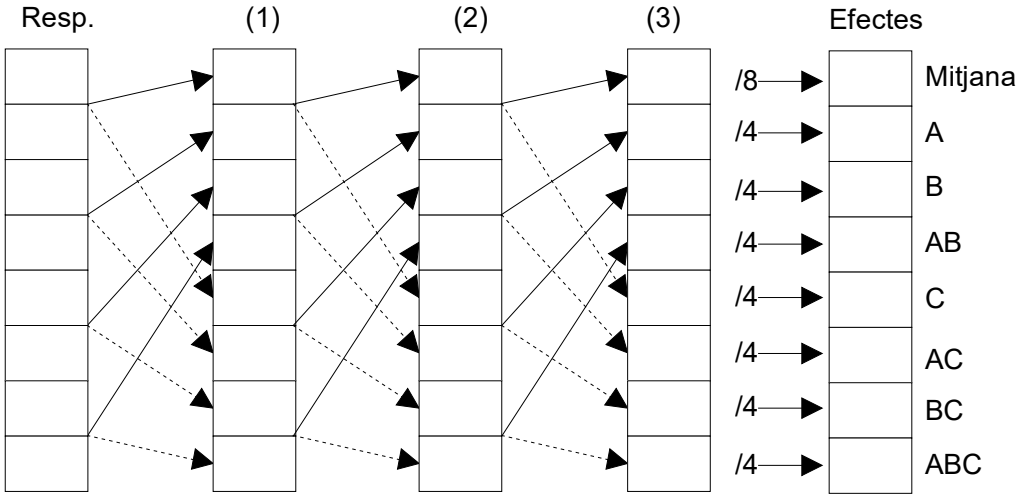
Once we have the data correctly taken for each scenario we can go further and apply Yates algorithm to determine eh interaction and the effects.

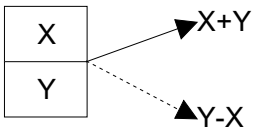
We:

1. Add the answer in the column "i" in the standard form of the matrix of the experimental design.
2. Add an auxiliary column as factors exists.
3. Add a new column dividing the first value of the last auxiliary column by the number of experimental conditions "E", and the others by the half of "E".

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 with two + on A and C corresponds to the interaction of AC, etc.





Life testing of weld-repaired

Consider a life testing of weld-repaired. The objective of the test is to identify the key factors that affect the life and to improve the product life. There are seven factors that may affect the life. A two-level full factorial design will require 2^7 = 128 runs. It will be time-consuming and costly.

For this example, the seven factors are:

Factor	Name	Level -	Level +
A	Initial Structure	as received	beta treat
B	Bead Size	small	large
C	Pressure Treat	none	HIP
D	Heat Treat	anneal	solution treat/age
E	Cooling Rate	slow	rapid
F	Polish	chemical	mechanical
G	Final Treat	none	peen

Compare the alternative of a full factorial design with other less costly alternatives. Discuss the pros and the cons of the considered alternatives.

A maximum of 64 scenarios can be analyzed.

Answer

Defining the table for this experimental full factorial 2^7 design we obtain:

Exp.	A	B	C	D	E	F	G	H
1	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	+	-
4	-	-	-	-	-	-	+	+
..	..	..	..	..	..	..	..	..
127	+	+	+	+	+	+	+	-
128	+	+	+	+	+	+	+	+

For each one of the different experiments it is needed to calculate the number of replications and depending on the resources needed for each replication this can be unfeasible.

To reduce the amount of experiments to be considered two alternatives can be done, a fractional design or a Plackett and Burman (PB) design.

For a fractional design, it is needed to define the "fraction" of the experiments that are going to be executed, and the confounding factors. Depending on the desired "resolution" of the experiment we are losing information related to the interaction between the different factors. The maximum resolution for this example needs 64 experiments and could be defined as follows:

Number of factors	**Fraction**	**Resolution**	**Experiments**	
7	2	VII	64	I=ABCDEFG

Hence:

$$64\left[2^{7-1}\right]^{7}$$

$$\pm 1=234567$$

PRO: we can reduce the number of experiments depending on the desired resolution, 2^{7-4}

CONS: we lose some interactions information.

For the PB design the table that we obtain is:

Config	**Input Parameters (factors)**							**Response**
	A	B	C	D	E	F	G	
1	+1	+1	+1	-1	+1	-1	-1	
2	-1	+1	+1	+1	-1	+1	-1	
3	-1	-1	+1	+1	+1	-1	+1	

4	+1	-1	-1	+1	+1	+1	-1	
5	-1	+1	-1	-1	+1	+1	+1	
6	+1	-1	+1	-1	-1	+1	+1	
7	+1	+1	-1	+1	-1	-1	+1	
8	-1	-1	-1	-1	-1	-1	-1	
Effect								

PRO: less experiments to be analyzed that in the previous alternative.

CONS: only the main effects are analyzed.

Improving the factory

We want to determine the effects of improving or changing several machines on a factory. Our goal is to minimize the time needed to produce a piece.

We have 3 machines that can be changed or improved in its mean service time.

1. Machine 1, we can change this machine, the time of the old machine is 10 seconds, the new machine needs 5 seconds. The price for the new machine is 8000€.
2. Machine 2, we can improve the software, in that case with the new software the machine needs 30 seconds to complete its operation, 50 seconds with the old software. The price to improve the software is 3000 €.
3. Machine 3, on this machine we can use more oil. Using more oil, we reduce the operation time 5 seconds, from 10 to 5 seconds to operation. The price of the oil is 10€ by day.

Define a DOE to determine what is the best scenario regarding our response variable, the total time needed to complete the piece.

We know that the function that defines the total time of operation is:

$$Time(piece) = \exp(M1_t) + \exp(M2_t)\exp(M3_t)$$

Where *exp* is an exponential distribution with the parameter defined by the factor.

$$x = \frac{\ln(1-r)}{-\alpha} = \frac{\ln(r)}{-\alpha}$$

Consider that (because you don't have enough time to calculate a full factorial design) you need to work with just 4 experiments.

How do you deal with the randomness of the experiment? what is the number of replications needed?

Considering that only 2 replications are enough, explain the results and calculate the investment needed to improve the industry results during a year.

Answer

For each one of the different factors in our system we have two levels. In the next table we summarize the levels represented with the '+' and '-' sign.

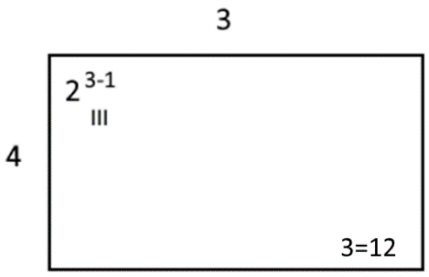
FACTOR	LEVEL '-'	LEVEL '+'
M1	10	5
M2	50	30

M2	10	5

Since each of the factors owns two levels, naturally we can propose a factorial design defined by the next table.

EXPERIMENT	**M1**	**M2**	**M3**
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

However, since we have time constrains we need to reduce the amount of experiments to 4, hence a fractional factorial design is proposed where M3=M1·M2, below the definition and the table.



EXPERIMENT	**M1**	**M2**	**M3=M1·M2**
1	-	-	+
2	+	-	-
3	-	+	-
4	+	+	+

How do you deal with the randomness of the experiment? what is the number of replications needed?

To calculate the number of replications we need to execute the experiments. The experiments are represented by this expression:

$$Time(piece) = \exp(M1_t) + \exp(M2_t)\exp(M3_t)$$

Where M1t will be the level used for M1 factor on the scenario to be executed

And exp is:

$$x = \frac{\ln(1-r)}{-\alpha} = \frac{\ln(r)}{-\alpha}$$

Where α is the parameter of the distribution (the value M1t, M2t or M3t depending on the scenario), and r is a random number (uniform distribution from [0,1)).

With this information we fill the table:

M1	M2	M3	VALUES			x1	x2	mean
-	-	+	10	50	5	69,68545	29,83393	49,75969
+	-	-	5	50	10	376,26	1994,57	1185,415
-	+	-	10	30	10	41,10304	323,2293	182,1662
+	+	+	5	30	5	28,68768	29,76822	29,22795

We now can calculate the replications needed for each experiment using:

$$n^* = n(\frac{h}{h^*})^2$$

Where:

- n = initial number of replications.
- n* = total replications needed.
- h = half-range of the confidence interval for the initial number of replications.
- h* = half-range of the confidence interval for all the replications (the desired half-range).

For the first experiment, we need (at 5%) more than 20.710,77 replications, for the second 60.178,91, 77.448,54 replications for the third, and only 44,13 for the last. The explanation of this huge number of replications is because the exponential distribution we use to represent the machines, and the very low number of replications used for the pilot test (only two). The process is iterative and converge while we increase the number of replications (if the system is stable).

Since they do not explain nothing regarding the system time constrains, we suppose that the factory closes at night; hence **independent repetitions** will be the method selected to deal with randomness.

Considering that only 2 replications are enough, explain the results and calculate the investment needed to improve the industry results during a year.

To calculate the cost, we analyze the money we need to invest in the best solution (we consider a usual year of 365 days).

M1	M2	M3	mean	Cost M1	Cost M2	Cost M3	Total cost
-	-	+	49,75969	0	0	3650	3650
+	-	-	1185,415	8000	0	0	8000
-	+	-	182,1662	0	3000	0	3000
+	+	+	29,22795	8000	3000	3650	14650

Considering that two replications are enough, the best option is the last one, hence the investment will be of 14.650€ (no constrains regarding the investment we can do).

Manufacturing process

Consider a manufacturing process for a new computer chipset. The objective is to improve the overall production process. Mangers detects four factors that are key elements on the process development.

Factor	Name	Unit	Low	High
A	Aperture Setting	-	Small: value 1	Large: value 2
B	Exposure Time	min	20	30
C	Develop Time	s	30	45
D	Mask quality	-	Average: value 10	Good: value 20

We want to determine the effects of these factors on the quality of the final chipset. This quality is measured in number of errors, measured on the experiment (less errors is better).

Define a DOE to determine what is the best scenario regarding our response variable. How do you deal with the randomness of the experiment? What are you going to apply to determine the best scenario? Justify your answers.

To obtain the answer of our experiments we can use this function:

$$error = \left(A * \frac{10}{B} * \frac{10}{C}\right) + {}^{1}/_{D} + \varepsilon$$

Answer

In that case we need to define a design that constrains the amount of experiments we ca perform, since Factor 1 and Factor 2 are real values.

We propose to define a 2^k factorial design with the next levels for the 4 factors we have.

With this the table we have is composed by $2^4 = 16$ experiments as is shown in the next table.

Experiment	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*	*Answer*
1	-	-	-	-	?
2	-	-	+	-	?
3	-	+	-	-	?
4	-	+	+	-	?
5	+	-	-	-	?
6	+	-	+	-	?
7	+	+	-	-	?
8	+	+	+	-	?
9	-	-	-	+	?
10	-	-	+	+	?
11	-	+	-	+	?
12	-	+	+	+	?
13	+	-	-	+	?
14	+	-	+	+	?

15	+	+	-	+	?
16	+	+	+	+	?

Since the answer depends on an experiment that deals with random, it is needed to replicate the scenario. In this case we are in a FINITE scenario, and we want to analyze the loading process, hence INDEPENDENT REPETITIONS will be the best technique to deal with randomness.

	Finite	No finite
Loading period needed	Independent repetitions	Independent repetitions
Loading period unneeded	Independent repetitions erasing the loading period/ Batch means	Batch means

To determine the number of replications needed in each experiment (row) it is needed to calculate the half range for each experiment, and the desired half range. We can apply the next expression to determine if the number of replications is enough.

$$n^* = n\left(\frac{h}{h^*}\right)^2$$

where:

- n = initial number of replications.
- n^* = total replications needed.
- h = half-range of the confidence interval for the initial number of replications.
- h^* = half-range of the confidence interval for all the replications (the desired half-range).

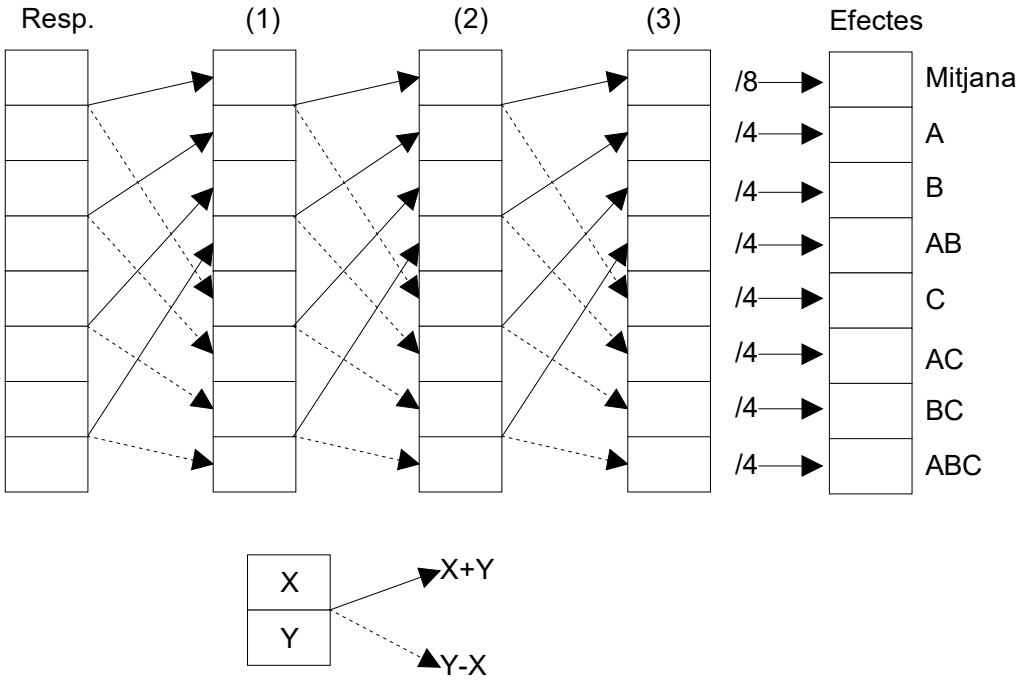
Once we have the data correctly taken for each scenario we can go further and apply Yates algorithm to determine eh interaction and the effects.

We:

1. Add the answer in the column “i” in the standard form of the matrix of the experimental design.
2. Add an auxiliary column as factors exists.
3. Add a new column dividing the first value of the last auxiliary column by the number of experimental conditions “E”, and the others by the half of “E”.

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 with two + on A and C corresponds to the interaction of AC, etc.



Once we have some best candidates we can apply an ANOVA test to assure that our best alternative is different from others and makes sense to be applied in the industry.

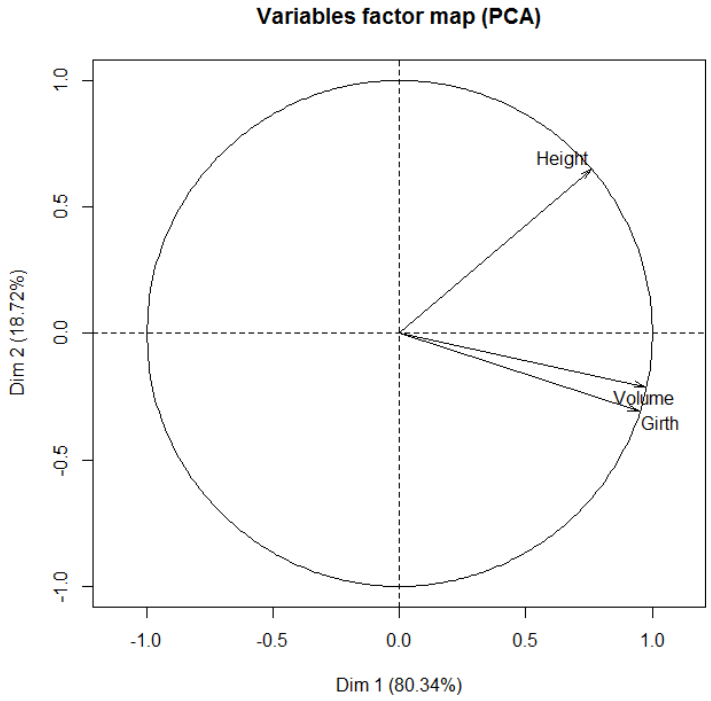
Sequoias

P1 (0,5 points). We want to analyze the relation of the Girth, Height and the volume in large trees (sequoias). Discuss the nature of the information we have. See the table.

trees

	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

P2 (1 point). Analyzing the next diagram, what can you understand of the Volume of a tree? Describe the information you can obtain form the analysis.



P3 (2 points). We want to conduct an experiment to understand if we can modify the Height of the trees considering other factors, like climatic condition (mainly humidity, sun radiation), and the exposure to some gases (mainly Co2 and No2). Define an experimental design to analyze all these factors (Volume, Girth, Height, Humidity, Sun radiation, Co2, NO2). The main idea is to plant different trees under the selected conditions and analyze the evolution of those trees during 5 years. We have the chance to control only **8 different environments** (greenhouses).

What will be the experimental design to be used? Justify your answers.

How are you going to deal with randomness?

Answer

P1. The dataset is composed by tree numerical variables. We see that we have more observations than tree, hence we can apply a PCA. Regarding this, just to mention that PCA is usually applied on large datasets. However, there is no limitation in that sense, hence here we can apply it.

P2. The first element to consider the variability that is explained on the Variables Factor Map chart. Analyzing the first and the second dimension, Dim 1(80.34) and Dim 2 (18.72), the amount of the variability explained is about 99.06. This is a very high value, meaning that a lot of the variability is explained by the chart.

Regarding the variables, we can detect that Volume is clearly related with Girth, while Height is more unrelated with the other two variables. We can assume that the Girth and the Volume depends one on the other, but we cannot conclude too much regarding the Height.

P3. Now the dataset will be composed by the next variables (we consider that we use also initial Girth and Volume in the analysis): Volume, Girth, Height, Humidity, Sun radiation, Co2, NO2. Height will be our answer variable, hence is not going to be considered as an experimental factor. Hence, we have now 6 factors. Considering that we can limit the values of those variables to two levels, the amount of experiments to be executed (on a two-factorial scenario) will be 2^6=64 scenarios.

However, we have only o different environments to test this, hence we need to reduce the number of scenarios to test.

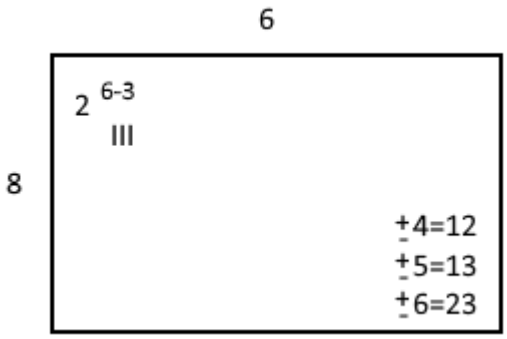
In that case, we need to define a fractional factorial with a maximum of 8 experiments: 2^{6-3}=8

To define the structure first, we define a table with all the factors we own.

Note that the factors will be numbered from 1 to 6 and F1= Girth, F2=Height, F3=Humidity, F4=Sun Radiation, F5=CO2 and F6=NO2.

	F1	F2	F3	F4= F1*F2	5= F1*F3	6= F2*F3
1	-	-	-	+	+	+
2	-	-	+	+	-	-
3	-	+	-	-	+	-
4	-	+	+	-	-	+
5	+	-	-	-	-	+
6	+	-	+	-	+	-
7	+	+	-	+	-	-
8	+	+	+	+	+	+

With this we can now define the fractional factorial experimental design as:



The resolution in this experiment will be III, defined by the length of the shortest word.

To deal with randomness, we review the experimental procedure we are going to follow. We have 8 different environments where we can control the Girth and the Height of the trees that initially we are going to plant, the Humidity, the Sun Radiation, the CO2 and NO2. Obviously is needed to assure that the amount of observation we have in each environment is enough, hence we recommend to plant as many trees as we can in each environment. Since we want to analyze the evolution of each tree, from the beginning and until 5 years, we can consider each tree as an "independent repetition".

Just as comment, in this kind of experiments is interesting to prior to develop the real experiment (that last for 5 years and usually implies the use of lots of resources) to develop a simulation model to test the assumptions of the experimental design, to, as an example, detect what is the minimum number of trees to plant in each environment to assure that the assumptions related with randomness can be fulfilled.

Tire pressure on fuel consumption

The treatment to be considered is the effect of tire pressure on fuel consumption Cox (1958, [COX1]). Four different tire pressures (levels) are to be tested, A, B, C, D. The design uses four buses and is carried out over four days. To eliminate variations between buses, and between days, these factors provide the row and column blocking elements of the design.

Answer

The basic design applied is then of the form shown in the table below:

4x4 Latin square	Bus 1	Bus 2	Bus 3	Bus 4
Day 1	A	B	C	D
Day 2	B	C	D	A
Day 3	C	D	A	B
Day 4	D	A	B	C

The pattern shown above is systematic, not randomized, and prior to conducting a trial of this type the design should be randomized. A simple random permutation of rows and columns would suffice. Thus, a random permutation for columns might yield the sequence 3,4,2,1, and for rows 1,3,2,4. The table utilized would then be:

4x4 Latin square	Bus 1	Bus 2	Bus 3	Bus 4
Day 1	C	D	B	A
Day 2	A	B	D	C
Day 3	D	A	C	B
Day 4	B	C	A	D

This 4x4 arrangement requires a total of 16 trials, as opposed to the full factorial experiment which would require 4x4x4=64 trials to obtain every possible combination of day, bus and treatment. It is thus a fractional experiment (and an unbalanced design) and as such loses some information that could be obtained through a complete experiment (the loss is in the interaction effects between the pairs of factors, e.g. days and buses) but it does retain all the main effects at a considerable saving in time and cost.

Latin square example

We want to test 4 treatments, we select for 4 cows and during 4 periods. To do so, we define a Latin square experiment as is shown next.

	Cow 1	*Cow 2*	*Cow 3*	*Cow 4*
Period 1	T4	T1	T3	T2
Period 2	T1	T4	T2	T3
Period 3	T3	T2	T1	T4
Period 4	T2	T3	T4	T1

After the randomizations and the executions, we obtain the values of the next table. We also add the means.

	Cow 1	*Cow 2*	*Cow 3*	*Cow 4*			
Period 1	192	195	292	249	232	T1	191
Period 2	190	203	218	210	205,25	T2	206,75
Period 3	214	139	245	163	190,25	T3	217
Period 4	221	152	204	134	177,75	T4	190,5
	204,25	172,25	239,75	189	201,3125		

Calculate the ANOVA table that allows to detect the effects of the experimental design.

Answer

The ANOVA table looks like the next table.

Source	*SS*	*df (N-1)=k*	*MS*	*F*
Row	SSROW	k-1	MSROW =SSROW /df	MSROW/MSE
Column	SSCOL	k-1	MSCOL =SSCOL/df	MSCOL/MSE
Treatments	SSTR	k-1	MSTR =SSTR/df	MSTR/MSE
Error	SSE	(k-1)(k-2)	MSE =SEE/df	

To fill this table, remember the expressions on the next figure.

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{k}(y_{ijt}-\bar{y})^2}_{\text{TSS}} = \underbrace{k\sum_{i=1}^{k}(\bar{y}_{i..}-\bar{y})^2}_{\text{SSROW}} + \underbrace{k\sum_{j=1}^{k}(\bar{y}_{.j.}-\bar{y})^2}_{\text{SSCOL}} + \underbrace{k\sum_{t=1}^{k}(\bar{y}_{..t}-\bar{y})^2}_{\text{SSTR}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{k}(y_{ijt}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..t}+2\bar{y})^2}_{\text{SSE}}$$

With this we can obtain this data. In that case, we cannot reject the null hypothesis.

Source	*SS*	*df (N-1)=k*	*MS*	*F*	*Pr>F*
Row	6539,188	3	2179,729	1,760686	0,254219
Column	9929,188	3	3309,729	2,673448	0,141025
Treatments	1995,688	3	665,2292	0,537342	0,673788
Error	7428	6	1238		

NZEBs

Here we want to define the best parameters for a building in order to **improve its energetic behavior**. The building is the (e)CO building that is going to compete in the prestigious solar Decathlon competition, see Figure 1.

Figure 1. (e)CO 2012 Solar Decathlon building.

QUESTION 1: What are the steeps that are going to be done in this analysis?

You will be able to define a model but, a detailed analysis makes clear that is needed to define a complex simulation model to obtain accurate results.

We work with four walls in two modules, **North Wall, West Wall, South Wall** and **East Wall** for north models and south modules, and the **ceiling**. Also, we discuss regarding the **external supply** use (Yes or Not). For the walls and the ceiling, we can use different materials that goes from **0.212** to **0.174** of **thermal insulation**.

We can build a simulation model that taking all those parameters provides an answer in about one hour (to complete a single replication for a specific parametrization).

If 3 replications are enough for the pilot test analysis…

QUESTION 2: What is the experimental design we need define to obtain results in less than 5 weeks? We have just a single computer. How are you going to deal with randomness (replications)? Discuss the proposed solution.

Once we have the experimental design to be executed completed, is needed to execute the simulation model that be able to obtain the answers for each scenario. This simulation model takes care of two elements, the arrival of a new energy demand to the building, and the time that this energy demand last.

We are going to consider only the factor that defines if we are using an external energy supply. If a new energy demand arrives to the building, we can use an external power supply or we can put this demand on a queue and wait.

QUESTION 3: Model the two scenarios using an event scheduling approach. The data to be used is on the next table. We consider that when a new demand arrives, the solar power of the building is enough, but if two demands arrives (at the same time) then, if the power supply exist we are going to use it, if not, this demand must wait to be served in a queue.

ELEMENT	DEMAND	UNTIL
1	1	1
2	2	2
3	2.5	2
4	3	1
5	6	2

Once you execute the model and you put the results in the DOE table you obtain the next answers (we consider only a subset of the scenarios).

North MODULE isolation

North Wall	**South Wall**	**External supply**	VALUES			External supply use
-	-	-	0,174	0,174	0	2,87
-	-	+	0,174	0,174	1	0,74
-	+	-	0,174	0,212	0	2,59
-	+	+	0,174	0,212	1	0,72
+	-	-	0,212	0,212	0	2,36
+	-	+	0,212	0,212	1	0,70
+	+	-	0,212	0,212	0	2,36
+	+	+	0,212	0,212	1	0,70

QUESTION 4: What is the effect of the insulation or the external power use in this NZEB?

Answer

Steps to be done in this analysis.

The specific steps to be done are:

7. **Goals of the analysis:** To identify the main factors that determine the behavior of a building from the **energetic** perspective. Minimize energy consumption.
8. **Design of the analysis**. It is needed to reduce the dimensionality of the problem. We can start with a PCA and clustering to define different subsets of the data to work. We can use first PCA to discard all variables that are not important in our analysis. Later we can define a simulation model to predict the energy consumption for all the scenarios. With this we are now selecting the variables to be considered on the analysis that are 10

	North MODULE isolation				South MODULE isolation				Other	
U (W/m2 K)	North Wall	West Wall	South Wall	East Wall	North Wall	West Wall	South Wall	East Wall	External supply	Ceiling material

9. **Hypotheses of the Analysis**. If we use regression analysis, normality, homoscedasticity and independence on the data must be assured. Since here we use simulation we must validate the assumptions following Simulation approaches. See Pau Fonseca i Casas, *Formal Languages for Computer Simulation: Transdisciplinary Models and Applications*, ed. Pau Fonseca i Casas (IGI Global, 2014), doi:10.4018/978-1-4666-4369-7. or Robert G Sargent, "Verification and Validation of Simulation Models," *Journal of Simulation* 7, no. 1 (February 7, 2013): 12–24, doi:10.1057/jos.2012.20..
10. **Analytical procedure.** We estimate the model and we evaluate the fit to the data. In this step may appear unusual observations (outliers) or influential whose influence on the estimates and the goodness of fit must be analyzed. If we use simulation, the simulation steeps must be followed.
11. **Interpretation of the results.** Such interpretations can lead to additional specifications or model variables with which you can return back to steps 3) and 4)
12. **Analysis Validation.** Is to establish the validity of the results obtained by analyzing whether the results, obtained with the sample, is generalized to the population from which it comes. This sample can be divided into several parts in which the model is re-estimated and the results are compared. If simulation is applied VV&A simulation techniques can be applied.

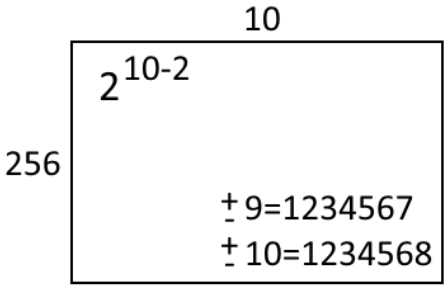
DOE proposed

In this case we have 10 factors, 2 levels by factor, 3 replications for each experiment, and we know that one simulation needs at least one hour.

North MODULE isolation	South MODULE isolation	Other

U (W/m2 K)	North Wall	West Wall	South Wall	East Wall	North Wall	West Wall	South Wall	East Wall	External supply	Ceiling material
s1	0,212	0,212	0,212	0,212	0,212	0,212	0,212	0,212	Yes	0,212
s2	0,212	0,212	0,212	0,212	0,212	0,212	0,212	0,212	Yes	0,212
s3	0,212	0,212	0,212	0,212	0,212	0,212	0,212	0,212	Yes	0,212
s4	0,212	0,212	0,212	0,212	0,212	0,212	0,212	0,212	Yes	0,174
...	...	...	...	...	...	...	...	...	...	...
S4096	0,174	0,174	0,174	0,174	0,174	0,174	0,174	0,174	No	0,174

Been optimistic, the number of hours needed to complete the experiment in a full factorial design is 3*2^10=3073 hours, but we want answers in 5 weeks, meaning that we have only 840 hours. It is needed to apply a fractional factorial design. With this proposed design:



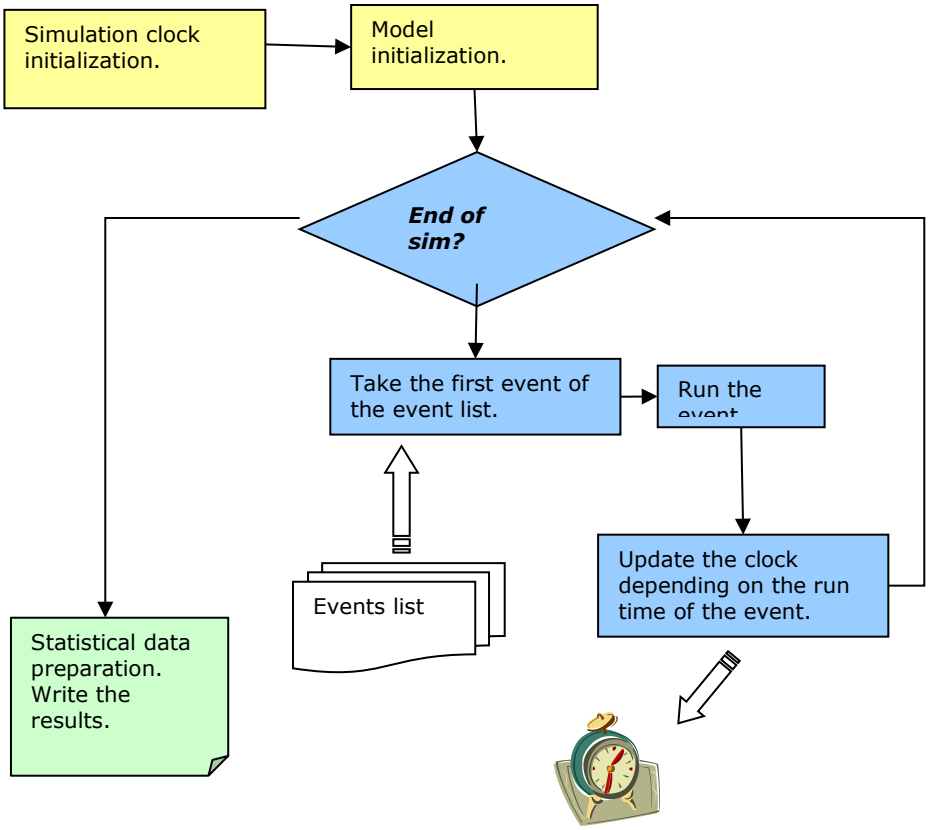
The amount of experiments to be conducted is reduced to 256*3=768, which fits on our requirements.

The replications will be executed using the **independent repetitions** approach.

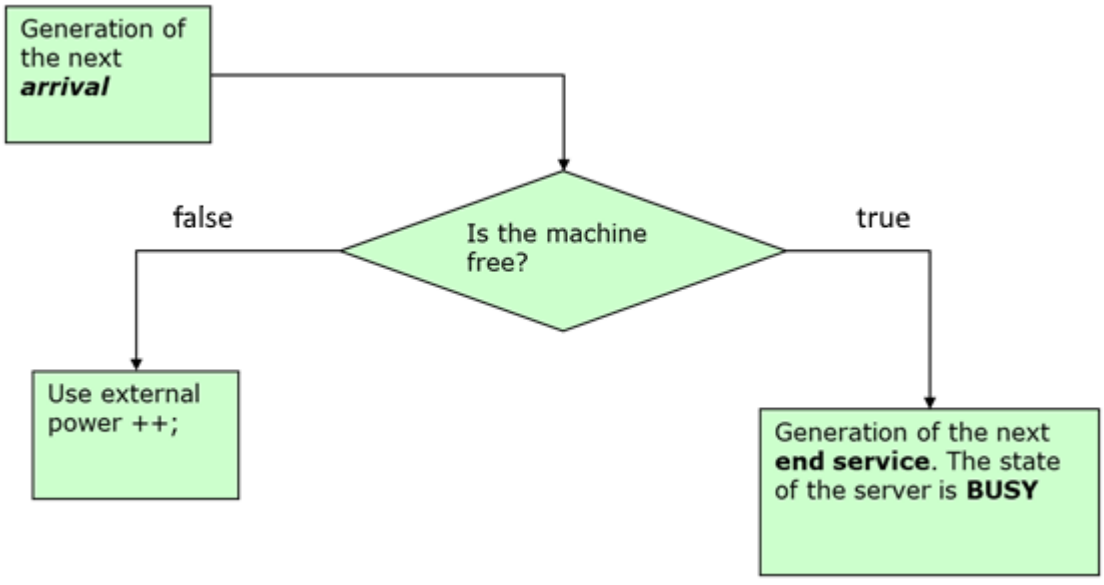
Simulation trace

Considering that our model has only the two events proposed, the behavior is quite similar to the one described for a queue system, see next the procedures that defines its behavior.

The simulation engine will follow:



For the arrival of a new demand we have:



And for the exit:

Statistics update.
Remove the element of the system.

Hence the trace will be like the one presented next.

Id	Time	Demand (Next Arrival)	Until (Next exit)	Solar cells (Server) state	Queue long	Is Arrival?	Is Exit?
0	0	1	-	0	0	0	0

Hence the trace for the first alternative (using an external power supply) is (the priority is for the Arrival event).

Id	Time	Next Arrival	Next exit	Server state	Using external power	Is Arrival?	Is Exit?
0	0	1	-	0	0	0	0
1	1	2	2	1	0	1	0
2	2	2.5	2	1	1	1	0
3	2	2.5	-	0	0	0	1
4	2.5	3	4.5	1	0	1	0
5	3	6	4.5	1	1	1	0
6	4.5	6	-	0	0	0	1
7	6	-	8	1	0	1	0
8	8	-	-	0	0	0	1

The trace using a queue is:

Id	Time	Next Arrival	Next exit	Server state	Queue long	Is Arrival?	Is Exit?
0	0	1	-	0	0	0	0

1	1	2	2	1	0	1	0
2	2	2.5	2	1	1	1	0
3	2	2.5	4	1	0	0	1
4	2.5	3	4	1	1	1	0
5	3	6	4	1	2	1	0
6	4	6	6	1	1	0	1
7	6	-	6	1	2	1	0
8	6	-	7	1	1	0	1
9	7	-	9	1	0	0	1
10	9	-	-	0	0	0	1

Yates calculus

North MODULE isolation

North Wall	**South Wall**	**External supply**	VALUES			External supply use	Yates				
-	-	-	0,174	0,174	0	2,87	3,62	6,93	13,05	**1,63**	Mean
-	-	+	0,174	0,174	1	0,74	3,31	6,12	-7,31	**-1,83**	External supply
-	+	-	0,174	0,212	0	2,59	3,06	-4,00	-0,30	**-0,08**	South Wall
-	+	+	0,174	0,212	1	0,72	3,06	-3,31	0,26	**0,07**	
+	-	-	0,212	0,212	0	2,36	-2,13	-0,30	-0,81	**-0,20**	North Wall
+	-	+	0,212	0,212	1	0,70	-1,87	0,00	0,69	**0,17**	
+	+	-	0,212	0,212	0	2,36	-1,66	0,26	0,30	**0,08**	
+	+	+	0,212	0,212	1	0,70	-1,66	0,00	-0,26	**-0,07**	

NZEB parameters

Consider a NZEB where we can change several parameters in order to improve the overall behavior (energy consumption) of the building.

The next table defines the energy needs depending on 3 factors.

Factor	Name	Unit	Low	High
A	Apertures (windows)	-	Small window: 1	Large window: 2
B	Roof insulation	range	20	30
C	Floor insulation	range	30	45
D	Wall insulation	range	10	25

We want to determine the effects of these factors on the energy consumption of the final building. This quality is measured in the energy consumption, measured on the experiment (less energy is better). To start we consider that we FIX the wall with the cheaper material (less insulation i.e 10 value)

Define a DOE to determine what is the best scenario regarding our response variable. What is the more important factor from the analyzed?

Justify your answers.

To obtain the answer of our experiments we can use this function (as usual this is a simplification, the expression will come from real experiments conducted on a real building, or as is done more a more frequently, from a simulator that is like a black box).

$$energy\ consumption = A * (\frac{10}{B} + \frac{10}{C} + {}^{1}/_{D})$$

Once we finish this analysis we want to see what the effect of D on the results FIXING all the other values is (with the best candidates). Can you define an expression to explain this *maybe* linear relation?

Discuss te results, considering that you have money to analyze 4 more scenarios.

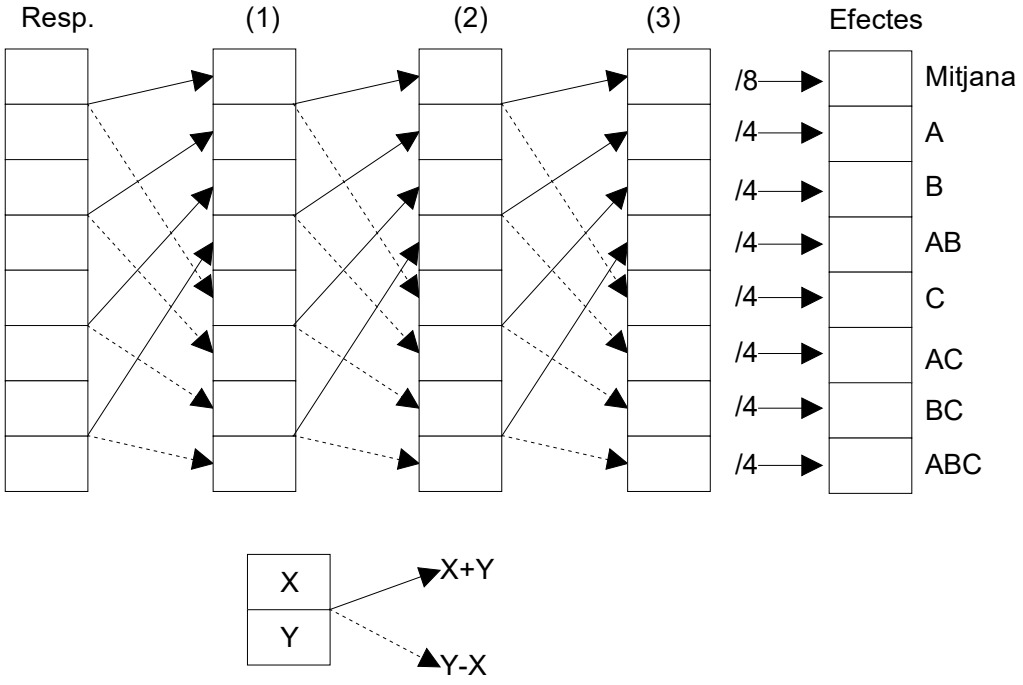
Solution

We define a factorial design and we use the expression to calculate the answers

					obs.
	A	B	C	D(FIXED)	----------

1	1	20	30	10	0,933333
2	1	20	45	10	0,822222
3	1	30	30	10	0,766667
4	1	30	45	10	0,655556
5	2	20	30	10	1,866667
6	2	20	45	10	1,644444
7	2	30	30	10	1,533333
8	2	30	45	10	1,311111

Once we have the data correctly taken for each scenario we can go further and apply Yates algorithm to determine eh interaction and the effects.



We:

1. Add the answer in the column “i” in the standard form of the matrix of the experimental design.
2. Add an auxiliary column as factors exists.
3. Add a new column dividing the first value of the last auxiliary column by the number of experimental conditions “E”, and the others by the half of “E”.

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 with two + on A and C corresponds to the interaction of AC, etc.

For B and C factors + represents less insulation, while + better insulation. For A factor – is 1 and + is 2.

					Estimated	Effect
	obs.	1	2	3	Effect	Name
	----------	---	------	------	------------	---------
1	0,933333	1,755556	3,177778	9,533333	1,191667	Mean
2	0,822222	1,422222	6,355556	-0,66667	-0,16667	C
3	0,766667	3,511111	-0,22222	-1	-0,25	B
4	0,655556	2,844444	-0,44444	0	0	BC
5	1,866667	-0,11111	-0,33333	3,177778	0,794444	A
6	1,644444	-0,11111	-0,66667	-0,22222	-0,05556	AC
7	1,533333	-0,22222	0	-0,33333	-0,08333	AB
8	1,311111	-0,22222	0	0	0	ABC

Clearly A (windows) is the more important factor, using 2 value (that means big windows) implies the use of more energy (worse scenario). B have a good impact on the energy saving and finally C.

We see that the best case is A=1, B=30 and C=45, as expected, since we know the linear expression that defines our system (is not a black box as it usually must be).

Now we analyze the results with D (that can take values from 10 to 20), we can select in this case 10, 15, 20 and 25. The parametrization of the experiment will be:

	A	B	C	D	
1	1	30	45	10	0,655556
2	1	30	45	15	0,622222

3	1	30	45	20	0,605556
4	1	30	45	25	0,595556

Applying the know expression we obtain the answers of the last column. Not surprisingly (because we know ow this expression) we can conclude that the wall insulation is also important, and follows a linear relation (we already know that).

We can calculate a LR model to see this relation:

D	Energy consumption
10	0,65555556
15	0,62222222
20	0,60555556
25	0,59555556

Confidence interval for the forecasted value

n	4	
df	2	= n - 2
mean(x)	17,50	= AVERAGE(x)
x_0	10	
$\hat{y}_0$	0,649222	= FORECAST(y,x,x_0)
s_{Res}	0,0084	= STEYX(y,x)
SS_x	125	= DEVSQ(x)
se	0,007028	= s_{Res}*SQRT(1/n+(x_0-$\bar{x}$)^2/SS_x)
t-crit	4,302653	= TINV(0.05,df)
lower	0,618984	= $\hat{y}_0$ - t-crit * se
upper	0,67946	= $\hat{y}_0$ + t-crit * se