Introduction to Experiment Design

Kauko Leiviskä

University of Oulu

Control Engineering Laboratory

2013

Table of Contents

The main source: W.J. Diamond. Practical Experiment Design for Engineers and Scientists. Lifetime Learning Publications, 1981.

http://www.itl.nist.gov/div898/handbook/

1. Introduction

1.1 Industrial Experiments

Industrial experiments are in principle comparative tests; they mean a comparison between two or more alternatives. One may want to compare the yield of a certain process to a new one, prove the effect of the process change compared to an existing situation or the effect of new raw materials or catalyser to the product quality or to compare the performance of an automated process with manually controlled one.

When we speak about systematic experimental design, we presume statistical interpretation of the results so that we can say that a certain alternative outperforms the other one with e.g. 95% probability or, correspondingly, that there is a 5% risk that our decision is erroneous. What is the best is that we can tell the statistical significance of the results before testing, or, just to put in another way round, we can define our test procedure so that it produces results with a required significance.

We can also experiment with some process aiming to optimize its performance. Then we have to know in advance what the available operation area is and design our experiments so that we by using them together with some mathematical software can search for the optimum operating point. The famous Taguchi method is a straightforward approach to optimize quality mainly by searching process conditions that produce the smallest quality variations. By the way, this is also the approach that control engineers most often use when speaking about stabilizing controls. Also in this case, the focus is in optimizing operational conditions using systematic experimental design.

There is also a large group of experiment design methods that are useful in optimizing nonlinear systems, namely response surface methods that we will be dealing with later on.

1.2 Matrix Designs

The conventional experiment design proceeds usually so that changes are made one variable at time; i.e. first the first variable is changes and its effect is measure and the same takes place for the second variable and so on. This is an inefficient and time-consuming approach. It cannot also find the probable interactions between the variables. Result analysis is straightforward, but care must be taken in interpreting the results and multi-variable modelling is impossible.

Systematic design is usually based on so called matrix designs that change several variables simultaneously according to the program decided beforehand. Changing is done systematically and the design includes either all possible combinations of the variables or at the least the most important ones.

E.g. in experimenting with three variables at two possible levels, there are eight possible combinations (2^3). If all combinations are included we can speak about 2-level, 3 variable case which requires 8 experiments. As mentioned before, statistical interpretation is needed and because of the exponential increase dimensional explosion is expected with more variables and levels.

Example. We want to test the effect of different factors on the yield in a chemical reactor: temperature (A), reaction time (B) and raw material vendor (C). We assume that testing at two levels of each variable is enough. This means that the process is assumed linear with respect to continuous variables. The levels are chosen as

Factor A:	Factor B:	Factor C:
(-)-level is 100 °C	(-)-level is 5 min.	(-)-level is vendor X
(+)-level is 150 °C	(+)-level is 10 min.	(+)-level is vendor Y

Using these denotations, the design matrix can be written as

Run number	A	B	C
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

So in the first experiment, the temperature is held at 100 °C, reaction time at 5 minutes and the raw material from vendor X is used, and so on. Note that this experiment design allows using both continuous and non-continuous variables in the same design matrix.

2. Basic Definitions

Linearity and interactions

Example. We continue testing the yield of the chemical reaction, but this time with two variables, only: the temperature and reaction time. Figure 1 below shows four possible cases; both linear and non-linear cases with and without interaction. The panels on the lkeft show linear and non-linear cases without interaction and, respectively, the panels on the rifgh-hand side picture cases with interaction.

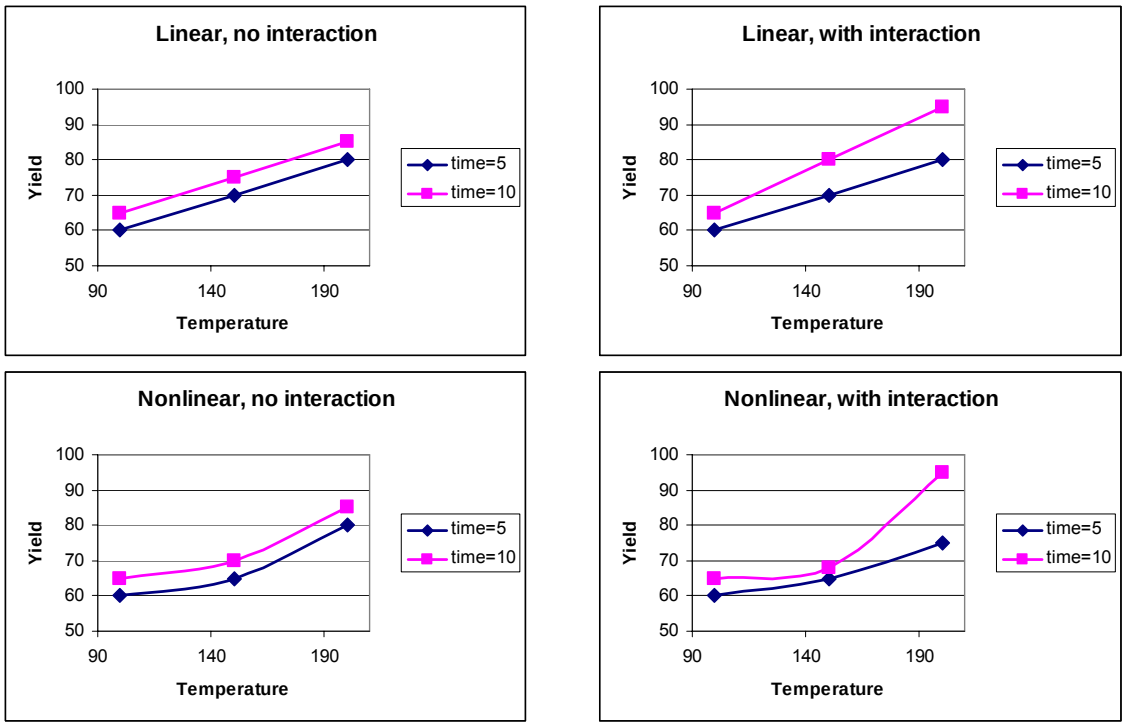


Figure 1.1. Graphs illustrating concepts of linearity and interaction.

Some conclusions can be drawn from the graphs:

-in non-interacting cases, the curves follow each other; i.e. the effect of the reaction time does not depend on the temperature

-in interactive case, the effect of the reaction time is stronger with higher temperature

- two-level designs can reveal only the linear behaviour

Effect

Experimental designs test, if a variable influences another. This influence is called “effect”. There are two different effects: the variable effects on another directly or via an interaction (or uses both mechanisms simultaneously). The calculation of the strength of an effect is

commented later. The significance of an effect is determined statistically with some probability (usually 95%) or risk (usually 5%).

Full factorial designs

These designs include all possible combinations of all factors (variables) at all levels. There can be two or more levels, but the number of levels has an influence on the number of experiments needed. For two factors at p levels, 2^p experiments are needed for a full factorial design.

Fractional factorial designs are designs that include the most important combinations of the variables. The significance of effects found by using these designs is expressed using statistical methods. Most designs that will be shown later are fractional factorial designs. This is necessary in order to avoid exponential explosion. Quite often, the experiment design problem is defined as finding the minimum number of experiments for the purpose.

Orthogonal designs

Full factorial designs are always orthogonal, from Hadamard matrices at 1800's to Taguchi designs later. Orthogonality can be tested easily with the following procedure:

In the matrix below, replace + and – by +1 and -1. Multiply columns pairwise (e.g. column A by column B, etc.). For the design to be orthogonal, the sum of the four products must be zero for all pairs.

Run number	A	B	C
1	+	+	-
2	+	-	+
3	-	+	+
4	-	-	-

Run number	AB	BC	AC
1	1	-1	-1
2	-1	-1	1
3	-1	1	-1
4	1	1	1
Sum	0	0	0

Condition number

Condition number is a measure of sphericity – orthogonality – of the design. It has emerged together with computerized experimental design methods. If we describe the design as a matrix *X* consisting of -1's and +1's, the condition number is the ratio between the largest

and smallest eigenvalue of *X'X* matrix. All factorial designs without centre points (the mid point between the + and – levels) have a condition number 1 and all points are located on a sphere (2D case). In MATLAB, the command *cond(X)* calculates the condition number for matrix *X*.

Contrast

The concept of the contrast column is easiest to clarify with an example. We take once again the earlier used matrix and denote + and – with +1 and -1. The sum of the columns must be zero.

Run number	A	B	C
1	1	1	-1
2	1	-1	1
3	-1	1	1
4	-1	-1	-1

In order to find the contrast column for columns A and B, we multiply column A by B. If there is now a column which has the opposite sign on all rows, it is the contrast column for A and B. Now it happens to be column C. This has a meaning in defining the effect of interactions later on.

Run number	AB		C
1	1		-1
2	-1		1
3	-1		1
4	1		-1

Resolution

The resolution of an experiment design tells, what kind of effects can be revealed with the design in question. There are three resolutions usually referred to:

-Resolution V or better: main effects and all two variable interactions

-Resolution IV: main effects and a part of two variable interactions

-Resolution III: only main effects.

3. On Statistical Testing

Hypotheses

In process analysis, we are often encountered with a situation where we are studying, if two populations are similar or different with respect to some variable; e.g. if the yield in the previous example is different at two reaction temperatures. In this comparison, there are two possibilities: the populations are either similar or different (statistically).

The comparison uses usually means or variances. We are testing, if the energy consumption of the new process is smaller (in average) than of the existing one or if the variation in some quality variable increases, if we take a new raw material into use.

In many cases it is advantageous to set formal hypotheses and do some tests to show, which is the actual situation. Statistically, there are two possible hypotheses:

Null hypothesis claims that there is no significant difference between the populations. It can be written for means of two populations as follows:

$$H_0 : \mu_1 = \mu_2$$

The alternative hypothesis says that two populations differ from each other. There are two possible alternative hypotheses, a: double-sided

$$H_a : \mu_1 \neq \mu_2$$

In this case the user is not interested, which one of the alternatives is better. The situation might be even so that the tester does not know to which direction the variable in question effect. In the opposite case, we can use one-sided hypothesis

$$H_a : \mu_1 > \mu_2$$

With this kind of hypothesis we can test the effect of the variable in a more detailed way: e.g. the energy consumption of a new process is smaller than in the existing one. We can also test only one population against some fixed (target, constraint) value by writing:

$$H_0 : \mu_1 = \mu_o$$

$$H_a : \mu_1 < \mu_o$$

For instance, we can test, if the conductivity of our waste liquor is smaller than the limit set in the environmental permission for the plant.

In the above definitions, the variance can be tested instead of the mean. Of course, there can be more than two populations tested. Note that the definitions above are no actual equations, but more or less a formal way to write linguistic hypotheses in a mathematical form.

Working with hypotheses proceeds usually so that the experimenter tries to show that the null hypothesis is wrong with high enough probability, meaning that the alternative hypothesis can be accepted. If the null hypothesis cannot be proved wrong, it must be accepted.

Risks

Risk in this connection describes the probability to make a wrong decision from test data; i.e. to choose the wrong hypothesis. It is mainly controlled by the sample size. There are two possible errors that the experimenter can do:

Alpha error (α): the experimenter accepts the alternative hypothesis, while the null hypothesis is true

Beta error (β): the experimenter accepts the null hypothesis, while the alternative hypothesis is true

Of course, both errors cannot be made simultaneously. Numerical values are given as 0...1 or 0...100%. Usually values 0.95 or 95% are used (meaning that the error takes place with 95% probability), but the selection of the value is subjective. Note that these values equal to 5% risk. One guideline might be that, if accepting the alternative hypothesis lead to heavy investments, the probability of α-error should be kept small. We will see later that the selection of accepted risk will influence on the number of experiments in matrix designs**.**

Example. It is claimed that with a new control system for pulp cooking, the variance of the Kappa number is decreased under 4 units with 95% probability. It can also be said that the corresponding alternative hypothesis is accepted with an alpha risk of 5% (or 0.05).

Criterion

Quite often the experimenter wants to know, if the change he is doing has the expected effect in the studied system. Before starting experiments, he has to define the required minimum change and the β-risk that minimizes the probability of not accepting the advantageous change. They are needed in statistical testing.

This is necessary, when the whole population cannot be tested, but sampling is needed. This criterion depends on the variance, the acceptable risk and the sample size.

Example. Let us assume that we are testing, if steel alloying improves the tensile strength or not. The existing mean value (μ_o) is 30000 units and the acceptable minimum change is δ=1500. All products cannot be measured. Decision is made from a sample of products.

The hypotheses are now

$$H_0 : \mu_1 = 30000$$

$$H_a : \mu_1 > 30000$$

Following decisions are easy: (a) If the mean of samples is equal or less than 30000, alloying is not reasonable and (b) if the mean is bigger than 31500, it is advantageous. The problem appears if (c) the mean is between 30000 and 31500; what would happen, if the number of samples taken would be increased?

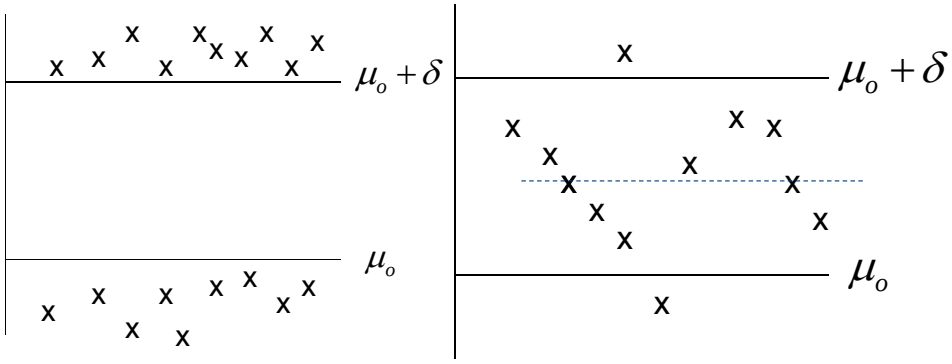


Figure 3.1. Situations (a) and (b) on the left and situation (c) on the right.

We need a criterion that depends on the variance, risk and sample size. In this case it tells how much bigger than 30000 the mean value must be so that we are on the safe side and can accept that the alloying is advantageous. Some thinking seems to tell that this value must be bigger with higher variance and it can be smaller, if more samples are taken. The smaller the α-risk we can take, the bigger the criterion must be. Based on this thinking we can write the general equation

$$\bar{X}^* = \mu_o + \frac{\sigma U_\alpha}{\sqrt{N}}$$

U_α depends on α-risk and the form of alternative hypothesis. For one-sided hypothesis and α=0.05 U_α=1.645. See statistical tables; on-line calculator is available for example in http://www.tutor-homework.com/statistics_tables/statistics_tables.html).

The alternative hypothesis is accepted, if

$$|\bar{X}| \geq \bar{X}^*$$

Null hypothesis is accepted, correspondingly, if

$$|\bar{X}| < \bar{X}^*$$

If β-risk is used, the equation becomes

$$\bar{X}^* = (\mu_o + \delta) - \frac{\sigma U_\beta}{\sqrt{N}}$$

Nest tables show examples on using both risks in this example. Remember that alpha risk means that the experimenter accepts the alternative hypothesis, while the null hypothesis is true.

α	σ	N	X^*
0.05	300	12	30142
0.05	1000	12	30475
0.05	300	24	30100
0.10	300	12	30111

β	σ	N	X^*
0.05	300	12	31358
0.05	1000	12	31025
0.05	300	24	31400
0.10	300	12	31389

If the samples are from two populations and the alternative hypothesis is written as

$$H_a : \mu_1 \neq \mu_2$$

the criterion is calculated as follows

$$(\bar{X}_1 - \bar{X}_2)^* = U_{\alpha/2}\,\sigma\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{1/2} \quad when \quad \bar{X}_1 > \bar{X}_2$$

$$(\bar{X}_1 - \bar{X}_2)^* = -U_{\alpha/2}\,\sigma\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{1/2} \quad when \quad \bar{X}_1 \leq \bar{X}_2$$

Sample size

The formula used in sample size calculations depends on the case; i.e. on the form of the hypotheses and if the variance is known.

H_o: $\mu_1 = \mu_o$; σ^2 known**:**

$N = (U_\alpha + U_\beta)^2(\sigma^2/\delta^2)$

H_o: $\mu_1 = \mu_2$; variances are known and **$\sigma_1^2 = \sigma_2^2$**:

$N = 2(U_\alpha + U_\beta)^2(\sigma^2/\delta^2)$

H_o: $\mu_1 = \mu_2$; variances are not equal

$$N_1 = \left(U_\alpha + U_\beta\right)^2 \frac{\sigma_1(\sigma_1 + \sigma_2)}{\delta^2}$$

$$N_2 = \left(U_\alpha + U_\beta\right)^2 \frac{\sigma_2(\sigma_1 + \sigma_2)}{\delta^2}$$

Example. The factory has prepared a light sensitive film for a longer time in the same process conditions. The mean of the film sensitivity is μ_o = 1.1 μJ/in^2. The factory wants to improve the sensitivity and it is believed that decreasing the film thickness from 20 mil (mil [=] 1/1000 inch) to 18 mil will give the right result. The variance is assumed to stay constant. s^2 = 0.01. Now in this case

H_o: $\mu_{18} = \mu_{20}$ = 1.1 μJ/in^2

H_a: μ_{18} < 1.1 μJ/in^2

α = 0.05, β = 0.10

δ= 0.10

$U_{0.05}$ = 1.645 and $U_{0.10}$ = 1.282

N = $(1.645 + 1.282)^2(0.01/0.01)$ = 8.567

This result means that 9 experiments must be done, if the given risk levels must be satisfied.

Example. The experimenter wants to test the similar product from two different vendors aiming to find out, if they have significant differences. Risks, criterion and variance are same as in the previous example. The hypotheses now are

H_o: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Because of two-sided alternative hypothesis, U_α must be taken from tables for two-sided distributions and

$U_{0.05} = 1.96$

U_β remains the same as in the previous example. The sample size is

$N = 2(1.96 + 1.282)^2(0.01/0.01) = 21.02$

This means that 21 runs are needed at minimum. With this number of tests the similarity of the products can be proved with the risks given before.

4. Two-level Hadamard Matrix Designs

This Section deals with Hadamard matrix for eight runs. It was originally developed by French mathematician Jacques Hadamard. Plackett ja Burman used it in experiment design 1945.

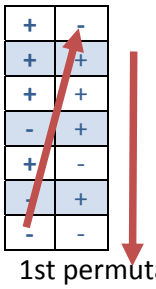
There are different Hadamard matrices (8x8-, 16x16-, 32x32, 64x64 and 128x128) developed from initial vectors by permutation. 8x8-matrix makes it possible to make 8 runs (T), for seven factors (T-1) at two levels (+,-).

Matrix generation

Initial vector consisting of seven elements is first written in a column and permutated six times

+
+
+
-
+
-
-

Initial vector



1st permutation

Other permutations follow the similar principle. This results in a matrix with seven columns and seven rows. Note that the order of elements in the initial vector can be different. It is essential that there are four plusses and three minuses. In the final matrix each variable will be four times at the plus-level and four times at the minus-level. This is guaranteed by writing a row of minuses as the eight row. The 8x8 matrix is completed by adding a column of plusses as the leftmost column. The columns are numbered starting from zero. Now the whole matrix is

0	1	2	3	4	5	6	7
+	+	-	-	+	-	+	+
+	+	+	-	-	+	-	+
+	+	+	+	-	-	+	-
+	-	+	+	+	-	-	+
+	+	-	+	+	+	-	-
+	-	+	-	+	+	+	-
+	-	-	+	-	+	+	+
+	-	-	-	-	-	-	-

This matrix is used in two level designs and seven factors can be tested at maximum. The calculated sample size must be 4 or less (each variables is tested four times at minus–level and four times at plus-level. Next, we will consider how it is used with different number of factors.

One factor

In this case, the experiment design for a factor (variable) A is red from column 1.

		A						
Run	0	1	2	3	4	5	6	7
1	+	+	-	-	+	-	+	+
2	+	+	+	-	-	+	-	+
3	+	+	+	+	-	-	+	-
4	+	-	+	+	+	-	-	+
5	+	+	-	+	+	+	-	-
6	+	-	+	-	+	+	+	-
7	+	-	-	+	-	+	+	+
8	+	-	-	-	-	-	-	-

Now, factor A is kept at the higher level in runs 1, 2, 3 and 5 and at the lower level in runs 4, 6, 7 and 8. The results from different runs are denoted later as response 1, response 2, etc. The effect (see the definition in Chapter 2) of factor A to the response is

$(\bar{X}_{A+} - \bar{X}_{A-})$ =(response 1 + response 2 + response 3 - response 4 + response 5 - response 6 - response 7 - response 8)/4

The selection of the criterion and the actual calculations are presented in following examples.

Two factors

In the two-factor case, the design matrix looks as follows. The experiment design is in columns 1 and 2. Column 4 is the contrast column for 1 and 2 and it is used in the calculations to reveal the effect of interaction between variables A ad B.

		A	B		-AB			
Run	0	1	2	3	4	5	6	7
1	+	+	-	-	+	-	+	+
2	+	+	+	-	-	+	-	+
3	+	+	+	+	-	-	+	-
4	+	-	+	+	+	-	-	+
5	+	+	-	+	+	+	-	-
6	+	-	+	-	+	+	+	-
7	+	-	-	+	-	+	+	+
8	+	-	-	-	-	-	-	-

Example. Copy machine should work in temperatures (A) between 100 – 200 degrees and with the relative humidity of air (B) between 30 - 80 % [Diamond, 1981]. Tests are done to define the effects of these two factors and their possible interactions. The output variable is the attachment of the colouring agent on the hot surfaces of the machine. Its variance is unknown.

The hypotheses now are

$H_{o1}: \mu_{100} = \mu_{200}$

$H_{a1}: \mu_{200} > \mu_{100}$

$H_{o2}: \mu_{30} = \mu_{80}$

$H_{a2}: \mu_{30} \neq \mu_{80}$

$H_{o3}: \mu_{AB} = 0$

$H_{a3}: \mu_{AB} > 0$

The risks and criterion are given by

α=0.1

β=0.1

δ=2.5σ

Note that the criterion is now given as a function of the variance that is actually unknown. We see the reason why later on. Next, the sample size is calculated. If we are going to use 8x8 matrix, it should be 4 at maximum. Now, instead of normal distribution, t distribution is used. We are expecting to have a small sample size! One of the alternative hypotheses is two-sided and therefore for α-risk a two sided t distribution is used.

$N_+ = N_- = 2(t_\alpha + t_\beta)^2\left(\sigma^2/\delta^2\right)$=4.2

$t(4,1-\alpha/2=0.95)=2.13$

$t(4,1-\beta=0.90)=1.53$

Note that "4" represents the assumed degrees of freedom in t distribution and statistical tables showing t-values as a function of degrees of freedom and the probability corresponding the risk in question are used. Using four runs results in a slight higher risk than required. The design matrix is as shown before. High and low levels for the variables are chosen as follows:

Variable	Low (-)	High (+)
Temperature (A)	100	200
Humidity (B)	30	80

After doing the test runs, the results look as follows

A	B	Result
200	30	16
200	80	32
200	80	28
100	80	15
200	30	14
100	80	17
100	30	9
100	30	12

Figure below shows the results graphically. High temperature and high moisture seem to lead to colour deposits in the hot surfaces of the machine. This seems logical. According to Chapter 2, there seems also to be interaction between these two variables. The question is, however, if these effects were statistically significant.

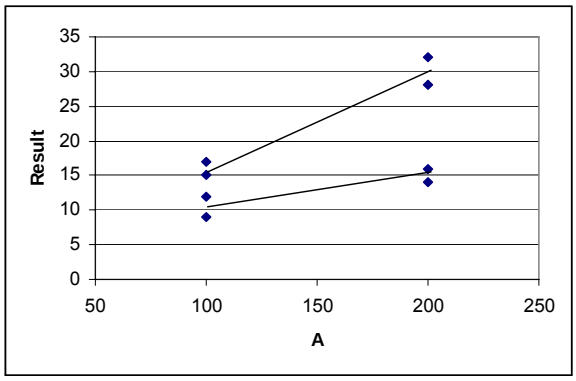


Figure 4.1. The results of the test runs with the copy machine. The lower line is for low humidity and, respectively, the upper line for high humidity.

Next, the effects are calculated for columns 1, 2 and 4 according to the same procedure as in one factor case:

	A	B	-AB
	+16	-16	+16
	+32	+32	-32
	+28	+28	-28
	-15	+15	+15
	+14	-14	+14
	-17	+17	+17
	-9	-9	-9
	-12	-12	-12
$\sum X_+ - \sum X_-$	+37	+41	-19

We see from here that the increase in temperature and humidity increases the response variable both directly and also through the interaction. Note that column 4 gives the negative effect of the interaction (-AB). The effects are now calculated by dividing the last row by 4; the sample size

X_{200}-X_{100} = 37/4 = 9.25

X_{80}-X_{30} = 41/4 = 10.25

X_{+AB}-X_{-AB} = -18/4 =-4.75

Next we need the criterion to which to compare the calculated effects. This requires variance of the response variable, but it is not given in this case. It could, however, be estimated with four degrees of freedom from four “free” columns (columns not reserved for any variable) 3, 5, 6, 7. It happens according to the same procedure as calculating the actual effects before:

	3	5	6	7
	-16	-16	+16	+16
	-32	+32	-32	+32
	+28	-28	+28	-28
	+15	-15	-15	+15
	+14	+14	-14	-14
	-17	+17	+17	-17
	+9	+9	+9	+9
	-12	-12	-12	-12
$\sum X_+ - \sum X_-$	-11	+1	-3	+1

In this way, we get four estimates for the variance

$S_3^2=(-11)^2/8=15.125$

$S_5^2=(1.0)^2(8=+0.125$

$S_6^2=(-3)^2/8=1.125$

$S_7^2=(1)^2/8=0.125$

The variance is now their average 4.125. There are both one-sided and two-sided hypotheses that both need their own criterion. Using the formula given before and respective α and β values we have for one-sided hypothesis 2.17 and for two-sided hypothesis 3.03 as the criterion. Comparing the above calculated effects (9.25, 10.25, -4.75) we see that their absolute values are bigger than the corresponding criteria. This means that all effects are statistically significant. As mentioned before, the risks are somewhat higher than required.

Three factors

In the three-factor case, all columns are reserved either for actual variables or their interactions.

		A	B	C	-AB	-BC	ABC	-AC
Run	0	1	2	3	4	5	6	7
1	+	+	-	-	+	-	+	+
2	+	+	+	-	-	+	-	+
3	+	+	+	+	-	-	+	-
4	+	-	+	+	+	-	-	+
5	+	+	-	+	+	+	-	-
6	+	-	+	-	+	+	+	-
7	+	-	-	+	-	+	+	+
8	+	-	-	-	-	-	-	-

Some conclusions can be drawn: All columns are in use; either for main effects or two-factor interactions. No columns are left for variance estimation. Replications are required for it. The more usual way, however, is to use centre point runs. All possible two-factor interactions can be evaluated (Resolution V), but if more factors are included, Resolution V does not realise. No interaction is in two columns and no column has been used for estimating two interactions.

From four to seven factors

If the fourth factor is included it is easy to realize that interactions cannot be reliably found. They must be assumed negligible or care and process knowledge must be practiced. Only main effects can be considered, but even then, be careful with the conclusions, because possible interactions disturb the analysis. One possibility to get over this is to repeat designs with the most important factors or use bigger matrix from the start.

	ABCD	BCD A	ACD B	ABD C	-AB -CD	-BC -AD	ABC D	-AC -BD
Run	0	1	2	3	4	5	6	7
1	+	+	-	-	+	-	+	+
2	+	+	+	-	-	+	-	+
3	+	+	+	+	-	-	+	-
4	+	-	+	+	+	-	-	+
5	+	+	-	+	+	+	-	-
6	+	-	+	-	+	+	+	-
7	+	-	-	+	-	+	+	+
8	+	-	-	-	-	-	-	-

Plackett-Burman screening design uses 8x8 Hadamard-matrix at Resolution III. It assumes no interactions and makes it possible to test seven variables with eight runs, if this assumption is valid. Screening here means testing to find the most important variables for actual testing.

Example. There are five variables influencing the production of a certain chemical [Diamond, 1981]. The quality of the chemical is described by the concentration of a side-product that should be minimized. The variables are

Code	Variable	+	-
A	Temperature	5 °C	15°C
B	Catalyser %	2.5 %	3.5 %
C	Mixing time	10 min	20 min
D	Solvent	acetone	toluene
E	Washing time	24 h	48 h

It is probable that there are interactions between at least two variables. The experiments are expensive; 2000 dollars each, and they take 3 days. They must also be accomplished in a sequence. The variance of the side product is 1.0 with 10 degrees of freedom. The target is to improve the process so that the concentration of the side product decreases from 13 % to only 1 %.

All alternative hypotheses are now two-sided

$(H_o)_A: \mu_5 = \mu_{15}$

$(H_a)_A: \mu_5 \# \mu_{15}$

Etc.

The risks and criterion are now

$\alpha = 0.10$

ß = 0.05

δ = 2.5 %

σ^2 = 1.0 and df = 10

Following table shows now, how the number of tests effects on the resolution, price and duration of the test.

Type	N	Resolution	Price, $	Duration, d
Full factorial	32	V+	64 000	96
Fractional f.	16	V	32 000	48
Fractional f.	8	III	16 000	24

Utilising the equation given before and the t test, the sample size is now 4.19. The last alternative is used. Note that all interactions cannot be found and the risks are a little higher than required. 8x8 Hadanard matrix is used. Variables D and E are now put in columns 4 and 5. The criterion with the given α-risk is now 1.27 (t test, df=10). The results are now

Run	Results (%)
1	15.5
2	2.5
3	12.0
4	8.0
5	13.5
6	7.0
7	12.0
8	13.6

Note that the value 1 % is nor achieved with any combination.

Following table shows the effects of each variable (A-E) and free columns (6-7).

Variable	A	B	C	D	E	6	7
Effect	0.75	-6.25	1.75	1	-3.5	2.25	-2

Negative effect means that the high value of the variable is better and v.v. If we compare the values with the criterion, we see that variables A and D are not significant. The high values of B and E and the low value of C are better. If we go back to the original Hadamard matrix, we see that runs 2 and 6 are done at these 'optimal' levels. Columns 6 and 7 show significance. In practice it means that there is some interactions effecting on the response variable. The problem is that it is impossible to tell exactly what interactions are in question. If you use the concept of contrast columns you can easily see that there are two interactions (for two variables) present in both columns 6 and 7.

One possibility to solve this problems is to repeat the whole design, but it would double the cost and time. There is, however, an alternative way:

Let's go back to look at the results of runs 2 and 6 which are done at the better levels of three significant variables. They, however, show very different results: 2.5 and 7 % (variance 1.0). This can be interpreted to be caused by some interactions. Next, two more tests are carried out. In these tests, B, C and E are kept at their 'optimal' levels, and other two combinations of A and D are tested:

Run	A	D	Result
2	+	-	2.5
6	-	+	7.0
9	-	-	0.7
10	+	+	10.1

The criterion for this case is 1.81. The effect for A is 2.45 and for D 6.95. The effect for AD is 0.65 so this interaction is not significant. This test tells that variables A and D are significant because of some interactions, but they could not tell which interactions they are.

More variables mean more runs

The following Table shows, how the number of factors tested increases when increasing the number of runs at different resolutions.

Number of runs	Resolution V	Resolution IV	Resolution III
16	1-4	5-8	9-15
32	1-6	7-16	17-31
64	1-8	9-32	33-53
128	1-11	12-64	65-127

5. Response Surface Methods

5.1 Introduction

Linear methods reveal main effects and interactions, but cannot find quadratic (or cubic) effects. Therefore they have limitations in optimization; the optimum is found in some edge point corresponding linear programming. They cannot model nonlinear systems; e.g. quadratic phenomena

$$Y = b_o + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2$$

In an industrial process even third-order models are highly unusual. Therefore, the focus will be on designs that are good for fitting quadratic models. Following example shows a situation where we are dealing with a nonlinear system and a two-level design does not provide us with the good solution.

Example. The yield in a chemical reactor as a function of the reaction time and temperature is studied with 2-level, 2 factor tests. Four runs give following results:

Time	Temperature	Yield
15	100	93
15	150	96
5	150	95
5	100	92

Figure 5.1. shows the results graphically. Higher temperature and longer reaction time give improved yield. The figure reveals no interaction between the variables.

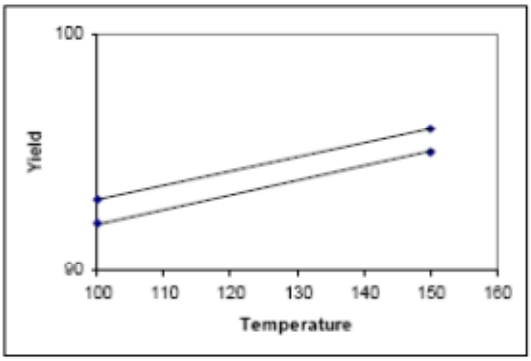


Figure 5.1. Yield versus temperature. The upper curve corresponds the longer reaction time.

There is, however, a chance that when the temperature increases, the reaction time improves the yield in a nonlinear fashion and there is an optimum point somewhere in the middle of the temperature range. Therefore, two more runs are done in the centre point with respect to the temperature:

Time	Temperature	Yield
15	100	93
15	150	96
5	150	95
5	100	92
15	125	98
5	125	93,5

Now, the relationship between the yield and temperature is no longer linear with the longer reaction time, but a clear optimum exists, when the temperature is 125 degrees and the reaction time is 15 minutes.

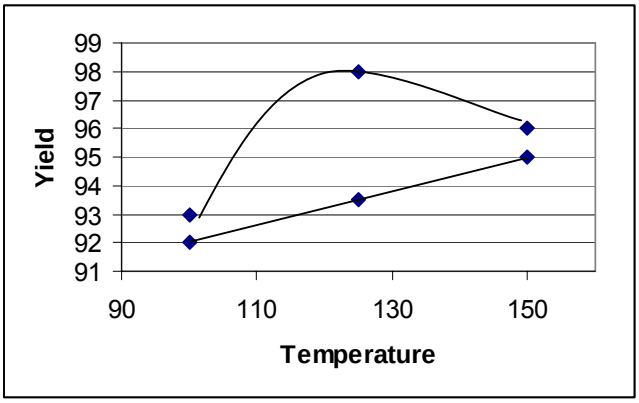


Figure 5.2. Graphical presentation with two centre point runs.

The example seems to point out that adding centre points into a two-level design would be enough. However, it cannot estimate individual pure quadratic effects, even though it can detect them effectively. Therefore, real three- (or higher) level designs should be used.

Including the third level in design means increasing the number of combinations of variable levels and, consequently, more experiments are needed. This is shown in the following table.

Number of factors	Combinations with three levels	Number of coefficients in a quadratic model
2	9	6
3	27	10
4	81	15
5	243	21
6	729	29

When nonlinearities are included in the design, the results give us an idea of the (local) shape of the response surface we are investigating. These methods are called response surface methods (RSM) designs. They are used in finding improved or optimal process

settings, in troubleshooting process problems, and in making a product or process more robust. Figure 5.3 shows an example of a response surface. It shows e.g. the price of the product as a function of the reaction temperature and pressure. The optimum lies in the centre of the region and it can be found numerically by modelling the response surface based on experimental data and using some optimization method (e.g. Nelder and Mead method, genetic algorithm, ect.) to locate point A numerically.

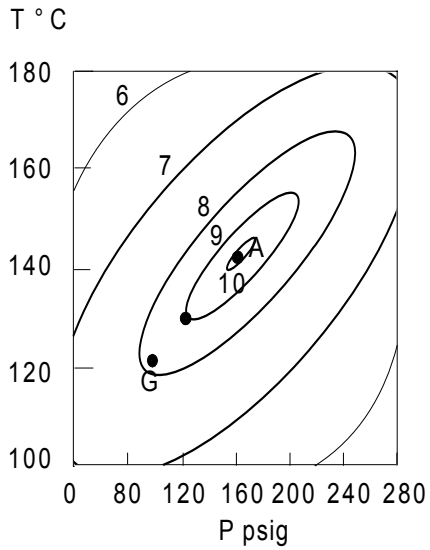


Figure 5.3. An example of the response surface.

5.2 (Box-Wilson) Central Composite Designs

Central Composite Design (CCD) has three different design points: edge points as in two-level designs (±1), star points at ±α; |α| |≥1 that take care of quadratic effects and centre points, Three variants exist: circumscribed (CCC), inscribed (CCI) and face centred (CCF)

CCC

CCC design is the original central composite design and it does testing at five levels. The edge points (factorial or fractional factorial points) are at the design limits. The star points are at some distance from the centre depending on the number of factors in the design. The star points extend the range outside the low and high settings for all factors. The centre points complete the design. Figure 5.4 illustrates a CCC design. Completing an existing factorial or resolution V fractional factorial design with star and centre points leads to this design.

CCC designs provide high quality predictions over the entire design space, but care must be taken when deciding on the factor ranges. Especially, it must be sure that also the star points remain at feasible (reasonable) levels.

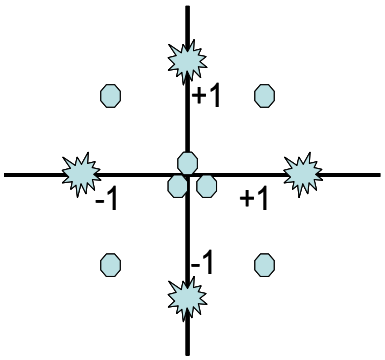


Figure 5.4. CCC design for two factors.

CCI

In CCI, the star points are set at the design limits (hard limits) and the edge points are inside the range (Figure 5.5). In a ways, a CCI design is a scaled down CCC design. It also results in five levels for each factor. CCI designs use only points within the factor ranges originally specified, so the prediction space is limited compared to the CCC.

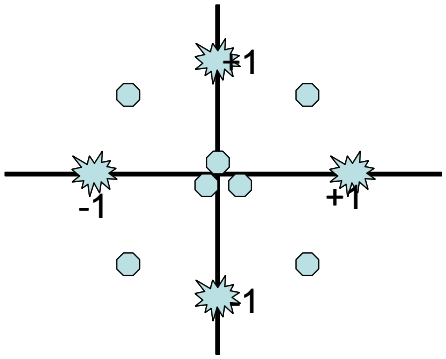


Figure 5.5. CCI design for two factors

CCF

In this design the star points are at the centre of each face of the factorial space, so $\alpha = \pm 1$ and only three levels are used (Figure 5.6). Complementing an existing factorial or resolution V design with appropriate star points can also produce this design. CCF designs provide relatively high quality predictions over the entire design range, but poor precision for estimating pure quadratic coefficients. They do not require using points outside the original factor range.

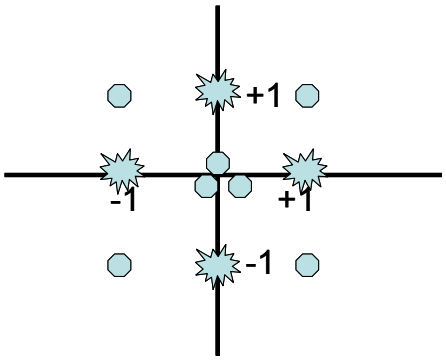


Figure 5.6. CCF design for two factors.

CCC with more than two variables

The following table shows the number of different points and the value for parameter α for some number of factors.

Factors	Edge points	Star points	Centre points	α
2	4	4	5	1.4142
3	8	5	6	1.63
4	16	8	7	2
5	16	10	6	2
6	32	12	9	2.378
7	64	14	14	2.828
8	128	16	20	3.364

Example. In this example, the casting strength is to be optimized for the casting time (A, 40-60 s) and temperature (B, 200-260 °C) [Diamond, 981]. Dependencies are supposed to be nonlinear and also interactions may exist. CCC design is shown in the next table.

Run	A	B
1	-1	-1
2	+1	-1
3	-1	+1
4	+1	+1
5	-α	0
6	+α	0
7	0	-α
8	0	+α
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0

Using the actual process values gives a table

Run	A	B
1	43	209
2	57	209
3	43	251
4	57	251
5	40	230
6	60	230
7	50	200
8	50	260
9-13	50	230

Following results are available after the experiments

Run	Strength
1	210
2	280
3	365
4	420
5	250
6	380
7	190
8	420
9	330
10	335
11	340
12	335
13	335

Below, the results are analysed using Minitab experiment design tool. An alternative way is given in [Diamond]. Figure 5.7 shows the surface plot of the strength as a function of time and temperature. A conventional regression program is used in fitting the parameters of the following quadratic model and the t test is applied in testing the significance of each parameter.

$$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_{12} AB + \beta_{11} A^2 + \beta_{22} B^2$$

The statistical analysis of these parameters is given as a Minitab print-out in Figure 5.8.

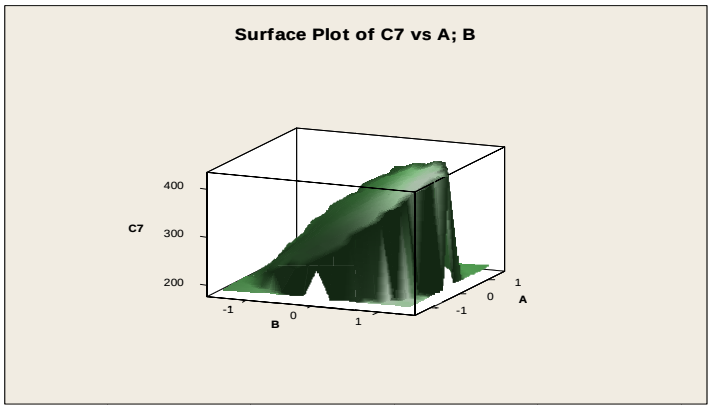


Figure 5.7. The surface plot of the casting strength.

•Estimated Regression Coefficients for C7

•Term	β	SE Coef	T	P
•Constant	335,000	4,631	72,342	0,000
•A	38,606	3,661	10,545	0,000
•B	77,534	3,661	21,179	0,000
•A*A	-7,813	3,926	-1,990	0,087
•B*B	-12,813	3,926	-3,264	0,014
•A*B	-3,750	5,177	-0,724	0,492

•S = 10,35 R-Sq = 98,8% R-Sq(adj) = 97,9%

Figure 5.8. The statistical analysis of regression coefficients. $T = \dfrac{\beta}{SE\ Coeff}$

The significance of each coefficient can be evaluated by looking T and P factors. Higher T and small P (≤0.1) mean a significant coefficient. In this case, both linear and quadratic effects are significant, but the interaction is not. Note that the model does not allow testing mixed quadratic-linear interactions of the type AB^2. It is also possible to use Anova and F-test in the same connection[1].

5.3 Box-Behnken Design

The Box-Behnken design is an independent quadratic design in that it does not contain an embedded factorial or fractional factorial design. In this design the treatment combinations are at the midpoints of edges of the process space and at the centre. These designs are

[1] http://doe.reliasoft.com/examples/doe_ex8/index.htm

rotatable[2] (or near rotatable) and require three levels of each factor. The designs have a limited capability for orthogonal blocking compared to the central composite designs.

These designs require fewer treatment combinations than a central composite design in cases involving 3 or 4 factors. Its "missing corners" may be useful when the experimenter should avoid combined factor extremes. This property prevents a potential loss of data in those cases. The design matrix for three factors is as follows:

A	B	C
-1	-1	0
-1	1	0
1	-1	0
1	1	0
-1	0	-1
-1	0	1
1	0	-1
1	0	1
0	-1	-1
0	-1	1
0	1	-1
0	1	1
0	0	0
0	0	0
0	0	0

Figure 5.9 shows the design graphically.

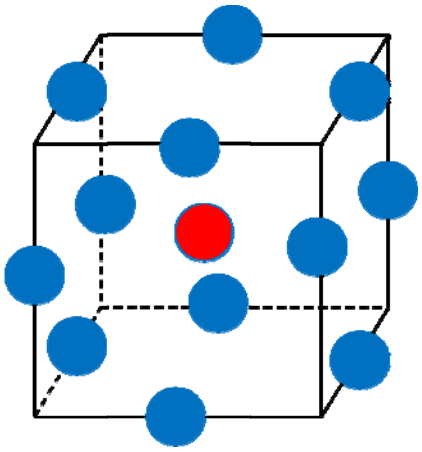


Figure 5.9. Box-Behnken design.

[2] In a rotatable design, the variance of the predicted values of *y* is a function of the distance of a point from the centre of the design and not a function of the direction the point lies from the centre]NIST].

5.4 D-optimal Designs

D-optimal designs are one form of design provided by a computer algorithm. These types of computer-aided designs are particularly useful when classical designs do not apply. Unlike standard classical designs such as factorials and fractional factorials, D-optimal design matrices are usually non-orthogonal and effect estimates are correlated. These types of designs are always an option regardless of the type of the model the experimenter wishes to fit (for example, first order, first order plus some interactions, full quadratic, cubic, etc.) or the objective specified for the experiment (for example, screening, response surface, etc.).

The optimality criterion results in minimizing the generalized variance of the parameter estimates for a pre-specified model; the 'optimality' of a given D-optimal design is model dependent. The experimenter must specify a model for the design and the total number of runs allowed and the computer algorithm chooses the optimal set of design runs from a candidate set. This candidate set usually consists of all possible combinations of various factor levels that one wishes to use in the experiment.

To put it in another way, the candidate set is a collection of treatment combinations from which the D-optimal algorithm chooses the treatment combinations to be included in the design. The computer algorithm generally uses a stepping and exchanging process to select the set of runs. Note that there is no guarantee that the design the computer generates is actually D-optimal.

The reasons for using D-optimal designs instead of standard classical designs generally fall into two categories: the standard factorial or fractional factorial design requires too many runs for the amount of resources or time allowed for the experiment or the design space is constrained; i.e. the process space contains factor settings that are not feasible or are impossible to run.

Example. Suppose that an industrial process has three design variables, and engineering judgment tells that the following model is an appropriate representation of the process.

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2$$

The levels being considered by the experimenter are (coded)

X_1: 5 levels (-1, -0.5, 0, 0.5, 1)

X_2: 2 levels (-1, 1)

X_3: 2 levels (-1, 1)

Due to resource limitations, only n = 12 runs can be done.

Given the experimental specifications, the first step in generating the design is to create a candidate set of runs. The candidate set is a data table with a row for each point (run) to be considered for the design, often a full factorial. For our problem, the candidate set is a full factorial in all factors containing 5*2*2 = 20 possible design runs. The table is omitted from here. It is available in [NIST][3]. The final design is shown in the Table below.

The optimality of D-optimal design is measured by D-efficiency. It is a function of the number of points in the design, the number of independent variables in the model, and the maximum standard error for prediction over the design runs. The best design is the one with the highest D-efficiency. The D-efficiency of the standard fractional factorial is 100 % (1), but it is not possible to achieve 100 % D-efficiency when pure quadratic terms are included in the model. In this case, D-efficiency is 0.68. The order of the design runs should be randomized.

Run	X1	X_2	X_3
1	-1	-1	-1
2	-1	-1	+1
3	-1	+1	-1
4	-1	+1	+1
5	0	-1	-1
6	0	-1	+1
7	0	+1	-1
8	0	+1	+1
9	+1	-1	-1
10	-1	-1	+1
11	+1	+1	-1
12	+1	+1	+1

Software packages may have different procedures and optimality criteria for generating D-optimal designs, so the final design may be different.

[3] http://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm

6. Some Experiment Design Programs

6.1 Matlab

Matlab[4] has all main experimental design programs available included in the Statistics Toolbox. Their use requires the basic skills in using Matlab. Matlab has versatile possibilities for results printing and presentation. The designs included are:

- Box-Behnken
- CCC (three algorithms)
- D-optimal (three algorithms)
- 2-factor full-factorial
- Fractional factorial
- Hadamard

Example. Some examples of Matlab commands:

d= fullfact([4 3]) designs a matrix for testing all combination of four machines and three operators in making some product. It gives possible combinations in a matrix of 12 rows:

M	O
1	**1**
2	**1**
3	**1**
4	**1**
1	**2**
2	**2**
3	**2**
4	**2**
1	**3**
2	**3**
3	**3**
4	**3**

d=ff2n(2) designs a four-run matrix for two variables at two levels

Run	A	B
1	0	0
2	0	1
3	1	0
4	1	1

d=fracfact('a b ab') creates a two-level fractional factorial design for two factors that takes also their interaction into account. It requires four runs.

[4] http://www.mathworks.se/products/matlab/

Run	A	B
1	-1	-1
2	-1	1
3	1	-1
4	1	1

d=hadamard(8) creates the eight-run Hadamard matrix shown already earlier.

You can learn how to use rsmtool in Matlab by running rsmdemo-prpgram. Thre are four commands available for doing D-optimal designs: cordexch, daugment, dcovary, and rowexch. For instance

settings = cordexch(2,9,'q')

creates a nine-run design for a fully quadratic model (parameter ‘q’)

Run	A	B
1	-1	1
2	1	1
3	0	1
4	1	-1
5	-1	-1
6	0	-1
7	1	0
8	0	0
9	-1	0

Analysis in Matlab can use all available Matlab tools: rstool(x,y) opens a GUI that can build linear, interactive and quadratic models, and nlintool(x,y, model,beta) is a corresponding GUI for response surface designs. Parameters can be transferred to workspace and normal model evaluation tools available: correlations, rmse, residuals, t-test for b-coefficients, etc, together with normal model evaluation tools

6.2 Minitab

Minitab is a versatile analysis tool that is based on a spreadsheet-like interface[5]. It makes it possible to explore data with graphs; e.g. normal plotting, histograms, scatter plots and printing. It has also statistical analysis tools available: descriptive statistics, ANOVA, control charts, and quality assessment tools.

For experiment design it offer the main tools like full factorial and fractional factorial designs, response surface designs, and Taguchi method. Minitab includes different

[5] http://www.minitab.com/en-FI/default.aspx

alternatives available for results analysis. They are based on effects analysis and modelling facilities.

6.3 Modde

Modde is an experiment design tool[6]. It is available for screening designs and response surface designs. All main methods are available and it has also versatile analysis tools. Results are shown mainly graphically as coefficient and effects plots, contour and surface plots together with a summary plot. It has also an efficient on-line help tool.

[6] http://www.umetrics.com/modde