5
Statistical Models in Chemical Engineering

The models based on the equations of transport phenomena and on stochastic models contain an appreciable quantity of mathematics, software creation, computer programming and data processing.

In many countries, a high level in mathematics is not a requirement for achieving a good knowledge in theoretical and practical chemistry or in chemical engineering, so, chemists or chemical engineers do not often have a deep knowledge of mathematics even though most areas of chemistry are often based upon quantitative measurements and computation. For example, the statistical validation of the techniques currently used in laboratories specializing in chemical analysis may be necessary to maintain the laboratory accreditation and/or for legal reasons In this case, the chemists or chemical engineers, who may have left formal training in mathematics 10 or 20 years before, could suddenly be faced with the need to brush up on statistics.

An important number of reference books on chemistry and chemical engineering statistics [5.1–5.11] have been published by specialists. The chemists and chemical engineers who intend to attend programs on statistical modelling of processes, must have a good basic knowledge in descriptive statistics, distribution of random variables and statistics hypotheses, and be able to carry out the experiments connecting the various measurements. These basic notions are therefore introduced in the following examples and discussions:

Descriptive statistics. A series of physical measurements can be described numerically. If, for example, we have recorded the concentration of 1000 different samples in a research problem, it is not possible to provide the user with a table giving all 1000 results. In this case, it is normal to summarize the main trends. This can be done not only graphically, but also by considering the overall parameters such as mean and standard deviation, skewness etc. Specific values can be used to give an overall picture of a set of data.

Distribution for random variables. The concept of distribution is fundamental to statistics. If a series of measurements is extracted from a great number of similar non-produced measurements (called population), we obtain a population sample. However, it is not possible to have the same mean characteristics for all the sam-

Chemical Engineering. Tanase G. Dobre and José G. Sanchez Marcano

ISBN: 978-3-527-30607-7

ples, because errors and noise influence the characterization properties of each sample. In fact, it is impossible for each sample to be identical. The distribution of measurements is often approximated by a normal distribution, although, in some cases, this does not represent an accurate model. If a sufficient number of measurements is taken during the analysis of samples, it is possible to see whether they fit into such a distribution.

If the number of samples with characteristics presenting a normal distribution is not significant, then we can have an error structure. This situation can also be due to outliers, i.e. samples that are atypical of the population or that might have been incorrectly labeled or grouped.

Statistics hypotheses and their testing. In many cases, the measurements are used to answer qualitative questions. For example, for the quality control of a batch of liquid products, a concentration analysis is carried out. If the analysis of a sample from the batch results in a higher concentration with respect to a reference value then we can reject the batch. In this case, we can use different tests to validate the rejection or acceptance of the batch. One example of such tests is the comparison of the mean values. Concerning the example described above, the measurements are realized by two groups of researchers, A and B. Group A has recorded twenty concentrations in a series of samples and has obtained a mean concentration value of 10 g/l and a deviation of 0.5 g/l. Group B, who monitors the same series of samples, has obtained a mean concentration value of 9.7 g/l and a deviation of 0.4 g/l. Then both mean values and deviations must be compared so as to answer the following questions: Are they actually different? What is the probability for both groups to have measured the same fundamental parameters? Is this difference in mean values simply caused by a different sampling or a variation in the measurement technique?

Relating measurements. Evaluating the relationships between the different types of measurements of the variables that are coupled or not to a process is fundamental in statistics. In the case of variables coupled to a process, the separation in the class of independent variables (x_i, I = 1, n) and dependent variables (y_j, j = 1, p) must be established based on the schematic representation of the process (see Fig. 1.1 in Chapter 1). The statistical models will be built based on experimental measurements. However, good models can be developed only if experimental results are obtained and processed from a statistical analysis. The analysis of neural networks processes, which are also statistical models, represents a modern and efficient research technique based on the experimental measurement of one actual process.

The first step for the analysis of a statistical modelling problem concerns the definition of the concept of statistical models. This definition is based on the diagram shown in Fig. 5.1 (which is a variation of Fig. 1.1 in Chapter 1). Statistical modelling contains all the statistical and mathematical procedures that use measured data of y_i (i = 1,P) and x_j (j = 1,N) simultaneously in order to obtain the mul-

tiple inter-dependences between dependent and independent variables. The relation (5.1) obtained on this basis represents the statistical model of a process:

$$y_i = f_i(x_1, x_2 \ldots x_n), i = 1, p \tag{5.1}$$

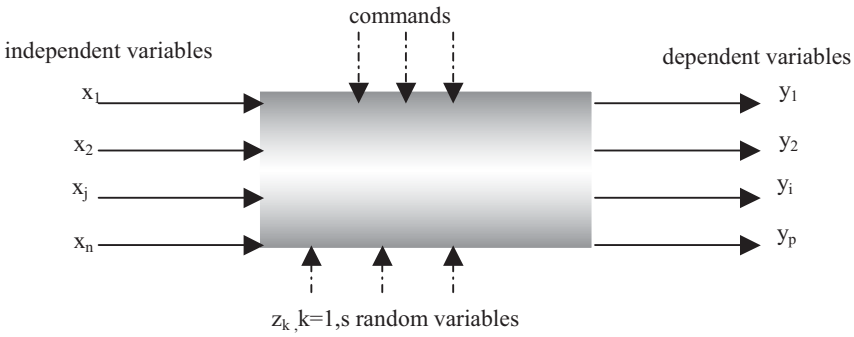


Figure 5.1 Schematic representation of a process.

5.1 Basic Statistical Modelling

The statistical modelling of a process can be applied in three different situations: (i) the information about the investigated process is not complete and it is then not possible to produce a deterministic model (model based on transfer equations); (ii) the investigated process shows multiple and complex states and consequently the derived deterministic or stochastic model will be very complex; (iii) the researcher's ability to develop a deterministic or stochastic model is limited.

The statistical modelling of a process presents the main advantage of requiring nothing but the inputs and outputs of the process (the internal process phenomena are then considered as hidden in a black box). We give some of the important properties of a statistical model here below:

1. As far as a statistical model has an experimental origin, it presents the property to be a model which could be verified (verified model).
2. Statistical models are strongly recommended for process optimization because of their mathematical expression and their being considered as verified models.
3. Classic statistical models cannot be recommended for the analysis of a dynamic process because they are too simple. Dynamic processes are better described by using the artificial neural network.

Two types of experiments can produce the data needed to establish statistical models. *Passive experiments* refer to the classical analysis of an experimental process investigation. They occur when the sets of experiments have been produced (in an industrial or in a pilot unit) either by changing the values of independent process variables one by one or by collecting the statistical materials obtained with respect to the evolution of the investigated process. *Active experiments* will be produced after the establishment of a working plan. In this case, the values of each of the independent variables of the process used for each planned experiment are obtained by specific fixed procedures.

To start the procedure of the statistical modelling of a process, we have to produce some initial experiments. These experiments will allow us:

1. to identify the domain of the value for each independent variable.
2. to identify the state of the dependent variables when the independent variables of the process increase.
3. to determine whether the state of the dependent variables of the process is affected by the interaction of the independent variables.

Dispersion and correlation analyses are used to process the data obtained in the preliminary experiments. The goal of these statistical analyses is to have qualitative or quantitative answers to points 2 and 3 mentioned above. Finally, when all the statistical data have been collected, a correlation and regression analysis will be used to obtain the inter-dependence relationships between the dependent and the independent variables of the process (see relation (5.1)).

In a process, when the value domain of each of the independent variables is the same in the passive and in the active experiments simultaneously, two identical statistical models are expected. The model is thus obtained from a statistical selection and its different states are represented by the response curves, which combine the input parameters for each of the output parameters.

Now, if the obtained model is used to produce output data and these are compared with the corresponding experimental results, some differences can be observed. This behavior is expected because the model has been extended outside the selection of its bases and this extension is only permissible if it is possible to take into account the confidence limits of the model.

Each of the independent variables x_1, x_2, x_3, ... x_N is frequently called a *factor* whereas the N-dimensional space containing the coordinates x_1, x_2, x_3, ... x_N is called *factorial space*; the *response surface* is the representation of one response function into N-dimensional space. A statistical model with a unique response surface would characterize the process that shows only one output (Fig. 5.1).

In a model, the number of response surfaces and the number of process outputs are the same. Figure 5.2 shows the surface response for a chemical reaction where the degree of transformation of the reactive species (dependent variables) in an expected product is controlled by their concentration and temperature (two independent variables). When we look at this figure, it is not difficult to observe

that the maximum degree of transformation can easily be established; so, ignoring the economic aspects of the process, the optimal states of the temperature and concentration are automatically given by the maximum conversion.

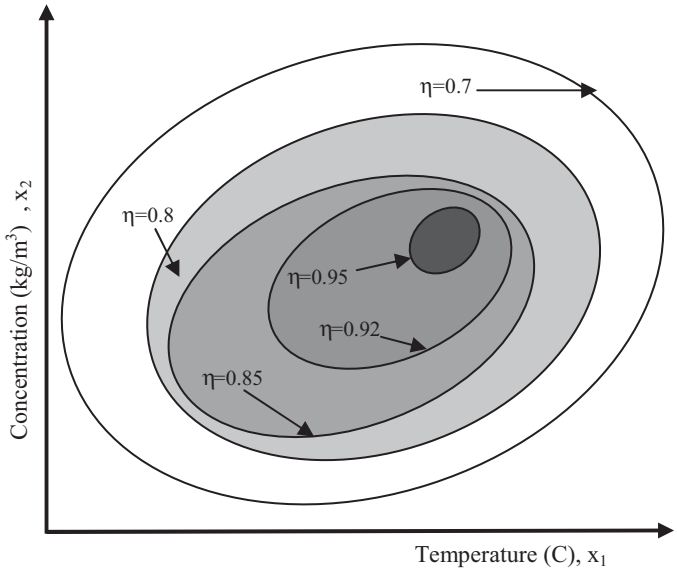


Figure 5.2 Reaction efficiency (y = η) versus temperature (x_1) and limiting reactant concentration (x_2).

The basis of the statistical model is given by the Taylor expansion of relation (5.1). It is established for the vicinity of the factors of the process where a fixed/an established value is given to the dependent variable (y_{i0}). In this expansion, y_{i0} results in the y_i value when the factors take the corresponding $x_{10}, x_{20}, \ldots x_{N0}$ values:

$$
\begin{aligned}
y_i = y_{i0} &+ \sum_{j=1}^{N} \left(\frac{\partial f_i}{\partial x_j}\right)_0 (x_j - x_{j0}) - \frac{1}{2!} \sum_{j=1}^{N} \sum_{k \neq J}^{N} \left(\frac{\partial^2 f_i}{\partial x_j \partial x_k}\right)_0 (x_j - x_{j0})(x_k - x_{k0}) - \\
&\frac{1}{2!} \sum_{j=1}^{N} \left(\frac{\partial^2 f_i}{\partial x_j^2}\right)_0 (x_j - x_{j0})^2 + \\
&\frac{1}{3!} \sum_{j=1}^{N} \sum_{k=1,k\neq j,k\neq 1}^{N} \sum_{l=1;l\neq j,l\neq k}^{N} \left(\frac{\partial^3 f_i}{\partial x_j \partial x_k \partial x_l}\right) (x_j - x_{j0})(x_k - x_{k0})(x_l - x_{l0}) + \\
&\frac{1}{3!} \sum_{j=1}^{N} \sum_{k=1,k\neq j}^{N} \left(\frac{\partial^3 f_i}{\partial^2 x_j^2 \partial x_k}\right) (x_j - x_{j0})^2 (x_k - x_{k0}) + \frac{1}{3!} \sum_{j=1}^{N} \left(\frac{\partial^3 f_i}{\partial x_j^3}\right) (x_j - x_{j0})^3 - \\
&\frac{1}{4!} \sum_{j=1}^{N} \sum_{k=1,k\neq j,j,m}^{N} \sum_{l=1,l\neq j,k,m}^{N} \sum_{m=1,m\neq j,k,l}^{N} \left(\frac{\partial^4 f_i}{\partial x_j \partial x_k \partial x_l \partial x_m}\right) \\
&(x_j - x_{j0})(x_k - x_{k0})(x_l - x_{l0})(x_m - x_{m0}) + \ldots..
\end{aligned}
\tag{5.2}
$$

It is not difficult to observe that the Taylor expression can be transposed as Eq. (5.3) where the index "i" has been extracted because it stays unchanged along the relation:

$$y^{(i)} = \beta_0^{(i)} + \sum_{j=1}^{N} \beta_j^{(i)} x_j + \sum_{j=1}^{N} \sum_{k=1,k\neq j}^{N} \beta_{jk}^{(i)} x_j x_k + \sum_{j=1}^{N} \beta_{jj}^{(i)} x_j^2 + \sum_{j=1}^{N} \sum_{k=1,k\neq j,l=1,l\neq j,k}^{N} \beta_{jkl}^{(i)} x_j x_k x_l + \sum_{j=1}^{N} \sum_{k=1,k\neq j}^{N} \beta_{jjk}^{(i)} x_j^2 x_k \sum_{j=1}^{N} \beta_{jjj}^{(i)} x_j^3 + \sum_{j=1}^{N} \sum_{k=1,k\neq j,j,m}^{N} \sum_{l=1,l\neq j,k,m}^{N} \sum_{m=1,m\neq j,k,l}^{N} \beta_{jklm}^{(i)} x_j x_k x_l x_m + \ldots\ldots \quad (5.3)$$

From the analysis of Eqs. (5.2) and (5.3) we can observe that each β coefficient has a specific expression. As an example, relation (5.4) shows the definition expression for $\beta_0^{(i)}$:

$$\beta_0^{(i)} = y_{i0} - \sum_{j=1}^{N} \left(\frac{\partial f_i}{\partial x_j}\right)_0 x_{j0} + \frac{1}{2!} \sum_{j=1}^{N} \sum_{k=1,k\neq j}^{N} \left(\frac{\partial^2 f_i}{\partial x_j \partial x_k}\right)_0 x_{j0} x_{k0} + \frac{1}{2!} \sum_{j=1}^{N} \left(\frac{\partial^2 f_i}{\partial x_j^2}\right)_0 x_{j0}^2 - \ldots \quad (5.4)$$

In Eq. (5.4) we can note that, in fact, the model describes the relationship between the process variables. Nevertheless, coefficients $\beta_0^{(i)}, \beta_j^{(i)}$, etc are still unknown because the functions $f_i(x_1,x_2,x_3,\ldots x_N)$ are also unknown.

A real process is frequently influenced by non-commanded and non-controlled small variations of the factors and also by the action of other random variables (Fig. 5.1). Consequently, when the experiments are planned so as to identify coefficients $\beta_0^{(i)}, \beta_j^{(i)}$, etc, they will apparently show different collected data. So, each experiment will have its own $\beta_0^{(i)}, \beta_j^{(i)}$, etc. coefficients. In other words, each coefficient is a characteristic random variable, which is observable by its mean value and dispersion.

Coefficients $\beta_0^{(i)}, \beta_j^{(i)}$, etc. (called *regression coefficients*) can be identified by means of an organised experiment. Since they have the quality to be the estimators of the real coefficients defined by relation (5.4), two questions can be formulated:

1. What is the importance of each coefficient in the obtained model?
2. What confidence can be given to each value of $\beta_0^{(i)}, \beta_j^{(i)}$, etc. when they are established as the result of programmed experiments?

The aim of statistical modelling is certainly not to characterize the relationship in a sample (experiment). So, after the identification of $\beta_0^{(i)}, \beta_j^{(i)}$.... etc., it is important to know what *confidence limits* can be given to the obtained model.

Each $\beta_0^{(i)}, \beta_j^{(i)}$..., etc. coefficient signification, is formally estimated. For instance, in this example, $\beta_0^{(i)}$ is the constant term for the regression relationship, $\beta_j^{(i)}$ corre-

sponds to the linear effects of the factors, $\beta_{jk}^{(i)}$ gives the effect of the interaction of x_j and x_k factors on the regression relationship, etc.

In relation (5.3), where $\beta_0^{(i)}, \beta_j^{(i)} \ldots$ etc. are the unknown parameters, we can observe that among the different methods to identify these parameters, the method of least-squares can be used without any restriction. So, the identification of $\beta_0^{(i)}, \beta_j^{(i)} \ldots$ etc. coefficients has been reduced to the functional minimisation shown in relation (5.5):

$$\Phi^{(i)}(\beta_0^{(i)}, \beta_j^{(i)}, \beta_{jk}^{(i)}, \ldots) = \sum_{i=1}^{Ne} \left(y_i^{(i),ex} - y_i^{(i),th} \right)^2 \tag{5.5}$$

where "Ne" gives the dimension of the experimental sample produced for the identification of the parameters; $y_i^{(i),ex}$ is the "i" experimental value of the output (i) and $y_i^{(i),th}$ is the "i" model-computed value of the output (i). This $y_i^{(i),th}$ is obtained using relation (5.3) and the numerical values of $x_{ji}, j = 1, N$. The dimension of the model (for the identification of the parameters) depends on the number of terms considered in relation (5.3). Table 5.1 gives the number of coefficients to be identified when the number of the factors of the process and the statistical model degree are fixed at the same time.

Table 5.1 Number of coefficients to be identified for the polynomial state of a statistical model.

Number of factors of the process	Statistical model with polynomial state (polynomial degree)			
	Number of identifiable coefficients			
	First degree	Second degree	Third degree	Fourth degree
2	3	6	10	15
3	4	10	20	35
4	5	15	35	70
5	6	21	56	126

In Table 5.1, where the statistical model is presented in a polynomial state, a rapid increase in the number of identifiable coefficients can be observed as the number of factors and the degree of the polynomial also increase. Each process output results in a new identification problem of the parameters because the complete model process must contain a relationship of the type shown in Eq. (5.3) for each output (dependent variable). Therefore, selecting the "Ne" volume and particularizing relation (5.5), allows one to rapidly identify the regression coefficients. When Eq. (5.5) is particularized to a single algebraic system we take only one input and one output into consideration. With such a condition, relations (5.3) and (5.5) can be written as:

$$y^{(1)} = y^{(1),th} = y = f_1(x_1, \beta_0^{(1)}, \beta_1^{(1)}, \beta_{11}^{(1)}; \beta_{111}^{(1)}....) = f(x, \beta_0, \beta_1, \beta_2, ...) \tag{5.6}$$

$$\Phi^{(1)}(\beta_0^{(1)}, \beta_1^{(1)}, \beta_{11}^{(1)}; \beta_{111}^{(1)}..) = \Phi(\beta_0, \beta_1, \beta_2, ..) = \sum_{i=1}^{Ne} (y_i - f(x_i, \beta_0, \beta_1, \beta_2.))^2 = \min \tag{5.7}$$

Now, developing the condition of the minimum of relation (5.7) we can derive relation (5.8). It then corresponds to the following algebraic system:

$$\begin{aligned} \frac{\partial \Phi(\beta_0, \beta_1, \beta_2, ...)}{\partial \beta_0} &= \frac{\partial \Phi(\beta_0, \beta_1, \beta_2, ...)}{\partial \beta_1} = \frac{\partial \Phi(\beta_0, \beta_1, \beta_2, ...)}{\partial \beta_2} = .. \\ &= \frac{\partial \Phi(\beta_0, \beta_1, \beta_2, ...)}{\partial \beta_n} = 0 \end{aligned} \tag{5.8}$$

Here β_n is the last coefficient from the function $f(x, \beta_0, \beta_1, \beta_2, ...)$. The different coefficients of this function multiply the x^n monomial, and "n" gives the degree of the polynomial that establishes the y–x relationship.

The computing of the derivates of relation (5.8) results in the following system of equations:

$$\begin{cases} \sum_{i=1}^{Ne} y_i \frac{\partial f(x_i, \beta_0, ..\beta_n)}{\partial \beta_0} - \sum_{i=1}^{Ne} f(x_i, \beta_0, \beta_1 ... \beta_n) \frac{\partial f(x_i, \beta_0, \beta_1, ...\beta_n)}{\partial \beta_0} = 0 \\ \sum_{i=1}^{Ne} y_i \frac{\partial f(x_i, \beta_0, ..\beta_n)}{\partial \beta_1} - \sum_{i=1}^{Ne} f(x_i, \beta_0, \beta_1 ... \beta_n) \frac{\partial f(x_i, \beta_0, \beta_1, ...\beta_n)}{\partial \beta_1} = 0 \\ \text{---} \\ \sum_{i=1}^{Ne} y_i \frac{\partial f(x_i, \beta_0, ..\beta_n)}{\partial \beta_n} - \sum_{i=1}^{Ne} f(x_i, \beta_0, \beta_1 ... \beta_n) \frac{\partial f(x_i, \beta_0, \beta_1, ...\beta_n)}{\partial \beta_n} = 0 \end{cases} \tag{5.9}$$

The system above contains N equations and consequently it will produce a single real solution for $\beta_0, \beta_1, ..., \beta_n$ (n unknowns). It is necessary to specify that the size of the statistical selection, here represented by Ne, must be appreciable. Moreover, whenever the regression coefficients have to be identified, Ne must be greater than n. This system (5.9) is frequently called: *system of normal equations* [5.4, 5.12–5.14].

In relation (5.5) we can see that $\Phi(\beta_0, \beta_1, \beta_2, ...)$ can be positive or null for all sorts of real $\beta_0, \beta_1, \beta_2, ...\beta_n$. As a consequence, it will show a minimal value for the identified $\beta_0, \beta_1, \beta_2, ...\beta_n$. Thus, the description of function $f(x, \beta_0, \beta_1, \beta_2, ...)$ results in a particularization of system (5.9).

If the number of independent variables is increased in the process, then the regression function will contain all the independent variables as well as their simple or multiple interactions. At the same time, the number of dependent variables also increases, and, for each of the new dependent variables, we have to consider the problem of identifying the parameters.

Coefficients $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ or $\beta_0^{(i)}, \beta_j^{(i)}, \beta_{jk}^{(i)}, \ldots$ can be considered as the estimators of the real coefficients of the Taylor expansion in the relationship between the variables of model (5.1). These coefficients are estimators of maximum confidence because their identification starts with minimizing the function which contains the square deviation between the observed and computed values of the output variables. The quality of the identified coefficient and, indirectly, the quality of the regression model depend firstly on the proposed regression function. Then, the quality of the regression function imposes the volume of experiments needed to produce the statistical model. Indeed, with a small number of experiments we cannot suggest a good regression function. However, in the case of a simple process, the regression function can be rapidly determined with only a few experiments. It is important to note that, after the identification of the coefficients, the regression model must be improved with a signification test. Only the coefficients that have a noticeable influence on the process will be retained and the model that contains the established coefficients will be accepted.

In Fig. 5.3 the different steps of the statistical modelling of a process are shown. These steps include the analysis of the variables, the planning and developing of experimental research and the processing of the experimental data needed to establish the model. We can observe that the *production of the statistical model of a process is time consuming and that the effort to bypass experimentation is considerable*. With respect to this experimental effort, it is important to specify that it is sometimes difficult to measure the variables involved in a chemical process. They include concentrations, pressures, temperatures and masses or flow rates. In addition, during the measurement of each factor or dependent variable, we must determine the procedure, as well as the precision, corresponding to the requirements imposed by the experimental plan [5.4]. When the investigated process shows only a few independent variables, Fig. 5.3 can be simplified. The case of a process with one independent and one dependent variable has a didactic importance, especially when the regression function is not linear [5.15].

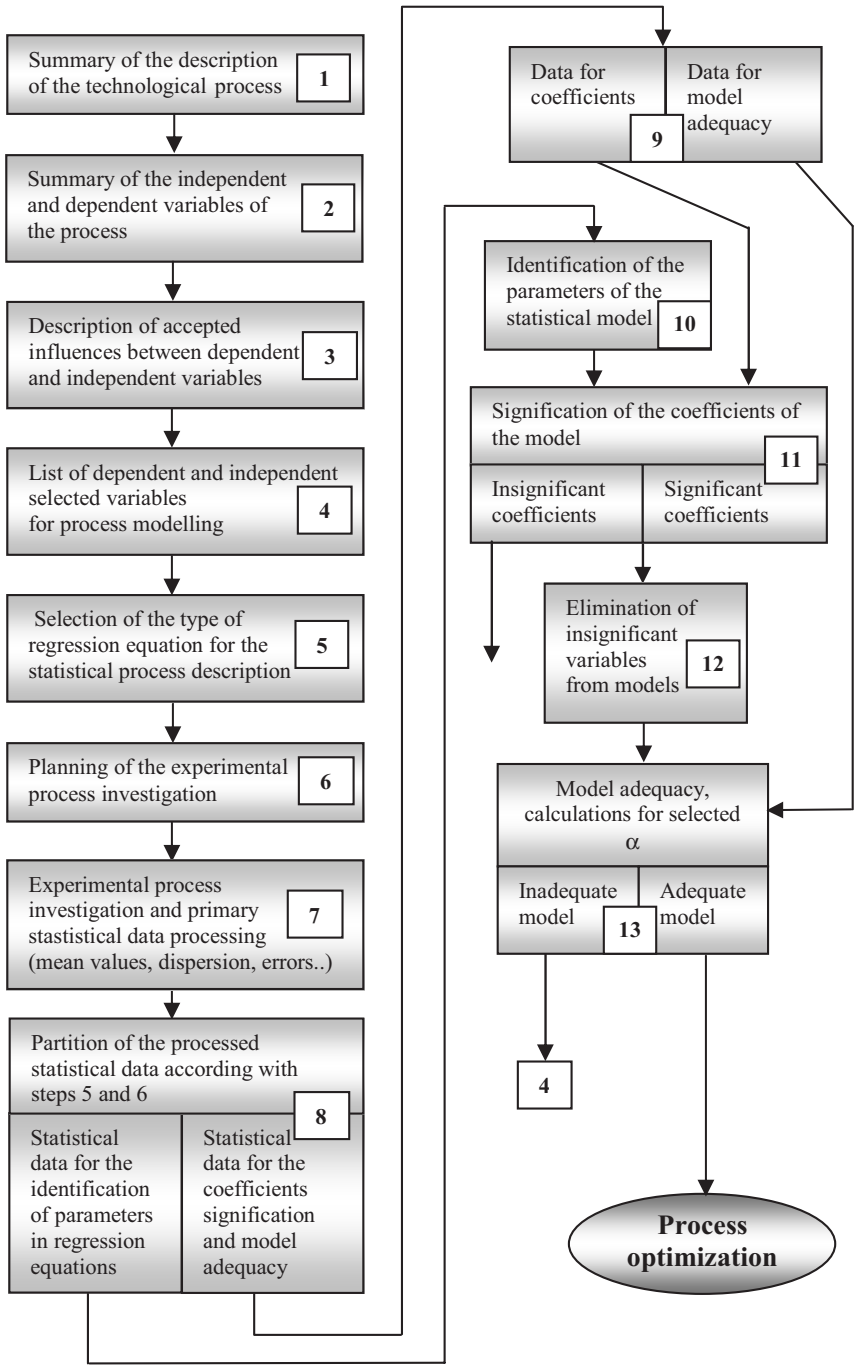


Figure 5.3 The different steps of statistical modelling.

5.2
Characteristics of the Statistical Selection

When we consider a process with only one input and one output variable, the experimental analysis of the process must contain enough data to describe the relationship between the dependent variable "y" and the independent variable "x". This relation can be obtained only if the data collected result from the evolution of one stationary process, and then supplementary experimental data can be necessary to demonstrate that the process is really in a stationary state.

As an actual process, we can consider the case of an isothermal and isobaric reactor working at steady state, where the input variable is the reactant's concentration and the output process variable (dependent variable) is the transformation degree. In this case, the values of the data collected are reported in Table 5.2. We can observe that we have the proposed input values (a prefixed set-point of the measurements) and the measured input values.

Table 5.2 Data for the characterization of y vs. x.

Current number for input	Proposed input value of x (set point)	Measured x value		Measured y value
		i	x_i	y_i
1	13.5	1	14.2	0.81
		2	13.5	0.75
		3	13.8	0.77
		4	14.3	0.75
		5	13.4	
2	20	1	20.5	0.66
		2	21.2	0.64
		3	19.8	0.63
		4	19.8	0.68
		5	19.5	0.65
		6		0.67
3	27	1	27.0	0.61
		2	27.4	0.59
		3	26.9	0.58

Table 5.2 Continued.

Current number for input	Proposed input value of x (set point)	Measured x value		Measured y value
		i	x_i	y_i
4	34	1	35.2	0.52
		2	34.7	0.49
		3	34.3	0.48
		4	35.1	0.55
		5	34.5	0.53
5	41	1	42.3	0.47
		2	42.6	0.43
		3	42.9	0.39
		4	41.8	0.46

In experimental research, each studied case is generally characterized by the measurement of x (x_i values) and y (y_i values). Each chain of x and each chain of y represents a statistical selection because these chains must be extracted from a very large number of possibilities (which can be defined as populations). However, for simplification purposes in the example above (Table 5.2), we have limited the input and output variables to only 5 selections. To begin the analysis, the researcher has to answer to this first question: "what values must be used for x (and corresponding y) when we start analysing of the identification of the coefficients by a regression function?" Because the normal equation system (5.9) requires the same number of x and y values, we can observe that the data from Table 5.2 cannot be used as presented for this purpose. To prepare these data for the mentioned scope, we observe that, for each proposed x value (x = 13.5 g/l, x = 20 g/l, x = 27 g/l, x = 34 g/l, x = 41 g/l), several measurements are available; these values can be summed into one by means of the corresponding mean values. So, for each type of x_i data, we use a mean value, where, for example, i = 5 for the first case (proposed x = 13.5 g/l), i = 3 for the third case, etc. The same procedure will be applied for y_i where, for example, i = 4 for the first case, i = 6 for the second case, etc.

With this method, we can create such couples as $(\overline{x}_1, \overline{y}_1), (\overline{x}_2, \overline{y}_2), ...(\overline{x}_5, \overline{y}_5)$ characterizing each case presented in Table 5.2. Thus, they can be used without any problem to solve the system of normal equations. Each class of finite data x_i or y_i with $i \succ 1$ represents a statistical selection.

The most frequently used statistical measure for a selection is the mean value. For a selection x_i with $i = 1, n$, the mean value ($\overline{x}$) will be computed by the following relation:

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_{n-1} + x_n) = \frac{1}{2}\sum_{i=1}^{n} x_i \tag{5.10}$$

In order to complete the selection characterization, we can use the variance or dispersion that shows the displacement of the selection values with respect to the mean value. Relations (5.11) and (5.12) give the definition of the dispersion:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{N}\left[\sum_{i=1}^{n} x_i\right]^2\right] \tag{5.11}$$

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{5.12}$$

It is often necessary to simplify the calculations by replacing the initial selection by another one, which presents the same mean value and dispersion [5.8, 5.9]. Therefore, if, for each value x_i, i = 1,n of the selection, we subtract the x_0 value, we obtain a new selection u_i, i = 1,n

$$u_i = x_i - x_0 \tag{5.13}$$

computing the mean value and the dispersion for this new selection we have:

$$\overline{u} = \tfrac{1}{n}\sum_{i=1}^{n}(x_i - x_0) = \overline{x} - x_0 \tag{5.14}$$

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \overline{u})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2 \tag{5.15}$$

Table 5.3 shows the values obtained after the calculation of the mean values and the dispersions respect to the statistical data presented in Table 5.2.

It is very important to pay attention to two important aspects: (i) the selection is a sample drawn from a population; (ii) the scope of the statistical analysis is to characterize the population by using one or more selections.

It is easily observable that each selection x_i and its associated y_i shown in Tables 5.2 or 5.3 correspond to a sample extracted from each type of population. In the current example we have 5 populations, which give the input reactant concentration, and 5 populations for the transformation degree of the reactant. In the tables, the first population associated to the input concentration corresponds to the experiment where the proposed concentration has the value 13.5 g/l.

During the experiment, the numerical characterization of the population is given by the concentration of the reactant associated to the flow of the material fed into the reactor. Therefore, this reactant's concentration and transformation degree are random variables. As has been explained above (for instance see Chapters 3 and 4), the characterization of random variables can be realized taking into account the mean value, the dispersion (variance) and the centred or non-centred momentum of various degrees. Indeed, the variables can be characterized by the following functions, which describe the density of the probability attached

Table 5.3 Mean values and dispersions for the statistical data given by Table 5.2.

Current number for input	Proposed input value for x	Measured x value				Measured y value			
		i	x_i	$\bar{x}$	s^2_x	y_i	$\bar{y}$	s^2_y	
1	13.5	1	14.2	13.86	$(14.2-13.86)^2 + (13.5-13.86)^2 + (13.8-13.86)^2 + (14.3-13.86)^2 + (13.4-13.86)^2 = 0.654$ $s^2 = 0.654/4 = 0.1635$	0.81	0.77	$(0.81-0.77)^2 + (0.75-0.77)^2 + (0.77-0.77)^2 + (0.75-0.77)^2 = 0.0024$ $s^2 = 0.0024/3 = 0.0008$	
		2	13.5			0.75			
		3	13.8			0.77			
		4	14.3			0.75			
		5	13.4						
2	20	1	20.5	20.16	0.473	0.66	0.655	0.00035	
		2	21.2			0.64			
		3	19.8			0.63			
		4	19.8			0.68			
		5	19.5			0.65			
		6				0.67			
3	27	1	27.0	27.1	0.07	0.61	0.593	0.0001015	
		2	27.4			0.59			
		3	26.9			0.58			
4	34	1	35.2	34.76	0.148	0.52	0.514	0.00083	
		2	34.7			0.49			
		3	34.3			0.48			
		4	35.1			0.55			
		5	34.5			0.53			
5	41	1	42.3	42.4	0.22	0.47	0.438	0.001291	

to the continuous random variable: repartition (5.16); mean value (5.17); variance (dispersion) (5.18); non-centred momentum of 'i' order (5.19); centred momentum of 'i' order (5.20):

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \tag{5.16}$$

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{5.17}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx \tag{5.18}$$

$$m_i = E(X^i) = \int_{-\infty}^{+\infty} x^i f(x)dx \tag{5.19}$$

$$M_i = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^i f(x)dx \tag{5.20}$$

The transposition from a selection to a population raises the following fundamental questions: *When a selection characterizes its original population? What is its procedure?* Until now, there has been no existing procedure able to prove whether or not a selection reproduces its original population identically. However, this fact can be improved if it is assumed that $\mu = \overline{x}$ and $\sigma^2 = s^2$. Nevertheless, we have to verify whether these identities are realistic using an acceptable confidence degree.

5.2.1
The Distribution of Frequently Used Random Variables

The distribution of a population's property can be introduced mathematically by the repartition function of a random variable. It is well known that the repartition function of a random variable X gives the probability of a property or event when it is smaller than or equal to the current value x. Indeed, the function that characterizes the density of probability of a random variable (X) gives current values between x and x + dx. This function is, in fact, the derivative of the repartition function (as indirectly shown here above by relation (5.16)). It is important to make sure that, for the characterization of a continuous random variable, the distribution function meets all the requirements. Among the numerous existing distribution functions, the normal distribution (N), the chi distribution (χ^2), the Student distribution (t) and the Fischer distribution are the most frequently used for statistical calculations. These different functions will be explained in the paragraphs below.

The famous normal distribution can be described with the following example: a chemist carries out the daily analysis of a compound concentration. The samples studied are extracted from a unique process and the analyses are made with identical analytical procedures. Our chemist observes that some of the results are

scarcely repeated whereas others are more frequently obtained. In addition, the concentration values are always found in a determined range (between a maximum and a minimum experimental result). By computing the apparition frequency of the results as a function of the observed apparition number and the total number of analysed samples, the chemist begins to produce graphic relationships between the apparition frequency of one result and the numerical value of the experiment.

The graphic construction of this computation is given in Fig. 5.4. Two examples are given: the first concerns the processing of 50 samples and the second the processing of 100 samples. When the mean value of the processed measurements has been computed, we can observe that it corresponds to the measurement that has the maximum value of apparition frequency. The differences observed between the two measurements are the consequence of experimental errors [5.16, 5.17]. Therefore, all the measurement errors have a normal distribution written as a density function by the following relation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.21}$$

Here μ and σ^2 are, respectively, the mean value and the dispersion (variance) with respect to a population. These characteristics establish all the integral properties of the normal random variable that is represented in our example by the value expected for the species concentration in identical samples. It is not feasible to calculate the exact values of μ and σ^2 because it is impossible to analyse the population of an infinite volume according to a single property. It is important to say that μ and σ^2 show physical dimensions, which are determined by the physical dimension of the random variable associated to the population. The dimension of a normal distribution is frequently transposed to a dimensionless state by using a new random variable. In this case, the current value is given by relation (5.21). Relations (5.22) and (5.23) represent the distribution and repartition of this dimensionless random variable. Relation (5.22) shows that this new variable takes the numerical value of "x" when the mean value and the dispersion are, respectively, $\mu = 0$ and $\sigma^2 = 1$.

$$u = \frac{x - \mu}{\sigma} \tag{5.22}$$

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{5.23}$$

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}} du = \mathrm{erf}(u) \tag{5.24}$$

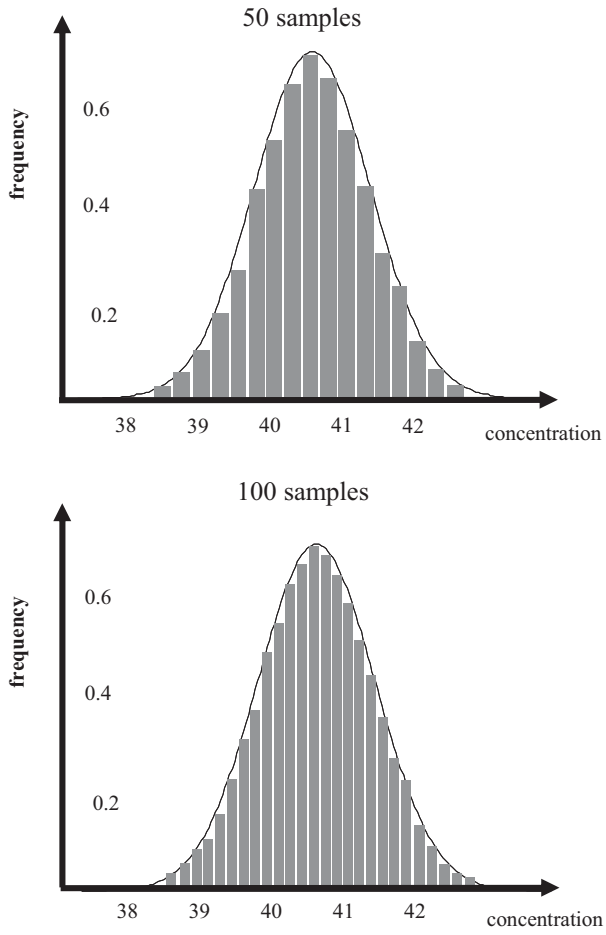


Figure 5.4 Graphic introduction of normal distribution.

Before presenting some properties of normal distribution, we have to present the relation (5.25) that gives the probability for which one random variable is fixed between a and b values (a<b), with the repartition function:

$$P(a < X < b) = F(b) - F(a) \tag{5.25}$$

The particularization of this general relation (5.25) to the dimensionless normal distribution results in the following observations:

1. the current value of the random variable is positioned within the interval $[\mu-\sigma, \mu+\sigma]$ with a probability equal to 0.684
2. the current value of the random variable is positioned within the interval $[\mu-2\sigma, \mu+2\sigma]$ with a probability equal to 0.955

3. the current value of the random variable is positioned within the interval [μ–3σ, μ+3σ] with a probability equal to 0.9975.

The observations mentioned above, which are graphically represented in Fig. 5.4, can also be demonstrated mathematically. For example, for the observation which gives $P(\mu - \sigma \prec x \prec \mu + \sigma) = 0.683$, we consider a = (x –μ)/σ = [(μ – σ) – μ]/σ = –1 and b = (x – μ)/σ = [(μ + σ)– μ]/σ = +1; with Eqs. (5.25) and (5.24) we can derive P(–1<u<+1) = erf(1) – erf(–1) = 0.8413 – 0.1586 = 0.6823.

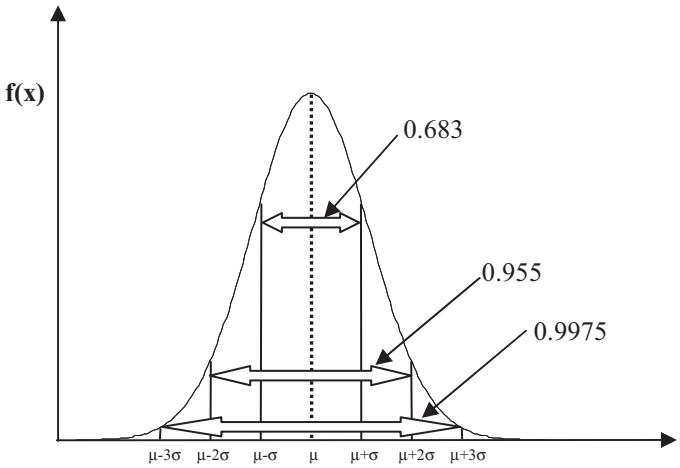


Figure 5.5 Some properties of a normal distribution (population (σ) or sample extracted (s)).

By using normal distribution, we can introduce other random variables, which are very important for testing the significance of $\beta_0, \beta_1, \beta_2, \beta_{12}$... coefficients as well as for testing the model confidence (see Fig. 5.3).

The first of these random variables is the chi distribution (χ^2). It is derived from relation (5.26), which defines the expression of the current random variable. Here μ and σ are the characteristics of a normal distribution; x_i is the current i value for the same normal distribution. It is easy to observe that a χ^2 distribution adds positive values, consequently $\chi^2 \in (0, \infty)$ and χ^2 is a dimensionless random variable. Relation (5.27) expresses the density of the χ^2 random variable. Here $\upsilon = n - 1$ represents the degrees of freedom of the χ^2 variable:

$$u_i = \frac{x_i - \mu}{\sigma} \approx \frac{x_i - \bar{x}}{\sigma}, \quad \chi^2 = \sum_{i=1}^{n} u_i^2 \tag{5.26}$$

$$f_\upsilon(\chi^2) = \frac{1}{2^{\frac{\upsilon}{2}}\sigma^\upsilon \Gamma\left(\frac{\upsilon}{2}\right)} (\chi^2)^{\frac{\upsilon}{2}-1} e^{-\frac{(\chi^2)}{2\sigma^2}} \tag{5.27}$$

For a rapid calculation, we can use the tabulated data values for the repartition function of the χ^2 variable $F_\upsilon(\chi^2)$. These tabulated data are obtained with Eq. (5.28):

$$F_\upsilon(\chi^2) = \int_0^{\chi^2_\alpha} f_\upsilon(\chi^2)d\chi^2 = 1 - \alpha \tag{5.28}$$

The second important random variable for statistical modelling is the Student (t) variable. It is derived from a normal variable, which is associated with "u" and χ^2 dimensionless random variables. Relation (5.29) introduces the current value of the Student (t) random variable:

$$t = \frac{u}{\sqrt{\frac{\chi^2}{\upsilon}}} \tag{5.29}$$

Equation (5.30), where $\Gamma(\upsilon)$ is given by relation (5.31) shows the probability to have a Student random variable with values between t and t + dt; so this relation gives the density function of the Student variable distribution:

$$f_\upsilon(t) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{(\upsilon+1)\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{t^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} \tag{5.30}$$

$$\Gamma(\upsilon) = \int_0^\infty t^{\upsilon-1}e^{-t}dt \tag{5.31}$$

The third random variable is the Fischer variable. It is defined by the use of two normal variables, each of which is expressed by a χ^2 random variable. The current Fischer variable is given by Eq. (5.32) where $\upsilon_1 = n - 1$ and $\upsilon_2 = m - 1$ represent the degrees of freedom associated, respectively, to random variables χ_1^2 and χ_2^2.

$$x = \frac{\chi_1^2}{\chi_2^2} \cong \frac{\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x})_1^2}{\sigma_1^2}\right]}{\left[\sum_{i=1}^{m}\frac{(x_i-\bar{x})_2^2}{\sigma_2^2}\right]} \cong \frac{\frac{\sigma_2^2}{\upsilon_2}}{\frac{\sigma_1^2}{\upsilon_1}} \tag{5.32}$$

The values of the Fischer variable are within the interval (0, ∞). The density of probability for this variable is given by Eq. (5.33):

$$f_{\upsilon_1,\upsilon_2}(x) = f_{\nu_1}(\chi_1^2)/f_{\nu_2}(\chi_2^2) \tag{5.33}$$

For a rapid calculation of this variable, we can use the tabulated values for the Fischer repartition function $F_{\upsilon_1,\upsilon_2}(x)$ corresponding to the confidence limits $\alpha = 0.05$ and $\alpha = 0.01$.

5.2.2
Intervals and Limits of Confidence

The paragraphs above show that μ and σ^2 are the most important characteristics for a random variable attached to a given population. Nevertheless, from a practical point of view, the main characteristics of μ and σ^2 remain unknown. Therefore, we have the possibility to draw one or more statistical selection(s) concerning the property considered by the associated random variable from a population. However, with this procedure, we cannot estimate μ and σ^2 for the whole population directly from the mean value and dispersion of the selection. The acceptance of the statement, which considers that the population mean value μ is placed in an interval containing the selection mean value ($\bar{x}$), must be completed with the observation that the placement of μ near $\bar{x}$ is a probable event. The probability of this event is recognized as the "confidence", "probability level" or "confidence level". A similar processing is carried out for σ^2 and s^2. If we define the probability by α, which shows that μ or σ^2 are not placed in a confidence interval, then, $1 - \alpha$ is the probability level or confidence level. α is frequently called the "significance limit". Figure 5.6 gives the graphic interpretation for α in the case of a normal repartition with $\mu = 0$ and $\sigma^2 = 1$.

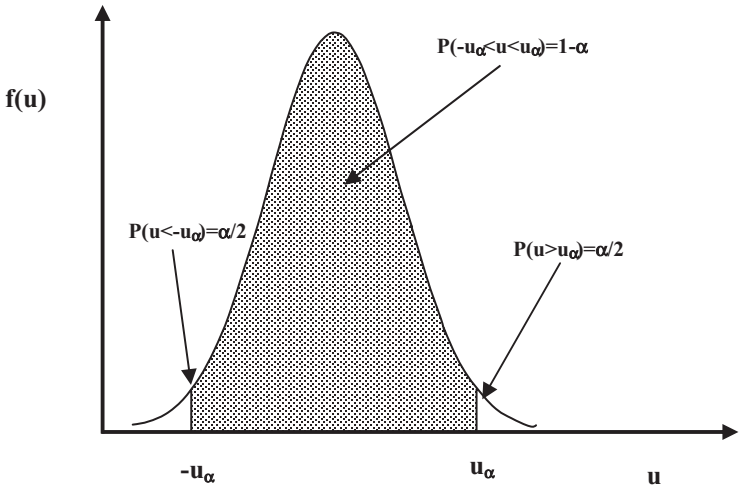


Figure 5.6 Diagram for the definition of the significance level.

Considering Fig. 5.6, we observe that, if we have a very high confidence level, then $1 - \alpha \rightarrow 1$ and the domain for the existence of parameters (μ, σ^2) is high. As far as our scope is to produce the relations between the population and the selection characteristics, i.e. between the couples (μ, σ^2) and $(\bar{x}, s^2)$, we can write Eq. (5.17) in a state that introduces the mean value ($\bar{x}$) and volume (n) of the selection. In relation (5.34) the population mean value has been divided into n parts. Now, if for each interval $a_{i-1} - a_i$, the population mean value is compared with the mean

value of the selection that has a similar volume, then relation (5.34) can be written as (5.35):

$$\mu = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{\alpha_1} xf(x)dx + \int_{\alpha_1}^{\alpha_2} xf(x)dx + ... + \int_{\alpha_n}^{+\infty} xf(x)dx = \mu_1 + \mu_2 + ... + \mu_n \tag{5.34}$$

$$\mu = \overline{x}_1 + \overline{x}_2 + \overline{x}_3 + ... + \overline{x}_n \tag{5.35}$$

Now if we consider the population variance (dispersion), each identical interval $a_{i-1} - a_i$ presents a dispersion which depends on the global σ^2, thus, we can write:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 = \left(\frac{\sigma^2}{n} + ... + \frac{\sigma^2}{n}\right) = n\left(\frac{\sigma^2}{n}\right) \tag{5.36}$$

The above relation shows that each of the n divisions of the population has the σ^2/n dispersion. Now, considering that a division $\overline{x} - \mu$ is a normal random variable and that the mean value of this variable is zero, we can transform relation (5.22) into relation (5.37) where u keeps its initial properties (mean value is zero and dispersion equal to unity):

$$u = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{5.37}$$

It is known that $P(a \le \chi^2 \le b) = P\left(a \le \frac{(n-1)s^2}{\sigma^2} \le b\right) = (1 - \alpha)$ and then, with an accepted significance limit, we can derive the confidence interval considering that $a = \chi^2_{1-\alpha/2}$ and $b = \chi^2_{\alpha/2}$. Thus, we obtain the following results:

$$P\left(\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2}\right) = 1 - \alpha$$

or:

$$\chi^2_{1-\alpha/2} \le \frac{n-1}{\sigma^2}s^2 \quad \text{and} \quad \chi^2_{\alpha/2} \ge \frac{n-1}{\sigma^2}s^2$$

and:

$$\sigma^2 \le \frac{n-1}{\chi^2_{1-\alpha/2}}s^2 \quad \text{and} \quad \sigma^2 \ge \frac{n-1}{\chi^2_{\alpha/2}}s^2 \tag{5.38}$$

The intersection of the expressions contained in Eq. (5.38) gives the expression for the confidence interval $I = \left(\frac{n-1}{\chi^2_{\alpha/2}}s^2; \frac{n-1}{\chi^2_{1-\alpha/2}}s^2\right)$. Here, for $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$, we use tabulated or computed values which correspond to the degrees of freedom $\nu = (n - 1)$ where n is the number of selected experiments.

When the selection contains a small number of measurements (for example n<25), the confidence interval for the mean value will be obtained by the use of the dimensionless Student variable given here by the current value (5.39):

$$t = \frac{\mu}{\sqrt{\frac{\chi^2}{\upsilon}}} = \frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\upsilon s^2}{\upsilon \sigma^2}}} = \frac{\bar{x} - \mu}{s}\sqrt{n} \tag{5.39}$$

Because $t \in (-\infty, +\infty)$ for a fixed significance level, we can write $P(-t_\alpha \le t \le t_\alpha) = 1 - \alpha$. Now the substitution of Eq. (5.39) into $P(-t_\alpha \le t \le t_\alpha) = 1 - \alpha$ results in the following relations:

$$P\left(-t_\alpha \le \frac{\bar{x} - \mu}{s}\sqrt{n} \le t_\alpha\right) = 1 - \alpha \tag{5.40}$$

or:

$$-t_\alpha \le \frac{\bar{x} - \mu}{s}\sqrt{n} \quad \text{and} \quad t_\alpha \ge \frac{\bar{x} + \mu}{s}\sqrt{n} \tag{5.41}$$

and:

$$\mu \le \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \quad \text{and} \quad \mu \ge \bar{x} - t_\alpha \frac{s}{\sqrt{n}} \tag{5.42}$$

The expressions from relation (5.42) show that the confidence interval for a mean value with a small number of measurements is:

$$I = \left(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}; \bar{x} + t_\alpha \frac{s}{\sqrt{n}}\right).$$

5.2.2.1 A Particular Application of the Confidence Interval to a Mean Value

The scope of this section is to show a practical application of the confidence interval to a mean value. The example below concerns the data given in Table 5.2. In order to verify the correctness of the data obtained, the chemist has carried out new measurements in case the proposed x should be near 20 g/l. Table 5.4 gives the new results obtained for the concentration of the reactant in the reactor feed. Concerning these data two questions are raised:

1. What is the confidence interval for the mean value of the population from which the selection in Table 5.4 has been extracted?
2. What is the difference between these new data and those given in Table 5.2?

Table 5.4. New values of the limiting reactant concentration in the reactor feed (Data equivalent to column 2 in Table 5.2).

Sample number (i)	Concentration x_i, g/l	Sample number (i)	Concentration x_i, g/l	Sample number (i)	Concentration x_i, g/l	Sample number (i)	Concentration x_i, g/l
1	19.4	9	21.2	17	18.4	25	21.6
2	22.2	10	18.7	18	18.1	26	20.4
3	21.9	11	19.3	19	18.9	27	18.5
4	23.2	12	18.7	20	22.0	28	20.8
5	19.8	13	23.5	21	18.5	29	18.8
6	21.3	14	22.5	22	20.5	30	22.1
7	17.8	15	18.9	23	18.7	31	20.7
8	23.2	16	19.3	24	21.1	32	19.2

The answers to the questions above are obtained numerically with the following procedure and the corresponding algorithm:

1. We compute the selection mean value ($\bar{x}$) and the dispersion (s^2) with the data from Table 5.4 and with Eqs. (5.10) and (5.12).
   Result: $\bar{x} = 20.3$ g/l; $s^2 = 3.86$; $s = 1.92$ g/l

2. We accept the equality between the population and the selection dispersion, i.e. $\sigma^2 = s^2$
   Result: $\sigma^2 = 3.86$; $\sigma = 1.92$ g/l

3. We establish the probability significance level (α).
   Result: $\alpha = 0.05$

4. Equation $\frac{1}{\sqrt{2\pi}} \int_{-u_\alpha}^{u_\alpha} e^{-\frac{u^2}{2}} du = 1 - \alpha$ is resolved in order to estimate u_α.
   Result: $u_\alpha = 1.96$.

   Observation: For this purpose we must use a computer program. Alternatively, we can also use the tabulated data of the normal u_α at various fixed α.

5. We obtain the mean value confidence with relation
   $I = (\bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}}; \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}})$

   Result: I = (19.5; 21)

6. We calculate the selection mean value ($\bar{x}$) and the dispersion (s^2) with the data from Table 5.2 column 2 and with Eqs. (5.10) and (5.12).
   Results: $\bar{x} = 20.16$ g/l; $s^2 = 0.473$; $s = 0.687$ g/l).

7. According to point 2 of the present algorithm, we accept the equality between the population and the selection dispersion $\sigma^2 = s^2$.
   Results: $\sigma^2 = 0.473$; $\sigma = 0.687$ g/l

8. We observe that for χ^2 variable $\nu = n - 1$.
   Result: $\nu = 4$

9. Equation $\int_{-t_\alpha}^{t_\alpha} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{t^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} dt = 1-\alpha$ is solved for t_α unknown.
   Result: $t_\alpha = 2.776$

10. We obtain the mean value confidence according to relation

$$I = \left(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}; \bar{x} - t_\alpha \frac{s}{\sqrt{n}}\right)$$

    Result: I = (19.307; 21.013)

11. Conclusion: The obtained results for the confidence intervals I = (19.5; 21) and I = (19.307; 21.013) show that the compared selections are almost the same or have a similar origin.

5.2.2.2 An Actual Example of the Calculation of the Confidence Interval for the Variance

The purpose of this section is to show the calculation of the confidence interval for the variance in an actual example. The statistical data used for this example are given in Table 5.3. In this table, the statistically measured real input concentrations and the associated output reactant transformation degrees are given for five proposed concentrations of the limiting reactant in the reactor feed. Table 5.3 also contains the values of the computed variances for each statistical selection. The confidence interval for each mean value from Table 5.3 has to be calculated according to the procedure established in steps 6–10 from the algorithm shown in Section 5.2.2.1. In this example, the number of measurements for each experiment is small, thus the estimation of the mean value is difficult. Therefore, we can compute the confidence interval for the dispersion $\left(I = \left(\frac{n-1}{\chi^2_{\alpha/2}}s^2; \frac{n-1}{\chi^2_{1-\alpha/2}}s^2\right)\right)$

for each experiment only if we establish the degrees of freedom ($\nu = n - 1$, where n is the number of experiments from each experimentation), and for a chosen α, we obtain the quintiles values $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$. These are the solutions of the following system of equations:

$$\begin{cases} \int_{\chi_{1-\alpha/2}}^{\chi^2_{\alpha/2}} f_\upsilon(\chi^2)d\chi^2 = 1 - \alpha \\ \int_{\chi_{1-\alpha/2}}^{\infty} f_\upsilon(\chi^2)d\chi^2 = 1 - \alpha/2 \end{cases} \tag{5.43}$$

Table 5.5 gives the results obtained for the mean value and dispersion intervals for a significance limit $\alpha = 0.05$.

Table 5.5 The confidence intervals of the mean value and dispersion for the data from Table 5.3.

Current number for input	n x/y	$\bar{x}$	t_α	I from Eq. (5.42)	s^2_x	$\chi^2_{1-\alpha/2}$	$\chi^2_{\alpha/2}$	I from Eq. (5.38)
1	5/4	13.86	2.571	13.67; 14.04	0.163	1.15	11.1	0.058; 0.265
2	5/6	20.16	2.571	20.43; 19.97	0.473	1.15	11.1	0.172; 1.641
3	3/3	27.10	3.182	27.23; 26.97	0.070	0.352	7.81	0.018; 0.398
4	5/5	34.76	2.571	34.58; 34.94	0.148	1.15	11.1	0.053; 0.514
5	4/4	42.40	2.776	42.09; 42.705	0.220	0.711	9.49	0.069; 0.928

Current number for input	$\bar{y}$	t_α	I from (5.48)	$s^2_y * 10^2$	$\chi^2_{1-\alpha/2}$	$\chi^2_{\alpha/2}$	I from (5.44)
1	0.77	2.776	0.772; 0.768	0.080	0.711	9.49	0.033; 0.331
2	0.655	2.447	0.654; 0.656	0.035	1.64	12.6	0.013; 0.106

Table 5.5 Continued.

Current number for input	$\bar{y}$	t_α	I from (5.48)	$s^2_y * 10^2$	$\chi^2_{1-\alpha/2}$	$\chi^2_{\alpha/2}$	I from (5.44)
3	0.593	3.182	0.592; 0.594	0.010	0.352	7.81	0.002; 0.036
4	0.514	2.571	0.513; 0.515	0.083	1.15	11.1	0.03; 0.288
5	0.438	2.776	0.434; 0.442	0.129	0.711	9.49	0.041; 0.544

5.2.3
Statistical Hypotheses and Their Checking

The introduction of the formulation of the statistical hypotheses and their checking have already been presented in Section 5.2.2.1 where we proposed the analysis of the comparison between the mean values and dispersions of two selections drawn from the same population. If we consider the mean values in our actual example, the problem can be formulated as follows: if $\bar{x}_1$ is the mean value calculated with the values in Table 5.4 and $\bar{x}_2$ is the mean value for another selection extracted from the same population (such as for example $\bar{x}_2$, which is the limiting reactant concentration at the reactor input for Table 5.2, column 2) we must demonstrate whether $\bar{x}_1$ is significantly different from $\bar{x}_2$.

A similar formulation can be established in the case of two different dispersions in two selections extracted from the same population. Therefore, this problem can also be extended to the case of two populations with a similar behaviour, even though, in this case, we have to verify the equality or difference between the mean values μ_1 and μ_2 or between the variances σ_1^2 and σ_2^2. We frequently use three major computing steps to resolve this problem and to check its hypotheses:

- First, we begin the problem with the acceptance of the zero or null hypothesis. Concerning two similar populations, the null hypothesis for a mean value shows that $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$. Thus, we can write $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 - \sigma_2^2 = 0$ for dispersion. We have $\bar{x}_1 = \bar{x}_2$ or $\bar{x}_1 - \bar{x}_2 = 0$ for both selections and $s_1^2 = s_2^2$ or $s_1^2 - s_2^2 = 0$ for the mean value and dispersion respectively.
- Then, we obtain the value of a random variable associated to the zero hypothesis and to the commonly used distributions, we establish the value of the correlated repartition function, which is in fact a probability of the hypothesis existence.

- Finally, we accept a confidence level and we compare this value with those given by the repartition function and we eventually accept or reject the null hypothesis according to this comparison.

Table 5.6 presents the statistical hypotheses frequently formulated and the tests used for their validation.

Table 5.6 Frequently formulated statistical hypotheses and their validation tests.

Current number for input	Comparison state	Zero hypothesis	Test used	Computed value for the random variable	Associated probability	Condition of rejection
1	Two populations and two selections. Parameters: μ_1, σ_1^2 population 1 μ_2, σ_2^2 population 2; $\bar{x}_1$,s_1^2 selection 1 $\bar{x}_2$,s_1^2 selection 2	$\mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$	u	$u = \frac{\mu_1 - \mu_2}{\sigma_1/\sqrt{n}}$ $u = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_1/\sqrt{n}}$	$P(X \le u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-u^2} du$	$P(X \le u) > 1 - \alpha$
2	Same as 1 but for selections with a small volume	$\mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$	t $\nu = n-1$	$t = \frac{\mu_1 - \mu_2}{\sigma_1}\sqrt{n}$ or $t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_1}\sqrt{n}$	$P(X \le t) = \int_{-\infty}^{t} f_{\upsilon}(t)dt$	$P(X \le t) > 1 - \alpha$
3	The n volume of selection and its population	$s_1^2 = \sigma^2$	χ^2 $\nu = n-1$	$\chi^2 = \frac{n-1}{\sigma^2} s^2$	$P(X \le \chi^2) =$	$P(X \le \chi^2) > 1 - \alpha$
4	Two selections of n_1 and n_2 volumes	$s_1^2 = s_2^2$ or $s_1^2 > s_2^2$	F $\nu_1 = n_1 - 1$ $\nu_1 = n_2 - 1$	$F = \frac{s_1^2}{s_2^2}$	$P(X \le F) = \int_{-\infty}^{F} f_{\upsilon 1, \upsilon 2}(F)dF$	$P(X \le F) > 1 - \alpha$

In order to clarify this conceptual discussion we will use the actual example we have been working on in this chapter. First, it is required to verify whether dispersion s_1^2, which characterizes the selection given in Section 5.5.2.1, is similar to dispersion s_2^2, established in Table 5.5, column 2. Indeed, it is known that these selections have been extracted from the same original population. The response to this question is obtained with the calculation methodology described above. This computation is organized according to the algorithmic rule proposed at the beginning of this paragraph, so:

- We write the actual H_0 hypothesis: $H_0 : s_1^2 = s_2^2$
- We compute the current value of the Fischer random variable associated to the dispersions s_1^2 and s_2^2: $F = s_1^2/s_2^2$;
  Result: F = 3.86/0.473 = 8.16
- We establish the degrees of freedom for the Fischer variable:
  $\upsilon_1 = n_1 - 1, \upsilon_2 = n_2 - 1$;
  Results $\nu_1 = 3$, $\nu_2 = 4$
- We obtain the probability of the current Fischer variable by computing the value of the repartition function:

$$P(X \leq 8.16) = \int_0^{8.16} f_{\upsilon 1,\upsilon 2}(F)dF;$$

Result: $P(X \leq 8.16) = \int_0^{8.16} f_{\upsilon 1,\upsilon 2}(F)dF = 0.97$

- We accept the most used significance level $\alpha = 0.05$
- We observe that $P(X \leq 8.16) = 0.97 \succ 1 - \alpha = 0.95$ and, as a consequence, we reject the zero hypothesis.

5.3 Correlation Analysis

When the preliminary steps of the statistical model have been accomplished, the researchers must focus their attention on the problem of correlation between dependent and independent variables (see Fig. 5.1). At this stage, they must use the description and the statistical selections of the process, so as to propose a model state with a mathematical expression showing the relation between each of the dependent variables and all independent variables (relation (5.3)). During this selection, the researchers might erroneously use two restrictions: Firstly, they may tend to introduce a limitation concerning the degree of the polynomial that describes the relation between the dependent variable $y^{(i)}$ and the independent variables x_j, j = 1,n; Secondly, they may tend to extract some independent variables or terms which show the effect of the interactions between two or more independent variables on the dependent variable from the above mentioned relationship.

The problem of simplifying the regression relationship can be omitted if, before establishing those simplifications, the specific procedure that defines the type of the correlations between the dependent and independent variables of the process, is applied on the basis of a statistical process analysis.

Classical dispersion analyses, dispersion analyses with interaction effects and especially correlation analyses can be used successfully to obtain the information needed about the form of an actual regression expression. Working with the statistical data obtained by the process investigation, the dispersion and the correlation analyses, can establish the independent process variables and the interactions of independent variables that have to be considered in a regression expression [5.18, 5.19].

For a process with one dependent variable and one independent variable, the statistical process analysis gives one chain with values of y_i, i = 1,n and another one with values of x_i, i = 1,n. Here, n is the number of the processed experiments. The correlation analysis shows that the process variables y and x are correlated if the indicator cov(x,y), given here by relation (5.44), presents a significant value:

$$\mathrm{cov}(y,x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \tag{5.44}$$

We observe that the covariance indicator (cov(x,y)) has an expression which is similar to the dispersion of a statistical selection datum near the mean value (Eq. (5.11)). It is important to specify that the notion of variance (or dispersion) differs completely from the notion of covariance.

If the multiplication $(x_i - \bar{x})(y_i - \bar{y})$ from the covariance definition (5.44) gives a positive number, then the figurative point (x_i, y_i) will be placed in the first or third quadrant of an x,y graphic representation, whereas, the figurative point (x_i, y_i) will be placed in the second or fourth quadrant. Now if the x and y variables are independent, then the placement probability of the figurative point is the same for all quadrants. So, in this case, we have the graphic representation from Fig. 5.7(a), and the sum $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ tends to zero or to a very small number. For the case when x and y are dependent, then the placement probability is not the same for all four quadrants and consequently the sum $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \neq 0$. This last situation is shown in Fig. 5.7(b).

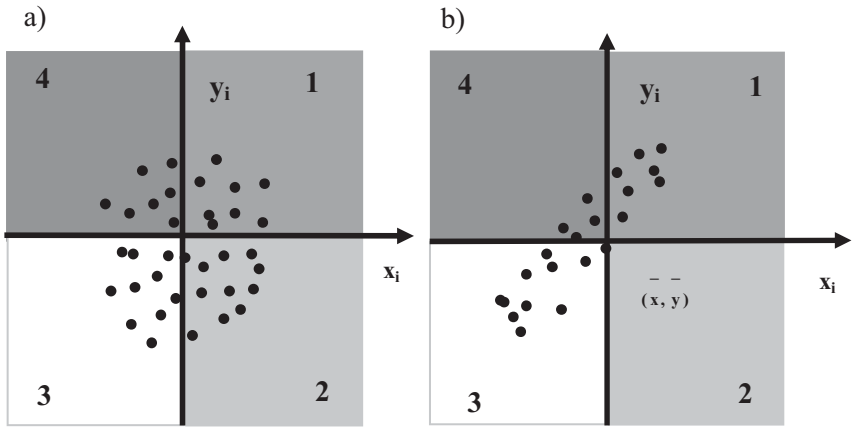


Figure 5.7 Graphic introduction of the correlation between statistical variables, (a) independent variables, (b) dependent variables.

The x and y covariance increases or decreases with the values of $(x_i - \bar{x})$ and $(y_i - \bar{y})$. Thus, if we repeat the statistical experiment in order to obtain the chains

of values x_i, y_i $i = 1, n$ and if we compute again the cov(x,y), this new cov value can be different from the cov initially calculated. This distortion is eliminated if we replace the covariance by the correlation coefficient of the variables:

$$r_{yx} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{\mathrm{cov}(y, x)}{s_x s_y} \tag{5.45}$$

It is easy to observe that the domain of the values of the correlation coefficient is placed between –1 and +1 and that $r_{yx} : R \rightarrow [-1, 1]$.

The following observations can also be made with respect to the correlation coefficient:

- If the value of the correlation coefficient approaches zero, then we can accept x and y variables to be independent. So, the variations on the dependent variable do not affect the independent variable;
- When the correlation coefficient takes a positive value, the independent and dependent variables increase simultaneously. The opposite case corresponds to a negative value of the correlation coefficient;
- The extreme values ($r_{yx} = 1; r_{yx} = -1$) for the correlation coefficient show that a linear relationship exists between the dependent and independent variables.

The discussion presented above for the case when the process has only one input can easily be extended to a process with more than one independent variable (many inputs). For example, when we have one dependent and two independent variables, we can compute the $r_{yx_1}, r_{yx_2}, r_{yx_1x_2}$.. coefficients. All the observations concerning r_{yx} stay unchanged for each $r_{yx_1}, r_{yx_2}, r_{yx_1x_2}$.. When this process involves two inputs, if we obtain $r_{yx_1} = 1, r_{yx_2} = -1, r_{yx_1x_2} = 1$ and if the other possible correlation coefficients approach zero, then dependence $y = \beta_0 + \beta_1 x_1 - \beta_2 x_2 + \beta_{12} x_1 x_2$ is recommended to build the statistical model of the process.

If we once more consider the example studied throughout this chapter, we can use the statistical data presented in Table 5.3 in order to compute the value of the correlation coefficient. However, before carrying out this calculation, we can observe an important dependence between variables x and y due to the physical meaning of the results in this table. The value obtained for the correlation coefficient confirms our a priori assumption because the cov has a value near unity. It shows that a linear relationship can be established between process variables. The results of these calculations are shown in Table 5.7.

Table 5.7 Calculation of the correlation coefficient between the reactant conversion degree and the input concentration (the statistical data used are from Table 5.3)

Current number for input	x_i	$\bar{x}$	$(x_i - \bar{x})$	y_i	$\bar{y}$	$(y_i - \bar{y})$	$(x_i - \bar{x}) * (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	13.86		–13.91	0.770		0.1762	–2.45975	193.655	0.03104
2	20.16		–7.616	0.655		0.0612	–0.46609	58.003	0.00374
3	27.70		–0.076	0.593	0.594	–0.0008	0.00006	0.0057	0.00006
4	34.76	27.78	6.984	0.514		–0.0798	–0.55732	48.776	0.00636
5	42.40		14.62	0.437		–0.1568	–2.29304	213.861	0.02496
$\sum_{i=1}^{5}$	138.88			2.969			5.776144	514.30	0.070672

The value of the correlation coefficient is: $r_{yx} = (5.7761/(514.3*0.0707)^2) = 0.957$.

We can eliminate all the false dependent variables from the statistical model thanks to the correlation analysis. When we obtain $r_{y_1y_2} = 1$ for a process with two dependent variables (y_1, y_2), we have a linear dependence between these variables. Then, in this case, both variables exceed the independence required by the output process variables. Therefore, y_1 or y_2 can be eliminated from the list of the dependent process variables.

5.4 Regression Analysis

Regression analysis is the statistical computing procedure that begins when the model regression equations have been established for an investigated process. The regression analysis includes [5.18, 5.19]:

- the system of normal equations for the particularizations to an actual case, in which the relationship between each dependent variable and the independent process variables is established on the basis of Eq. (5.3);
- the calculations of the values of all the coefficients contained in the mathematical model of the process;
- the validation of the model coefficients and of the final statistical model of the process.

The items described above have already been introduced in Fig. 5.3 where the steps of the development of the statistical model of a process are presented. It should be pointed out that throughout the regression analysis, attention is commonly concentrated on the first and second aspects, despite the fact that virgin

statistical data are available for the third aspect. Normally, this new non-used data (see Fig. 5.3) allows the calculation of the reproducibility variance (s^2_{rp}) as well as the residual variance, which together give the model acceptance or rejection. In fact, this aspect contains the validation of the hypothesis considering that $s^2_{rp} = s^2_{rz}$; it is clear that the use of the Fischer test (for instance, see Table 5.6) is crucial in this situation. The following paragraphs contain the particularization of the regression analysis to some common cases. It is important to note that these examples differ from each other by the number of independent variables and the form of their regression equations.

5.4.1 Linear Regression

A linear regression occurs when a process has only one input (x) and one output variable (y) and both variables are correlated by a linear relationship:

$$y^{th} = y = f(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x \tag{5.46}$$

This relation is a particularization of the general relation (5.3). Indeed, polynomial regression presents the limitation of being first order. In accordance with Eq. (5.46), the system of equations (5.9) results in the following system for the identification of β_0 and β_1:

$$\begin{cases} \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i) = 0 \\ \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i) x_i = 0 \end{cases} \tag{5.47}$$

which is equivalent to:

$$\begin{cases} N\beta_0 + \beta_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} (x_i)^2 = \sum_{i=1}^{N} y_i x_i \end{cases} \tag{5.48}$$

Now it is very simple to obtain coefficients β_0 and β_1 as the Cramer solution of system (5.39). The following expressions for β_0 and β_1 are thus obtained:

$$\beta_0 = \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i x_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2} \tag{5.49}$$

$$\beta_1 = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \tag{5.50}$$

After the calculation of β_1 we can extract β_0 from relation (5.51) where $\overline{x}$ and $\overline{y}$ are the mean values of variables x and y respectively. Otherwise, this relation can also be used to verify whether β_0 and β_1 are correctly obtained by relations (5.50) and (5.51):

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \tag{5.51}$$

The next step in developing a statistical model is the verification of the significance of the coefficients by means of the Student distribution and the reproducibility variance.

The problem of the significance of the regression coefficients can be examined only if the statistical data take into consideration the following conditions [5.19]:

1. The error of the measured input parameter (x) must be minor. In this case, any error occurring when we obtain "y" will be the consequence of the non-explicit input variables. These non-explicit variables are input variables which have been rejected or not observed when the regression expression was proposed.
2. When the measurements are repeated, the results of the output variable must present random values with a normal distribution (such samples are shown in Table 5.2).
3. When we carry out an experimentation in which "N" is the dimension of each experiment and where each experiment is repeated "m" times, the variances $s_1^2, s_2^2, ...s_N^2$, which are associated to the output variable, should be homogeneous.

The testing of the homogeneity of variances concerns the process of primary preparation of the statistical data. It is important to note that this procedure of homogeneity testing of the output variances is in fact a problem which tests the zero hypothesis, i.e.: $H_0 : s_1^2 = s_2^2 = ... = s_N^2$. For this purpose, we comply with the following algorithm:

1. We compute the mean values of samples with respect to the output process variable:

$$\overline{y}_i = \sum_{k=1}^{m} y_{ik}/m \quad i = 1, 2, 3, ...N \tag{5.52}$$

2. With the mean values and with each one of the experiments we establish the variables $s_1^2 = s_2^2 = ... = s_N^2$ as well as their maximum values:

$$s_i^2 = \frac{\sum_{i=1}^{m} (y_{ik} - \overline{y}_i)^2}{m - 1} \tag{5.53}$$

3. We proceed with the calculation of the sum of the variances that give the value of the testing process associated to the Fischer random variable:

$$s^2 = \sum_{i=1}^{N} s_i^2 \qquad F = \frac{s_{max}^2}{s^2} \tag{5.54}$$

4. At this point we identify the values which have the same degrees of freedom as variable F and thus we obtain an existence probability of this random variable between 0 and the computed value of point c):

$$\upsilon_1 = N, \upsilon_2 = m - 1 \; P(X \leq F) = \int_{-\infty}^{F} f_{\upsilon 1, \upsilon 2}(F) dF \tag{5.55}$$

5. For a fixed significance level α, all the variances $s_1^2 = s_2^2 = ... = s_N^2$ will be accepted as homogenous if we have:

$$P(X \leq F) \leq 1 - \alpha \tag{5.56}$$

6. When the homogeneity of the variances has been tested, we continue to compute the values of the reproducibility variance with relation (5.64):

$$s_{rp}^2 = s^2/N \tag{5.57}$$

In statistics, the reproducibility variance is a random variable having a number of degrees of freedom equal to $\upsilon = N(m - 1)$. Without the reproducibility variances or any other equivalent variance, we cannot estimate the significance of the regression coefficients. It is important to remember that, for the calculation of this variance, we need to have new statistical data or, more precisely, statistical data not used in the procedures of the identification of the coefficients. This requirement explains the division of the statistical data of Fig. 5.3 into two parts: one significant part for the identification of the coefficients and one small part for the reproducibility variance calculation.

The significance estimation of β_0 and β_1 coefficients is, for each case, a real statistical hypothesis, the aim of which is to verify whether their values are null or not. Here, we can suggest two zero hypotheses ($H_{01} : \beta_0 = 0$ and $H_{02} : \beta_1 = 0$) and by using the Student test (see Table 5.6), we can find out whether these hypotheses are accepted or rejected.

In a more general case, we have to carry out the following calculations:

- firstly: the values of the t_j variable using relation (5.58) where β_j is the j regression coefficient, and s_{β_j} represents the corresponding β_j mean square root of variance $s^2_{\beta_j}$:

$$t_j = \frac{|\beta_j|}{s_{\beta_j}} \tag{5.58}$$

- secondly: the existence probability of the t_j value of the Student variable, where υ is the number of the degrees of freedom respect to the calculation of the t_j value:

$$P_j(X \leq t_j) = \int_{-\infty}^{t_j} f_\upsilon(t)dt$$

- finally: if we have $P_j(X \leq t_j) > 1 - \alpha$, the zero hypothesis for β_j so that $H_{0j} : \beta_j = 0$ will be rejected. In this case, β_j is an important coefficient in the relationship between the regression variables. The opposite case corresponds to the acceptance of the H_{0j} hypothesis.

It is then important to show that, in case of generalization, the mean square root of the variances with respect to the mean β_j value as well as its variances have the quality to respect the law of the accumulation of errors [5.13, 5.16, 5.19]. As a result, the mean square root of the variances will have a theoretical expression, which is given by:

$$s_{\beta_j} = \sqrt{\sum_{i=1}^{N} \left(\frac{\partial \beta_j}{\partial y_i}\right) s_i^2} \tag{5.59}$$

Because, in a normal case, we have the homogenous variances $s_1^2 = s_2^2 = ... = s_N^2 = s_{rp}^2$, then for the case of a linear regression, we can particularize relation (5.59) in order to obtain the following relations:

$$s_{\beta_0} = \sqrt{\frac{s_{rp}^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}} \tag{5.60}$$

$$s_{\beta_1} = \sqrt{\frac{s_{rp}^2 N}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}} \tag{5.61}$$

After estimation of the significance of the coefficients, each non-significant coefficient will be excluded from the regression expression and a new identification can

be made for all the remaining coefficients. This new calculation of the remaining regression coefficients is a consequence of the fact that these regression coefficients are in an active interrelated state. Before ending this problem, we must verify the model confidence, i.e. we must check whether the structure that remains after the testing of the significance coefficients, is adequate or not. For the example discussed above, the model is represented by the final expression of regression. Its confidence can thus be verified using the Fischer test the orientation of which is to verify the statistical hypothesis: $H_{0m} : s^2_{rz} = s^2_{rp}$ suggesting the equality of the residual and reproducibility variances. The Fischer test begins with the calculation of the Fischer random variable value: $F = s^2_{rz}/s^2_{rp}$. Here the degrees of freedom have the values $\upsilon_1 = N - 1, \upsilon_2 = N - n_\beta$, where N is the number of statistical data used to calculate s^2_{rz} in υ_1, as well as in υ_2. Here, n_β introduces the number of regression coefficients that remain in the final form of the regression expression. For a process with only one output (only one dependent variable) the residual variance measures the difference between the model computed and the mean value of the output:

$$s^2_{rz} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2}{N - n_\beta} \tag{5.62}$$

$$\hat{y}_i = \beta_0 + \beta_1 x_i \quad i = 1, N \tag{5.63}$$

After calculating the value of the random variable F, we establish the reproducibility variances and carry out the test according to the procedure given in Table 5.6. Exceptionally, in cases when we do not have any experiment carried out in parallel, and when the statistical data have not been divided into two parts, we use the relative variance for the mean value (s^2_y) instead of the reproducibility variance. This relative variance can be computed with the statistical data used for the identification of the coefficients using the relation (5.64):

$$s^2_y = \frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N - 1} \tag{5.64}$$

In this case, the value of N for υ_1 and υ_2 is the same and it is equal to the number of experiments accepted for the statistical calculations. Coming back to the problem of the model adequacy, it is clear that the zero hypothesis has been transformed into the following expression: $H_{0m} : s^2_{rz} = s^2_y$.

5.4.1.1 Application to the Relationship between the Reactant Conversion and the Input Concentration for a CSR

The statistical data shown in Table 5.2 were obtained for an isothermal continuously stirred reactor (CSR) with a spatial time of 1.5 h. With these experimental data, we can formulate a relationship between the reactant conversion (y) and the input concentration (x). For the establishment of a statistical model based on a

linear regression, we have a coefficient of regression close to 1 (found in Table 5.7 which contains the values obtained with the same statistical data). However, we did not have any additional experiments carried out in parallel and consequently we cannot establish a real reproducibility variance. The correlation coefficient from Table 5.7, sustains the proposal of a linear dependence between the conversion (y) and the input concentration of the reactant (x): $y = \beta_0 + \beta_1 x$. Table 5.8 shows the statistical data and the results of some calculations needed for the determination of β_0 and β_1.

Table 5.8 The statistical data and calculated parameters for the estimation of β_0 and β_1.

i =	x_i	y_i	$(x_i)^2$	$(y_i x_i)$	$\bar{x}$	$\bar{y}$
1	13.86	0.77	194.8816	10.6722		
2	20.16	0.655	406.4256	13.2048		
3	27.70	0.593	767.29	16.4261	27.776	0.5938
4	34.76	0.514	1208.2576	17.86664		
5	42.40	0.437	1797.76	18.5288		
$\sum_{i=1}^{N}$	138.88	2.969	4374.6	76.67		

Thus, for β_0 and β_1 we obtain:

$$\beta_1 = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{5 * 76.67 - 138.88 * 2.969}{5 * 43754.6 - (138.88)^2} = -0.0112 \ ;$$

$$\beta_0 = 0.0112 * 27.76 + 0.5938 = 0.92692$$

The significance estimation of β_0 and β_1 is made by computing the residual variance and the variance relative to the mean value of the dependent variable. The results corresponding to these calculations are shown in Table 5.9.

Table 5.9 Computed values of the residual and relative variances.

Number (N)	1	2	3	4	5	$\sum_{i=1}^{N}$
x_i	13.86	20.16	27.70	34.76	42.40	138.88
y_i	0.77	0.655	0.593	0.514	0.437	2.969
$y_i - \overline{y}$	0.1762	0.0612	–0.0008	–0.0798	–0.1568	
$\hat{y} = \beta_0 - \beta_1 x_i$	0.884	0.701	0.617	0.538	0.452	
$\hat{y} - \overline{y}$	0.2902	0.1072	0.0232	–0.0558	–0.1418	
Variance	$s_{rz}^2 = (0.2902^2 + 0.1072^2 + 0.0232^2 + 0.0558^2 + 0.1418^2)/4 = 0.02986675$			$s_{\hat{y}}^2 = (0.1762^2 + 0.0612^2 + 0.0008^2 + 0.0798^2 + 0.1568^2)/4 = 0.0164367$		

Now, we can obtain the variances due to β_0 and β_1 by using relations (5.60) and (5.61):

$$s_{\beta_0} = \sqrt{\frac{s_{rp}^2 \sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}} = \sqrt{\frac{0.01644 * 4374.6}{5 * 4374.6 - (138.88)^2}} = 0.1667$$

$$s_{\beta_1} = \sqrt{\frac{s_{rp}^2 N}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}} = \sqrt{\frac{0.01644 * 5}{5 * 4374.6 - (138.88)^2}} = 0.0018$$

The estimations of the β_0 and β_1 significance are computed by the procedure given in Table 5.6. The results are shown in Table 5.10.

Table 5.10 The significance of β_0 and β_1 coefficients estimated by the Student test.

Hypothesis	ν	T	P(X<t), relation (5.30)	1–α	Conclusion
$\beta_0 = 0$	4	0.92692/0.1667 = 5.5	0.87	0.95	β_0 important
$\beta_1 = 0$	4	0.0112/0.0018 = 6.2	0.91	0.95	β_1 important

At this point, we have to think about the problem of the model confidence. For this purpose we have to consider that:

- the value of the Fischer variable is F = 0.0298/0.0164 = 1.817;

- for $1-\alpha = 0.95$, we obtain $F = F_{0.05} = 3.24$ by solving the equation $1 - \alpha = \int_0^F f_{4,4}(F)dF$;
- we accept the zero hypothesis $H_{0m} : s_{rz}^2 = s_y^2$ because $F_{0.05} = 3.24 \succ F = 1.817$.

In other words, the reactant transformation degree (η), depends on the input reactant concentration (c_0), according to the following relation: $\eta = 0.92692 - 0.0112c_0$. The results obtained here show that physical and chemical processes occurring in the reactor of this case under study are not simple. It is well known that for a reaction occurring in a CSR with a simple kinetics, the degree of transformation is not significantly dependent on the input reactant concentration. For example, if a first order reaction occurs in a CSR, η will depend only on the residence time and the kinetic reaction constant $\eta = k_r\tau_s/(k_r\tau_s + 1)$.

5.4.2 Parabolic Regression

If the regression expression is a polynomial, then, by applying the method of least squares to identify the coefficients and compute the values of the coefficients, we obtain a simple linear system. If we particularize the case for a regression expression given by a polynomial of second order, the general relation (5.3) is reduced to:

$$y^{th} = y = f(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x + \beta_{11} x^2 \tag{5.65}$$

By computing the derivatives of the system of normal equations $\frac{\partial f(x, \beta_0, \beta_1, \beta_{11})}{\partial \beta_0} = 1, \frac{\partial f(x, \beta_0, \beta_1, \beta_{11})}{\partial \beta_1} = x, \frac{\partial f(x, \beta_0, \beta_1, \beta_{11})}{\partial \beta_{11}} = x^2$, we establish the system of equations which is necessary to calculate the values of $\beta_0, \beta_1, \beta_{11}$:

$$\begin{cases} \beta_0 N + \beta_1 \sum_{i=1}^{N} x_i + \beta_{11} \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i \\ \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 + \beta_{11} \sum_{i=1}^{N} x_i^3 = \sum_{i=1}^{N} y_i x_i \\ \beta_0 \sum_{i=1}^{N} x_i^2 + \beta_1 \sum_{i=1}^{N} x_i^3 + \beta_{11} \sum_{i=1}^{N} x_i^4 = \sum_{i=1}^{N} y_i x_i^2 \end{cases} \tag{5.66}$$

The same procedure is used if we increase the polynomial degree given by the regression equation. In this case, the tests of the coefficient significance and model confidence are implemented as shown in the example developed in Section 5.4.1.1. It is important to note that we must use relation (5.59) for the calculation of the variances around the mean value of β_j.

5.4.3
Transcendental Regression

For statistical samples of small volume, an increase in the order of the polynomial regression of variables can produce a serious increase in the residual variance. We can reduce the number of the coefficients from the model but then we must introduce a transcendental regression relationship for the variables of the process. From the general theory of statistical process modelling (relations (5.1)–(5.9)) we can claim that the use of these types of relationships between dependent and independent process variables is possible. However, when using these relationships between the variables of the process, it is important to obtain an excellent ensemble of statistical data (i.e. with small residual and relative variances).

It is well known that using an exponential or power function can also describe the portion of a polynomial curve. Indeed, these types of functions, which can represent the relationships between the process variables, accept to be developed into a Taylor expansion. This procedure can also be applied to the example of the statistical process modelling given by the general relation (5.3) [5.20].

In this case, the calculation of the coefficients for the transcendental regression expression can be complicated because, instead of a system of normal equations (5.9), we obtain a system of non-linear equations. However, we can simplify the calculation by changing the original variables of the regression relationship. In fact, changing the original variables results in the mathematical application of one operator to the expression of the transcendental regression. As an example, we can consider the relations (5.67)–(5.69) below, where the powers or an exponential transcendental regression are transformed into a linear regression:

$$y^{th} = y = f(x, \beta_0, \beta_1) = \beta_0 \beta_1^x \tag{5.67}$$

$$y^{th} = y = f(x, \beta_0, \beta_1) = \beta_0 x^{\beta_1} \tag{5.68}$$

$$\lg y = \lg \beta_0 + x \lg \beta_1 \ , \quad z = \lg y \ , \ \beta_0' = \lg \beta_0 \ , \ \beta_1' = \lg \beta_1 \ , \quad z = \beta_0' + \beta_1^x \tag{5.69}$$

Coefficients β_0', β_1' can easily be obtained by using the method of least squares. Nevertheless, the interest is to have the original coefficients of the transcendental regression. To do so, we apply an inverse operator transformation to β_0' and β_1'. Here, we can note that β_0' and β_1' are the bypassed estimations for their correspondents β_0 and β_1.

5.4.4
Multiple Linear Regression

When the studied case concerns obtaining a relationship for the characterization of a process with multiple independent variables and only one dependent variable, we can use a multiple linear regression:

$$y^{th} = y = f(x_1..x_k, \beta_0, \beta_1...\beta_k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (5.70)$$

It is clear that Eq. (5.70) results from the general relation (5.3). In this case, when k = 2, we have a regression surface whereas, when k>2, a hypersurface is obtained. For surface or hypersurface constructions, we have to represent the corresponding values of the process parameters (factors and one dependent variable) for each axis of the phase's space. The theoretical starting statistical material for a multiple regression problem is given in Table 5.11.

Table 5.11 The starting statistical material for a multiple regression.

i	x_1	x_2	x_3	.....	x_k	y
1	x_{11}	x_{21}	x_{31}	.....	x_{k1}	y_1
2	x_{12}	x_{22}	x_{32}	.....	x_{k2}	y_2
3	x_{13}	x_{23}	x_{33}	.....	x_{k3}	y_3
.	...	...	...	.....	...	..
.	...	...	...	.....	...	..
N	x_{1N}	x_{2N}	x_{3N}	.....	x_{kN}	y_N

The starting data are frequently transformed into a dimensionless form by a *normalization* method in order to produce a rapid identification of the coefficients in the statistical model. The dimensionless values of the initial statistical data ($y_i^0$and x_{ji}^0) are computed using Eqs. (5.71) and (5.72), where s_y, s_{xj} are the square roots of the correspondent variances:

$$y_i^0 = \frac{y_i - \overline{y}}{s_y} \quad , \quad x_{ji}^0 = \frac{x_{ji} - \overline{x}_j}{s_{xj}} \quad , \; i = 1, N \; ; \; j = 1, k \quad (5.71)$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N-1}} \, , \, s_{xj} = \sqrt{\frac{\sum_{i=1}^{N} (x_{ji} - \overline{x}_j)^2}{N-1}} \quad (5.72)$$

At this step of the data preparation, we can observe that each column of the transformed statistical data has a zero mean value and a dispersion equal to one. A proof of these properties has already been given in Section 5.2 concerning a case of normal random variable normalization.

Then, considering the statistical data from Tables 5.11 and 5.12 and using the statistical correlation aspects (see Section 5.3), we can observe that the correlation coefficients are the same for variables y, x_j and y_i^0, x_{ji}^0 (relation (5.73)). This observation remains valid for the correlations concerning x_j and x_l.

$$r_{yx_j} = r_{y^0x_j^0} = \frac{1}{N-1}\sum_{i=1}^{N} y_i^0 x_{ji}^0 \tag{5.73}$$

$$r_{x_jx_l} = r_{x_j^0x_l^0} = \frac{1}{N-1}\sum_{i=1}^{N} x_{ji}^0 x_{li}^0 \ , \quad j \neq l \quad j,l = 1,2,...k \tag{5.74}$$

Table 5.12 The dimensionless statistical data for a multiple regression.

i	x_1^0	x_2^0	x_3^0	.....	x_k^0	Y^0
1	x^0_{11}	x^0_{21}	x^0_{31}	.....	x^0_{k1}	$y_1{}^0$
2	x^0_{12}	x^0_{22}	x^0_{32}	.....	x^0_{k2}	$y_2{}^0$
3	x^0_{13}	x^0_{23}	x^0_{33}	.....	x^0_{k3}	$y_3{}^0$
.	...	...	...	.....	...	..
.	...	...	...	.....	...	..
N	x^0_{1N}	x^0_{2N}	x^0_{3N}	.....	x^0_{kN}	$y_N{}^0$

The observations mentioned above are important because they will be used in the following calculations. As was explained above, the mean value of the dependent normalized variable is zero, consequently the regression expression with the normalized variables can be written as:

$$y^{0\,th} = f^0(x_1^0, ..x_k^0, \alpha_1, ..\alpha_k) = \alpha_1 x_1^0 + \alpha_2 x_2^0 + ... + \alpha_k x_k^0 \tag{5.75}$$

It is evident that, for the identification of the α_j coefficients, we have to determine the minimum of the quadratic displacement function between the measured and computed values of the dependent variable:

$$\Phi(\alpha_1, \alpha_2, ...\alpha_k) = \sum_{i=1}^{N} (y_i^0 - y_i^{0\,th})^2 = \min \tag{5.76}$$

thus, we obtain the minimum value of function $\Phi(\alpha_1, \alpha_2...., \alpha_k)$ when we have:

$$\frac{\partial\Phi(\alpha_1..\alpha_k)}{\partial\alpha_1} = \frac{\partial\Phi(\alpha_1..\alpha_k)}{\partial\alpha_2} = ... = \frac{\partial\Phi(\alpha_1..\alpha_k)}{\partial\alpha_k} = 0 \tag{5.77}$$

the relation above can be developed as follows:

$$\begin{cases} \alpha_1 \sum_{i=1}^{N} (x_{1i}^0)^2 + \alpha_2 \sum_{i=1}^{N} (x_{2i}^0 x_{1i}^0) + \ldots\ldots\ldots\ldots + \sum_{i=1}^{N} (x_{Ni}^0 x_{1i}^0) = \sum_{i=1}^{N} (y_1^0 x_{1i}^0) \\ \alpha_1 \sum_{i=1}^{N} (x_{1i}^0 x_{2i}^0) + \alpha_2 \sum_{i=1}^{N} (x_{2i}^0)^2 + \ldots\ldots\ldots\ldots + \sum_{i=1}^{N} (x_{Ni}^0 x_{2i}^0) = \sum_{i=1}^{N} (y_1^0 x_{2i}^0) \\ \text{---} \\ \alpha_1 \sum_{i=1}^{N} (x_{1i}^0 x_{Ni}^0) + \alpha_2 \sum_{i=1}^{N} (x_{2i}^0 x_{Ni}^0) + \ldots\ldots\ldots\ldots + \sum_{i=1}^{N} (x_{Ni}^0)^2 = \sum_{i=1}^{N} (y_1^0 x_{Ni}^0) \end{cases} \tag{5.78}$$

The system (5.80) for the identification of the α_j coefficients is obtained after multiplying each term of system (5.78) by 1/(N–1) and after coupling this system with relations (5.73), (5.74) and (5.79):

$$\frac{1}{N-1} \sum_{i=1}^{N} (x_{ji}^0)^2 = s_{x_j^0}^2 = 1 \tag{5.79}$$

$$\begin{cases} \alpha_1 + \alpha_2 r_{x_1x_2} + \alpha_3 r_{x_1x_3} + \ldots\ldots\ldots + \alpha_k r_{x_1x_k} = r_{yx_1} \\ \alpha_1 r_{x_2x_1} + \alpha_2 + \alpha_3 r_{x_2x_3} + \ldots\ldots\ldots + \alpha_k r_{x_2x_k} = r_{yx_2} \\ \text{---------------------------------------} \\ \alpha_1 r_{x_kx_1} + \alpha_2 r_{x_kx_2} + \alpha_3 r_{x_kx_3} + \ldots\ldots\ldots + \alpha_k = r_{yx_k} \end{cases} \tag{5.80}$$

Considering the commutability property of the correlations of coefficients ($r_{x_jx_i} = r_{x_ix_j}$) we can solve the above system. After solving it with unknown $\alpha_1, \alpha_2 \ldots \alpha_k$, we can determine the value of the correlation between the coefficients of the process variables by using Eq. (5.81):

$$R_{yx_j} = \sqrt{\alpha_1 r_{yx_1} + \alpha_2 r_{yx_2} + \ldots\ldots + \alpha_k r_{yx_k}} \tag{5.81}$$

When the statistical sample is small, the multiple linear correlation coefficient must be corrected. The correction is imposed by the fact that, in this case, the small number of degrees of freedom ($\upsilon = N - n_\beta$ is small) adds errors systematically. Therefore, the most frequently used correction is given by:

$$R_{yx_j}^c = \sqrt{1 - (1 - R_{yx_j}^2) \frac{N-1}{N-n_\beta}} \tag{5.82}$$

At this point, we have to consider coefficients $\alpha_1, \alpha_2 \ldots \alpha_k$ according to the dimensional relationship between the process variables (5.70). For this purpose, we must transform α_j into β_j, and j = 1,k. Indeed, these changes can take place using the following relations: $\beta_j = \alpha_j s_y / s_{xj}$, $j = 1, 2, \ldots k$, $j \neq 0$, $\beta_0 = \overline{y} - \sum_{i+}^{N} \beta_j \overline{x}_j$.

Now we have to estimate the reproducibility of the variance, to carry out the confidence tests for the coefficients so as to establish the final model.

5.4.4.1 Multiple Linear Regressions in Matrix Forms

The regression analysis, when the relationship between the process variables is given by a matrix, is frequently used to solve the problems of identification and confidence of the coefficients as well as the problem of a model confidence. The matrix expression is used frequently in processes with more than two independent variables which present simultaneous interactive effects with a dependent variable. In this case, the formulation of the problem is similar to the formulation described in the previous section. Thus, we will use the statistical data from Table 5.11 again in order to identify the coefficients with the following relation:

$$y^{th} = y = f(x_1..x_k, \beta_0, \beta_1...\beta_k) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{5.83}$$

The first step in this discussion concerns the presentation of the matrix of the independent variables (X), the experimental observation vector of the dependent variable (Y) and the column matrix of the coefficients (B) as well as the transposed matrix of the independent variables (X^T). All these terms are introduced by relation (5.84). A fictive variable x_0, which takes the permanent value of 1, has been considered in the matrix of the independent variables:

$$X = \begin{bmatrix} x_{01} & x_{11} & . & . & x_{k1} \\ x_{02} & x_{12} & . & . & x_{k2} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{0N} & x_{1N} & . & . & x_{NN} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{bmatrix} \quad B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix} \tag{5.84}$$

$$X^T = \begin{bmatrix} x_{01} & x_{02} & . & . & x_{0N} \\ x_{11} & x_{12}^T & . & . & x_{1N} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{k1} & x_{k2} & . & . & x_{kN} \end{bmatrix}$$

The particularization of the system of the normal equations (5.9) into an equivalent form of the relationship between the process variables (5.83), results in the system of equations (5.85). In matrix forms, the system can be represented by relation (5.86), and the matrix of the coefficients is given by relation (5.87). According to the inversion formula for a matrix, we obtain the elements for the inverse matrix of the matrix multiplication (XX^T), where (X^T) is the transpose matrix of the matrix of independent variables.

Relation (5.89) gives the value of each element of matrix (XX^T), where the symbol d_{jk} represents a current element, as shown in relation (5.88),

$$\begin{cases} \beta_0 \sum_{i=1}^{N} (x_{0i})^2 + \beta_1 \sum_{i=1}^{N} (x_{0i}x_{1i}) + \ldots\ldots\ldots\ldots + \beta_k \sum_{i=1}^{N} (x_{0i}x_{ki}) = \sum_{i=1}^{N} (y_i x_{0i}) \\ \beta_0 \sum_{i=1}^{N} (x_{1i}x_{0i}) + \alpha_2 \sum_{i=1}^{N} (x_{1i})^2 + \ldots\ldots\ldots\ldots + \sum_{i=1}^{N} (x_{1i}x_{ki}) = \sum_{i=1}^{N} (y_i x_{1i}) \\ \text{---} \\ \beta_0 \sum_{i=1}^{N} (x_{ki}x_{0i}) + \alpha_2 \sum_{i=1}^{N} (x_{1i}x_{ki}^0) + \ldots\ldots\ldots\ldots + \sum_{i=1}^{N} (x_{ki})^2 = \sum_{i=1}^{N} (y_1^0 x_{Ni}^0) \end{cases} \tag{5.85}$$

$$X^T X B = XY \tag{5.86}$$

$$B = (X^T X)^{-1} XY \tag{5.87}$$

$$(X^T X)^{-1} = \begin{bmatrix} d_{00} & d_{01} & d_{02} & . & d_{0k} \\ d_{10} & d_{11} & d_{12} & . & d_{1k} \\ d_{20} & d_{21} & d_{22} & . & d_{2k} \\ . & . & . & . & . \\ d_{k0} & d_{k1} & d_{k3} & . & d_{kk} \end{bmatrix} \tag{5.88}$$

$$d_{jk} = \frac{\left(\sum_{i=1}^{N} x_{ki}x_{ji} \right)'}{\Delta} \tag{5.89}$$

It is not easy to compute the cofactors $\left(\sum_{i=1}^{N} x_{ki}x_{ji} \right)'$ and the determinant of the $(X^T X)$ matrix multiplication. Therefore, the computation depends on the (X) matrix dimension and more specifically on the number of the independent variables of the process as well as on the number of experiments produced during the process of the statistical investigation. Frequently, the computation software of the problem cannot produce a solution for the coefficient matrix, even if we have carefully prepared data (controlled and verified). To overcome this situation, we must verify whether the inverse matrix of the (X) and (X^T) matrix multiplication presents a degenerated state. This undesirable situation appears when one or more correlation(s) exist(s) between the independent variables of the process. For this reason, when we have two factors with a strong correlation in the ensemble of independent variables, one of them will be excluded before developing the calculation algorithm to determine the correlation coefficients.

In order to obtain the value of the residual variance, we first define the matrix of the expected observations $\hat{Y} = XB$ and then we observe that the quadratic displacement between the measured and computed output values of the variables can be written as:

$$\left[Y - \hat{Y}\right]\left[Y - \hat{Y}\right]^T = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Now we introduce the matrix of the theoretical coefficients of the regression (coefficients of the relation (5.4)) here symbolized by Br. Therefore, the coefficient matrix B, defined above, is an estimation of the Br matrix and we can consequently write that the mean value of matrix B is matrix Br: $M(B) \to Br$ or $M[B - Br] \to 0$.

If we apply the concept of mean value to the matrix obtained from the multiplication of [B–Br] and $[B–Br]^T$, and using the definition for the variance and covariance of two variables, we obtain the result given by matrix (5.90):

$$M[(B - Br)(B - Br)^T] = \begin{bmatrix} \sigma^2_{\beta_0} & cov(\beta_0\beta_1) & cov(\beta_0\beta_2) & . & cov(\beta_0\beta_k) \\ cov(\beta_1\beta_0) & \sigma^2_{\beta_1} & cov(\beta_1\beta_2) & . & cov(\beta_1\beta_k) \\ cov(\beta_2\beta_0) & cov(\beta_2\beta_1) & \sigma^2_{\beta_2} & . & cov(\beta_2\beta_k) \\ . & . & . & . & . \\ cov(\beta_k\beta_0) & cov(\beta_k\beta_1) & cov(\beta_k\beta_2) & . & \sigma^2_{\beta_k} \end{bmatrix} \tag{5.90}$$

It should be mentioned that the diagonal components of this matrix contain the theoretical variances of coefficients $\beta_j, j = 1, ..N$. Moreover, these variances are necessary to test the significance of the coefficients of the model. Indeed, when matching a model with an experimental study, matrix (5.90) is fundamental for testing the significance of the coefficients. Now, we have to consider the differences between the measured $y_i, i = 1, ..N$ and the expected mean values of the measurements introduced through the new vector column (Y_{ob}):

$$Y_{ob} = Y - M(Y) = \begin{bmatrix} y_1 - m(y_1) \\ y_2 - m(y_2) \\ . \\ y_N - m(y_{N)} \end{bmatrix} \tag{5.91}$$

Thus, the replacement of $B = (X^TX)^{-1}XY$(5.94) in the left-hand side of relation (5.90) results in: $M[(B - Br)(B - Br)^T] = M[[(XX^T)^{-1}X^TY_{ob}][(X^TX)^{-1}X^TY_{ob}]^T]$. Here, we can observe that (X^TX) is a diagonal symmetric matrix and, for that reason, we can write that $[(X^TX)^{-1}]^T = [(X^TX)^T]^{-1}$. Therefore, the relation $M[(B - Br)(B - Br)^T]$ can be written as $M[(B - Br)(B - Br)^T] = (X^TX)^{-1}M(Y_{ob}Y_{ob}^T)$. Because we generally have $\sigma^2_{y1} = \sigma^2_{y2} = ... = \sigma^2_{yN} = \sigma^2_y$ and due to the statistical independence of errors, we have $cov[(y_i - m(y_i))(y_l - m(y_l))]$ as zero for all $i \neq l$ and thus we can write the matrix $M(Y_{ob}Y_{ob}^T)$ as follows:

$$M(Y_{ob}Y_{ob}^T) = \begin{bmatrix} \sigma^2_{y1} & 0 & 0 & . & 0 \\ 0 & \sigma^2_{y2} & 0 & . & 0 \\ 0 & 0 & \sigma^2_{y3} & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & . & \sigma^2_{yN} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & . & 0 \\ 0 & 1 & 0 & . & 0 \\ 0 & 0 & 1 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & . & 1 \end{bmatrix} \sigma^2_y \tag{5.92}$$

With this last observation, the calculation for $M[(B - Br)(B - Br)^T]$ results in:

$$M[(B - Br)(B - Br)^T] = (X^T X)^{-1}\sigma_y^2 \tag{5.93}$$

This result is very important because it shows how we compute the values of the elements of the matrix of mean errors $M[(B - Br)(B - Br)^T]$. These elements allow the calculation of the dispersions (variances) that characterize each β_j model coefficient of the process as shown in relation (5.94), which results from combining relations (5.93), (5.90) and (5.89):

$$\sigma_{\beta_j}^2 = d_{jj}\sigma_y^2 \;\; ; \;\; \text{cov}(\beta_j\beta_k) = d_{jk}\sigma_y^2 \tag{5.94}$$

From a practical point of view, we should draw the readers' attention to the following significant and important specifics:

1. The $(X^T X)^{-1}$ matrix is the most important to identify the coefficients of the model and to estimate the mean values of errors associated to each β_j coefficient. This matrix is currently called the correlation matrix or error matrix.
2. This matrix does not have the state of a diagonal matrix and consequently, all the regression coefficients are in mutual correlation. So we cannot develop a different significance test for each of the coefficients. From this point of view it is not possible to use the t_j values given by relation (5.95) as the base of a procedure for the process factor arrangement:

$$t_j = \frac{|\beta_j|}{\sigma_y\sqrt{d_{jj}}} = \frac{|\beta_j|}{s_y\sqrt{d_{jj}}} \tag{5.95}$$

3. We can use the t_j values to start a heuristic procedure, which can be obtained from the regression expression of the non-significant coefficients. For this purpose, the following algorithm is used:
   a) the factor with the smallest t_j value is eliminated.
   b) if the residual variance decreases, then the exclusion is correct and thus, a new identification for the coefficients can be carried out. The opposite case shows that the excluded factor is important.
   c) new values for t_j will be obtained and a new elimination procedure can start.
   d) we close the procedure when is not possible to decrease the residual variance.
   e) the final remaining coefficients are the based estimations of the true coefficients.

Until now, no other procedures have been available for the enhancement of an initial proposed relationship between the regression variables.

5.4.5
Multiple Regression with Monomial Functions

In a multiple regression with monomial functions, the particularization of the relationship between the general process variables (5.3) gives the relation written below, where $f_j(x_j)$ is a continuous function:

$$y^{th} = y = f(x_1, x_2, .., \beta_0, \beta_1, ...\gamma) = \gamma f_1(x_1) f_2(x_2) f_3(x_3) .... f_k(x_k)$$

This type of relationship between the dependent and all independent variables was first reported by Brandon [5.21]. In this form of function, we observe that the index (i) does not have a random position; thus, for $i = 1$, the function of the factor has a strong influence on the process, whereas, for $i = k$, the function of the factor has a slight influence on the process.

The algorithm that allows the identification of the functions and the γ constant can be described as follows:

1. An empirical regression line will be processed for the $y–x_1$ dependence with the statistical data from Table 5.11.
2. Thus, the dependence of $y_{x_1} = f_1(x_1)$ can now be appreciated and, using the classical least squares method, we can identify all the unknown coefficients.
3. A new set of values for the dependent variables of the process will be produced by dividing the old values by the corresponding $f_1(x_1)$ values, so that $y_1 = y/f_1(x_1)$. This new set of values of dependent variables are independent of factor x_1 and, as a consequence, we can write:
   $y_1^{th} = \gamma f_2(x_2) f_3(x_3) .... f_k(x_k)$.
4. The first point of the algorithm can be repeated with respect to the $y_1–x_2$ interdependence. Consequently, we can write:
   $y_{x_2} = f_2(x_2)$;
5. We compute the coefficients of function $f_2(x_2)$ by the procedures recommended in item 2. and we build a new set of values for the dependent variables
   $y_2 = y_1/f_2(x_2) = y/[f_1(x_1) f_2(x_2)]$. These new values are independent with respect to x_1 and x_2;
6. The procedure continues with the identification of $f_3(x_3), ... f_k(x_k)$ and we finally obtain the set of the last dependent variables as

   $$y_k = \frac{y_{k-1}}{f_k(x_k)} = \frac{y}{f_1(x_1) f_2(x_2) ... f_k(x_k)}.$$

   It is easy to observe that vector y_k, gives its value to constant γ:
   $\hat{y}_k = \gamma = \frac{1}{N} \sum_{i=1}^{N} y_{ki}$ because it is absolutely independent.

5.5
Experimental Design Methods

For all researchers, and especially for those working in experimental domains, a frequent requirement is summarized by the following phrase: *a maximum of information with a minimum of experiments*. This expression considers not only saving the researcher's time but also expensive reactants and energy. The use of experimental design or planning methods can guarantee not only to greatly reduce the number of experiments needed in an actual research but also to maintain the maximum information about the process. At the same time this technique gives the mathematical procedures of data processing for the complete characterization of the statistical model of a process [5.1, 5.13, 5.21–5.24].

The methodology of experimental design uses a terminology which is apparently different from the vocabulary frequently used in this chapter. Therefore, we call experimental conditions *factors* (or *factor* when we have only one); in fact, in Fig. 5.1, the experimental conditions are entirely included in the class of independent variables of the process. The word *level* (or *levels* when we have more than one), introduces here the values taken by the factors (factor). The term *response* is used to quantitatively characterize the observed output of the process when the levels of the factors are changed.

If we consider a process with k factors and if we suggest N_1 changes for the first factor, N_2 changes for the second factor, etc, then the total number of experiments will be $N_{ex} = N_1N_2....N_k$. In fact, then, $N_1,N_2,.....N_k$ represent each factor level. The most frequent situation is to have $N_1 = N_2 = N_3 = ... = N_k = 2$ and in this case, we obtain the famous 2^k *method* for experimental planning. In fact, the method represents an optimal plan to describe the experiments using two levels for each process factor.

5.5.1
Experimental Design with Two Levels (2^k Plan)

The experimental research of a process with k factors and one response can be carried out considering all the combinations of the k factors with each factor at both levels. Thus, before starting the experimental research, we have a plan of the experiments which, for the mentioned conditions, is recognized as a *complete factorial experiment* (CFE) or 2^k *plan*. The levels of each of the various factors establish the frontiers of the process-investigated domain.

This abstract definition will be explained with the actual example of gaseous permeation through a zeolite/alumina composite membrane. Here, we must investigate the effect of the five following factors on the rate of permeation: the temperature (T) when the domain is between 200 and 400 °C, the trans-membrane pressure (Δp) when the domain is between 40 and 80 bar, the membrane porosity (ε) ranging from 0.08 to 0.18 m^3/m^3, the zeolite concentration within the porous structure (c_z) from 0.01 to 0.08 kg/kg and the molecular weight of the permeated gas (M) which is between 16 and 48 kg/kmol. With respect to the first

factor (T), we can easily identify the value of the maximum level $z_1^{max} = 400\ ^\circ C$, the value of the minimum level $z_1^{min} = 200\ ^\circ C$, the value of the intermediate level $z_1^0 = 300\ ^\circ C$ and the factor (temperature) displacement which is considered as $\Delta z_1 = 100\ ^\circ C$. We can observe that

$$z_1^0 = \frac{z_1^{min} + z_1^{max}}{2} \quad \text{and} \quad \Delta z_1 = \frac{z_1^{max} - z_1^{min}}{2}$$

If we switch this observation to a general case we can write:

$$z_j^0 = \frac{z_j^{max} + z_j^{min}}{2} \quad , \quad \Delta z_j = \frac{z_j^{max} - z_j^{min}}{2} \tag{5.96}$$

Here $z_j,\ j = 1, k$ introduce the original values of the factors. The point with coordinates $(z_1^0, z_2^0, ...z_k^0)$ is recognized as the *centre of the experimental plan* or *fundamental level*. Δz_j introduces the *unity* or *variation interval* respect to the axis $z_j,\ j = 1, k$. At this point, we have the possibility to transform the dimensional coordinates $z_1, z_2, ...z_k$ to the dimensionless ones, which are introduced here by relation (5.97). We also call these relations *formulas*.

$$x_j = \frac{z_j - z_j^0}{\Delta z_j} \quad , \quad j = 1, 2, ..k \tag{5.97}$$

It is not difficult to observe that, by using this system of dimensionless coordinates for each factor, the upper level corresponds to +1, the lower level is –1 and the fundamental level of each factor is 0. Consequently, the values of the coordinates of the experimental plan centre will be zero. Indeed, the centre of the experiments and the origin of the system of coordinates have the same position. In our current example, we can consider that the membrane remains unchanged during the experiments, i.e. the membrane porosity (ε) and the zeolite concentration (c_z) are not included in the process factors.

Therefore, we have to analyse the variation of the rate of permeation according to the temperature (z_1), the trans-membrane pressure difference (z_2) and the gas molecular weight (z_3). Then, we have 3 factors each of which has two levels. Thus the number of experiments needed for the process investigation is $N = 2^3 = 8$. Table 5.13 gives the concrete plan of the experiments. The last column contains the output "y" values of the process (flow rates of permeation). Figure 5.8 shows a geometric interpretation for a 2^3 experimental plan where each cube corner defines an experiment with the specified dimensionless values of the factors. So as to process these statistical data with the procedures that use matrix calculations, we have to introduce here a fictive variable x_0, which has a permanent +1 value (see also Section 5.4.4).

Table 5.13 The matrix for a 2^3 experimental plan (example of gas permeation).

Natural values of factors				Dimensionless values of factors			Response values
Experiment number	z_1	z_2	z_3	x_1	x_2	x_3	**Permeation flow rates $y * 10^6$ (kg/s)**
1	200	40	16	–1	–1	–1	8
2	400	40	16	+1	–1	–1	11
3	200	80	16	–1	+1	–1	10
4	400	80	16	+1	+1	–1	18
5	200	40	44	–1	–1	+1	3
6	400	40	44	+1	–1	+1	5
7	200	80	44	–1	+1	+1	4
8	400	80	44	+1	+1	+1	7

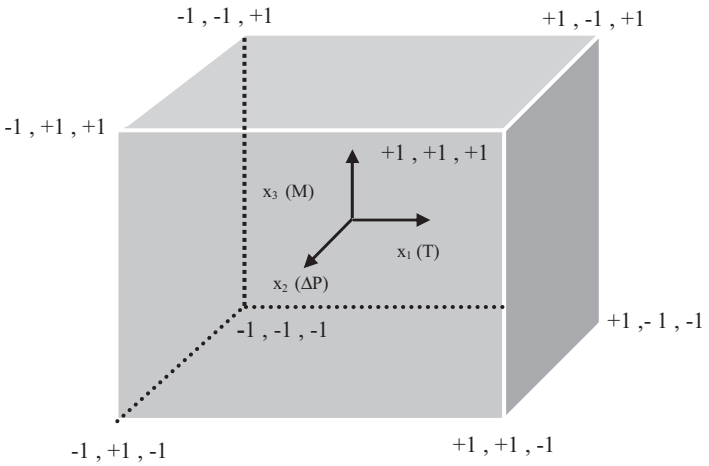


Figure 5.8 Geometric interpretation of a 2^3 experimental plan.

From a theoretical point of view, if we transform the matrix according to the 2^3 experimental plan, we obtain the state form shown in Table 5.14. This matrix has two important properties: the first is its orthogonality, the mathematical expression of which is:

$$\sum_{i=1}^{N} x_{li} x_{ju} = 0 \quad \forall \; l \neq j \;, \; l, u = 0, 1, \ldots k \tag{5.98}$$

The second is recognized as the normalization property, which shows that the sum of the dimensionless values of one factor is zero; besides, the sum of the square values of one factor is equal to the total number of experiments. Relations (5.99) and (5.100) give the mathematical expression of the norm property:

$$\sum_{i=1}^{N} x_{ji} = 0 \quad j \neq 0\ ,\ j = 1, 2 \ldots k \tag{5.99}$$

$$\sum_{i=1}^{N} x_{ji}^2 = N \quad j = 0, 1, \ldots . k \tag{5.100}$$

Table 5.14 Matrix for a 2^3 experimental plan with x_0 as fictive factor. Each line x_1,x_2,x_3 corresponds to one point of Fig. 5.8.

i	x_0	x_1	x_2	x_3	y
1	+1	–1	–1	–1	y_1
2	+1	+1	–1	–1	y_2
3	+1	–1	+1	–1	y_3
4	+1	+1	+1	–1	y_4
5	+1	–1	–1	+1	y_5
6	+1	+1	–1	+1	y_6
7	+1	–1	+1	+1	y_7
8	+1	+1	+1	+1	y_8

The orthogonality of the planning matrix, results in an easier computation of the matrix of regression coefficients. In this case, the matrix of the coefficients of the normal equation system (X^TX) has a diagonal state with the same value N for all diagonal elements. As a consequence of the mentioned properties, the elements of the inverse matrix $(X^TX)^{-1}$ have the values $d_{jj} = 1/N,\ d_{jk} = 0, j \neq k$.

In these conditions, we obtain the coefficients of the regression equation according to very simple relations as can be observed in the following matrix expression:

$$B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 1/N & 0 & 0 & . & 0 \\ 0 & 1/N & & . & 0 \\ 0 & 0 & 1/N & . & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & . & 1/N \end{bmatrix} . \begin{bmatrix} \sum_{i=1}^{N} x_{0i} y_i \\ \sum_{i=1}^{N} x_{1i} y_i \\ . \\ . \\ \sum_{i=1}^{N} x_{ki} y_i \end{bmatrix}$$

$$= \begin{bmatrix} (\sum_{i=1}^{N} x_{0i} y_i)/N \\ (\sum_{i=1}^{N} x_{1i} y_i)/N \\ . \\ . \\ (\sum_{i=1}^{N} x_{0i} y_i)/N \end{bmatrix} \tag{5.101}$$

Each coefficient β_j of the regression relationship is given by the scalar multiplication and summation of the y column and the x_j column; a final multiplication by 1/N closes the β_j calculation ($\beta_j = \frac{1}{N} \sum_{i=1}^{N} x_{ji} y_i, j = 0, k$). Now, with the help of the experimental planning from Table 5.13, we can compute the multiple linear regression given by relation (5.102). Physically, this calculation corresponds to the assumption that the flow rate of permeation through a membrane depends linearly on the temperature, trans-membrane pressure and molecular weight of permeated gas.

$$y^{th} = y = f(x_1, x_2, x_3, \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2 \tag{5.102}$$

$$\begin{bmatrix} x_{1i} \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \\ +1 \end{bmatrix} * \begin{bmatrix} y_i \\ 9 \\ 11 \\ 10 \\ 18 \\ 3 \\ 5 \\ 4 \\ 7 \end{bmatrix} = \begin{bmatrix} x_{1i} y_i \\ -9 \\ +11 \\ -10 \\ +18 \\ -3 \\ +5 \\ -4 \\ +7 \end{bmatrix} \qquad \sum_{i=1}^{8} x_{1i} y_i = 15 \quad \beta_1 = \left(\sum_{i=1}^{8} x_{1i} y_i \right) / N = 15/8 = 1.86$$

Table 5.15 contains the calculation results for all the coefficients of relation (5.102).

Table 5.15 The coefficients of relationship (5.102) according to the data from Table 5.13.

i	x_{0i}	x_{1i}	x_{2i}	x_{3i}	$x_{0i}y_i$	$x_{1i}y_i$	$x_{2i}y_i$	$x_{3i}y_i$	β_j
1	+1	–1	–1	–1	9	–9	–9	–9	
2	+1	+1	–1	–1	11	+11	–11	–11	$\beta_0 = 67/8 = 8.375$
3	+1	–1	+1	–1	10	–10	+10	–10	
4	+1	+1	+1	–1	18	+18	+18	–18	$\beta_1 = 15/8 = 1.875$
5	+1	–1	–1	+1	3	–3	–3	+3	
6	+1	+1	–1	+1	5	+5	–5	+5	$\beta_2 = 11/8 = 1.375$
7	+1	–1	+1	+1	4	–4	+4	+4	
8	+1	+1	+1	+1	7	+7	+7	+7	$\beta_3 = -29/8 = -3.625$
$\sum_{i=1}^{N=8}$	8	0	0	0	67	15	11	–29	

For a 2^3 plan, when we consider a more complete regression relationship in which the factors interact, we can write:

$$f(x_1, x_2, x_3, \beta_0, ..\beta_3, \beta_{12}, ..\beta_{23}, \beta_{123}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 \quad (5.103)$$

Here, β_{12}, β_{13}, β_{23} correspond to the effect of double interactions (factor 1 with factor 2, etc) and β_{123} introduces the effect of triple interaction. Table 5.16 completes the values shown in Table 5.13 with the values needed to calculate the considered interactions whereas Table 5.17 shows the synthesized calculations of the interactions between the coefficients.

Table 5.16 The operation matrix for double and triple interaction effects.

i	x_{0i}	x_{1i}	x_{2i}	x_{3i}	$x_{1i}x_{2i}$	$x_{1i}x_{3i}$	$x_{2i}x_{3i}$	$x_{1i}x_{2i}x_{3i}$	y_i
1	+1	–1	–1	–1	+1	+1	+1	–1	9
2	+1	+1	–1	–1	–1	–1	+1	+1	11
3	+1	–1	+1	–1	–1	+1	–1	+1	10
4	+1	+1	+1	–1	+1	–1	–1	–1	18
5	+1	–1	–1	+1	+1	–1	–1	+1	3
6	+1	+1	–1	+1	–1	+1	–1	–1	5
7	+1	–1	+1	+1	–1	–1	+1	–1	4
8	+1	+1	+1	+1	+1	+1	+1	+1	7

Table 5.17 Calculation of the interaction coefficients for model (5.110).

i	$x_{1i}x_{2i}$	$x_{1i}x_{3i}$	$x_{2i}x_{3i}$	$x_{1i}x_{2i}x_{3i}$	y_i	$x_{1i}x_{2i}y_i$	$x_{1i}x_{3i}y_i$	$x_{2i}x_{3i}y_i$	$x_{1i}x_{2i}x_{3i}y_i$	$\beta_{12,etc}$
1	+1	+1	+1	–1	9	+9	+9	+9	–9	$\beta_{12} = 7/8 = 0.875$
2	–1	–1	+1	+1	11	–11	–11	+11	+11	
3	–1	+1	–1	+1	10	–10	+10	–10	+10	$\beta_{13} = -5/8 = -0.625$
4	+1	–1	–1	–1	18	+18	–18	–18	–18	
5	+1	–1	–1	+1	3	+3	–3	–3	+3	$\beta_{23} = -5/8 = -0.625$
6	–1	+1	–1	–1	5	–5	+5	–5	–5	
7	–1	–1	+1	–1	4	–4	–4	+4	–4	$\beta_{123} = -5/8 = -0.625$
8	+1	+1	+1	+1	7	+7	+7	+7	+7	
$\sum_{i=1}^{8}$	0	0	0	0	67	7	–5	–5	–5	

If one or more parallel trials are available for the data from Table 5.13, then for the pleasure of statistical calculation, we can compute new values for the given coefficients and consequently we can investigate their statistical behaviour. A real residual variance can then be established. Unfortunately, we do not have the repeated data for our problem of gaseous permeation through a porous membrane. It is known that the matrix $(X^TX)^{-1}$ has the values $d_{jj} = 1/N,\ d_{jk} = 0, j \neq k$ and that, consequently, the regression coefficients will not be correlated. In other words, they are independent of each other. Two important aspects are noticed from this observation: (i) we can test the significance of each coefficient in the regression relationship separately; (ii) the rejection of a non-significant coefficient from the regression relationship does not have any consequence on the values of the remaining coefficients.

Coefficients $\beta_j, \beta_{jl}, \beta_{jlm}$, $j \neq l,\ j \neq m$, j and l and m $= 1, 2, \ldots k$ obtained with the help of a CFE have the quality to be absolutely correct estimators of the theoretical coefficients as defined in relation (5.4). It is important to repeat that the value of each coefficient quantifies the participation of the corresponding factor to the response construction.

Because the diagonal elements of the correlation matrix $(X^TX)^{-1}$ have the same value, we can conclude (please see the mentioned relation) that they have been determined with the same precision. Indeed, we can write that all the square roots of the coefficient variances have the same value:

$$s_{\beta_j} = s_{\beta_{jl}} = s_{\beta_{jlm}} = \frac{s_{rp}}{\sqrt{N}} \tag{5.104}$$

Let us now go back to the problem of gaseous permeation and more precisely to the experimental part when we completed the data from Table 5.13 with the values

from the permeation flow rate. These values are obtained from three experiments for the centre of plan 2^3:

$$y_1^0 = 10.5.10^{-6}\text{kg/(m}^2\text{s)};\ y_2^0 = 11.10^{-6}\text{kg/(m}^2\text{s)};\ y_3^0 = 10.10^{-6}\text{kg/(m}^2\text{s)}.$$

We obtain all the square roots of the variances needed to test the significance of the coefficients with these data:

$$\bar{y}^0 = \sum_{i=1}^{3} y_i^0/3 = 10.5 \ ; \ s_{rp}^2 = \sum_{i=1}^{3} (y_i^0 - \bar{y}^0)^2/2 = 0.25 \ ; \ s_{rp} = 0.5 \ ; \ s_{\beta_j} = s_{\beta_{jl}}$$
$$= s_{\beta_{jlm}} = s_{rp}/\sqrt{N} = 0.5/\sqrt{8} = 0.177.$$

Table 5.18 contains the calculation concerning the significance of the regression coefficients from relation (5.110). However, respect to table 5.6, the rejection condition of the hypothesis has been changed so that we can compare the computed t value (t_j) with the t value corresponding to the accepted significance level ($t_{\alpha/2}$).

Table 5.18 The significance of the coefficients for the statistical model (5.103).

n	H_0	Student variable value: t_j	$t_{\alpha/2}$ for $\nu = 2$	t_j and $t_{\alpha/2}$	Verdict
1	$\beta_0 = 0$	$t_0 = \lvert\beta_0\rvert/s_{\beta_0} = 8.37/0.17 = 47.2$	4.3	$t_0 > t_{\alpha/2}$	rejected
2	$\beta_1 = 0$	$t_1 = \lvert\beta_1\rvert/s_{\beta_1} = 1.86/0.17 = 10.5$	4.3	$t_1 > t_{\alpha/2}$	rejected
3	$\beta_2 = 0$	$t_2 = \lvert\beta_2\rvert/s_{\beta_2} = 2.75/0.17 = 15.5$	4.3	$t_2 > t_{\alpha/2}$	rejected
4	$\beta_3 = 0$	$t_3 = \lvert\beta_3\rvert/s_{\beta_3} = 3.62/0.17 = 20.45$	4.3	$t_3 > t_{\alpha/2}$	rejected
5	$\beta_{12} = 0$	$t_{12} = \lvert\beta_{12}\rvert/s_{\beta_{12}} = 0.875/0.17 = 4.94$	4.3	$t_{12} > t_{\alpha/2}$	rejected
6	$\beta_{13} = 0$	$t_{13} = \lvert\beta_{13}\rvert/s_{\beta_{13}} = 0.625/0.17 = 3.5$	4.3	$t_{13} < t_{\alpha/2}$	accepted
7	$\beta_{23} = 0$	$t_{23} = \lvert\beta_{23}\rvert/s_{\beta_{23}} = 0.625/0.17 = 3.5$	4.3	$t_{23} < t_{\alpha/2}$	accepted
8	$\beta_{123} = 0$	$t_{123} = \lvert\beta_{123}\rvert/s_{\beta_{123}} = 0.625/0.17 = 3.5$	4.3	$t_{123} < t_{\alpha/2}$	accepted

The calculation from Table 5.18 shows that coefficients β_{13}, β_{23}, β_{123} have no importance for the model and can consequently be eliminated. From these final observations, the remaining model of gaseous permeation, can be represented in a dimensionless form by the relation (5.105). We must notice that, in these calculations, the values of the y column have been multiplied by 10^6.

$$\hat{y} = 8.37 + 1.85x_1 + 2.75x_2 - 3.632x_3 + 0.875x_1x_2 \tag{5.105}$$

At the end of the process of the statistical modelling, we have to test the significance of the model. Here is the case of the model for gaseous permeation through a porous membrane for which we compute:

- the value of the residual variance:

$$s_{rz}^2 = \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2\right)/(N - n_\beta) = 7.14/3 = 1.78;$$

- the numerical value of the associated Fischer variable:
  $F = s_{rz}^2/s_{rp}^2 = 1.78/0.177 = 10$;
- the theoretical value of the associated Fischer variable corresponding to this concrete case:
  $\alpha = 0.05, \nu_1 = 3, \nu_2 = 2$ and $F_{3,2,0.05} = 19.16$.

Thanks to the assigned significance level, we can acknowledge the model to be adequate because we have $F < F_{3,2,0.05}$ (10 < 19.6).

5.5.2
Two-level Experiment Plan with Fractionary Reply

Each actual experimental research has its specificity. From the first chapter up to the present paragraph, the process modelling has been requiring more and more statistical data. With an excess of statistical data we have a better residual and reproducibility in the calculation of variances and thus coefficients can be identified more precisely. Nevertheless, this excess is not absolutely necessary and it is known that reducing the volume of statistical data saves money. When we use a CFE in our research, we first assume that each process model regression relationship is a polynome in which the interactions of the factors are considered. For example, if the relationship of the variables of the model can be limited to the linear approximation then, to develop the model, it is not necessary to use an experimental investigation made of a complete CFE. We can indeed use only one part of a CFE for experimental investigation; this part of the CFE is recognized as a fractionary factorial experiment (FFE). Because an FFE must be orthogonal, we start from the next CFE below; from this start we make sure that the number of experiments in the regression relationship remains greater than the number of unknown coefficients. We consider that the purpose of a process including three factors is to obtain a linear approximation between the process variables because we assume that this process gives a good characterization of an interesting part of the response surface. Therefore, for this part of the response surface, we can write:

$$y^{th} = f(x_1, x_2, x_3, \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{5.106}$$

To solve this problem where we have 3 unknowns, we can chose a type 2^2 CFE in which the x_1x_2 column will be the plan for x_3. Table 5.19 gives CFE 2^2 whereas Table 5.20 shows the transformation of our problem into an FFE plan. Thus, from

an initial number of $2^3 = 8$ experiments we will produce only 2^2 experiments; more generally we can say that, when we use a type 2^{k-1} FFE, we halve the initial minimum required number of experiments.

Table 5.19 The type 2^2 CFE matrix.

i	x_0	x_1	x_2	x_1x_2	y
1	+1	+1	+1	+1	y_1
2	+1	+1	–1	–1	y_2
3	+1	–1	–1	+1	y_3
4	+1	–1	+1	–1	y_4

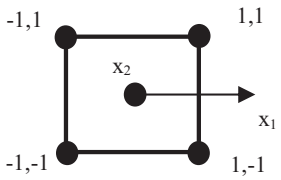


Table 5.20 The FFE plan from a type 2^2 CFE plan.

i	x_0	x_1	x_2	x_3	y
1	+1	+1	+1	+1	y_1
2	+1	+1	–1	–1	y_2
3	+1	–1	–1	+1	y_3
4	+1	–1	+1	–1	y_4

Using the experimental plan from Table 5.20 it is possible to estimate the constant terms and the three coefficients related to the linear terms from the regression relationship.

Practically, we cannot a priori postulate the nullity of the effects of the interaction. Indeed, we can accept the fact that some or all of the effects of the interaction are insignificant according to the linear effects but these are present. Then, from a practical point of view, when the coefficients corresponding to the effects of interaction are not zero and when we have the coefficients obtained by a 2^{3-1} plan, it is clear that these last coefficients include the participation of interactions on the major linear participations into the process response. The estimators of the general or theoretical coefficients are: $\beta_1^{th}, \beta_2^{th}, \beta_3^{th}, \beta_{12}^{th}, \beta_{13}^{th}, \beta_{23}^{th}$ and consequently, we can write:

$$\beta_1 \rightarrow \beta_1^{th} + \beta_{23}^{th} \quad \beta_2 \rightarrow \beta_2^{th} + \beta_{13}^{th} \quad \beta_3 \rightarrow \beta_3^{th} + \beta_{12}^{th} \tag{5.107}$$

In order to complete the FFE we can add a new column which contains the multiplication x_1x_3 to Table 5.20. However, we observe that the elements of this multi-

plication and the elements of the x_2 column are the same; so we cannot complete the FFE. Thus we can also use the fact that, in Table 5.20, we have:

$$x_3 = x_1 x_2 \tag{5.108}$$

If we multiply the relation above by x_3, we obtain $x_3^2 = x_1 x_2 x_3$ or $1 = x_1 x_2 x_3$, which is recognized as the contrast of the FFE plan. Now multiplying this contrast by x_1, x_2, x_3 yields the relations (5.109). These relations explain the relationships described in Eq. (5.107).

$$x_1 = x_1^2 x_2 x_3 = x_2 x_3 \qquad x_2 = x_1 x_3 \qquad x_3 = x_1 x_2 \tag{5.109}$$

When we decide to work with an FFE plan and when we have more than three factors, a new problem appears because we then have more possibilities to build the plan. For an answer to the question that requires a choice of most favourable possibility, we use the resolution power of each one of the options. So we generate the first possibility for an FFE plan by choosing the production (generation) relation. We can then go on with the contrast relation through which we obtain all the actual relations that are similar to those given in (5.107).

This procedure will be repeated for all the possibilities of building an FFE plan. The decision will be made according to the researcher's interest as well as to the need to obtain as much information as possible about the investigated process.

We will complete this abstract discussion with the concrete case of a process with $k = 4$ factors taking CFE 2^3 as a basis for an FFE plan. To this end we have:

$$x_4 = x_1 x_2 x_3 \tag{5.110}$$

or one out of the next three relations as a production relation:

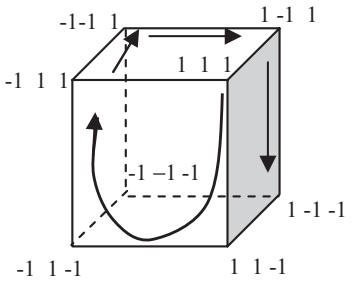
$$x_4 = x_1 x_2 \qquad x_4 = x_1 x_3 \qquad x_4 = x_2 x_3 \tag{5.111}$$

Table 5.21 gives the FFE matrix that is associated with the production relation (5.110). According to the procedure described above (showing the development of relations (5.109)), we produce the formal (5.112) system. It shows the correlation between the obtainable and theoretical coefficients of the regression relationships.

$$\begin{cases} x_1 = x_2 x_3 x_4 \rightarrow \beta_1 = \beta_1^{th} + \beta_{234}^{th} \\ x_2 = x_1 x_3 x_4 \rightarrow \beta_2 = \beta_2^{th} + \beta_{134}^{th} \\ x_3 = x_1 x_2 x_4 \rightarrow \beta_3 = \beta_3^{th} + \beta_{124}^{th} \\ x_4 = x_1 x_2 x_3 \rightarrow \beta_4 = \beta_4^{th} + \beta_{123}^{th} \\ x_1 x_2 = x_3 x_4 \rightarrow \beta_{12} = \beta_{12}^{th} + \beta_{34}^{th} \\ x_1 x_3 = x_2 x_4 \rightarrow \beta_{13} = \beta_{13}^{th} + \beta_{24}^{th} \\ x_1 x_4 = x_2 x_3 \rightarrow \beta_{14} = \beta_{14}^{th} + \beta_{23}^{th} \end{cases} \tag{5.112}$$

Table 5.21 The FFE matrix from 2^3 plan and $x_4 = x_1x_2x_2$ as a production relation.

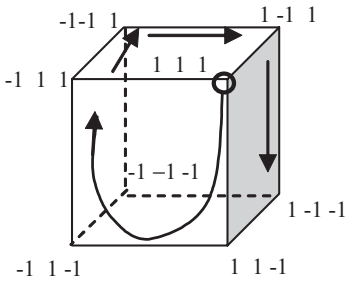
i	x_0	x_1	x_2	x_3	x_4	y
1	+1	+1	+1	+1	+1	y_1
2	+1	+1	+1	–1	–1	y_2
3	+1	–1	+1	–1	+1	y_3
4	+1	–1	+1	+1	–1	y_4
5	+1	–1	–1	+1	+1	y_5
6	+1	+1	–1	+1	–1	y_6
7	+1	+1	–1	–1	+1	y_7
8	+1	–1	–1	–1	–1	y_8



Considering the formal system (5.112), we observe that the triple interaction is indirectly considered here. It is doubtful that the actual results could confirm this class of interaction but if we can prove that they are present, then the plan from Table 5.21 can be suggested. Table 5.22 shows the FFE plan from the case when the first relation from the assembly (5.111) is the production relation. It is important to observe that all the binary interactions are indirectly considered in the formal system of the correlations of the obtainable and theoretical coefficients (5.113). Therefore, if the interest is to keep all the binary interactions of factors in the process model relationship, this FFE plan can be used successfully.

Table 5.22 The FFE matrix from a 2^3 plan and $x_4 = x_1x_2$ as a production relation.

i	x_0	x_1	x_2	x_3	x_4	y
1	+1	+1	+1	+1	+1	y_1
2	+1	+1	+1	–1	+1	y_2
3	+1	–1	+1	–1	–1	y_3
4	+1	–1	+1	+1	–1	y_4
5	+1	–1	–1	+1	+1	y_5
6	+1	+1	–1	+1	–1	y_6
7	+1	+1	–1	–1	–1	y_7
8	+1	–1	–1	–1	+1	y_8



$$\begin{cases} x_1 = x_2x_4 \rightarrow \beta_1 = \beta_1^{th} + \beta_{24}^{th} \\ x_2 = x_1x_4 \rightarrow \beta_2 = \beta_2^{th} + \beta_{14}^{th} \\ x_3 = x_1x_2x_3x_4 \rightarrow \beta_3 = \beta_3^{th} + \beta_{1234}^{th} \\ x_4 = x_1x_2 \rightarrow \beta_4 = \beta_4^{th} + \beta_{12}^{th} \\ x_1x_3 = x_2x_3x_4 \rightarrow \beta_{13} = \beta_{13}^{th} + \beta_{234}^{th} \\ x_2x_3 = x_1x_3x_4 \rightarrow \beta_{23} = \beta_{23}^{th} + \beta_{134}^{th} \\ x_3x_4 = x_1x_2x_4 \rightarrow \beta_{34} = \beta_{34}^{th} + \beta_{124}^{th} \end{cases} \quad (5.113)$$

High-level FFEs such as, for example 1/4 or 1/8 from complete factorial experiments (CFEs) can be used for complex processes, especially if the effect on the response of some factors is the objective of the research. It is not difficult to decide that, if we have a problem with the k factors where the p linear effects compensate the effects of interaction, then the 2^{k-p} FFE can be used without any restriction.

The plan 2^{k-p} FFE keeps the advantages of the CFE 2^k plan, then:

- it is an orthogonal plan and consequently simple calculation for $\beta_0, \beta_1, \ldots$ is used;
- all the regression coefficients keep their independence;
- each coefficient is computed as a result of all N experiments;
- the same minimal variance characterizes the determination of all regression coefficients.

We can then add a new "spherical" property to the properties of CFE 2^k and FFE 2^{k-p}. This new property can be used to characterize the quantity of planning information. To show the content of this property, by means of the independence of the regression relationship coefficients and according to the law governing the addition of variance for a linear regression, we can write:

$$s_{\hat{y}}^2 = s_{rz}^2 = s_{\beta_0}^2 + x_1^2 s_{\beta_1}^2 + x_2^2 s_{\beta_2}^2 + \ldots\ldots + x_k^2 s_{\beta_k}^2 \quad (5.114)$$

Because $s_{\beta_j}^2 = s_{rp}^2/N$, relation (5.114) becomes:

$$s_{\hat{y}}^2 = s_{rz}^2 = \frac{s_{rp}^2}{N}(1 + x_1^2 + x_2^2 + \ldots\ldots + x_k^2) = \frac{s_{rp}^2}{N}(1 + \gamma^2) \qquad \gamma^2 = \sum_{j=0}^{N} x_j^2 \quad (5.115)$$

Here, from a geometric viewpoint, γ is a sphere radius for the space of k dimension. When γ is significant, the residual variance $s_{\hat{y}}^2$ is also significant and, consequently, only a small quantity of information characterizes the process model.

5.5.3

Investigation of the Great Curvature Domain of the Response Surface: Sequential Experimental Planning

Figure 5.9 shows the response surface that gives the correlation between the dependent variable y (or η) and two independent factors (with values z_1 (or t) and z_2 (or c) respectively). The problem of this example concerns a chemical reaction

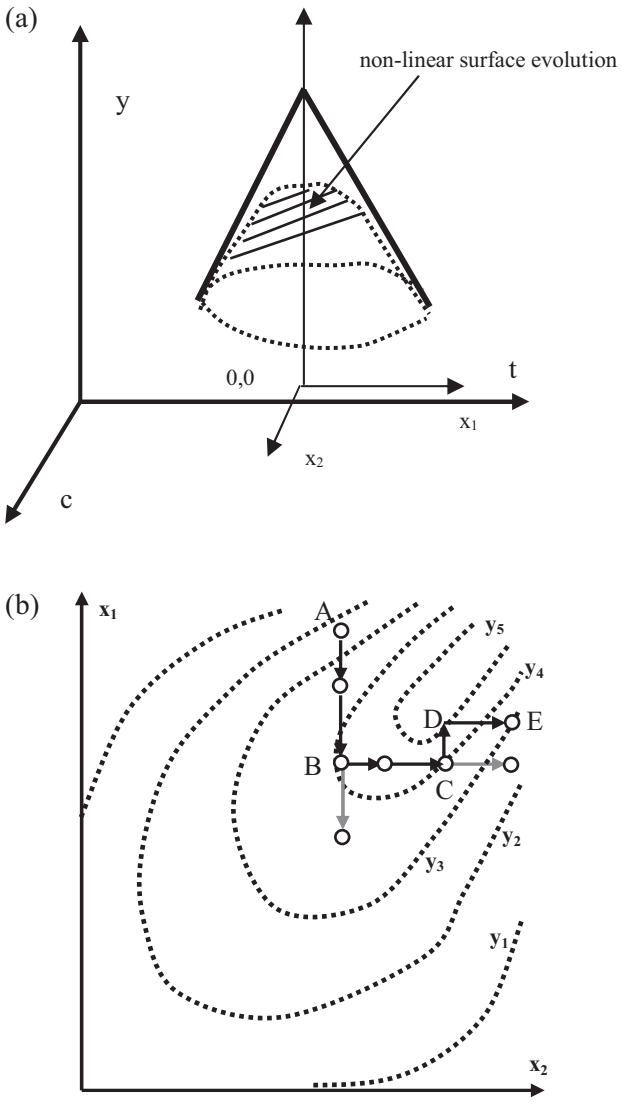


Figure 5.9 (a) Response surface for k = 2. (b) Sections of the response surface and of the gradual displacement towards the domain of the great surface curvature.

where the conversion (η) is a function of the concentration (c) and temperature (t) of the reactant. For more details, please see the data from Table 5.2. Considering Fig. 5.9(a), we can easily identify the two different domains: the first domain corresponds to the cases when y is linearly dependent on x_1 and x_2 (or near to a linear dependence); the second domain corresponds to the height of the curvature surface where the effects of the quadratic factors are significant.

We have the possibility, from a theoretical as well as from a practical point of view, to plan an experimental research so as to investigate this domain. Figure 5.9 (b) shows how we can gradually carry out these experiments.

We begin the experiments from an a priori starting point and according to the y variations. First, we keep the x_2 value unchanged and increase or decrease x_1. If y begins to decrease (point B from Fig. 5.9 (b)), we stop decreasing x_1 and go on increasing x_2 while keeping x_1 fixed. Then, we get to point C where we find out that x_2 must be maintained and x_1 changed.

It is clear that we can thus determine a way to the extreme point of the response surface curvature. At the same time, it is not difficult to observe that the ABCDE way is not a gradient. Despite its triviality, this method can be extended to more complex dependences (more than two variables) if we make amendments. It is important to note that each displacement required by this procedure is accomplished through an experiment; here the length of displacement is an apparently random variable since we cannot compute this value because we do not have any analytical or numerical expression of the response function. The response value is available at the end of the experiments.

The example shown above, introduces the necessity for a statistical investigation of the response surface near its great curvature domain. We can establish the proximity of the great curvature domain of the response surface by means of more complementary experiments in the centre of the experimental plan ($x_1 = 0, x_2 = 0, \ldots x_k = 0$). In these conditions, we can compute $\overline{y}_0$, which, together with β_0(computed by the expression recommended for a factorial experiment $\beta_0 = \left(\sum_{i=1}^{N} x_{0i} y_i\right)/N = \left(\sum_{i=1}^{N} y_i\right)/N$), gives relative information about the curvature of the surface response through relation (5.116).

$$\beta_0 - \overline{y}^0 \to \sum_{j=1}^{k} \beta_{jj}^{th} \tag{5.116}$$

It is well known that the domains of the great curvature of the response surface are characterized by non-linear variable relationships. The most frequently used state of these relationships corresponds to a two-degree polynomial. Thus, to express the response surface using a two-degree polynomial, we must have an experimental plan which considers one factor and a minimum of three different values. A complete factorial 3^k experiment requires a great number of experiments ($N = 3^k$; $k = 3$ $N = 27$; $k = 4$ $N = 81$). It is obvious that the reduction of the number of experiments is a major need here. We can consequently reduce the number of experiments if we accept the use of a composition plan (sequential

plan) [5.25] for the experimental process. The core of a sequential plan is a CFE 2^k plan with k<5 or an FFE plan with k>5. If, by means of CFE or FFE plans, the regression analysis results in an inadequate regression relationship, we can carry out new experiments for the plan. To be classified as sequential plans, these supplementary experiments require:

1. The addition of a 2^k number of experiments (uniformly disposed on the system axes) to the 2^k CFE plan. The coordinates of these points will be $(\pm\alpha, 0, 0....0), (0, \pm\alpha, 0...0)$, $...(0, 0, 0...0, \pm\alpha)$ where α is the dimensionless distance from the plan centre to an additional point.
2. An increase in the number of experiments in the centre of the experimental plan (n_0).

For a process with k factors and one response, relation (5.117) can be used to estimate the number of experiments needed by a sequential plan:

$$\begin{aligned} N &= 2^k + 2k + n_0 \quad \text{for} \quad k < 5 \\ N &= 2^{k-1} + 2k + n_0 \quad \text{for} \quad k > 5 \end{aligned} \tag{5.117}$$

The construction of a sequential plan with k = 2 is shown in Fig. 5.10. Points A B C D are the components of the 2^2 CFE and points A′ B′ C′ D′ are the components added to the basis plan. The notation n_0 in the centre of the plan shows that we must repeat the experiments. The recommended values of a dimensionless α, corresponds to the situation when the obtained composition plan keeps almost all the properties of the CFE plan.

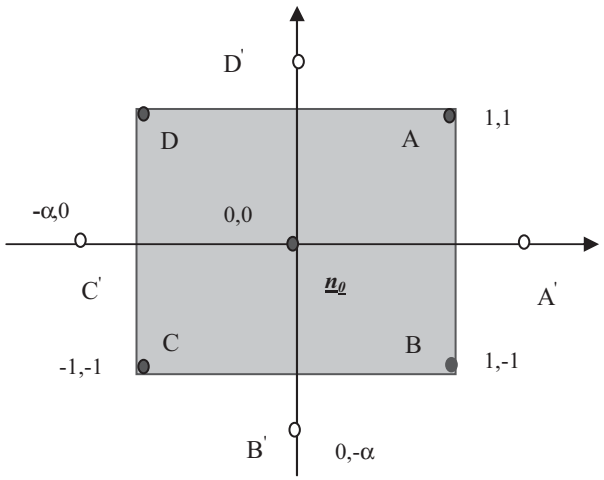


Figure 5.10 Composition of the plan based on a 2^2 CFE.

5.5.4
Second Order Orthogonal Plan

When we select the good value of the dimensionless α, then the corresponding sequential plan remains orthogonal like its CFE basic plan. At the same time, if we do not have any special request concerning a sequential plan, the number of experiments to determine fundamental factors can be drastically reduced to $n_0 = 1$. With $n_0 = 1$ and k = 2, we obtain the sequential plan shown in Table 5.23. However, with this general state this plan is not orthogonal because we have

$$\sum_{i=1}^{N} x_{0i}x_{ji}^2 \neq 0 \ , \ \sum_{i=1}^{N} x_{ji}^2 x_{li} \neq 0 \tag{5.118}$$

Table 5.23 Sequential plan for a 2^2 CFE.

i	x_0	x_1	x_2	x_1x_2	x_1^2	x_2^2	y
1	+1	+1	+1	+1	+1	+1	y_1
2	+1	+1	−1	−1	+1	+1	y_2
3	+1	−1	−1	+1	+1	+1	y_3
4	+1	−1	+1	−1	+1	+1	y_4
5	+1	+α	0	0	α^2	α^2	y_5
6	+1	−α	0	0	α^2	α^2	y_6
7	+1	0	+α	0	0	0	y_7
8	+1	0	−α	0	0	0	y_8
9	+1	0	0	0	0	0	y_9

In order to comply with the othogonality property, we have to transform the plan described in Table 5.23. For this purpose, we carry out the quadratic transformations of the data given in Table 5.23 by:

$$x_j' = x_j^2 - \frac{\sum_{i=1}^{N} x_{ji}^2}{N} = x_j^2 - \overline{x}_j^2 \tag{5.119}$$

With these transformations, we observe that:

$$\sum_{i=1}^{N} x_{0i}x_{ji}' = \sum_{i=1}^{N} x_{ji}^2 - N\overline{x}_j^2 = 0 \ , \ \sum_{i=1}^{N} x_{ji}'x_{li}' \neq 0 \tag{5.120}$$

which is a fundamental approach to the orthogonal matrix of the planned experiments. Once the quadratic transformations have been carried out, we have to

complete the orthogonal matrix. As far as we have a multiple equation system with α as unique unknown, we need to have a correlation matrix $(X^TX)^{-1}$ where all the non-diagonal elements are null. Table 5.24 has been obtained subsequent to the modified Halimov procedure [5.26]. This table gives the α values for the various factors and for a 2^{k-1} type basic CFE plan.

Table 5.24 Computed α values for a second order orthogonal plan.

	Number of independent factors				
	2	**3**	**4**	**5**	**6**
CFE basic plan	2^2	2^3	2^4	2^{5-1}	2^{6-1}
α	1	1.215	1.414	1.547	1.612

For k = 2, a second order orthogonal matrix plan is the state shown in Table 5.25. Due to the orthogonality of the matrix plan, the regression coefficients will be computed one after the other as follows:

$$\beta_j = \frac{\sum_{i=1}^{N} x_{ji} y_i}{\sum_{i=1}^{N} x_{ji}^2} \tag{5.121}$$

The relation (5.104) can be particularized to the general case of the second order orthogonal plan when we obtain the following relation for coefficients variances:

$$s_{\beta_j}^2 = s_{rp}^2 / \sum_{i=1}^{N} x_{ji}^2 \tag{5.122}$$

So the regression coefficients have been calculated for an orthogonal composition matrix and as a consequence, for the quadratic effect, we obtain the next expressions:

$$\begin{aligned} \hat{y} = \beta_0' + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \beta_{12} x_1 x_2 + ...+ \\ \beta_{k-1k} x_{k-1} x_k + \beta_{11}(x_1^2 - \overline{x}_1^2) + ... + \beta_{kk}(x_k^2 - \overline{x}_k^2) \end{aligned} \tag{5.123}$$

Therefore, the classic form of the regression relationship derives from calculating β_0 with relation (5.124):

$$\beta_0 = \beta_0' - \beta_{11}\overline{x}_1^2 - \beta_{22}\overline{x}_2^2 - \beta_{33}\overline{x}_3^2 - ..... - \beta_{kk}\overline{x}_k^2 \tag{5.124}$$

The associated variance β_0 is thus taken into account from the addition law as follows:

$$s_{\beta_0}^2 = s_{\beta_0'}^2 = (\overline{x}_1^2) s_{\beta_{11}}^2 + (\overline{x}_2^2) s_{\beta_{22}}^2 + (\overline{x}_3^2) s_{\beta_{33}}^2 + ... + (\overline{x}_k^2) s_{\beta_{kk}}^2 \tag{5.125}$$

The use of the reproducibility variance allows the significance test of the coefficients of the final regression relationship:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + \beta_{12} x_1 x_2 + ... + \beta_{k-1k} x_{k-1} x_k + \beta_{11} x_1^2 + ...\beta_{kk} x_k^2 \quad (5.126)$$

Finally, we conclude this analysis with the estimation of the model confidence. For this purpose, we use a classical procedure which consists in calculating $F = s_{rz}^2/s_{rp}^2$, establishing $\upsilon_1 = N - n_\beta$, $\upsilon_2 = N - n_0$ and calculating the significance level (α) and $F_{\upsilon_1,\upsilon_2,\alpha}$ as well as comparing F and $F = s_{rz}^2/s_{rp}^2$ before making a decision.

It is important to emphasize that in the case of an orthogonal composition plan, as shown by relation (5.122), the various regression coefficients are not calculated with similar precisions.

Table 5.25 Orthogonal composition matrix from a CFE 2^2.

i	x_0	x_1	x_2	x_1x_2	$x_1^1 = x_1^2 - \bar{x}_1^2$	$x_2^1 = x_2^2 - \bar{x}_2^2$	y
1	+1	+1	+1	+1	+1/3	+1/3	y_1
2	+1	+1	–1	–1	+1/3	+1/3	y_2
3	+1	–1	–1	+1	+1/3	+1/3	y_3
4	+1	–1	+1	–1	+1/3	+1/3	y_4
5	+1	+1	0	0	+1/3	–2/3	y_5
6	+1	–1	0	0	+1/3	–2/3	y_6
7	+1	0	+1	0	–2/3	+1/3	y_7
8	+1	0	–1	0	–2/3	+1/3	y_8
9	+1	0	0	0	–2/3	–2/3	y_9

5.5.4.1 Second Order Orthogonal Plan, Example of the Nitration of an Aromatic Hydrocarbon

The presentation of this example has two objectives: (i) to solve a problem where we use a second order orthogonal plan in a concrete case; (ii) to prove the power of statistical process modelling in the case of the non-continuous nitration of an aromatic hydrocarbon.

The initial step is the description of the process. Indeed, the nitration of the aromatic hydrocarbon occurs in a discontinuous reactor in a perfectly mixed state. The reaction takes place by contacting an aqueous phase containing nitric and sul-

furic acids with an organic phase which initially contains the aromatic hydrocarbon. The aromatic hydrocarbon transformation degree (y) depends on the following factors of the process:

- the temperature of the reaction (t associated to z_1, $\prec t \succ =$ °C);
- the time for reaction lasts (reaction time) (τ associated to z_2, $\prec \tau \succ =$ min);
- the concentration of the sulfonitric mixture according to the total reaction mass (c_{sn} associated to z_3, $\prec c_{sn} \succ = \%$ g/g);
- the concentration of the nitric acid in the sulfonitric mixture (c_a associated to z_4, $\prec c_a \succ = \%$ g/g).

The fundamental level of the factors and their variation intervals have been established and are given in Table 5.26. We accept that the factors' domains cover the great curvature of the response surface. Consequently, a regression relationship with interaction effects is a priori acknowledged.

Table 5.26 Fundamental level and intervals of variation of the factors (example 5.5.4.1)

	z_1	z_2	z_3	z_4
z_j^0	50	40	60	40
Δz_j	25	20	20	15

To solve this problem we have to use a second order orthogonal plan based on a 2^4 CFE plan. According to Table 5.24, we can establish that, for α dimensionless values of factors, we can use the numerical value $\alpha = 1.414$. Table 5.27 contains all the data that are needed for the statistical calculation procedure of the coefficients, variances, confidence, etc., including the data of the dependent variables of the process (response data).

The transformation of the dimensional z_j into the dimensionless x_j has been made using relations (5.96) and (5.97). For the reproducibility variance, four complementary experiments are available in the centre of the plan. The degrees of transformation measured at the centre of the plan are: $y_1^0 = 61.8\%$, $y_2^0 = 59.3\%$, $y_3^0 = 58.7\%$, $y_4^0 = 69\%$.

Table 5.27 Composition matrix for a 2^4 CFE (Statistical data for the example 5.5.4.1).

n^0	x_0	x_1	x_2	x_3	x_4	x_1'	x_2'
1	+1	+1	+1	+1	+1	0.2	0.2
2	+1	–1	–1	+1	+1	0.2	0.2
3	+1	+1	–1	–1	+1	0.2	0.2
4	+1	–1	+1	–1	+1	0.2	0.2
5	+1	+1	–1	+1	–1	0.2	0.2
6	+1	–1	+1	+1	–1	0.2	0.2
7	+1	+1	+1	–1	–1	0.2	0.2
8	+1	–1	–1	–1	–1	0.2	0.2
9	+1	+1	–1	+1	+1	0.2	0.2
10	+1	–1	+1	+1	+1	0.2	0.2
11	+1	+1	+1	–1	+1	0.2	0.2
12	+1	–1	–1	–1	+1	0.2	0.2
13	+1	+1	+1	+1	–1	0.2	0.2
14	+1	–1	–	+1	–1	0.2	0.2
15	+1	+1	–1	–1	–1	0.2	0.2
16	+1	–1	+1	–1	–1	0.2	0.2
17	+1	0	0	0	0	–0.8	–0.8
18	+1	1.414	0	0	0	1.2	–0.8
19	+1	–1.4.14	0	0	0	1.2	–0.8
20	+1	0	1.414	0	0	–0.8	1.2
21	+1	0	–1.4.14	0	0	–0.8	1.2
22	+1	0	0	1.414	0	–0.8	–0.8
23	+1	0	0	–1.414	0	–0.8	–0.8
24	+1	0	0	0	1.414	–0.8	–0.8
25	+1	0	0	0	–1.414	–0.8	–0.8

Table 5.27 Continued.

I	X_3'	X_4'	X_1X_2	X_1X_3	X_1X_4	X_2X_3	X_2X_4	X_3X_4	y
1	0.2	0.2	+1	+1	+1	+1	+1	+1	86.9
2	0.2	0.2	+1	–1	–1	–1	–1	+1	40.0
3	0.2	0.2	–1	–1	+1	+1	–1	–1	66.0
4	0.2	0.2	–1	+1	–1	–1	+1	–1	34.4
5	0.2	0.2	–1	+1	–1	–1	+1	–1	76.6
6	0.2	0.2	–1	–1	+1	+1	–1	–1	55.7
7	0.2	0.2	+1	–1	–1	–1	–1	+1	91
8	0.2	0.2	+1	+1	+1	+1	+1	+1	43.6
9	0.2	0.2	–1	+1	+1	–1	–1	+1	74.1
10	0.2	0.2	–1	–1	–1	+1	+1	+1	52.0
11	0.2	0.2	+1	–1	–1	–1	+1	–1	74.5
12	0.2	0.2	+1	+1	+1	+1	–1	–1	29.6
13	0.2	0.2	+1	+1	–1	+1	–1	–1	94.8
14	0.2	0.2	+1	–1	+1	–1	+1	–1	49.6
15	0.2	0.2	–1	–1	–1	+1	+1	+1	68.6
16	0.2	0.2	–1	+1	+1	–1	–1	+1	51.8
17	–0.8	–0.8	0	0	0	0	0	0	61.8
18	–0.8	–0.8	0	0	0	0	0	0	95.4
19	–0.8	–0.8	0	0	0	0	0	0	41.7
20	–0.8	–0.8	0	0	0	0	0	0	79.0
21	–0.8	–0.8	0	0	0	0	0	0	42.4
22	1.2	–0.8	0	0	0	0	0	0	77.6
23	1.2	–0.8	0	0	0	0	0	0	58.0
24	–0.8	1.2	0	0	0	0	0	0	45.6
25	–0.8	1.2	0	0	0	0	0	0	52.3

The various steps of the statistical calculation are:

1. We begin with calculating the reproducibility:

$$\bar{y}^0 = \left(\sum_{i=1}^{4} y_i^0\right)/4 = 60.95 \text{ and } s_{rp}^2 = \left(\sum_{i=1}^{4} (y_i^0 - \bar{y}^0)^2\right)/3 = 5.95.$$

2. We use Eq. (5.121) to calculate the regression coefficients and Eq. (5.122) to determine the variances of the coefficients. The calculation is explained below and the results given in Table 5.28.

Table 5.28 Values of the coefficients and their variances (example 5.5.4.1).

	β_0'	β_1	β_2	β_3	β_4	β_{11}	β_{22}	
β_j	61.54	17.37	6.42	4.7	–4.37	4.5	1.3	
$s^2_{\beta j}$	0.245	0.245	0.245	0.245	0.245	0.746	0.746	
$s_{\beta j}$	0.545	0.545	0.545	0.545	0.545	0.864	0.864	
	β_{33}	β_{44}	β_{12}	β_{13}	β_{14}	β_{23}	β_{24}	β_{34}
β_j	4.09	–5.34	2.18	0.2	1.2	0.56	0.76	1.9
$s^2_{\beta j}$	0.746	0.746	0.372	0.373	0.372	0.372	0.372	0.372
$s_{\beta j}$	0.864	0.864	0.61	0.61	0.61	0.61	0.61	0.61

$$\beta_1 = \left(\sum_{i=1}^{25} x_{1i} y_i\right) / \left(\sum_{i=1}^{25} x_{1i}^2\right)$$

$$= \frac{86.9 - 40 + 66 - 34.4 + 76.6 - 55.7 + 91 - 47.6 + 74.1 - 52 + 74.5 - 29.6 + 94.8 - 49.6 + 68.6 - 51.8 + 1.414 * 95.4 - 1.141 * 41.7}{1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+2+2}$$

$$= 17.37$$

$$\beta_{11} = \left(\sum_{i=1}^{25} x_{1I}' y_i\right) / \left(\sum_{i=1}^{25} (x_{1i}')^2\right)$$

$$= \frac{0.2(86.9 + 40 + 66 + 34.4 + 76.6 + 55.7 + 91.0 + 47.6 + 74.1 + 52 + 74.5 + 29.6 + 91.1 + 49.6 + 68.6 + 51.8}{16 * (0.2)^2 + 7 * (0.8)^2 + 2 * (1.2)^2}$$

$$+ \frac{0.8 * 61.8 + 1.2 * 95.4 + 1.2 * 41.7 - 0.8 * 79 - 0.8 * 42.4 - 0.8 * 77.6 - 0.8 * 58.0 - 0.8 * 45.6 - 0.8 * 52.3}{16 * (0.2)^2 + 7 * (0.8)^2 + 2 * (1.2)^2}$$

$$= 4.5$$

$$\beta_{12} = \left(\sum_{1}^{25} (x_1 x_2)_i y_i\right) / \left(\sum_{1}^{25} (x_1 x_2)_i^2\right)$$

$$= \frac{86.9 + 40 - 66 - 34.4 - 76.6 - 55.7 + 91 + 47.6 - 74.1 - 52 + 74.5 + 29.6 + 94.8 + 49.6 - 68.6 - 51.8}{16}$$

$$= 2.18$$

$$s^2_{\beta 1} = s^2_{rp} / \left(\sum_{i=1}^{25} x^2_{1i}\right) = 5.95/20 = 0.245 \quad ,$$

$$s^2_{\beta 11} = s^2_{rp} / \left(\sum_{i=1}^{25} x_{1i}{}'\right)^2 = 5.95/(16 * 0.2^2 + 7 * 0.8^2 + 2 * 1.2^2) = 0.746 \ ,$$

$$s^2_{\beta 12} = s^2_{rp} / \left(\sum_{i=1}^{25} x_1 x_2\right)^2_i = 5.95 : /16 = 0.372 \quad .....$$

3. We verify the significance of each coefficient of the model with the Student test. In Table 5.29 the results of the tests are given. We can then observe that coefficients β_{22}, β_{14}, β_{23} and β_{34} are non-significant.

Table 5.29 Results of the Student test for the significance of the coefficients (example 5.5.4.1).

	t_0	t_1	t_2	t_3	t_4	t_{11}	t_{22}	t_{33}
$t_j = \lvert\beta_j\rvert / s_{\beta j}$		31.9	11.7	8.64	8.04	5.2	1.5	4.73
$t_{3,0.05}$	3.09	3.09	3.09	3.09	3.09	3.09	3.06	3.09
Conclusion	S	S	S	S	S	S	NS	S

	t_{44}	t_{12}	t_{13}	t_{14}	t_{23}	t_{24}	t_{34}
$t_j = \lvert\beta_j\rvert / s_{\beta j}$	6.22	3.57	3.18	1.97	0.91	1.25	3.8
$t_{3,0.05}$	3.09	3.09	3.09	3.09	3.09	3.09	3.09
Conclusion	S	S	S	NS	NS	NS	S

4. After elimination of the non-significant coefficients, we write the new model expression and then, we compute the residual variance:

$$\hat{y} = 61.54 + 17.37x_1 + 6.4x_2 + 4.7x_3 - 4.37x_4 + 2.18x_1x_2 + 1.9x_3x_4$$
$$= 4.5(x_1^2 - 0.8) + 4.9(x_2^2 - 0.8) - 5.34(x_4^2 - 0.8)$$

$$\hat{y} = 58.9 + 17.37x_1 + 6.4x_2 + 4.7x_3 - 4.37x_4 + 2.18x_1x_2 + 1.9x_3x_4 + 4.5x_1^2 + 4.09x_3^2 - 5.31x_4^2$$

$$s_{rz}^2 = \left(\sum_{i=1}^{25}(y_i - \hat{y}_i)^2\right)/(N - n_\beta) = \left(\sum_{i=1}^{25}(y_i - \hat{y}_i)^2\right)/(25 - 10) = 396.77/15$$
$$= 26.4$$

5. We check whether the obtained model is adequate or not:
   - $F = s^2_{rz}/s^2_{rp} = 26.4/5.95 = 4.4$
   - $F_{v1,v2,\alpha} = F_{15,3,0.05} = 8.6$
   - Since we have $F_{v1,v2,\alpha} > F$, we admit that the established equation for the degree of transformation of the aromatic hydrocarbon is satisfactory.
6. Finally, we come back to the dimensional state of the factors. The result that can be used for the process optimization is:

$$\hat{y} = 64.87 - 21.68z_1 - 4.04z_2 - 34.31z_3 + 20.53z_4 + 0.00436z_1z_2 + 0.00633z_3z_4 + 0.25z_1^2 + 0.2045z_3^2 - 0.354z_4^2$$

5.5.5
Second Order Complete Plan

Even though the second order orthogonal plan is not a rotatable plan (for instance see Eqs. (5.114) and (5.115)), the errors of the experimental responses (from the response surface) are smaller than those coming from the points computed by regression. It is possible to carry out a second order rotatable plan using the Box and Hunter [5.23, 5.27] observation which stipulates that the conditions to transform a sequential plan into a rotatable plan are concentrated in the dimensionless α value where $\alpha = 2^{k/4}$ for $k<5$ and $\alpha = 2^{(k-1)/4}$ for $k>5$ respectively. Simultaneously, the number of experiments at the centre of the experimental plan (n_0) must be increased in order to make it possible to stop the degeneration of the correlation matrix $(X^TX)^{-1}$. Table 5.30 contains the required values of dimensionless α and of n_0 for a second order rotatable plan.

Table 5.30 Values of dimensionless α and n_0 for a rotatable plan with k factors.

	Number of process factors							
	2	**3**	**4**	**5**	**6**	**6**	**6**	**7**
CFE basic plan	2^2	2^3	2^{4-1}	2^{5-1}	2^{5-1}	2^{6-1}	2^{6-1}	2^{7-1}
α	1.414	1.682	2.00	2.378	2.00	2.828	2.378	3.33
n_0	5	6	7	10	6	15	9	21

For k = 2 the values of a rotatable planning matrix of the second order are given in Table 5.31. This table derives from Table 5.23. It is important to observe that the complete second order-planning matrix is not orthogonal because we have $\sum_{i=1}^{N} x_{0i}x_{ji} \neq 0$ and $\sum_{i=1}^{N} x_{ji}x_{li} \neq 0$ (see relation (5.98) for the orthogonality property).

Table 5.31 Second order complete matrix from a 2^2 CFE.

i	x_0	x_1	x_2	x_1x_2	x_1^2	x_2^2	**y**
1	+1	+1	+1	+1	+1	+1	y_1
2	+1	+1	–1	–1	+1	+1	y_2
3	+1	–1	–1	+1	+1	+1	y_3
4	+1	–1	+	–1	+1	+1	y_4
5	+1	+1.414	0	0	+2	0	y_5
6	+1	–1.414	0	0	+2	0	y_6
7	+1	0	+1.414	0	0	+2	y_7
8	+1	0	–1.414	0	0	+2	y_8
9	+1	0	0	0	0	0	y_9
10	+1	0	0	0	0	0	y_{10}
11	+1	0	0	0	0	0	y_{11}
12	+1	0	0	0	0	0	y_{12}

As a consequence, the β_{jj} coefficients will be linked with other coefficients and with the β_0 constant term. Moreover, to solve the problem of coefficients we must resolve the normal equation system by computing the inverse $(X^TX)^{-1}$ of the characteristic matrix (X^TX). As already noted in Section 5.4, the matrix of the coefficients and their associated variances are computed as follows:

$$B = (X^TX)^{-1}XY,\ s^2_{\beta_j} = d_{jj}s^2_{rp} \tag{5.127}$$

For this case of second order complete plan, the specificity of the matrix of the coefficients results in an assembly of relations directly giving the regression values of the coefficients. In this example, where the complete second order plan is based on a 2^k CFE, these relations are written as follows:

$$\beta_0 = \frac{A}{N}\left[2\lambda_k^2(k+2)\sum_{i=1}^{N}x_{0i}y_i - 2\lambda_k C\sum_{j=1}^{K}\sum_{i=1}^{N}x_{ji}^2y_i\right] \tag{5.128}$$

$$\beta_j = \frac{C}{N}\sum_{i=1}^{N}x_{ji}y_i \tag{5.129}$$

$$\beta_{jj} = \frac{A}{N}\left[C^2[(k+2)\lambda_k - k]\sum_{i=1}^{N}x_{ji}^2y_i + C^2(1-\lambda_k)\sum_{j=1}^{k}\sum_{i=1}^{N}x_{ji}^2y_i - 2\lambda_k C\sum_{i=1}^{N}x_{0i}y_i\right] \tag{5.130}$$

$$\beta_{lj} = \frac{C^2}{N\lambda_k}\sum_{i=1}^{N}x_{ji}x_{li}y_i \tag{5.131}$$

$$s^2_{\beta_0} = \frac{2A\lambda_k^2(k+2)}{N}s^2_{rp} \tag{5.132}$$

$$s^2_{\beta_{jj}} = \frac{A[(k+1)\lambda_k - (k-1)C^2]}{N}s^2_{rp} \tag{5.133}$$

$$s^2_{\beta_{lj}} = \frac{C^2}{\lambda_k N}s^2_{rp} \tag{5.134}$$

$$C = C_j = \frac{N}{\sum_{i=1}^{N}x_{ji}^2} \tag{5.135}$$

$$A = A_k = \frac{1}{2\lambda_k[(k+2)\lambda_k - k]} \tag{5.136}$$

$$\lambda_k = \frac{Nk\sum_{i=1}^{s}n_i\gamma_i^4}{(k+2)\left(\sum_{i=1}^{s}n_i\gamma_i^2\right)^2} \tag{5.137}$$

It should be mentioned that, in the calculation of parameter λ_k, s represents the number of spheres circumscribed to the experimental centre plan, γ_i is recognized as the radius of each i circumscribed sphere (see relation (5.115)) and n_i is the number of experimental points for the i sphere. It is evident that $\sum_{i=1}^{s} n_i = N$, where N gives the total number of experiments in the plan. When we use a complete second order plan, it is not necessary to have parallel trials to calculate the reproducibility variance, because it is estimated through the experiments carried out at the centre of the experimental plan. The model adequacy also has to be examined with the next procedure:

1. We begin with calculating the sum of residual squares
   $S^2_{rp} = \sum_{i=1}^{n_0} (y_i^0 - \overline{y}^0)^2$ with the following degrees of freedom:
   $\nu_1 = N - n_\beta = N - \frac{(k+1)(k+2)}{2}$;
2. We then compute the sum of the reproducibility squares with the experimental centre plan: $S^2_{rp} = \sum_{i=1}^{n_0} (y_i^0 - \overline{y}^0)^2$, where the degrees of freedom are: $\nu_2 = n_0 - 1$.
3. We define $S^2_{na} = S^2_{rz} - S^2_{rp}$ with $\nu_{na} = \nu_1 - \nu_2$ degrees of freedom as the sum of non-adequacy squares;
4. Finally, for a selected significance level, the computed Fischer variable value $F = (S^2_{na}/\nu_{na})/(S^2_{rp}/\nu_2)$ determines whether the model is adequate or not by comparison with the theoretical Fischer variable value $F_{\nu_{na},\nu_2,\alpha}$; when $F < F_{\nu_{na},\nu_2,\alpha}$ we agree to have an adequate model.

According to the testing of the significance of the model coefficients, we use the Student test where variances $s^2_{\beta_j}$ (relation (5.127)) are in fact S^2_{rp}/ν_2. Due to the fact that the coefficients are linked, if one or more coefficients are eliminated, then a new determination can be carried out.

5.5.6
Use of Simplex Regular Plan for Experimental Research

The simplex regular plan can be introduced here with the following example: a scientist wants to experimentally obtain the displacement of a y variable towards an optimal value for a $y = f(x_1,x_2)$ dependence. When the analytical expression $y = f(x_1,x_2)$ is known, the problem becomes insignificant, and then experiments are not necessary. Figure 5.11A shows that this displacement follows the way of the greatest slope. In the actual case, when the function $f(x_1,x_2)$ is unknown, before starting the research, three questions require an answer: (i) How do we select the starting point? (ii) Which experimental and calculation procedure do we use to select the direction and position of a new point of the displacement? (iii) When do we stop the displacement?

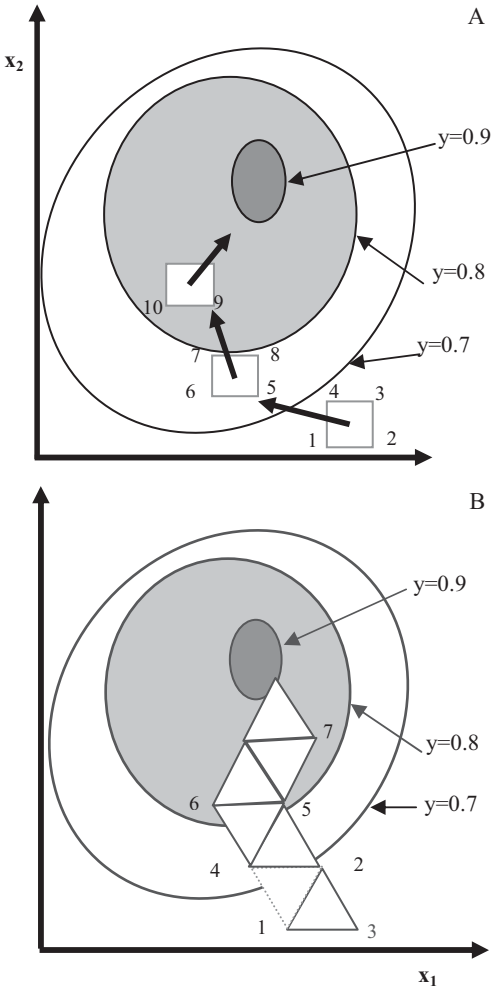


Figure 5.11 Representation of the displacement to the great curvature domain. A, according to the greatest slope method; B, according to the regular simplex method.

Question (ii) is certainly the most crucial. A possible answer to this question will be developed in the next section. The research has to begin with a small or local plan of experiments in order to describe the first movements from the starting point: when the first point of these previously planned experiments has been completed, the most non-favourable experiment will be rejected and it will be replaced by another experiment; thus we obtain, the displacement of the local group of experiments.

For a process with k factors, an abstract presentation of this procedure can be given as follows:

- We define a regular simplex plan as an assembly of k+1 equidistant points; for k = 1, the simplex is a segment; for k = 2, it is a triangle; for k = 3, we are faced with a regular tetrahedron, etc.
- Each simplex has a geometric centre placed at one point.
- When we replace the point rejected out of the group, in order to maintain a number of k+1 points, the next point will be the mirror image of the rejected point relative to the opposite face of the simplex.
- After the replacement of the rejected point, the simplex is rebuilt with a new geometric centre; only the experiments corresponding to the new point can be carried out to start the procedure (displacement and elimination) over again.

This procedure guarantees that, on the one hand, displacement towards the optimum point through the elimination of the less favourable points and, on the other hand, displacement through the maximum curvature of the response surface. For the example of a process with two factors, Fig. 5.9 B shows schematically the described procedure. The starting point of the regular simplex is triangle 123; point 3 is the less favourable response "y" and, consequently, it must be rejected; point 4 is the mirror image of point 3 according to the opposite face 12 of the simplex; thus, triangle 421 is the new simplex regular; here point 1 results in the less favourable y and as described above, we then choose point 5 which is the reflected image of point 4 (with respect to the opposite face of simplex 421).

If the dimensionless factors of an investigated process are distributed in a planning matrix (5.138) where x_j values are obtained using relation (5.139), then we can prove that the points of the matrix are organized as a regular simplex. Relation (5.140) corresponds to the distance from a point to its opposite face.

$$X = \begin{bmatrix} x_1 & x_2 & . & x_J & . & x_{k-1} & x_k \\ -x_1 & x_2 & . & x_J & . & x_{k-1} & x_k \\ 0 & -2x_2 & . & . & . & x_{k-1} & x_k \\ 0 & 0 & . & x_j & . & x_{k-1} & x_k \\ . & . & . & -jx_J & . & x_{k-1} & x_k \\ . & . & . & . & . & -(k-1)x_{k-1} & x_k \\ 0 & 0 & 0 & 0 & 0 & () & -kx_k \end{bmatrix} \tag{5.138}$$

$$x_j = \sqrt{\frac{1}{2j(j+1)}} \tag{5.139}$$

$$h_j = \frac{j+1}{\sqrt{2j(j+1)}} \tag{5.140}$$

For k factors, the number of experiments required by the simplex regular matrix is N = k+1. So, the class of saturated plans contains the simplex regular plan where the number of experiments and the number of the unknowns' coefficients are the same. For the process characterization in this example, we can only use the relationships of the linear regression. Concerning the simplex regular matrix

(5.138), we observe that it is an orthogonal matrix because we have $\sum_{i=1}^{N} x_{ji}x_{li} = 0\ ,\ \forall\ j \neq l\ ,\ j,l = 1,2,..k\ \ ;\ \sum_{i=1}^{N} x_{ji} = 0$. However, in this case, we observe that the conditions $\sum_{i=1}^{N} x_{ji}^2 = N$ are missing. Moreover, we can notice that:

$$\sum_{i=1}^{N} x_{ji}^2 = j\frac{1}{2j(j+1)} + j^2\frac{1}{2j(j+1)} = 0.5 \tag{5.141}$$

consequently, the correlation matrix of the regression coefficients can be written as follows:

$$(X^TX)^{-1} = \begin{bmatrix} 1/N & 0 & . & . & 0 \\ 0 & 2 & . & . & . \\ . & . & 2 & . & . \\ . & . & . & 2 & 0 \\ 0 & 0 & . & 0 & 2 \end{bmatrix} \tag{5.142}$$

then, the correlation matrix of the coefficients of the regression becomes:

$$\beta_0 = \left(\sum_{j=1}^{N} y_i\right)/N\ \ ,\ \ \beta_j = 2\sum_{i=1}^{N} x_{ji}y_i \tag{5.143}$$

In the previous sections we have shown that the variances relative to the β_j coefficients for the orthogonal plans are: $s_{\beta_j}^2 = s_{rp}^2/\left(\sum_{i=1}^{N} x_{ji}^2\right)$, and that, for a simple regular plan, these variances become $s_{\beta_j}^2 = s_{rp}^2/0.5 = 2s_{rp}^2$. This fact shows that the precision of a CFE plan is higher than the equivalent regular plan.

For practical use, the simplex regular plan must be drafted and computed before starting the experiment. For k process factors, this matrix plan contains k columns and k+1 lines; in the case of k = 6 the matrix (5.151) gives the following levels of the factors:

$$X = \begin{bmatrix} 0.5 & 0.289 & 0.204 & 0.158 & 0.129 & 0.109 \\ -0.5 & 0.289 & 0.204 & 0.158 & 0.129 & 0.109 \\ 0 & -0.578 & 0.204 & 0.158 & 0.129 & 0.109 \\ 0 & 0 & -0.612 & 0.158 & 0.129 & 0.109 \\ 0 & 0 & 0 & -0.632 & 0.129 & 0.109 \\ 0 & 0 & 0 & 0 & -0.645 & 0.109 \\ 0 & 0 & 0 & 0 & 0 & -0.654 \end{bmatrix} \tag{5.144}$$

The next observations will complete the understanding of this method when it is applied to the experimental scientific investigation of a real process:

1. When the experiments required by the initial simplex regular plan are completed then we eliminate the point that produces the most illogical or fool response values; by building the image of this point according to the opposite face of the simplex, we obtain the position of the new experimental point.

2. The position (coordinates) of the new experimental point can be determined as follows: (a) the j^{th} coordinates of the new point $x_j^{(k+2)}$ are computed by relation (5.145), where $x_j^{(e)}$ is the j^{th} coordinate of the rejected point and $x_j^{(c)}$ is the j^{th} corresponding coordinate of the opposite face of the rejected point; (b) the j^{th} coordinates of the centre of the opposite face of the excluded point are given by relation (5.146):

$$x_j^{(k+2)} = 2x_j^{(c)} - x_j^{(e)} \tag{5.145}$$

$$x_j^{(c)} = \left(\sum_{i=1, i \neq e}^{k+1} x_{ji} \right) / k \tag{5.146}$$

3. After each experiment a regression relationship can be obtained and analyzed using relation (5.143).
4. We can stop the experiments when the displacements of the factors do not result in a significant change in the process output.

To conclude this section, it is important to mention that the method of simplex regular plan is an open method. So, during its evolution, we can produce and add additional factors. This process can thus result in a transformation from a simplex regular plan with k columns and k+1 lines to a superior level with k+1 columns and k+2 lines. The concrete case described in the next section shows how we use this method and how we introduce a new factor into a previously established plan.

5.5.6.1 SRP Investigation of a Liquid–Solid Extraction in Batch

This example concerns a discontinuous (batch) liquid–solid extraction process. Here, the quantity of extracted species (y, $\prec y \succ =$ kgA/kg liq, A = type of species) depends on the following factors: the ratio of mixing phases (m_l/m_s-associated to z_1; $\prec m_l/m_s \succ =$ kg liq /kg solid), the contact time (τ associated to z_2; $\prec \tau \succ =$ min); the mixing rate ($w_a = \pi n d_a$-associated to z_3, $\prec w_a \succ =$ m/s, n-rotation speed, d_a – mixer diameter); the mean concentration of one species carrier, which is placed in the liquid phase (c_{sA}-associated to z_4, $\prec c_{sA} \succ =$ kg carrier/kg liq); the diameter of the solid particles (d-associated to z_5, $\prec d \succ =$ m). The temperature can be another important factor in the process, but initially we can consider that it is constant. Nevertheless, it will be considered as an additional factor in a second step of this analysis. The experiments are carried out with a solid containing 0.08 kg A/kg solid.

The fundamental levels of the factors and the variation intervals are shown in Table 5.32.

Table 5.32 Fundamental levels and variation intervals for the factors of the process.

	z_1	z_2	z_3	z_4	z_5
z_j^0	3	50	1.2	0.01	1.5×10^{-3}
Δz_j	1	20	0.6	0.004	0.5×10^{-3}

The objective of the problem is to obtain the values of the factors that correspond to a maximum concentration of the species (A) in the liquid phase.

To solve this problem we use the simplex regular method. For k = 5, the dimensionless matrix of experiments is obtained with relation (5.138). Thus, the matrix of the dimensionless factors is transformed into dimensional values with relations (5.96) and (5.97). Table 5.33 corresponds to this matrix, the last column of which contains the values of the process response. According to this table, the point placed in position 4 was found to be the least favourable for the process. However, before rejecting it, we have to build the coordinates of the new point by means of the image reflection of point number 4 (this point will be calculated to be number 7 from k+1+1). For this purpose, we use relations (5.145) and (5.146).

Table 5.33 Simplex regular plan with natural values of the factors (example 5.5.6.1).

i	z_1	z_2	z_3	z_4	$z_5 * 10^5$	y kgA/kg lq
1	3.5	55.7	1.32	0.0106	1.55	0.029
2	2.5	55.7	1.32	0.0106	1.55	0.042
3	3	39.4	1.32	0.0106	1.55	0.026
4	3	50	0.83	0.0106	1.55	0.023
5	3	50	1.2	0.0075	1.55	0.028
6	3	50	1.2	0.01	1.177	0.031

Now we show the results of these calculations, which began with the computation of the coordinate of the opposite face of the remaining simplex:

$$x_1^{(c)} = \left(\sum_{i=1, i \neq 4}^{6} x_{1i} \right) / 5 = (0.5 - 0.5 + 0 + 0 + 0)/5 = 0$$

$$x_2^{(c)} = \left(\sum_{i=1,i\neq 4}^{6} x_{2i}\right)/5 = (0.289 + 0.289 - 0.528 + 0 + 0)/5 = 0$$

$$x_3^{(c)} = \left(\sum_{i=1,i\neq 4}^{6} x_{3i}\right)/5 = (0.204 + 0.204 + 0.204 + 0 + 0)/5 = 0.612/5 = 0.122$$

$$x_4^{(c)} = \left(\sum_{i=1,i\neq 4}^{6} x_{4i}\right)/5 = (0.158 + 0.158 + 0.158 - 0.632 + 0)/5 = -0.158/5$$

$$= -0.0317$$

$$x_5^{(c)} = \left(\sum_{i=1,i\neq 4}^{6} x_{5i}\right)/5 = (0.129 + 0.129 + 0.129 + 0.129 - 0.645)/5 = 0.129/5$$

$$= -0.026$$

Now, the current dimensionless coordinate of the new point is obtained (see relation (5.145)) as follows:

$$x_1^{(k+2)} = x_1^{(7)} = 2x_1^{(c)} - x_1^{(e)} = 2x_1^{(c)} - x_1^{(4)} = 2 * 0 - 0 = 0$$

$$x_2^{(k+2)} = x_2^{(7)} = 2x_2^{(c)} - x_2^{(e)} = 2x_2^{(c)} - x_2^{(4)} = 2 * 0 - 0 = 0$$

$$x_3^{(k+2)} = x_3^{(7)} = 2x_3^{(c)} - x_3^{(e)} = 2x_3^{(c)} - x_3^{(4)} = 2 * 0.122 - (-0.612) = 0.8.56$$

$$x_4^{(k+2)} = x_4^{(7)} = 2x_4^{(c)} - x_4^{(e)} = 2x_4^{(c)} - x_4^{(4)} = -2 * 0.037 - 0.129 = -0.203$$

$$x_5^{(k+2)} = x_5^{(7)} = 2x_5^{(c)} - x_5^{(e)} = 2x_5^{(c)} - x_5^{(4)} = -2 * 0.026 - 0.109 = -0.164$$

Moreover, this new point is added to the remaining points and a new simplex (123567) will then be obtained. It is given in Table 5.34 where the factors are given in natural values. Relations (5.96) and (5.97) have been used to transform $x_1^{(k+2)}$....$x_5^{(k+2)}$ into natural values.

Table 5.34 Simplex regular plan with values of natural factors (second step of example 5.5.6.1).

n^0	z_1	z_2	z_3	z_4	$z_5 * 10^5$	y kgA/kg lq
1	3.5	55.7	1.32	0.0106	1.55	0.029
2	2.5	55.7	1.32	0.0106	1.55	0.042
3	3	39.4	1.32	0.0106	1.55	0.026
5	3	50	1.2	0.0075	1.55	0.028
6	3	50	1.2	0.01	1.177	0.031
7	3	50	1.54	0.009	1.41	0.0325

In this table, we can observe that the 7th experiment has been produced and its corresponding y value has been given. We can notice that, in simplex 123567, point number 3 is the less favourable point for the process (in this case it is the point with the lowest yield). It should therefore be eliminated. Now we can proceed with the introduction of the temperature as a new process factor. In the previous experiments, the temperature was fixed at $z_6^0 = 45$ °C. Initially, we consider that $z_6^0 = 45$ °C and we select the variation interval to be $\Delta z_6 = 15$ °C. In this situation, if we apply Eq. (5.95), we have $x_6 = \dfrac{z_6 - 45}{15}$ and obviously $x_6^{(0)} = 0$. In order to develop the 6-dimensions simplex we use relation (5.140) and then we obtain $h_6 = (6+1)/\sqrt{2 * 6 * (6+1)} = 0.764$. At this point, we can establish the values of the factors for the 8th experiment. For the first five factors the values are derived from the coordinates of the geometric centre of the simplex with 5 dimensions. These dimensionless values $x_1^{(8)}, x_2^{(8)}, \ldots x_5^{(8)}$ corroborate the procedure used for the calculation of the coordinates of a new point but, here, we consider that the coordinates of the rejected point are zero. The results of these computations are as follows:

$$x_1^{(c)} = \left(\sum_{i=1, i \neq 4}^{7} x_{1i} \right) /6 = (0.5 - 0.5 + 0 + 0 + 0 + 0)/6 = 0$$

$$x_2^{(c)} = \left(\sum_{i=1, i \neq 4}^{7} x_{2i} \right) /5 = (0.289 + 0.289 - 0.528 + 0 + 0 + 0)/6 = 0$$

$$x_3^{(c)} = \left(\sum_{i=1, i \neq 4}^{7} x_{3i} \right) /6 = (0.204 + 0.204 + 0.204 + 0 + 0 + 0.856)/6 = 1.468/6$$

$$= 0.245$$

$$x_4^{(c)} = \left(\sum_{i=1,i\neq 4}^{7} x_{4i}\right)/6 = (0.158 + 0.158 + 0.158 - 0.632 + 0 - 0.203)/6$$
$$= 0.158/5 = -0.060$$

$$x_5^{(c)} = \left(\sum_{i=1,i\neq 4}^{7} x_{5i}\right)/6 = (0.129 + 0.129 + 0.129 + 0.129 - 0.655 - 0.164)/6$$
$$= 0.109/5 = -0.048$$

$$x_1^{(k+2)} = x_1^{(8)} = 2x_1^{(c)} - x_1^{(e)} = 2x_1^{(c)} - x_1^{(4)} = 2 * 0 - 0 = 0$$

$$x_2^{(k+2)} = x_2^{(8)} = 2x_2^{(c)} - x_2^{(e)} = 2x_2^{(c)} - x_2^{(4)} = 2 * 0 - 0 = 0$$

$$x_3^{(k+2)} = x_3^{(8)} = 2x_3^{(c)} - x_3^{(e)} = 2x_3^{(c)} - x_3^{(4)} = 2 * 0.245 - 0 = 0.49$$

$$x_4^{(k+2)} = x_4^{(8)} = 2x_4^{(c)} - x_4^{(e)} = 2x_4^{(c)} - x_4^{(4)} = 2 * (-0.06) - 0 = -0.12$$

$$x_5^{(k+2)} = x_5^{(8)} = 2x_5^{(c)} - x_5^{(e)} = 2x_5^{(c)} - x_5^{(4)} = -2 * 0.048 - 0.109 = -0.096$$

For $z_6^{(8)}$ we obtain $z_6^{(8)} + z_6 x_6^{(8)} = z_6^{(8)} + z_6(x_6^{(0)} + h_6) = 45 + 15(0 + 0.764) = 52.2\,°C$. The 8th experiment together with the 123567 points gives the simplex 1235678, which is written with the values of the dimensional factors given in Table 5.35.

Table 5.35 Simplex matrix plan after the introduction of a new factor (example 5.5.6.1).

i	z_1	z_2	z_3	z_4	$z_5 * 10^5$	z_6	y kgA/kg lq
1	3.5	55.7	1.32	0.0106	1.55	45	0.029
2	2.5	55.7	1.32	0.0106	1.55	45	0.042
3	3	39.4	1.32	0.0106	1.55	45	0.026
5	3	50	1.2	0.0075	1.55	45	0.028
6	3	50	1.2	0.01	1.177	45	0.031
7	3	50	1.54	0.009	1.41	45	0.0325
8	3	50	1.49	0.0095	1.45	52.2	*waited* !

After carrying out the concrete experiment required by the 8th simplex point, the process analysis continues according to the exemplified procedure, which will stop when y cannot be increased anymore.

5.5.7 On-line Process Analysis by the EVOP Method

On-line investigation methods for statistical analysis are used when the performances of a continuous process carried out in a pilot unit or in an apparatus, have to be improved. The Evolutionary Operation Process (EVOP) method [5.7, 5.27, 5.28, 5.31] is the most famous method for on-line process analysis. The name of this method comes from its analogy with biological evolution. This analogy is based on the observation of the natural selection process in which a small variation in independent life factors is responsible for genetic mutations and thus for the evolution of species.

The objective of the EVOP method is to obtain changes in the factors of the process so as to get a more favourable state of the process outputs by means of on-line process investigation. This research is made up of small changes and programmed step-by-step. Due to the small changes in the factors, it is possible to have situations in which the effects on the output process variables can be difficult to detect because they are covered by the random effects (see the Fig. 5.1). To compensate for this difficulty in the EVOP method, the process analysis is carried out from one stage (phase) to another under a condition that imposes more iterative cycles for each phase. For each cycle, a variable number of experiments with unchanged values for the factors is important for controlling the propagation of errors. At each phase, all the experiments produced correspond to an a priori selected CFE or FFE plan. After completing the experiments required by one stage, we process their statistical data and make the necessary decision concerning the position and starting conditions for the next phase of the process analysis.

The number of cycles for each stage must be thoroughly selected because then the interest is to observe the small changes occurring simultaneously with the permanent random fluctuations in the process output. The data from a cycle are transferred to the next cycle to complete the new phase by calculation of the mean values and variances. It is well known that the errors in the mean value of n independent observations are $\sqrt{n}$ smaller than the error of an isolated measure. Therefore, this fact sustains the transfer of data from one cycle to the next one.

Figure 5.12 gives a graphic introduction to the EVOP method for the example of a process with two factors; it is important to notice that, in a real case, the displacements of the factors have to be smaller than those suggested in Fig. 5.12. Despite its apparent freedom, the EVOP method imposes some strict rules; some of them are described below:

1. For each phase, the number of cycles is not imposed by a rule or by a mathematical relation.
2. At the beginning of the second cycle, the calculation of the total effects of the analysis is obligatory; the total effects in

the example given by Fig. 5.12 are calculated by the following relation:

$$ET^a = \frac{1}{5}\left(\overline{y}_2^a + \overline{y}_3^a + \overline{y}_4^a + \overline{y}_4^a - 4\overline{y}_1^a\right) \tag{5.147}$$

Here, each $\overline{y}_i^a$ value is the mean value for all the cycles carried out with respect to the actual phase of the analysis.

3. At the beginning of the second cycle of each phase, each mean value $\overline{y}_i^a$ will be completed by its confidence intervals.
4. The cycles are stopped when the intervals of confidence for all $\overline{y}_i^a$, remain unchanged.
5. The analysis of the chain of the ET^a values completed with the split up of the total effect can help in selecting the next phase.

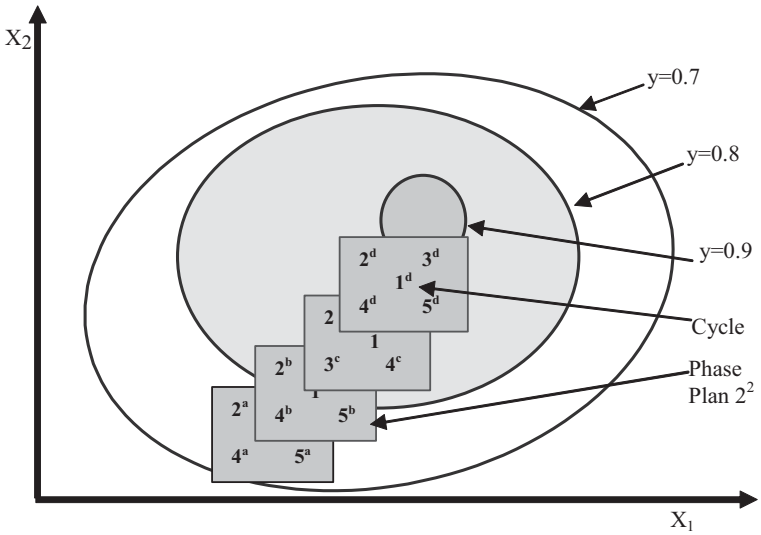


Figure 5.12 EVOP method particularized for a process with two variables.

This research method can be better illustrated by a concrete example. The investigated process example described in the next section is an organic synthesis, which takes place in a perfectly mixed reactor.

5.5.7.1 EVOP Analysis of an Organic Synthesis

We consider the case of a discontinuous organic synthesis, which occurs in a liquid medium undergoing intensive agitation; the temperature is controlled by an external heating device. The process efficiency is characterized by the conversion defined here as the ratio between the quantity of the useful species obtained and the theoretical quantity of the same species. This last value is fixed by the thermo-

dynamics and the reaction conditions. When using the EVOP method, we mean to observe the effects of the temperature and of the reaction time on the conversion. We can consider that all the other factors of the process, such as the mixing intensity, the concentrations of the reactants and catalyst, etc. remain constant, which is required by the technological considerations of the process.

We assume that the standard temperature and reaction time are fixed to 85 °C and 180 min. but small changes (± 5 °C and 10 min) have been observed to affect the process efficiency. However, these variations do not affect the process drastically. Moreover, to begin the analysis we can observe a similitude between this concrete case and the example shown in Fig. 5.12. Indeed, the working plan is a CFE 2^2 which is noted as $1^a2^a3^a4^a5^a$ in Fig. 5.12. The superscript ^{a} indicates that we are in the first phase of the EVOP procedure. The dimensionless coordinates for each point of the CFE2^2 plan are: $1^a(0,0)$, $2^a(-1,1)$, $3^a(1,1)$, $4^a(1,-1)$, $5^a(-1,-1)$. We can identify the first coordinate of 1^a to 5^a point of the CFE 2^2 plan which is $x_1 = (t - 85)/5$ and the second point coordinate is $x_2 = (\tau - 180)/10$. Table 5.36 contains the results for the first four cycles of the first phase of the particular EVOP method.

Table 5.36 Reaction conversion for four cycles of the first phase of the EVOP method.

EC	1^a	2^a	3^a	4^a	5^a
1	59.6	65.1	65.3	62.0	62.1
2	62.1	61.3	67.6	65.5	65.8
3	63.5	61.7	62.6	67.9	62.8
4	63.7	60.5	67.2	63.2	62.8

If we consider the coordinates of the points of the CFE plan, we observe that points 3^a and 4^a are the maximum values of x_1, whereas points 3^a and 5^a have the maximum values for x_2. Consequently, the effects of the factors and of their interactions will be written as follows:

$$EA^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_4 - \bar{y}_2 - \bar{y}_5) \tag{5.148}$$

$$EB^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_5 - \bar{y}_2 - \bar{y}_4) \tag{5.149}$$

$$EAB^a = \frac{1}{2}(\bar{y}_2 + \bar{y}_3 - \bar{y}_4 - \bar{y}_5) \tag{5.150}$$

We frequently use the concepts of mean values and variances in the application of the EVOP method. Before showing the concrete computations of this actual application, we need to recall here the expression for the confidence interval of a mean value: $\mu = \bar{x} \pm t_\alpha s/\sqrt{n}$ where s is the variance, $\bar{x}$ is the mean value of the selection, n gives the selection dimension and t_α is the value of the Student random variable with a significance level equal to α and with $\nu = n - 1$ degrees of freedom. Table 5.37 shows the EVOP evolution from one cycle to another respect to the data given in Table 5.36. The computations from Table 5.37 show that:

- The succession of cycles produces an important reduction in the mean deviation values and, at the same time, the confidence intervals tend to reach a final stable state.
- The effect of each factor and of its interactions on the process response (the conversion in our case) begins to be observable after running a suitable number of cycles.
- At the end of the fourth cycle, an increase in the conversion caused by the increase in temperature occurs; this observation is sustained by the positive values of the confidence interval for the mean effect of the temperature: $\mu_A = EA^a(+/-)ts/(n_c)^{0.5} = (3.6, 7.9)$.
- The positive total effect recorded after the fourth cycle, cannot sustain a further increase in the reaction time because the confidence interval for the mean effect of this factor contains a negative and a positive value: $\mu_B = EB^a(+/-)ts/(n_c)^{0.5} = (-1.0, 3.0)$; moreover we can observe that the interaction of both studied factors (temperature and reaction time) has a negative effect: $\mu_{AB} = EAB^a(+/-)ts/(n_c)^{0.5} = (-4, 0)$.
- with the situation given by the data from Table 5.37, we have two possibilities for the evolution of the research: (i) we can start with a new phase where the temperature will be increased; (ii) or we can increase the number of cycles in the actual phase so as to obtain more confidence with respect to the positive effect of the temperature.

Table 5.37 Calculation sheet for the analysis of an EVOP process (example 5.5.7.1). o.d – old deviations, n.d – new deviations.

Calculation elements Cycle = 1, $n_c = 1$		**Experiment conversion mean value**					**Mean deviations (m.d)**
		1^a	2^a	3^a	4^a	5^a	Sum of o.d: $S_a = xx$
1	Sum of old cycles S	–	–	–	–	–	Precedent m.d: $s_a = xx$
2	Mean value of the previous cycles M	–	–	–	–	–	Sum of n.d: $S_n = xx$

Table 5.37 Continued.

Calculation elements Cycle = 1, n_c = 1		**Experiment conversion mean value**					**Mean deviations (m.d)**
3	New results N	59.6	64.1	65.8	62.0	62.1	Mean value of n.d: s_n = xx
4	Differences (2) – (3) D	–	–	–	–	–	$S = \sqrt{\sum_{i=1}^{n} D_i^2}$, $s = S/\sqrt{n-1}$
5	New sum (1) + (3) SN	59.6	64.1	65.3	62.0	62.1	
6	New mean value (SN)/n_c						

Calculations of the mean effects	**The confidence intervals**
$EA^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_4 - \bar{y}_2 - \bar{y}_5) = 0.55$	$\mu_A = EA^a(+/-)ts/(n_c)^{0.5} = (xx, xx)$
$EB^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_5 - \bar{y}_2 - \bar{y}_4) = 0.65$	$\mu_B = EB^a(+/-)ts/(n_c)^{0.5} = (xx, xx)$
$EAB^a = \frac{1}{2}(\bar{y}_2 + \bar{y}_3 - \bar{y}_4 - \bar{y}_5) = 2.65$	$\mu_{AB} = EAB^a(+/-)ts/(n_c)^{0.5} = (xx, xx)$
$ET^a = \frac{1}{5}(\bar{y}_2^a + \bar{y}_3^a + \bar{y}_4^a + \bar{y}_4^a - 4\bar{y}_1^a) = 3.02$	$t = t_{5n_c-1,\alpha} =$

Calculation elements Cycle = 2 , n_c = 2		**Experiment conversion mean value**					**Mean deviations (m.d)**
		1^a	2^a	3^a	4^a	5^a	Sum of a.d: S_a = xx
1	Sum of the previous cycles S	59.6	64.1	65.8	62.0	62.1	Precedent m.d: s_o = xx
2	Mean value of previous cycles M	59.6	64.1	65.8	62.0	62.1	Sum of n.d: S_n = 6.72
3	New results N	62.1	61.3	67.6	65.5	65.8	Mean value of n.d: s_n = 3.36
4	Differences (2) – (3) D	–2.5	2.8	–2.3	–3.5	–3.7	$S = \sqrt{\sum_{i=1}^{n} D_i^2}$, $s = S/\sqrt{n-1}$
5	New sum (1)+(3) SN	121.7	125.4	132.9	127.5	127.9	
6	New mean value (SN)/n_c	60.8	62.7	66.4	63.7	63.9	

Table 5.37 Continued.

Calculations of the mean effects	The confidence intervals
$EA^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_4 - \bar{y}_2 - \bar{y}_5) = 1.75$	$\mu_A = EA^a(+/-)ts/(n_c)^{0.5} = (-2.17, 5.67)$
$EB^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_5 - \bar{y}_2 - \bar{y}_4) = 1.95$	$\mu_B = EB^a(+/-)ts/(n_c)^{0.5} = (-1.97, 5.89)$
$EAB^a = \frac{1}{2}(\bar{y}_2 + \bar{y}_3 - \bar{y}_4 - \bar{y}_5) = 0.75$	$\mu_{AB} = EAB^a(+/-)ts/(n_c)^{0.5} = (-3.17, 4.67)$
$ET^a = \frac{1}{5}(\bar{y}_2^a + \bar{y}_3^a + \bar{y}_4^a + \bar{y}_4^a - 4\bar{y}_1^a) = 2.70$	$t = t_{5n_c-1,\alpha} = t_{9,0.05} = 3.69$

Calculation elements Cycle = 3 , n_c = 3		Experiment conversion mean value					Mean deviations (m.d)
		1[a]	2[a]	3[a]	4[a]	5[a]	Sum of o.d: $S_o = 9.31$
1	Sum of the previous cycles S	121.7	125.4	132.9	127.5	127.9	Precedent m.d: $s_o = 3.36$
2	Mean value of previous cycles M	60.8	62.7	66.4	63.7	63.9	Sum of n.d: $S_n = 6.44$
3	New results N	63.5	61.7	62.6	67.9	62.8	Mean value of n.d: $s_n = 3.1$
4	Differences (2) – (3) D	–2.7	1	3.8	–4.2	1.1	$S = \sqrt{\sum_{i=1}^{n} D_i^2}$, $s = S/\sqrt{n-1}$
5	New sum (1)+(3) SN	185.2	187.1	195.5	195.4	190.7	
6	New mean value $(SN)/n_c$	61.7	62.4	65.2	65.1	63.6	

Calculations of the mean effects	The confidence intervals
$EA^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_4 - \bar{y}_2 - \bar{y}_5) = 2.15$	$\mu_A = EA^a(+/-)ts/(n_c)^{0.5} = (-0.69, 4.99)$
$EB^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_5 - \bar{y}_2 - \bar{y}_4) = 0.65$	$\mu_B = EB^a(+/-)ts/(n_c)^{0.5} = (-2.19, 3.49)$
$EAB^a = \frac{1}{2}(\bar{y}_2 + \bar{y}_3 - \bar{y}_4 - \bar{y}_5) = -0.55$	$\mu_{AB} = EAB^a(+/-)ts/(n_c)^{0.5} = (-3.39, 2.29)$
$ET^a = \frac{1}{5}(\bar{y}_2^a + \bar{y}_3^a + \bar{y}_4^a + \bar{y}_4^a - 4\bar{y}_1^a) = 1.90$	$t = t_{5n_c-1,\alpha} = t_{14,0.05} = 3.32$

Table 5.37 Continued.

Calculation elements Cycle = 4 , n_c = 4		Experiment conversion mean value					Mean deviations (m.d)
		1[a]	2[a]	3[a]	4[a]	5[a]	Sum of o.d: S_o = 12.0
1	Sum of the previous cycles S	185.2	187.1	195.5	195.4	190.7	Precedent m.d: s_o = 3.1
2	Mean value of previous cycles M	61.7	62.4	65.2	65.1	63.6	Sum of n.d: S_n = 3.982
3	New results N	63.7	60.5	67.2	63.2	62.8	Mean value of n.d: s_n = 2.82
4	Differences (2) – (3) D	–2.0	1.9	–2.0	1.9	0.8	$S = \sqrt{\sum_{i=1}^{n} D_i^2}$, $s = S/\sqrt{n-1}$
5	New sum (1) + (3) SN	248.7	247.6	262.7	258.6	253.5	
6	New mean value (SN)/n_c	62.2	61.9	65.2	65.7	63.4	

Calculations of the mean effects	The confidence intervals
$EA^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_4 - \bar{y}_2 - \bar{y}_5) = 5.6$	$\mu_A = EA^a(+/-)ts/(n_c)^{0.5} = (3.6, 7.9)$
$EB^a = \frac{1}{2}(\bar{y}_3 + \bar{y}_5 - \bar{y}_2 - \bar{y}_4) = 1.0$	$\mu_B = EB^a(+/-)ts/(n_c)^{0.5} = (-1.0, 3.0)$
$EAB^a = \frac{1}{2}(\bar{y}_2 + \bar{y}_3 - \bar{y}_4 - \bar{y}_5) = -2$	$\mu_{AB} = EAB^a(+/-)ts/(n_c)^{0.5} = (-4, 0)$
$ET^a = \frac{1}{5}(\bar{y}_2^a + \bar{y}_3^a + \bar{y}_4^a + \bar{y}_4^a - 4\bar{y}_1^a) = 7.4$	$t = t_{5n_c-1,\alpha} = t_{19,0.05} = 3.17$

5.5.7.2 Some Supplementary Observations

The example presented above successfully illustrates how we develop and use the EVOP method for a discontinuous process. When we have a continuous process, it is suggested to transform it artificially into a discontinuous process. For this purpose, we must take into consideration all the factors of the process representing flow rates according to a fixed period of time. With these transformations we can control the effect of the random factors that influence the continuous process. If, for example, we consider the case of a continuous reactor, then, the conversion can be obtained from the analysis of 5 to 6 samples (each selected at a fixed period of time), when the corresponding input and output quantities are related to the

reactor. Additionally, the other reactor factors are not different from those of the discontinuous process. The case of the continuous reactor can easily be extended to all separation apparatuses or pilot units working continuously.

In all experimental process investigations, where the final decision is the result of the hypotheses based on a comparison of the variances, we must know whether the observed variances are related to the process or to the experimental analysis procedure. Indeed, it is quite important to determine, when an experimental research is being carried out, whether we have to use a method or an instrument of analysis that produces an artificially high variance on the measured parameters.

Before the era of modern computers, the EVOP process investigation was used successfully to improve the efficiency of many chemical engineering processes. Now its use is receding due to the competition from process mathematical modelling and simulation. However, biochemical and life processes are two large domains where the use of the EVOP investigation can still bring spectacular results.

5.6
Analysis of Variances and Interaction of Factors

The objective of the statistical analysis of variances is to separate the effects produced by the dependent variables in the factors of the process. At the same time, this separation is associated with a procedure of hypotheses testing what allows to reject the factors (or groups of factors) which do not significantly influence the process. The basic mathematical principle of the analysis of variances consists in obtaining statistical data according to an accepted criterion. This criterion is complemented with the use of specific procedures that show the particular influence or effects of the grouping criterion on dependent variables.

Besides, after identifying the effects, it is necessary to compare variances of the process produced by the variation of the factors and the variances of the process produced by the random factors [5.5, 5.8, 5.29–5.31].

The number of criteria that determines the grouping of the data is strictly dependent on the number of the factors of the process accepted for the investigation.

These abstract concepts will be illustrated in the next section with the example of a catalytic chemical reaction in which we consider that different type of catalysts are available to perform the reaction and where the conversion for a fixed contact time is the dependent variable of the process. If we consider that all the other factors of the process stay unchanged, then, we can take into account a single variable factor of the process: the type of catalyst. The basis of the mono factor variance analysis concerns the collected data containing the maximum number of conversion measurements respect to each type of catalyst. Now, if the temperature is also considered as an independent variable (factor), for each fixed temperature, the collected data must show the conversion values for each catalyst. Now, we can

arrange the data in order to start the analysis of the variances of two factors. Obviously, this example can be generalized to the case of k factors (analysis of the variances with k-factors). If the residual variance increases from one experiment to the other, the effect of each factor is not summative, then we can claim that, in this case, we have an effect of the interaction factors.

For a process with more than two factors, we can consider the interactions of different factors theoretically. However, in real cases only two and a maximum of three factors interactions are accepted. All the examples selected in what follows consider the same major problem: how do we reject the non-significant factors out of the large range of factors of the process.

5.6.1 Analysis of the Variances for a Monofactor Process

The analysis of the variances of a monofactor process can be used for the indirect testing of both mean values obtained when the process factors take m discrete values. Table 5.38 introduces the preparation of the data for the analysis of the variances of a monofactor process. We can note that each value of the factor must produce m measurements of the process response.

The data arrangement shown in Table 5.38 can hint that the observable differences from one value to the other, from one column to the other are caused by the factor changes and by the problems of reproducibility.

Table 5.38 Experimental data arrangement for starting the analysis of the variances of a monofactor.

Factor value	$x = \alpha_1$	$x = \alpha_2$	$x = \alpha_j$	.....	$x = \alpha_m$
Trial					
1	v_{11}	v_{21}	v_{j1}	.....	v_{m1}
2	v_{12}	v_{22}	v_{j2}	.....	v_{m2}
3	v_{13}	v_{23}	v_{j3}	.....	v_{m3}
.....	.....	.....	.....	.....	.....
i	v_{1i}	v_{2i}	v_{ji}	.....	v_{mi}
.....	.....	.....	.....	.....	.....
n	v_{1n}	v_{2n}	v_{jn}	.....	v_{mn}
Total	v_1	v_2	v_j	.....	v_m

In the table, the differences between the columns result from the change in the values of the factors and the differences between the lines give the reproducibility problems of the experiments. The total variance (s^2) associated to the table data, here given by relation (5.151), must be divided according to its components: the variances of inter-lines (or reproducibility variances) and variances of inter-columns (or variances caused by the factor).

$$s^2 = \frac{\left[\sum_{j=1}^{m}\sum_{i=1}^{n}\left(v_{ji} - v^{=}\right)^2\right]}{mn-1} = \frac{\left[\sum_{j=1}^{m}\sum_{i=1}^{n} v_{ji}^2 - \frac{1}{mn}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} v_{ij}\right)^2\right]}{mn-1} \tag{5.151}$$

The result of this division is given in Table 5.39 where the starting data to complete the table have been obtained using the sums S_1, S_2 and S_3:

- the sum of all the squares of all observations (S_1):

$$S_1 = \sum_{j=1}^{m}\sum_{i=1}^{n} v_{ji}^2 \tag{5.152}$$

- the sum of the squares of the total of each column divided by the number of observations (S_2):

$$S_2 = \frac{\sum_{j=1}^{m} v_j^2}{n} \tag{5.153}$$

- the sum of the squares of the all added experimental observations divided by the total number of observations (S_3):

$$S_3 = \frac{\left(\sum_{j=1}^{m}\sum_{i=1}^{n} v_{ij}\right)^2}{mn} = \frac{\left(\sum_{j=1}^{m} v_j\right)^2}{mn} \tag{5.154}$$

Table 5.39 Analysis of the variances for a monofactor.

Variance origin	Sums of the differences	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable
Between the columns	$S_2 - S_3$	$m - 1$	$s_1^2 = \frac{S_2 - S_3}{m-1}$	$F = \frac{s_1^2}{s_2^2}$	$F_{m-1,m(n-1),\alpha}$
Between the lines	$S_1 - S_2$	$m(n-1)$	$s_2^2 = \frac{S_1 - S_2}{m(n-1)}$		
Total	$S_1 - S_3$	$mn - 1$			

Table 5.39 also contains the indications and calculations required to verify the zero hypothesis. This hypothesis considers the equality of the variance containing the effect of the factor on the process response (s_1^2) with the variance that shows the experimental reproducibility (s_2^2).

According to the aspects of the statistical hypothesis about the equality of two variances (see also Section 5.3) we accept the zero hypothesis if the computation shows that $F<F_{m-1,m(n-1),\alpha}$. If we refuse the zero hypothesis, then we accept that the considered factor of the process has an important influence on the response.

The numerical application described above, concerns the catalytic oxidation of SO_2 where six different catalysts are tested. The main purpose is to select the most active catalysts out of the six given in this table. All the other parameters that characterize the reaction have been maintained constant during the experiments and eight measurements have been produced for each type of catalyst. Table 5.40 presents the SO_2 transformation degrees obtained. Before reaching a conclusion about these results, we have to verify whether the different transformation degrees obtained with the six catalysts are significant or not.

Table 5.40 SO_2 transformation degree for six different catalysts. Integral reactor $l/d = 50$, $l = 1$ m, $c_{SO2} = 8\%v/v$, $c_{o2} = 10\%v/v$, N_2 inert gas, $d_p = 0.003$ m, $w_f = 0.1$ m/s.

Catalyst m	**1**	**2**	**3**	**4**	**5**	**6**
Trial number n						
1	25.1	22.8	25.5	24.5	25.5	24.7
2	27.0	23.8	27.9	25.2	28.7	27.1
3	29.6	27.1	28.8	27.7	26.2	26.0
4	26.6	22.7	26.9	26.9	25.7	26.2
5	25.2	22.8	25.4	27.1	27.2	25.7
6	28.3	27.4	30.0	30.6	27.9	29.2
7	24.7	22.2	29.6	26.4	25.6	28.0
8	25.1	25.1	23.5	26.6	28.5	24.4
Total	211.6	193.9	217.6	215.0	215.3	211.2
Mean value	26.5	24.1	27.2	26.9	26.9	26.4

To begin the analysis, we consider the zero hypothesis (in which the degrees of transformation reached with the different catalysts are similar) and to verify it, we make the computations required in Table 5.39. Then we have: $S_1 = 33\,511.11$, $S_2 = 33\,368.53$, $S_3 = 33\,322.20$, $S_2 - S_3 = 46.33$, $S_1 - S_2 = 142.58$, $S_1 - S_3 = 188.91$,

$m - 1 = 5$, $m(n-1) = 42$, $mn = 47$, $s_1^2 = 46.33/5 = 9.27$, $s_2^2 = 144.58/42 = 3.16$, $F = 9.27/3.16 = 2.93$.

The theoretical value of the Fischer random variable corresponding to the confidence level $\alpha = 0.05$ is 2.44 (it is a solution of the equation $1 - \alpha = \int_0^{F_{m-1,m(n-1),\alpha}} f_{m-1,m(n-1)}(F)dF$). Now we can observe that $F = 2.93 \succ F_{5,35,0.05} = 2.44$ and consequently we can reject the zero hypothesis, which suggests the equality of the reproducibility variance and of the variance due to the change in catalyst. In other words, we can claim that each catalyst tested has a different influence on the SO_2 transformation degree.

5.6.2
Analysis of the Variances for Two Factors Processes

When we investigate the effect of two factors on a process response, then the collected data will be as shown in Table 5.41. Here the differences between the observed values along one line present the effect of the change of x_1 from α_1 to α_m, whereas the differences between the observed values along one column are the result of the change of x_2 from β_1 to β_n. Each value of the table represents an observation that corresponds to a grouping of factors. Here, we can have one or more measurements of the process response, but frequently only one measurement is used.

Table 5.41 Arrangement of the experimental data to start the analysis of two-factor variances.

Values for the first factor	$x_1 = \alpha_1$	$x_1 = \alpha_2$	$x_1 = \alpha_j$	.....	$x_1 = \alpha_m$	**total**
Values for the second factor						
$x_2 = \beta_1$	v_{11}	v_{21}	v_{j1}	.....	v_{m1}	v_{l1}
$x_2 = \beta_2$	v_{12}	v_{22}	v_{j2}	.....	v_{m2}	v_{l2}
$x_2 = \beta_3$	v_{13}	v_{23}	v_{j3}	.....	v_{m3}	v_{l3}
	.....	.....	.....	.....	.....	
$x_2 = \beta_i$	v_{1i}	v_{2i}	v_{ji}	.....	v_{mi}	v_{li}
	.....	.....	.....	.....	.....	
$x_2 = \beta_n$	v_{1n}	v_{2n}	v_{jn}	.....	v_{mn}	v_{lm}
Total	v_{c1}	v_{c2}	v_{cj}	.....	v_{cm}	

In this case, conversely to the residual variance, we can propose two zero hypotheses: the first is H_{10}: "the variance of the response values determined by the change of factor x_1 has the same value as the residual variance"; the second one is H_{20}: "the variance of the response values (when x_2 factor changes) is similar to the residual variance". With these hypotheses we indirectly start the validation of two others assumptions: (i) the equality of the mean values of the lines (related to H_{10}), (ii) the equality of the mean values of the columns (related to H_{20}°).

The splitting of the total variance into parts associated to Table 5.41 follows a procedure similar to that for the analysis of the variances of a monofactor process, as previously explained. In this case, we introduce the sums of the squares S_1, S_2, S_3, S_4, S_r that are defined using Eqs. (5.155)–(5.159). Then, we compute the variances of the data of the lines (s_1^2), the variances of the data of the columns(s_2^2) and the residual variance of all data (s_{rz}^2). Then, the sums for the computation of the analysis of the variances of two factors processes are:

- the sum of all squares for all experimental data:

$$S_1 = \sum_{j=1}^{m} \sum_{i=1}^{n} v_{ji}^2 \tag{5.155}$$

- the sum of the squares of all the added columns divided by the number of observations from a column:

$$S_2 = \frac{\sum_{j=1}^{m} v_{cj}^2}{n} \tag{5.156}$$

- the sum of the squares of all the added lines divided by the number of observations from a line:

$$S_3 = \frac{\sum_{i=1}^{n} v_{li}^2}{m} \tag{5.157}$$

- the sum of the squares of all added experimental observations divided by the number of total observations:

$$S_4 = \frac{\left(\sum_{j=1}^{m} \sum_{i=1}^{n} v_{ij}\right)^2}{mn} = \frac{\left(\sum_{j=1}^{m} v_{c}j\right)^2}{mn} \tag{5.158}$$

- the sum of the residual squares:

$$S_r = S_1 + S_4 - S_2 - S_3 \tag{5.159}$$

It is not difficult to observe, when we compare this example with the analysis of variances of a monofactor processes, that sum S_3 is the only one to be completely new. The other sums, such as S_1 and S_2, remain unchanged or are named differently (here, S_4 is similar to the S_3 of the analysis of variances for a monofactor process). The corresponding number of degrees of freedom is attached to S_2, S_3

and S_{rz}. They are respectively m – 1 for S_2, n – 1 for S_3 and (m – 1)(n – 1) for S_{rz}. These degrees of freedom will be associated to the Fischer random variable while the proposed hypotheses are being tested. Using the same principle as used for S_2, S_3 and S_{rz}, we can establish that mn – 1 corresponds to the number of degrees of freedom for sum S_4. With these observations, we can completely synthesize the analysis of variances for two factors processes, as shown in Table 5.42. The hypotheses $H_{10} : \sigma_1^2 = \sigma_{rz}^2 \Leftrightarrow s_1^2 = s_{rz}^2$ and $H_{20} : \sigma_2^2 = \sigma_{rz}^2 \Leftrightarrow s_2^2 = s_{rz}^2$ will be accepted when $F_1 < F^{(1)}_{(m-1),(n-1)(m-1),\alpha}$ and $F_2 < F^{(2)}_{(n-1),(n-1)(m-1),\alpha}$. It is possible to have situations where we accept one hypothesis and reject the second one. In this last case, we have to accept that both considered factors play an important role in the process response.

The analysis of the catalytic oxidation of SO_2 developed previously in this chapter, can be completed as follows: (i) the experiments with catalysts number 2 and number 6 are eliminated; (ii) new experiments are introduced in order to consider the temperature as a process factor. All the other factors of the catalytic process keep the values from Table 5.40. In Table 5.43 we present a new set of experimental results in order to obtain more knowledge of the effect of the type of catalyst and the temperature on the degree of oxidation. The correspondence between the different types of catalysts reported in Tables 5.43 and 5.40 are respectively: 1→1, 2→3, 3→4, 4→5. As has been explained above, the inlet gas composition, the gas flow rate and the length of the catalytic bed remain unchanged for all experiments, the last limitation is imposed in order to obtain the smallest errors in the measurements for the process response [5.32].

Table 5.42 Synthesis of the analysis of the variances of two factors.

Origin of the variance	Differences of sums	Number of degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Between the columns	$S_2 - S_4$	m – 1	$s_1^2 = \frac{S_4 - S_2}{m-1}$	$F_1 = s_1^2/s_r^2$	$F^{(1)}_{(m-1),(n-1)(m-1),\alpha}$	$F_1<F^{(1)}$ accept H_{10}
Between the lines	$S_3 - S_4$	n – 1	$s_2^2 = \frac{S_3 - S_4}{n-1}$	$F_2 = s_2^2/s_r^2$	$F^{(2)}_{(n-1),(n-1)(m-1),\alpha}$	$F_2<F^{(2)}$ accept H_{20}
Residual	$S_{rz} = S_1+S_4 - S_2 - S_3$	(m – 1)(n –1)	$s_{rz}^2 = \frac{S_{rz}}{(m-1)(n-1)}$	$H_{10} : \sigma_1^2 = \sigma_{rz}^2 \Leftrightarrow s_1^2 = s_{rz}^2$ $H_{20} : \sigma_2^2 = \sigma_{rz}^2 \Leftrightarrow s_2^2 = s_{rz}^2$		
Total	$S_1 - S_4$	mn –1				

Table 5.43 Comparison of SO_2 oxidation degree with different catalysts and at various temperatures.

Catalyst type	$x_1 = 1$	$x_2 = 2$	$x_3 = 3$	$x_4 = 4$	**Total of each line**
Temperature of reaction					
$x_2 = 440\,°C$	25	28	22	24	$v_{l1} = 99$
$x_2 = 450\,°C$	27	29	23	23	$v_{l2} = 102$
$x_2 = 460\,°C$	30	32	26	29	$v_{l3} = 117$
Total of each column	$v_{c1} = 82$	$v_{c2} = 89$	$v_{c3} = 71$	$v_{c4} = 76$	

With the experimental data from Table 5.43, we intend to show whether both the type of catalyst and the temperature have an important influence on the oxidation degree of sulfur dioxide. We begin with calculating the sums from Table 5.42. Then, we have:

$S_1 = (25^2+27^2+....+23^2+29^2) = 8538$, $S_2 = (82^2+89^2+71^2+76^2) = 8487.3$, $S_3 = (99^2+102^2+117^2) = 8473.5$, $S_4 = (25+27+...+23+29)^2 = 8427.0$, $S_r = S_1 + S_4 - S_2 - S_3 = 8538 + 8427.0 - 8487.3 - 8473.5 = 4.2$, $S_2 - S_4 = 60.3$ with $m - 1 = 3$ degrees of freedom, $S_3 - S_4 = 46.5$ with $n - 1 = 2$ degrees of freedom, $S_{rz} = 4.2$ with $(m - 1)(n - 1) = 6$ degrees of freedom, $s^2_1 = 60.3/3 = 20.1$, $s^2_2 = 46.5/2 = 23.3$, $s^2_{rz} = 4.2/6 = 0.7$, $F_1 = s^2_1/s^2_r = 20.1/0.7 = 28.8$, $F_2 = s^2_2/s^2_r = 23.3/0.7 = 33.3$, $F^{(1)}_{m-1,(m-1)(n-1),\alpha} = F^{(1)}_{3,6,0.05} = 4.786$, $F^{(2)}_{n-1,(m-1)(n-1),\alpha} = F^{(1)}_{2,6,0.05} = 5.14$.

The results of the computations are given in Table 5.44, which is a particularization of the general Table 5.42. The last three columns of Table 5.44, give the testing calculations for H_{10} and H_{20} showing that these hypotheses are rejected. We can thus observe that there are important differences between the residual variance and the variance due to the change in the type of catalyst and temperature. In other words, both factors are important factors in this process. It should be mentioned that the analysis of variances does not give a quantitative response detailing the exact type of catalyst or/and the temperature to be used for the best yield.

When the investigated process shows a small residual variance we can consider that the variance results from the action of small random factors. At the same time, this small variance is a good indication of an excellent reproducibility of the experimental measurements. Conversely, a great residual variance can show that the measurements are characterized by poor reproducibility. However, this situation can also result from one or more unexpected or unconsidered factors; this situation can be encountered when the interactions between the factors (parameters) have been neglected. In these cases, the variance of the interactions represents an important part of the overall residual variance.

Table 5.44 Synthesis of the analysis of variances for two factors – Example 5.6.2.

Origin of the variance	Differences of sums	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Between the columns	$S_2 - S_4 = 60.3$	$m - 1 = 3$	$s_1^2 = \frac{S_4 - S_2}{m-1} = 20.1$	$F_1 = s_1^2/s_r^2 = 28.7$	$F^{(1)}_{(m-1),(n-1)(m-1),\alpha} = 4.76$	$F_1 > F^{(1)}$ Reject H_{10}
Between the lines	$S_3 - S_4 = 46.2$	$n - 1 = 2$	$s_2^2 = \frac{S_3 - S_4}{n-1} = 23.3$	$F_2 = s_2^2/s_r^2 = 33.3$	$F^{(2)}_{(n-1),(n-1)(m-1),\alpha} = 5.14$	$F_2 > F^{(2)}$ Reject H_{20}
Residual	$S_r = S_1 + S_4 - S_2 - S_3 = 4.2$	$(m-1)(n-1) = 6$	$s_{rz}^2 = \frac{S_{rz}}{(m-1)(n-1)} = 0.7$	$H_{10} : \sigma_1^2 = \sigma_{rz}^2 \Leftrightarrow s_1^2 = s_{rz}^2$ $H_{20} : \sigma_2^2 = \sigma_{rz}^2 \Leftrightarrow s_2^2 = s_{rz}^2$		
Total	$S_1 - S_4 = 110$	$mn - 1 = 11$				

5.6.3
Interactions Between the Factors of a Process

To illustrate the interaction of factors in a concrete process, we will consider the example of a process with two factors which are called A and B. The experimental investigation of the considered process is made using a CFE 2^2 plan. Both parameters, A and B will present the levels A_1 and A_2, B_1 and B_2, respectively, and, consequently, the process response has four values which are a_1, a_2, b_1, b_2, (the subscripts 1 and 2 indicate the higher and lower level of the factor). With these four values, we can develop the analysis of variances for two factors. First, we have to divide the residual variance into two parts: the first shows that the differences between the measured values of the responses are due to the experimental problems of the reproducibility; and the second indicates the action of the interaction of the factors on the responses of the process. For this separation we need a great number of measurements for each grouping of factors. So, for point A_1B_1, where the values of the dimensionless factors are $x_1 = -1$ and $x_2 = -1$, we obtain more values of the process response; moreover, we have the same problem for the other following points: A_1B_2 ($x_1 = -1$, $x_2 = 1$), A_2B_1 ($x_1 = 1$, $x_2 = -1$), A_2B_2 ($x_1 = 1$, $x_2 = 1$). The solution to this problem will result in the possibility to compute the variance caused by the reproducibility. In other words, we will be able to appreciate the effect of small random factors on the process response.

The data concerning this example are shown in Table 5.45. For the development of the analysis of variances, we use sums S_1, S_2, S_3, S_4 which have already been introduced with the analysis with two factors. The supplementary sum S_5 (Eq.

(5.160)), which is the total sum of the squares sum of the repeated values for each experimental point, is also considered here:

$$S_5 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\sum_{k=1}^{p} v_{ij}^{(k)}\right)^2}{mn-1} \tag{5.160}$$

Table 5.45 Data for the analysis of variances of two factors with interaction effects.

Values of factor A	$x_1 = \alpha_1$	$x_1 = \alpha_2$	.....	$x_1 = \alpha_j$	.....	$x_1 = \alpha_m$	**Total**
Values of factor B							
$x_2 = \beta_1$	$v_{11}^{(1)}$ $v_{11}^{(p)}$	$v_{12}^{(1)}$ $v_{12}^{(p)}$	....	$v_{1j}^{(1)}$ $v_{1j}^{(p)}$	....	$v_{1m}^{(1)}$ $v_{1m}^{(p)}$	v_{l1}
$x_2 = \beta_2$	$v_{21}^{(1)}$ $v_{21}^{(p)}$	$v_{22}^{(1)}$ $v_{22}^{(p)}$	....	$v_{2j}^{(1)}$ $v_{2j}^{(p)}$	....	$v_{2m}^{(1)}$ $v_{2m}^{(p)}$	v_{l2}
.....	.....	.....	.....	.....	.....	.....	.....
$x_2 = \beta_i$	$v_{i1}^{(1)}$ $v_{i1}^{(p)}$	$v_{i2}^{(1)}$ $v_{i2}^{(p)}$	....	$v_{ij}^{(1)}$ $v_{ij}^{(p)}$	....	$v_{im}^{(1)}$ $v_{im}^{(p)}$	v_{li}
.....	.....	.....	.....	.....	.....	.....	.....
$x_2 = \beta_n$	$v_{n1}^{(1)}$ $v_{n1}^{(p)}$	$v_{n2}^{(1)}$ $v_{n2}^{(p)}$	....	$v_{nj}^{(1)}$ $v_{nj}^{(p)}$	....	$v_{nm}^{(1)}$ $v_{nm}^{(p)}$	v_{lm}
Total	v_{c1}	v_{c2}	.....	v_{cj}	.....	v_{cm}	

Table 5.46 contains a summary to analyze these variances. Here the basic problem is the testing of the following statistical hypotheses: $H_{10} : \sigma_1^2 \equiv \sigma_A^2 = \sigma_{rz}^2$, $\Leftrightarrow s_1^2 = s_{rz}^2$,$H_{20} : \sigma_2^2 \equiv \sigma_B^2 = \sigma_{rz}^2 \Leftrightarrow s_2^2 = s_{rz}^2$, $H_{120} : \sigma_{12}^2 \equiv \sigma_{AB}^2 = \sigma_{rz}^2 \Leftrightarrow s_{12}^2 = s_{rz}^2$ these hypotheses can be described as follows:

- the variance of the data produced by changes in factor A and the variance of the residual data are similar, then, all data represent the same population (H_{10});
- the variance of the data produced by changes in factor B and the variance of the residual data are similar, then, factor B is not significant for the evolution of the process (H_{20});
- the variance of the data produced by the interactions between factors A and B and the residual data variance are similar, then, the interaction factor has no effect on the process output (H_{120}).

Table 5.46 Summary of the analysis of variances for two factors with interaction effects.

Origin of variance	Differences of sums	Number of freedom degrees	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Between the columns	$S_2 - S_4$	m –1	$s_1^2 = \frac{S_2 - S_4}{m-1}$	$F_1 = s_1^2/s_r^2$	$F^{(1)}_{(m-1),(n-1)(m-1),\alpha}$	$F_1<F^{(1)}$ Accept H_{10}
Between the lines	$S_3 - S_4$	n – 1	$s_2^2 = \frac{S_3 - S_4}{n-1}$	$F_2 = s_2^2/s_r^2$	$F^{(2)}_{(n-1),(n-1)(m-1),\alpha}$	$F_2<F^{(2)}$ Accept H_{20}
Interaction AB	$S_{12} = S_5 + S_4 - S_2 - S_3$	(m – 1). (n – 1)	$s_{12}^2 = \frac{S_{12}}{(m-1)(n-1)}$	$F_{12} = s_{12}^2/s_r^2$	$F^{(12)}_{(m-1)(n-1),mp(n-1),\alpha}$	$F_{12}<F^{(12)}$ Accept H_{120}
Residual	$S_{rz} = S_1 - S_5$	mp(n – 1)	$s_{rz}^2 = \frac{S_{rz}}{mp(n-1)}$			

We can observe in Table 5.43 that the maximum yield is obtained with catalyst number two ($x_2 = 2$), the response obtained with this catalyst can be analyzed deeply with respect to other process parameters such as the input reactor gas flow rate and the temperature. Two different values or levels of these parameters will be considered whereas other parameters or factors will remain constant (Table 5.40). Table 5.47 gives the experimental data after the arrangement required by Table 5.45 together with the partial and total mean values of SO_2 oxidation degree.

Table 5.47 Data for the analysis of variances for two factors with interaction effects. Example of SO_2 oxidation factors: temperature (T) and flow rate (G).

Flow rate	$G_1 = 0.1 m^3/(m^3 cat\ s)$		$G_2 = 0.14 m^3/(m^3 cat\ s)$		Total	Mean
Temperature						
$T_1 = 450$	21.2		22.65			
	21.5	63.75	22.55	68.4	132.15	22.02
	21.05	21.27	23.20	22.8		
$T_2 = 470$	21.65		22.3			
	21.95	65.90	22.2	67.1	133.0	22.15
	22.30	21.76	22.7	22.36		
Total	129.65		135.20		265.15	
Mean	21.60		22.58			

The interest here is to verify whether the temperature, the flow rate and their interactions produce changes in the SO_2 oxidation degree. For the case when factors interact, it is interesting to determine what the favourable direction for factors variation is.

The problem is firstly investigated by making the necessary calculations to analyze the effect of the factors and the interactions with the following procedure:

- we identify: $m = 2$, $n = 2$, $p = 3$;
- we compute: $S_1 = (21.2^2 + 21.5^2 + ... + 22.2^2 + 22.7^2) = 5862.27$, $S_2 = (129.6^2 + 135.2^2)/6 = 5864.12$, $S_3 = (132.15^2 + 133^2)/6 = 5858$, $S_4 = (265.15^2)/12 = 5858.7$, $S_5 = (63.75^2 + 65.9^2 + 68.4^2 + 67.1^2)/3 = 5887.61$, $S_2 - S_4 = 5.41$, $S_3 - S_4 = 0.05$, $S_{12} = S_{AB} = S_5 + S_4 - S_3 - S_2 = 23.44$, $S_1 - S_5 = 3.56$, $s^2_1 = 5.41/(2-1) = 5.42$, $s^2_2 = 0.05/(2-1) = 0.05$, $s^2_{12} = 23.41/(1*1) = 23.41$, $s^2_{rz} = 3.56/(2*3*1) = 0.59$, $F_1 = 5.41/0.59 = 9.18$, $F_2 = 0.05/0.59 = 0.08$, $F_{12} = 23.41/0.59 = 39.9$, $F_{1,6,0.05} = 5.99$.
- we compute all the data for Table 5.48 where we verify hypotheses H_{10}, H_{20}, H_{120};
- we identify that the change in flow rate and the interaction temperature–flow rate are important for the sulphur dioxide oxidation degree.

Table 5.48 Numerical example introduced in Table 5.47.

Variances and degrees of freedom	Hypotheses	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
$s^2_1 = 5.41$, $\nu_1 = 1$	$H_{10} : \sigma_1^2 \equiv \sigma_A^2 = \sigma_{rz}^2 \Leftrightarrow s_1^2 = s_{rz}^2$	$F_1 = s^2_1 / s^2_{rz} = 9.18$	$F_{1,6,0.05} = 5.99$	Refuse
$s^2_2 = 0.05$, $\nu_2 = 1$	$H_{20} : \sigma_2^2 \equiv \sigma_B^2 = \sigma_{rz}^2 \Leftrightarrow s_2^2 = s_{rz}^2$	$F_2 = s^2_2/s^2_{rz} = 0.084$	$F_{1,6,0.05} = 5.99$	Accept
$s^2_{12} = 23.44$, $\nu_{12} = 1$ $s^2_{rz} = 0.59$, $\nu_1 = 6$	$H_{120} : \sigma_{12}^2 \equiv \sigma_{AB}^2 = \sigma_{rz}^2 \Leftrightarrow s_{12}^2 = s_{rz}^2$	$F_{12} = s^2_{12}/s^2_{rz} = 39.9$	$F_{1,6,0.05} = 5.99$	Refuse

The analysis of the effects of the interaction contains the calculation of the confidence interval with respect to the increase in SO_2 conversion when – for temperature T_1 – the flow rate varies between G_1 and G_2 and when – for flow rate G_1 – the temperature varies between T_1 and T_2. If d_i is the mean value of the increase in the SO_2 conversion degree for case i ($i = 1 \rightarrow T_1 =$ constant and the flow rate changes between and G_2), then the confidence interval for this mean value will be:

$$I_i = \left\langle d_{i-} - t_{p+p,\alpha} s_{rz}/\sqrt{2p} \;\; , \;\; d_{i-} + t_{p+p,\alpha} s_{rz}/\sqrt{2p} \right\rangle \tag{5.161}$$

where $t_{p+p,\alpha}$ is the student random variable value with 2p degrees of freedom and $1 - \alpha$ of confidence level.

In our case $t_{2p,\alpha} = t_{6,0.05} = 4.317$; $s_{rz} = (0.59)^{1/2} = 0.77$; the calculation of the intervals of confidence from Table 5.49 shows that we do not have a complete argument to suggest the variation of factors. This conclusion is sustained by the fact that we have negative and positive values for each confidence interval.

Table 5.49 The confidence intervals for the increase of the SO_2 oxidation degree.

Flow rate	$G_1 = 0.1$ m³/(m³cat s)	$G_1 = 0.1$ m³/(m³cat s)	$\bar{d}_i$	I_i
Temperature	mean value	mean value		
$T_1 = 450$ °C	21.27	22.8	1.53	(–0.57, 3.63)
$T_2 = 470$ °C	27.76	22.35	0.6	(–1.5, 2.7)

5.6.3.1 Interaction Analysis for a CFE 2^n Plan

When we use a CFE 2^2 plan to determine the interaction effects, we introduce associated variances that can be easily used to produce answers to the aspects concerning the interaction between the factors of the process [5.33, 5.34].

It is known that the analysis of variances shows which factors and interactions must be kept and which must be rejected. At the same time, the analysis of the significance for the coefficients of the statistical model of the process gives the same results: rejection of the non-significant factors and interactions from the model and consequently from the experimental process analysis. Here, apparently, we have two competitive statistical methods for the same problem. In fact, the use of the analysis of the variances before starting the regression analysis, guarantees an excellent basis to select the relationship between the variables of the process. Otherwise, a previous analysis of the dispersion (variances) drives the regression analysis to the cases when its development is made with non-saturated plans. After these necessary explanations, we can start the problem of detecting the interactions of the factors for a concrete process by showing the terminology used. For the example of the process with factors A, B and C, this terminology is given in Table 5.50. Here the values of the dependent variable of the process (process responses) are symbolically particularized according to the higher states (levels) of the factors.

Table 5.50 Terminology used for the interaction analysis using a CFE 2^3 plan.

C levels	C_1				C_2			
B levels	B_1		B_2		B_1		B_2	
A levels	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
Response values	(1)	a	b	ab	c	ac	bc	abc

In order to obtain the effect on the response values of factor A when it varies from level A_1 to A_2, we must extract the results obtained with A_1 from the results obtained with A_2. According to Table 5.50, we can write the following relations:

$$EA = (a - (1)) + (ab - b) + (ac - c) + (abc - bc)$$

$$EA = abc + ab + ac + a - bc - b - c - (1) \tag{5.162}$$

It is easy to observe that we subtract all results from the sum of responses that contain symbol "a". By the same procedure, we can write the effect of factors B and C. It results in:

$$EB = abc + ab + bc + b - ac - a - c - (1) \tag{5.163}$$

$$EC = abc + ac + bc + c - ab - a - b - (1) \tag{5.164}$$

The interaction effect AB is obtained by subtracting the effect of A at the level B_1 from the effect of factor A at level B_2. This is written mathematically as follows:

$$EAB = [(abc - bc) + (ab - b)] - [(ac - c) + (a - (1))]$$

$$EAB = abc + ab + c + (1) - ac - bc - a - b \tag{5.165}$$

The remaining interaction effects AC and BC are written using the same definition. Then, we obtain the following relations:

$$EAC = abc + ac + b + (1) - ab - bc - a - c \tag{5.166}$$

$$EBC = abc + bc + a + (1) - ab - ac - b - c \tag{5.167}$$

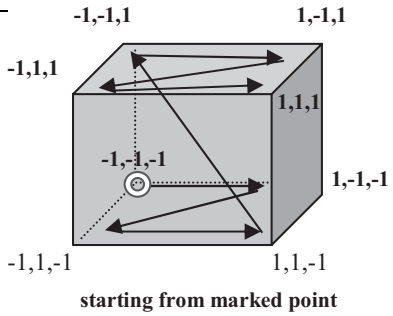
If we consider a formal vector E which includes all the effects on the process response, then we can build relation (5.168) which includes all the relations from (5.162) to (5.167):

$$[E] = \begin{bmatrix} EA \\ EB \\ EC \\ EAB \\ EAC \\ EBC \\ EABC \end{bmatrix} = \begin{bmatrix} -(1) + a - b + ab - c + ac - bc + abc \\ -(1) - a + b + ab - c - ac + bc + abc \\ -(1) - a - b - ab + c + ac + bc + abc \\ +(1) - a - b + ab + c - ac - bc + abc \\ +(1) - a + b - ab - c + ac - bc + abc \\ +(1) + a - b - ab - c - ac + bc + abc \\ -(1) + a + b - ab + c - ac - bc + abc \end{bmatrix} \tag{5.168}$$

In our example, we can keep the order of the values given in Table 5.50. However, if we change this order, then the expressions for relations (5.162)–(5.167) must agree with this change. Relation (5.168) can easily be written using the 2^3 matrix plan. Nevertheless, here, we have to consider the first point of the plan with negative coordinates. Table 5.51 shows the variation inside the factorial cube which is at the origin of relation (5.168). It is observable that the multiplication of the response column with columns A, B,..., ABC gives the corresponding partial effects EA, EB,....,EABC.

Table 5.51 Use of the CFE 2^3 for the development of relation (5.168).

I	A	B	C	AB	AC	BC	ABC	y_i
1	–1	–1	–1	+1	+1	+1	–1	$y_1 = (1)$
2	+1	–1	–1	–1	–1	+1	+1	$y_2 = a$
3	–1	+1	–1	–1	+1	–1	+1	$y_3 = b$
4	+1	+1	1	+1	–1	–1	–1	$y_4 = ab$
5	–1	–1	+1	+1	–1	–1	+1	$y_5 = c$
6	+1	–1	+1	–1	+1	–1	–1	$y_6 = ac$
7	–1	+1	+1	–1	–1	+1	–1	$y_7 = bc$
8	+1	+1	+1	+1	+1	+1	+1	$y_8 = abc$



As has been developed above, the analysis of variances imposes the calculation of the variances due to the changes and interactions between the factors. In addition, we also have to verify the next seven hypotheses where $\sigma_{rz}^2 = s_{rz}^2$ is assumed to be $\sigma_{ABC}^2 = s_{ABC}^2$:

$$H_A : \sigma_A^2 = \sigma_{rz}^2 \Leftrightarrow s_A^2 = s_{rz}^2, H_B : \sigma_B^2 = \sigma_{rz}^2 \Leftrightarrow s_B^2 = s_{rz}^2, H_C : \sigma_C^2 = \sigma_{rz}^2 \Leftrightarrow s_C^2 = s_{rz}^2$$

$$H_{AB} : \sigma_{AB}^2 = \sigma_{rz}^2 \Leftrightarrow s_{AB}^2 = s_{rz}^2, H_{AC} : \sigma_{AC}^2 = \sigma_{rz}^2 \Leftrightarrow s_{AC}^2 = s_{rz}^2, H_{BC} : \sigma_{BC}^2 = \sigma_{rz}^2$$
$$\Leftrightarrow s_{BC}^2 = s_{rz}^2$$

The acceptance of a hypothesis from those mentioned above corresponds to accepting the fact that the factor or interaction linked to the hypothesis is not important in the investigated process. In this example, the sums of the squares used for the production of the analysis of variances, is made with a CFE 2^3 plan (Table 5.52), they are expressed using the partial effects as follows: $S_A = (EA)^2/8$, $S_B = (EB)^2/8$, $S_C = (EC)^2/8$, $S_{AB} = (EAB)^2/8$, $S_{AC} = (EAC)^2/8$, $S_{BC} = (EBC)^2/8$, $S_{ABC} = (EABC)^2/8$.

Table 5.52 Synthesis of the analysis of variances for a CFE 2^3 plan.

Origin of the variance	Sums of the squares	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Change of factor A	S_A	1	$s^2_A = S_A/1$	$F_A = s^2_A/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_A<F_{1,1,\alpha}$ Accept H_A
Change of factor B	S_B	1	$s^2_B = S_B/1$	$F_B = s^2_B/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_B<F_{1,1,\alpha}$ Accept H_B
Change of factor C	S_c	1	$s^2_c = S_c/1$	$F_C = s^2_B/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_C<F_{1,1,\alpha}$ Accept H_c
Interaction A B	S_{AB}	1	$s^2_{AB} = S_{AB}/1$	$F_{AB} = s^2_{AB}/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_{AB}<F_{1,1,\alpha}$ Accept H_{AB}
Interaction A C	S_{AC}	1	$s^2_{AC} = S_{AC}/1$	$F_{AC} = s^2_{AC}/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_{AC}<F_{1,1,\alpha}$ Accept H_{AC}
Interaction B C	S_{BC}	1	$s^2_{BC} = {}_{BC}/1$	$F_{BC} = s^2_{BC}/ s^2_{rz}$	$F_{1,1,\alpha}$	$F_{BC}<F_{1,1,\alpha}$ Accept H_{BC}
Interaction A BC (residual)	S_{ABC}	1	$s^2_{ABC} = S_{ABC}/1$	$s^2_{ABC} = S_{ABC}/1 = s^2_{rz}$		
Total	$S_T = S_1 - S_2$	7	xxxx	xxxx		

The analysis of variances using a CFE 2^n plan in which, for each experimental point, we produce only one measurement, frequently presents an important residual variance. This result is a consequence of the fact that each point is the result of a particular combination of interaction effects. If, for each experimental point of the plan, we produce more experiments, then we have the normal possibility to compute a real residual variance (5.169). In this situation, the sum is successfully used as shown in Table 5.52 for the residual variance computation.

$$S_{rz} = \frac{\sum_{i=1}^{2^n} \sum_{k=1}^{r-1} d_{ik}^2}{r} \tag{5.169}$$

In relation (5.169), d_{ik} represents the difference between two values from the total values produced at point "i" (k = 1,r).

In the following example the application of this computation procedure is developed. The analysis of variances is carried out for the air oxidation of an aromatic hydrocarbon. In this process, where air is bubbled in the reaction vessel, we obtain two products: a desired compound and a secondary undesired compound. Here, it is important to know how the transformation degree of the hydrocarbon evolves towards the by-product when different process parameters (factors) are varied as follows:

- the catalyst concentration (A) varies from $A_1 = 0.1\%$ g/g to $A_2 = 0.4\%$ g/g
- the bubbling time (B) for air flow (0.01 m^3/m^3_{liquid} s) varies from $B_1 = 60$ min to $B_2 = 70$ min
- the reaction temperature(C) varies from $C_1 = 50\,°C$ to $C_2 = 60\,°C$

Table 5.53 gives the experimental results of the hydrocarbon conversion in a by-product. With the data below, we can characterize the particular effect of each parameter on the process output (hydrocarbon oxidation degree in an undesired compound) and the conclusion expected here is to suggest a proposal for the enhancement of the efficiency of the process.

Table 5.53 Analysis of the variances made for a 2^2 plan for an aromatic hydrocarbon oxidation in an undesired by-product.

$C_1 = 50\,°C$				$C_2 = 60\,°C$			
$B_1 = 60$ min		$B_2 = 70$ min		$B_1 = 60$ min		$B_2 = 70$ min	
$A_1 = 0.1\%$	$A_2 = 0.4\%$	$A_1 = 0.1\%$	$A_2 = 0.4\%$	$A_1 = 0.1\%$	$A_2 = 0.4\%$	$A_1 = 0.1\%$	$A_2 = 0.4\%$
12.6	13.5	13.4	14.9	13.2	17.7	15.9	19.2
13.1	12.0	12.4	13.4	15.7	18.2	16.4	18.7
S 25.7	S 25.5	S 25.8	S 28.3	S 28.9	S 35.9	S 32.3	S 37.9
(1)	a	b	ab	c	ac	bc	abc

The necessary computations for this example are organized as follows:

1. We compute the values of the particular effects with relations (5.168). The results are: EA = 14.9, EB = 8.3, EAB = 1.3, EC = 29.7, EAC = 10.3, EBC = 2.5, EABC = –4.5;

2. The associated sums of squares have the values: $S_A = 14.9^2/(2*8) = 3.937$, $S_B = 8.3^2/(2*8) = 4.305$, $S_C = 29.7^2/(2*8) = 55.13$, $S_{AB} = 1.3^2/(2*8) = 0.105$, $S_{AC} = 10.3^2/(2*8) = 6.630$, $S_{BC} = 2.5^2/(2*8) = 0.39$, $S_{ABC} = 4.5^2/(2*8) = 1.050$, $S_1 = (13.5^2 + 12.0^2 + ... + 19.2^2 + 18.7^2) = 3696.82$, $S_2 = ((13.5 + 12.0 + ...+ 19.2 + 18.7)^2)/16 = 3609$, $S_T = S_1 - S_2 = 87.82$;
3. The sum of the residual squares has been computed according to relation (5.169): $S_{rz} = (0.5^2 + 1.5^2 + 1^2 + 0.5^2 + 2.5^2 + 0.5^2 + 0.5^2 + 0.5^2)/2 = 6.35$.
4. All the values of the sums S_A, S_B,....S_{ABC} are one (1) for the associated number of degrees of freedom. So variances s_A, s_B,...s_{ABC} have the same values as the corresponding sums; the sum of residual squares associates value $\nu = 8$ to the number of the degrees of freedom. This fact gives value $s^2_{rz} = S_r/8 = 0.797$ for the residual variances.
5. The computed values for the associated Fischer variable (see also Table 5.52) for the variances of the factors and their interactions present the next values: $F_A = 17.4$, $F_B = 5.4$, $F_A = 17.4$, $F_C = 69.2$, $F_{AB} = 0.13$, $F_{AC} = 8.3$, $F_{BC} = 0.49$, $F_{ABC} = 1.3$.
6. The theoretical value of the Fischer random variable associated to this actual case is $F_{1,8,0.05} = 5.32$; By comparing this value with the computed values of the Fischer variable given here, we can decide that factors A, B, C as well as interaction AC determine the hydrocarbon oxidation degree in the undesired product.
7. Because we observe that factor B has an independent influence on the output of the process and considering the data from Table 5.54, we can assert that, in order to obtain small values of the conversion to by-product, it is not recommended to increase the value of B. We can compute the change in the degree of hydrocarbon oxidation in the undesired product when factors A and C increase. Indeed, this computation can result in a recommendation concerning the increase in A and C. The next mean values of the output variable are thus obtained in the points where we have only A and C, namely: A_1C_1, A_1C_2, A_2C_1, A_2C_2: $m_{A1C1} = (12.6 + 13.1 + 13.4 + 12.4)/4 = 12.875$, $m_{A1C2} = 15.30$, $m_{A2C1} = 13.45$, $m_{A2C2} = 18.45$. The changes in the oxidation degree associated to these mean values are: $d_1 = m_{A2C1} - m_{A1C1} = 0.575$ and $d_2 = m_{A2C2} - m_{A2C1} = 3.15$. Then, the mean value of the oxidation degree change is $d = (d_1 + d_2)/2 = 1.86$. This value is included within confidence interval $I = (0.4, 3.32)$ according to relation (5.161). Then, if we increase A or C or A and C, we will increase the conversion of the aromatic hydrocarbon in the undesired by-product.

5.6.3.2 Interaction Analysis Using a High Level Factorial Plan

Sometimes we may encounter situations requiring the analysis of the effects of the factors on the output variables of a process by working with more than two levels for one or more factors. The analysis of variances for this type of process is associated with a difficult methodology of data processing and interpretation. However, the method can be simplified if, at the starting point, we split the primary experimental data table into different tables in which each factor presents only two levels. We then analyze each table according to the methods presented in the previous paragraphs. The splitting up procedure is explained with a concrete example.

A small perfectly mixed discontinuous reactor is used at laboratory scale to conduct the Friedel–Crafts reaction $Ar–H + RCl \xrightarrow{AlCl_3} Ar–R + HCl$. Three factors and two or more levels of each parameter have been used in an experimental plan in order to separate and compare their influence on the aromatic hydrocarbon conversion. The following factors and levels have been used:

- reaction time (A) which has two levels: $A_1 = 10$ h, $A_2 = 7$ h;
- the particular time when the catalyst is introduced into the reactor or "timing" (B) with three levels: $B_1 = 2$ h, $B_2 = 3$ h, $B_3 = 4$ h;
- the mixing intensity (C) given here by the rotation speed of the mixer driver, which has been modified according to the following rotation levels: $C_1 = 10$ rot/min, $C_2 = 15$ rot/min, $C_3 = 20$ rot/min, $C_4 = 25$ rot/min.

The measurements of the hydrocarbon transformation are given in Table 5.54. Before using these measurements, we need to obtain data showing the interactions and the combination of factors producing the best process efficiency. Before beginning the analysis, we will divide the initial data into different fractional tables, each one with two factors. The data translation is very simply done by subtracting a constant number (such as 60 for example) from each value of the table. Then, the new table of data (Table 5.55) will be split up using the following algorithm:

1. The first variable factor (factor C) is taken from Table 5.54 by summing its values (all different levels) into only one which will give the new value of the process for the two other factors.
2. The obtained table will be noted with the interaction name of the non-rejected factors; so if C is rejected, the name of the partial table will be AB;
3. We repeat steps 1 and 2 for factor B, then we obtain the partial table AC. For factor A, table BC is produced.

Table 5.54 Friedel–Crafts reaction efficiency in an experimental plan with 3 factors and 4 levels.

	A_1			A_2		
	B_1	B_2	B_3	B_1	B_2	B_3
C_1	74.3	68.7	65.1	67.7	68.2	70.5
C_2	73.6	65.9	65.7	66.5	69.3	71.0
C_3	72.3	65.5	66.9	65.6	69.8	71.3
C_4	70.4	65.3	67.8	65.3	71.0	71.1

Table 5.55 Translation of data from Table 5.54.

	A_1			A_2		
	B_1	B_2	B_3	B_1	B_2	B_3
C_1	14.2	8.7	5.1	7.7	8.2	12.5
C_2	13.6	5.9	5.7	6.5	9.3	11.0
C_3	12.3	5.5	6.9	5.6	9.8	11.3
C_4	10.4	5.3	7.8	5.3	11.0	11.1

The computation for the division of Table 5.55 is:

- Elimination of factor C: partial table AB. For each point of the partial table AB (2*3 points), we compute the value of the response. Then, we have:
  $A_1B_1 = 14.3 + 13.6 + 12.3 + 10.4 = 50.6$, $A_1B_2 = 8.7 + 5.9 + 5.5 + 5.3 = 25.4$, $A_1B_3 = 5.1 + 5.7 + 6.9 + 7.8 = 25.5$, etc.;
- Elimination of factor B: partial table AC. In this case, with the same procedure used for partial table AB, we obtain:
  $A_1C_1 = 14.3 + 8.7 + 5.1 = 20.1$, $A_2C_1 = 7.7 + 8.2 + 10.5 = 26.4$, $A_1C_2 = 13.6 + 5.9 + 5.7 = 25.2$, etc.
- Elimination of factor A: partial table BC: as explained above,
  $B_1C_1 = 14.3 + 7.7 = 22.0$, $B_1C_2 = 13.6 + 6.5 = 20.1$, $B_1C_3 = 12.3 + 5.6 = 17.9$, etc.

The results of these calculation are summarized in Table 5.56, which is composed of three different partial tables: AB, AC, BC. This new set of data will be used for the final analysis of variances. For each partial table, the analysis of the variances of two factors will be carried out. Additionally, the values of the sums of the squares needed by the procedure of analyzing the variances (see Table 5.42) will

be computed. As far as each value in the partial tables is the result of an addition of many original data, all the sums of the squares for each of these tables, will be divided by the number of data used to produce the values. For example, partial table AB results from the elimination of factor C, because the C factor has four levels then all the sums of the squares associated to this table will be divided by four (number of factor levels). The addition that characterizes the interaction is obtained by the difference between the sum of the total squares and the sum of the squares containing the main effects.

Table 5.56 Division of Table 5.55 into three tables.

Two factors, table AB

	B_1	B_2	B_3	total
A_1	50.6	25.4	25.5	101.5
A_2	25.1	38.3	43.9	107.3
total	75.7	63.7	69.4	208.8

Two factors, table AC

	C_1	C_2	C_3	C_4	total
A_1	28.1	25.2	24.7	23.5	101.5
A_2	26.4	26.8	26.7	27.4	107.3
total	54.5	52.0	51.4	50.9	207.8

Two factors, table BC

	C_1	C_2	C_3	C_4	total
B_1	22.0	20.1	17.9	15.7	75.7
B_2	16.9	15.2	15.3	16.3	63.7
B_3	15.6	16.7	18.2	18.9	69.4
total	54.5	52.0	51.4	50.9	206.8

Now we can compute sums S_1, S_2, S_3, S_4 (see Table 5.42), which specifically concern partial table AB from Table 5.56. So we have: $S_{1(AB)} = (50.6^2 + 25.4^2 + ...+ 43.9^2)/(4*1) = 1971.6$, $S_{2(AB)} = (101.5^2 + 107.3^2)/(4*3) = 1817.96$, $S_{3(AB)} = (75.7^2 + 63.7^2 + 69.7^2)/(4*2) = 1826.07$, $S_{4(AB)} = (50.6 + 25.4 + ...+ 43.9)/(4*6) = 1816.47$.

Then we can calculate the following sums (see Table 5.42): $S_A = S_{2(AB)} - S_{4(AB)} = 1.49$, $S_B = S_{3(AB)} - S_{4(AB)} = 9.58$, $S_{T(AB)} = S_{1(AB)} - S_{4(AB)} = 155.20$ and so $S_{AB} = S_{T(AB)} - (S_A + S_B) = 144.13$.

By a similar procedure, we obtain the sums of squares S_1, S_2, S_3, S_4 when factor B (three levels) has been eliminated. These are: $S_{1(AC)} = (28.1^2 + 26.4^2 + ...+ 23.3^2 + 27.4^2)/(3*1) = 1821.94$, $S_{2(AC)} = S_{2(AB)} = (101.5^2 + 107.3^2)/(4*3) = 1817.96$, $S_{3(AC)} = (54.5^2 + 52.0^2 + 51.4^2 + 50.9^2)/(3*2) = 1817.83$, $S_{4(AC)} = S_{4(AB)} = (50.6 + 25.4 + ..+ 43.9)/(4*6) = 1816.47$, $S_C = S_{2(AC)} - S_{4(AC)} = 1.36$, $S_{T(AC)} = S_{1(AC)} - S_{4(AC)} = 5.47$. For the sum of squares that characterizes interaction AC, we have: $S_{AC} = S_{T(Ac)} - (S_A + S_C) = 2.62$. For the third partial table, the computations of these sums give: $S_{1(BC)} = (22.0^2 + 16.9^2 + ...+ 16.3^2 + 18.9^2)/(2*1) = 1841.82$, $S_{2(BC)} = (75.7^2 + 63.7^2 + 69.4^2)/(4*2) = 1826.07$, $S_{3(BC)} = S_{3(AC)} = (54.5^2 + 52.0^2 + 51.4^2 + 50.9^2)/(3*2) = 1817.83$, $S_{4(BC)} = S_{4(AC)} = S_{4(AB)} = 1816.47$. Whereas, for the sum of the squares that characterizes the BC interaction, we have: $S_{BC} = S_{T(BC)} - (S_B + S_C) = 14.41$.

For this application, the residual sum of squares is obtained by eliminating the sums of squares for A, B, C, AB, AC, BC from the total sum of the squares $S_T = S_1 - S_4$ where S_1 and S_4 are computed with the data from the original table (Table 5.55). Therefore, we obtain $S_1 = 14.3^2 + 8.7^2 + ...11.3^2 + 11.1^2 = 2000.6$, $S_4 = (14.3 + 8.7 + ..+ 11.2 + 11.1)^2/24 = 1816.47$, $S_T = 184.13$. Consequently, the residual sum of squares and their associated degrees of freedom will be: $S_{rz} = S_T - (S_A + S_B + S_C + S_{AB} + S_{AC} + S_{BC}) = 10.90$, $\nu = (2-1)(3-1)(4-1) = 6$.

Now we have all the necessary sums for the development of the analysis of variances. However, we first have to verify the following hypotheses:

- there is no significant difference between the variance due to the action of factor A and the residual variance

$$H_A : \sigma_A^2 = \sigma_{rz}^2 \Leftrightarrow s_A^2 = s_{rz}^2$$

- there is no significant difference between the variance due to the action of factor B and the residual variance

$$H_B : \sigma_B^2 = \sigma_{rz}^2 \Leftrightarrow s_B^2 = s_{rz}^2$$

- there is no significant difference between the variance due to the action of factor C and the residual variance

$$H_C : \sigma_C^2 = \sigma_{rz}^2 \Leftrightarrow s_C^2 = s_{rz}^2$$

- the interaction between factors A and B cannot lead to a new different statistical population

$$H_{AB} : \sigma_{AB}^2 = \sigma_{rz}^2 \Leftrightarrow s_{AB}^2 = s_{rz}^2$$

- the interaction between factors A and C cannot lead to a new different statistical population

$$H_{AC} : \sigma^2_{AC} = \sigma^2_{rz} \Leftrightarrow s^2_{AC} = s^2_{rz}$$

- the interaction of the factors B and C cannot lead to a new different statistical population

$$H_{BC} : \sigma^2_{BC} = \sigma^2_{rz} \Leftrightarrow s^2_{BC} = s^2_{rz}$$

Table 5.57 contains the synthesis of the analysis of variances for the problem of the Friedel–Crafts reaction. It is easy to observe that hypotheses H_A, H_B, H_C, H_{AC} and H_{BC} have been accepted. So, with respect to the specified state of the factors, the efficiency of the Friedel–Crafts reaction depends only on interaction AB (reaction time and timing (B)).

Table 5.57 Analysis of variances, example 5.6.3 (dependence of Friedel–Crafts reaction efficiency on temperature, reaction time and particular time of introduction of the catalyst).

Origin of variance	Sums for variance	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Change of factor A	$S_A = 1.49$	1	$s^2_A = S_A/1 = 1.49$	$F_A = s^2_A/s^2_{rz} = 1.49$	$F_{1,6,\alpha} = 5.99$	$F_A<F_{1,6,\alpha}$ Accept H_A
Change of factor B	$S_B = 9.58$	2	$s^2_B = S_B/2 = 4.49$	$F_B = s^2_B/s^2_{rz} = 2.63$	$F_{2,6,\alpha} = 5.14$	$F_B<F_{2,6,\alpha}$ Accept H_B
Change of factor C	$S_c = 1.36$	3	$s^2_c = S_c/3 = 0.45$	$F_c = s^2_B/s^2_{rz} = 0.25$	$F_{3,6,\alpha} = 4.76$	$F_C<F_{3,6,\alpha}$ Accep H_c
Interaction A B	$S_{AB} = 144.13$	2	$s^2_{AB} = S_{AB}/2 = 72.06$	$F_{AB} = s^2_{AB}/s^2_{rz} = 39.56$	$F_{2,6,\alpha} = 4.76$	$F_{AB}>F_{2,6,\alpha}$ Refuse H_{AB}
Interaction A C	$S_{AC} = 2.62$	3	$s^2_{AC} = S_{AC}/3 = 0.87$	$F_{AC} = s^2_{AC}/s^2_{rz} = 0.48$	$F_{3,6,\alpha} = 4.76$	$F_{AC}<F_{3,6,\alpha}$ Accept H_{AC}
Interaction B C	$S_{BC} = 14.41$	6	$s^2_{BC} = S_{BC}/6 = 2.41$	$F_{BC} = s^2_{BC}/s^2_{rz} = 1.32$	$F_{6,6,\alpha} = 4.28$	$F_{BC}<F_{6,6,\alpha}$ Accept H_{BC}
Residual	$S_r = 10.90$	6	$s^2_{rz} = S_{rz}/6 = 1.82$			

The interval of confidence of the variation of the reaction efficiency can be calculated using Table 5.56. Then, considering the AB interaction, we can compute the

mean efficiency of the reaction for positions A_1B_1, A_1B_2, A_1B_3; we consequently have $m_{A_1B_1} = 50.5/4 = 12.65$, $m_{A_1B_2} = 25.4/4 = 6.35$, $m_{A_1B_3} 25.5/4 = 6.36$ and therefore the variations associated to the efficiency of the reaction are: $d_{12} = m_{A_1B_1} - m_{A_1B_2} = 6.3$, $d_{13} = m_{A_1B_1} - m_{A_1B_3} = 6.29$. The mean value of the variation of the reaction efficiency will be $d_B = (d_{12} + d_{13})/2 = 6.295$ and we compute the confidence interval for this mean value. The calculation of the theoretical value of the Student variable for $\alpha = 0.05$ and $\upsilon = 6$(this is the number of degrees of freedom associated to the residual variance) is $1 - \alpha = \int_o^t f_\vartheta(s)ds$ and $t = t_{6,0.05} = 2.47$ and therefore now, using relation (5.161), we can compute the confidence interval for this variation of reaction efficiency. The result is I_B = (6.295 – 2.47*(1.82)$^{0.5}$/(2*3), 6.295 + 2.47*(1.82)$^{0.5}$/(2*3)) = (4.95, 7.65). In other words, we can say that, for reaction time A_1 = 10 h, if we change the timing of introduction of the catalyst from B_1 = 2 h to B_2 = 4 h, then we obtain a variation of the reaction efficiency between 4.95 and 7.65%. The case when factor A has level A_2 and factor B changes between B_1 and B_3 can be approached by the same procedure. The final conclusion of this analysis shows that the level A_1 for factor A, the level B_1 for factor B and any level for factor C are enough to ensure the conditions for the most favourable reaction efficiency.

5.6.3.3 **Analysis of the Effects of Systematic Influences**

The external systematic influence is common in experimental research when the quality of the raw materials and of the chemicals undergo minor changes and/or when the first data were obtained in one experimental unit and the remaining measurements were carried out in a similar but not identical apparatus.

In these situations, we cannot start the analysis of data without separating the effect of the external systematic influence from the unprocessed new data. In other words, we must separate the variations due to the actions of some factors with systematic influences from the original data. For this purpose, the methods of Latin squares and of effects of unification of factors have been developed in the plan of experiments.

In the method of Latin squares, the experimental plan, given by the matrix of experiments, is a square table in which the first line contains the different levels of the first factor of the process whereas the levels for the second factor are given in the first column. The rest of the table contains capital letters from the Latin alphabet, which represent the order in which the experiments are carried out (example: for pressure level P_1, four experiments for the temperature levels T_1, T_2, T_3, T_4 occur in the following sequence: A, B, C, A where A has been established as the first experiment, B as the second experiment, etc). The suffixes of these Latin capital letters introduce the different levels of the factors. Table 5.58 presents the schema of a plan of Latin squares. We can complete the description of this plan showing that the values of the process response can be written in each letter box once the experiment has been carried out. Indeed, we utilize three indexes for the theoretical utterance of a numerical value of the process response (v). For exam-

ple, for $v_{ij}^{(A)}$, the i index shows that the level of the P factor is P_i, the j index gives level T_j for factor T and the final superscript A shows that the progression of the experiments must be A. Considering Table 5.58, it is important to observe that the experiments complete each box placed in an intersection between a line and a column with only a single value.

Table 5.58 Data for the Latin squares method for a process with three factors.

First factor T (temperature)	**T_1**	**T_3**	**T_3**	**T_4**
Second factor P (pressure)				
P_1	A	B	C	D
P_2	B	C	D	A
P_3	C	D	A	B
P_4	D	A	B	C

The correct use of the Latin squares method imposes a completely random order of execution of the experiments. As far as the experiment required in the box table is randomly chosen and as a single value of the process response is introduced into the box, we guarantee the random spreading of the effect produced by the factor which presents a systematic influence.

Once the levels of the factors have been selected, we can begin to write the plan introducing the order of the experiments by using: (i) the random changes between lines or between columns; (ii) the line variations using a random number generator; (iii) the extraction from a black box. Using one of these procedures to select the order of the experiments allows one to respect the conditions imposed by the random spreading of the effect produced by the factors.

The variance analysis for a plan with Latin squares is not different from the general case previously discussed in Section 5.6.3. Therefore we must compute the following sums:

S_1 – sum of the squares of all individual observations;
S_2 – sum of the squares of the sums of the columns divided by the number of observations in a column;
S_3 – sum of the squares of the sums of the lines divided by the number of observations in a line;
S_4 – sum of the squares of the sums of observations with the same Latin letter divided by the number of the observations having the same letter;
S_5 – the square of the sum of all observations divided by the number of observations.

Indeed, these sums allow the calculation of the variances due to each of the factors introducing the columns, the lines and the letter in the plan. We can introduce the statistical hypotheses about the effect of the factors on the process response using the variances of the factors with respect to the residual variance. This residual variance is computed by $s_{rz}^2 = (S_1 + S_5 - S_2 - S_3)/[(n-1)(n-1)]$ where n is the box number in a line (or a column). We can then verify the following hypotheses:

- The effect on the process response of the factor which changes the columns of the plan is not important. Mathematically we can write:

  $H_C : \sigma_C^2 = \sigma_{rz}^2 \Leftrightarrow s_c^2 = s_{rz}^2$;

- The effect on the process response of the factor which changes the lines in a plan is not important. Therefore, we can write:

  $H_L : \sigma_L^2 = \sigma_{rz}^2 \Leftrightarrow s_L^2 = s_{rz}^2$;

- The factor which changes the letter in the plan does not have a considerable influence on the process response. Then, according to the updated cases, we can write:

  $H_A : \sigma_A^2 = \sigma_{rz}^2 \Leftrightarrow s_A^2 = s_{rz}^2$.

Now every reader knows that to check a hypothesis in which we compare two variances, we have to use the Fischer test. Here the computed value of a Fischer random variable is compared with its theoretical value particularized by the concrete degrees of freedom (ν_1, ν_2) and the confidence level $1 - \alpha$. Table 5.59 presents the synthesis of the analysis of variances for this case of the Latin squares method.

The following example will illustrate this method. The reaction considered is the chlorination of an organic liquid in a small laboratory scale reactor which works under agitation and at a constant chlorine pressure; the temperature of the reactor is controlled by a liquid circulating in a double-shell. The analysis of the reactor product shows the presence of some undesired components. The concentrations of the desired product and by-products are determined by the temperature, the chlorination degree (more precisely the reaction time) and by the catalyst concentration. Various procedures can be used for the addition of the catalyst: the whole catalyst is poured in one go, by fractions diluted with reactants, etc; the objective is to obtain a catalyst concentration between 0.1 and 0.3% g/g. The addition of the catalyst can be considered as an example of systematic influence and then its effect on the concentration of the by-products can be analyzed by the Latin squares method. Five levels are selected for the temperature and for the chlorination degree, which are considered as factors which do not have a systematic influence. The catalyst addition procedure and its concentration with respect to the reaction mixture can be introduced as a process factor with systematic influence by a group of five letters: A, B, C, D, E. Table 5.60 gives the factorial program obtained after the experiments have been extracted from a black box. This table contains all the measured concentrations of undesired products after each experiment.

Table 5.59 Analysis of the variances for the case of Latin squares method.

Origin of the variance	Sums of the squares	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Effect of a factor that changes the columns	$S_2 - S_5 = S_C$	$n - 1$	$s^2_C = S_C/(n-1)$	$F_C = s^2_C/s^2_{rz}$	$F_{n-1,(n-1)(n-2),\alpha}$	$F_C < F_{n-1,(n-1)(n-2),\alpha}$ H_C accepted
Effect of a factor that changes the lines	$S_3 - S_5 = S_L$	$n - 1$	$s^2_L = S_L/(n-1)$	$F_L = s^2_L/s^2_{rz}$	$F_{n-1,(n-1)(n-2),\alpha}$	$F_L < F_{n-1,(n-1)(n-2),\alpha}$ H_L accepted
Effect of a factor that changes the letter	$S_4 - S_5 = S_A$	$n - 1$	$s^2_A = S_A/(n-1)$	$F_A = s^2_A/s^2_{rz}$	$F_{n-1,(n-1)(n-2),\alpha}$	$F_A < F_{n-1,(n-1)(n-2),\alpha}$ H_c accepted
Residual	$S_r = S_1 + S_5 - (S_2 + S_4)$	$(n-1)*(n-2)$	$s^2_{rz} = S_r/[(n-1)(n-2)]$		Without the power enabling one to identify interaction effects	
Total	$S_1 - S_5$	$n^2 - 1$				

Table 5.60 Factorial plan for the Latin squares method – case of chlorination of an organic liquid.

Temperature	90	70	50	60	80	Total L	Total letter
Chlorination degree							
40	B r	A	D	C	E		
	39.5	28.9	11.6	13.9	22.6	116.5	A = 89.4
35	E	C	B r	D	A		
	32.2	25.5	10.0	12.5	14.6	94.8	B = 129.7
45	D	E	C	A	B r		
	57.6	41.0	12.1	14.7	25.9	151.3	C = 136.8
30	A	D	E	B	C		
	19.6	21.2	10.3	9.8	12.1	72.9	D = 138.3
50	C	B r	A	E r	D		
	73.2	44.5	11.7	19.7	35.4	184.5	E = 125.8
Total C	222.0	161.1	55.7	70.6	110.6	620	

The data in the table above will first be used to determine whether the addition procedure as well as the major factors of the process influence the dependent process variable (concentration of the undesired components in the reaction product). With the purpose being to obtain the real residual variance from the experiments, Table 5.60 (the boxes of which contain an r), have been repeated. These results are shown in Table 5.61.

Table 5.61 Results of repeated experiments, example 5.6.3.

Position (T,D)	(90,40)	(70,50)	(50,35)	(60,50)	(80,45)
Old value	39.5	44.5	10.0	19.7	25.9
New value	36.9	42.4	11.3	20.5	24.1

As previously explained, the first step to solve this application is the computation of the sums required by Table 5.59. Then we obtain:

S_1 – the sum of the squares of the individual observations:
$S_1 = 39.5^2 + 28.9^2 + \ldots\ldots + 19.7^2 + 35.4^2 = 21717.42$;

S_2 – the sum of the squares of the sums of the columns divided by the number of observations of the column:
$S_2 = (222^2 + 161^2 + 55.7^2 + 70.6^2 + 110.6^2)/5 = 19104.84$;

S_3 – the sum of the squares of the sums of the lines divided by the number of observations of the line:
$S_3 = (116.5^2 + 94.8^2 + 151.3^2 + 72.9^2 + 184.5^2)/5 = 16961.13$;

S_4 – the sum of the squares of sums with the same letter divided by the number of observations which have the same letter:
$S_4 = (89.4^2 + 129.7^2 + 156.8^2 + 138.3^2 + 125.8^2)/5$
$= 15696.28$

S_5 – the square of the sum of all observations divided by the total number of observations:
$S_5 = (39.5 + 28.9 + \ldots + 19.7 + 35.4)^2/25 = 620.0^2/25$
$= 15376$.

The results of the analysis using the data of Table 5.59 are given in Table 5.62. Considering the decision column, we conclude that two zero hypotheses have been refused and one has been accepted.

Table 5.62 Analysis of the variances for the Latin squares method, example 5.6.3.

Origin of the variance	Sums of squares	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of Fischer variable	Decision
Effect of factor that change the columns	$S_2 - S_5 = S_C$ $S_C = 3728.4$	$n - 1$ $n - 1 = 4$	$s^2_C = S_C/(n-1)$ $s^2_C = 933.82$	$F_C = s^2_C/s^2_{rz}$ $F_C = 15.82$	$F_{n-1,(n-1)(n-2),\alpha}$ $F_{4,12,0.05} = 3.26$	$F_C >$ $F_{n-1,(n-1)(n-2),\alpha}$ H_C rejected
Effect of a factor that changes the lines	$S_3 - S_5 = S_L$ $S_L = 1585.3$	$n - 1$ $n - 1 = 4$	$s^2_L = S_L/(n-1)$ $s^2_L = 396.28$	$F_L = s^2_L/s^2_{rz}$ $F_L = 6.71$	$F_{n-1,(n-1)(n-2),\alpha}$ $F_{4,12,0.05} = 3.26$	$F_L >$ $F_{n-1,(n-1)(n-2),\alpha}$ H_L rejected
Effect of a factor that changes the letter	$S_4 - S_5 = S_A$ $S_A = 320.24$	$n - 1$ $n - 1 = 4$	$s^2_A = S_A/(n-1)$ $s^2_A = 80.06$	$F_A = s^2_A/s^2_{rz}$ $F_A = 1.35$	$F_{n-1,(n-1)(n-2),\alpha}$ $F_{4,12,0.05} = 3.26$	$F_A <$ $F_{n-1,(n-1)(n-2),\alpha}$ H_c accepted
Residual	$S_{rz} = S_1 + S_5 - (S_2 + S_4)$ $S_r = 707.20$	$(n-1)*(n-2) = 12$	$s^2_{rz} = S_{rz}/[(n-1)(n-2)] = 58.91$		It is not possible to identify the effects of double interactions.	
Total	$S_1 - S_5$	$n^2 - 1 = 24$				

It is important to note that the effect of the factor that changes the letter in the Latin squares table is negligible. Then, for the investigated chlorination reaction both the concentration of the catalyst (between 0.1 and 0.3% g/g) and its process of addition do not have any effect on the concentration of the by-products. Nevertheless, this conclusion cannot be definitive because we can find from Table 5.62 that we have a high residual variance. In this case, we can suggest that the interaction effects are certainly included in the residual variance.

The real residual variance frequently named "reproducibility variance" can be determined by repeating all the experiments but this can turn out to be quite expensive. The Latin squares method offers the advantage of accepting the repetition of a small number of experiments with the condition to use a totally random procedure for the selection of the experiments. With the data from Table 5.61 and using the relation $s^2_{rz} = (\sum_{i=1}^{n_C} \sum_{j=1}^{n_L - 1} d^2_{ij})/[(n_C(n_L - 1))]$, where d_{ij} are the differences between the observed values for all the n_c columns and n_L lines (where the new experiments can be found), we obtain: $s^2_{rz} = (2.6^2 + 1.3^2 + 1.8^2 + 2.1^2 + 0.8^2)/(5 * 1) = 1.56$. Five degrees of freedom characterize this new computed variance.

Now, it is clear that the residual variance from Table 5.62 contains one or more interaction effects. Moreover, for this application or, more precisely, for the data given for the particularization of the Latin squares method, a partial response has

been obtained. Consequently, a new research plan must be suggested in order to answer our problem.

The method of the effects of the unification of factors considers that, for a fixed plan of experiments, we can produce different groups where each contains experiments presenting the same systematic influence [5.8, 5.13, 5.23, 5.35, 5.36]. To introduce this method, we can consider the case of a process with three factors analyzed with a CFE 2^3 plan of experiments. In our example, we will take into account the systematic influence of a new factor D. To begin this analysis, we will use the initial plan with eight experiments with the condition to separate these experiments into two blocks or groups:

- the first block is bound with the first level of the factor of systematic influence and the second block corresponds to the next level of the factor of systematic influence;
- we accept both blocks to be related by a triple interaction variance (s^2_{ABC}).

For this case of separation into two groups or blocks, it is important to determine the experiments from the 2^3 plan which are contained in block D_1 and those contained in D_2.

Table 5.63 shows the detailed separation of the experiments into groups. Each experiment corresponding to a different block is identified by a current name and by a code. The experiments with the sign + in the ABC column correspond to the block D_1, the remaining experiments to block D_2.

Table 5.63 The division of a CFE 2^3 plan into two blocks.

i	A	B	C	AB	AC	BC	ABC	y_i	
1	–1	–1	–1	+1	+1	+1	–1	$y_1 = (1)$	Block D_1
2	+1	–1	–1	–1	–1	+1	+1	$y_2 = a$	Experiments: 2, 3, 5, 8
3	–1	+1	–1	–1	+1	–1	+1	$y_3 = b$	Codified names: a, b, c, abc
4	+1	+1	–1	+1	–1	–1	–1	$y_4 = ab$	
5	–1	–1	+1	+1	–1	–1	+1	$y_5 = c$	Block D_2
6	+1	–1	+1	–1	+1	–1	–1	$y_6 = ac$	Experiments: 1, 4, 6, 7
7	–1	+1	+1	–1	–1	+1	–1	$y_7 = bc$	Codified names: (1), ab, ac, bc
8	+1	+1	+1	+1	+1	+1	+1	$y_8 = abc$	

When we have the possibility to obtain the real residual variances (2–3 experiments repeated in the D_1 and D_2 blocks), we can suggest to validate the following hypothesis: $H_{ABC} : \sigma^2_{ABC} = \sigma^2_{rz} \Leftrightarrow s^2_{ABC} = s^2_{rz}$ and if it is rejected, we can conclude

an important or crucial effect on the process response of the factor which shows a systematic influence.

The justification for our consideration showing that the action of a factor with systematic influence is concentrated in the relation which binds the blocks (frequently named contrast) is sustained by the following observations:

- if we accept that block D_1 increases the process response, then, with respect to the D_2 block, the results will be:
  (a + d), (b + d),
  (c + d), (abc + d);
- with Eq. (5.168) we obtain
  $EA = -(1) + (a + d) - (b + d) + ab - (c + d) + ac -$
  $bc + (abc + d) = -(1) + a - b + ab - c + ac - bc + abc$. Similar expressions are thus obtained for the effects EB, EC, EAB, EAC, EBC; these effects are not affected by the increase of the response in block D_1.
- for the contrast we obtain EABC = –(1) + (a + d) + (b + d) – ab + (c + d) – ac – bc + (abc + d) = –(1) + a + b – ab + c – ac – bc + abc + 4d; this result shows a displacement with 4d; so the variance due to this interaction is the only variance obtained when we utilize a two block division for a CFE 2^3 plan.

A division into four blocks made from two unification relations, is also possible with a CFE 2^3 plan where the systematic influence of one or more factors is considered. If interactions AB and AC give the unification relations, then, by using the block division procedure used above (Table 5.63), the following blocks will be obtained:

Block 1 or block + +: experiments 1 and 8 with code names (1) and abc;
Block 2 or block – –: experiments 2 and 7 with code names a and bc;
Block 3 or block – +: experiments 3 and 6 with code names b and ac;
Block 4 or block + –: experiments 4 and 5 with code names ab and c.

In this division example, if interactions AB and AC influence the process response, we can conclude that the displacement of the process response contains the effect of a systematic influence.

The examples where a CFE 2^3 plan has been divided into two or four blocks are not explicit enough to develop the idea that the relations of the unification of blocks are selected randomly. In the next example, a CFE 2^4 plan is developed with the purpose being to show the procedures to select the unification relations of inter-blocks. In this plan, the actions showing a systematic influence will be divided into two blocks or into four blocks with, respectively, eight experiments or four experiments per block. We start this new analysis by building the CFE 2^4 plan. Table 5.64 contains this CFE 2^4 plan and also gives the division of the two blocks when we use the ABCD interaction as a unification relation.

Table 5.64 The separation of a CFE 2^4 into two blocks.

i	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD	y_i
1	–1	–1	–1	–1	+1	+1	+1	+1	+1	+1	–1	–1	–1	–1	+1	$y_1 = (1)$
2	+1	–1	–1	–1	–1	–1	–1	+1	+1	+1	+1	+1	+1	–1	–1	$y_2 = a$
3	–1	+1	–1	–1	–1	+1	+1	–1	–1	+1	+1	+1	–1	+1	–1	$y_3 = b$
4	+1	+1	–1	–1	+1	–1	–1	–1	–1	+1	–1	–1	+1	+1	+1	$y_4 = ab$
5	–1	–1	+1	–1	+1	–1	+1	–1	+1	–1	+1	–1	+1	+1	–1	$y_5 = c$
6	+1	–1	+1	–1	–1	+1	–1	–1	+1	–1	–1	+1	–1	+1	+1	$y_6 = ac$
7	–1	+1	+1	–1	–1	–1	–1	+1	–1	–1	–1	+1	+1	–1	+1	$y_7 = bc$
8	+1	+1	+1	–1	+1	+1	+1	+1	–1	–1	+1	–1	–1	–1	–1	$y_8 = abc$
9	–1	–1	–1	+1	+1	+1	–1	+1	–1	–1	–	+1	+1	+1	–1	$y_9 = d$
10	+1	–1	–1	+1	–1	–1	+1	+1	–1	–1	+1	–1	–1	+1	+1	$y_{10} = ad$
11	–1	+1	–1	+1	–1	+1	–1	–1	+1	–1	+1	–1	+1	–1	+1	$y_{11} = bd$
12	+1	+1	–1	+1	+1	–1	+1	–1	+1	–1	–1	+1	–1	–1	–1	$y_{12} = abd$
13	–1	–1	+1	+1	+1	–1	–1	–1	–1	+1	+1	+1	–1	–1	+1	$y_{13} = cd$
14	+1	–1	+1	+1	–1	+1	+1	–1	–1	+1	–1	–1	+1	–1	–1	$y_{14} = acd$
15	–1	+1	+1	+1	–1	–1	–1	+1	+1	+1	–1	–1	–1	+1	–1	$y_{15} = bcd$
16	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	$y_{16} = abcd$
Block E_1 or Block E_+: Experiences: 1, 4, 6, 7, 10, 11, 13, 16 Codes: (1), ab, ac, bc, ad, bd, cd, abcd								Block E_2 or Block E_-: Experiences: 2, 3, 5, 8, 9, 12, 14, 15 Codes: a, b, c, abc, d, abd, acd, bcd							ABCD unification	

If we now suppose that the aim is to divide the CFE plan 2^4 into four blocks, we can select one of the following unification relations: (i) ABCD coupled with one from the three order interactions (ABC, ACD, BCD, etc.); (ii) ABCD coupled with one from the two-order interactions (AB, AC, AD, etc.); (iii) two interactions of three-order, etc. To establish which coupling is the most favourable, it is necessary to know what type of information disappears in each case. For this purpose we show here some of the multiplications of the ABCD interaction relations with their possible coupling interaction relations where $A^2 = B^2 = C^2 = D^2 = 1$.

$$\begin{aligned} &\text{I. } ABCD * BCD = A \quad \text{II. } ABCD * ABC = D \quad \text{III. } ABCD * ACD = B \\ &\text{IV. } ABCD * AB = CD \quad \text{V. } ABCD * AC = BD \quad \text{VI. } ABCD * AD = BC.... \end{aligned} \tag{5.170}$$

From this result we can then conclude that: (i) for a *four–three* coupling in the data processing, the information about the effect of the direct factor (A, B, C, D) action on the process response disappears; it is obvious that, for actual cases, it is difficult to accept this situation; (ii) when a *four–two* coupling occurs, the information that shows the effect of one interaction of order two disappears, however, this situation can sometimes be accepted in actual cases; (iii) it is not difficult to show that for a *three–three* coupling we obtain the case of the *four–two* coupling.

The division of the CFE 2^4 plan into four blocks by means of the *four–two* couple is useful to identify the weakest order two interactions that can be used with the order four interactions as unification relations. At the same time, we can also analyze the *three–three* couple obtained with the most non-important order three interactions. In fact, it is easy to accept that, for an investigated process, the effects on the process response of the order three interactions are non-important for most actual situations. Indeed, when for a CFE 2^4 plan, the ABC and BCD interactions are the weakest, these interactions can be selected as relations for the unification of the inter-blocks. Then, we can rapidly produce the division into four blocks: $E_1 = E_{--}, E_2 = E_{-+}, E_3 = E_{+-}, E_4 = E_{++}$. Table 5.65 shows the blocks and the corresponding experiments with their usual numbers and codes.

Table 5.65 The blocks repartition of a CFE 2^4 plan using the contrasts ABC, BCD.

ABC	**–1**	**+1**	**+1**	**–1**	**+1**	**–1**	**–1**	**+1**	**–1**	**+1**	**+1**	**–1**	**+1**	**–1**	**–1**	**+1**
BCD	**–1**	**–1**	**+1**	**+1**	**+1**	**+1**	**–1**	**–1**	**+1**	**+1**	**–1**	**–1**	**–1**	**–1**	**+1**	**+1**
y_i	(1)	a	b	ab	c	ac	bc	abc	d	ad	bd	abd	cd	acd	bcd	abcd

Block E1 = E–		Block E2 = E–+		Block E3 = E+–		Block E1 = E+	
(1)	number 1	ab	number 4	a	number 2	b	number 2
bc	number 7	ac	number 6	abc	number 8	c	number 5
abd	number 12	d	number 9	bd	number 11	ad	number 10
acd	number 14	bcd	number 15	cd	number 13	abcd	number 16

For the cases of $2^5$and 2^6 CFE plans, the division into blocks must respect the principles previously shown for a 2^4 plan. Considering a 2^5 plan, the recommended contrast couplings are of the *three–three–four* type. If the coupling chain is ABC–ADE–BCDE, then the main block F_1 (analogue to E_1 for the case of a CFE 2^4) will contain experiments (1)–bc–de–abd–acd–abc–ace–bcde.

We establish the repartition of the experiments for the remaining blocks by multiplying the F_1 chain by a, b and c; for these products we have $a^2 = b^2 = c^2 = d^2 = 1$. So, the F_2 will contain the next chain of experiments: a–abc–ade–bd–cd–bc–ce–abde. The following application presents an actual case for a CFE 2^4 plan where the separation has been obtained according to the contrasts ABC and BCD.

Numerical application. This application concerns the conversion of one reactant by an esterification reaction occurring in a discontinuous and stirred reactor. It is a function of the temperature (factor A), the alcohol–acid molar ratio (factor B), the reaction time (factor C) and the catalyst concentration (factor D). A CFE 2^4 plan is used to investigate the different effects of the factors. The levels of the factors have been established in order to obtain a good reactant conversion. These levels are: temperatures: $A_1 = 110\,°C$, $A_2 = 130\,°C$; alcohol–acid molar ratio: $B_1 = 1.2$, $B_2 = 1.5$; reaction time: $C_1 = 3$ h, $C_2 = 4$ h; catalyst concentration: $D_1 = 1\%$ g/g, $D_2 = 2\%$ g/g.

Three different qualities of alcohol have been used: recycled, distilled or rectified. It is easy to observe that the quality of the alcohol introduces a systematic influence towards factor B in the esterification reaction. Indeed, the development of the experimental research is made with a 2^4 plan with four blocks. The ABC and BCD have the contrasts considered for the blocks division. For the experiments grouped in block E_1, the first type of alcohol has been used. Distilled alcohol is the reactant used in the experiments of the second block (E_2) and the rectified alcohol for the experiments of the last two blocks (E_3, E_4). Table 5.66 presents the initial data where the division of the blocks is not visible.

Table 5.66 The conversion for an esterification reaction in a CFE 2^4 plan.

Esterification reaction 4 blocks	**D_1**				**D_2**			
	C_1		**C_2**		**C_1**		**C_2**	
	B_1	**B_2**	**B_1**	**B_2**	**B_1**	**B_2**	**B_1**	**B_2**
A_1	(1)	b	c	bc	d	bd	cd	bcd
	28	31	26	32	33	33	36	38
A_2	a	ab	ac	abc	ad	abd	acd	abcd
	20	24	20	30	24	24	37	31

Table 5.67 shows the conversions characterizing each block and the corresponding columns of the sums, these data are necessary to compute the variance due to the division into blocks. Indeed, these sums will be used for the computation of the square sums showing the differences in the reaction conversion produced by the alcohol quality (S_M).

Table 5.67 Presentation of the blocks in the example of an esterification reaction.

Block E_1		Block E_2		Block E_3		Block E_4	
Experiment code	Conversion	Experiment code	Conversion	Experiment code	Conversion	Experiment code	Conversion
(1)	28	a	20	b	31	d	33
abd	24	bd	33	ad	24	ab	24
acd	37	cd	36	abcd	31	ac	20
bc	32	abc	30	c	26	bcd	38
Total	121	Total	119	Total	112	Total	115

The computation for the analysis of the variances is carried out following the procedure described in Section 5.6.3.1. When we begin to complete Table 5.52 as recommended by this procedure, we can observe that we must add the effect of the D factor as well as its interactions. Nevertheless, in this table, we cannot add the unification of the interactions accepted by the data provided by the division into blocks. In addition to the data described here, we have to realize the following computations in order to complete Table 5.52:

- the sum of the squares of the sums of the conversion obtained for each block divided by the number of blocks:

  $S_1 = (121^2 + 119^2 + 112^2 + 117^2)/4 = 13642.75;$

- the sum of the squares of each conversion divided by the total number of the experiments:

  $S_2 = (28^2 + 31^2 + 26^2 + \ldots\ldots + 24^2 + 37^2 + 31^2)/16 = 13630.56;$

- the sum of the squares showing the differences due to the alcohol quality:

  $S_M = S_1 - S_2 = 12.19;$

- the sum of squares due to the unification of the interactions by using the following algorithm:
  (a) we compute EABC and EBCD using the general procedure particularized to the data in Table 5.64:
  EABC = –(1) + a + b – ab + c – ac – bc + abc – d + ad + bd – abd + cd – acd – bcd + abcd = –28 + 20 + 31 – 24 + 26 – 20 – 32 + 30 – 33 + 24 + 33 – 24 + 36 – 30 – 38 + 37 = 0;

EBCD = – (1) – a + b + ab + c + ac – bc – abc + d + ad – bd – abd – cd – acd + bcd + abcd = –28 – 20 + 31 + 24 + 26 + 20 – 32 –30 + 33 + 24 – 33 – 24 – 36 – 37 + 38 + 31 = –13;

(b) we calculate S_{ABC} and S_{BCD} using the values of the effects EABC and EBCD:
$S_{ABC} = (EABC)^2/16 = 0$, $S_{BCD} = (EBCD)^2/16 = 13^2/16 = 10.65$;

(c) we finish the algorithm by computing the squares sum of S_{ABC} and S_{BCD}: $S_{INT} = S_{ABC} + S_{BCD} = 10.65$.

At this point, we have to verify the correctness of the selection of the unification relations. When $S_M \cong S_{INT}$ we can conclude that our selection for the unification relations is good; in this case, we can also note that the calculations have been made without errors. Otherwise, if computation errors have not been detected, we have to observe that the selected interactions for the unification of blocks are strong and then they cannot be used as unification interactions. In this case, we have to carry out a new experimental research with a new plan. However, part of the experiments realized in the previous plan can be recuperated. Table 5.68 contains the synthesis of the analysis of the variances for the current example of an esterification reaction. We observe that, for the evolution of the factors, the molar ratio of reactants (B) prevails, whereas all other interactions, except interaction AC (temperature–reaction time), do not have an important influence on the process response (on the reaction conversion). This statement is sustained by all zero hypotheses accepted and reported in Table 5.68. It should be mentioned that the alcohol quality does not have a systematic influence on the esterification reaction efficiency. Indeed, the reaction can be carried out with the cheapest alcohol. As a conclusion, the analysis of the variances has shown that conversion enhancement can be obtained by increasing the temperature, reaction time and, catalyst concentration, independently or simultaneously.

Table 5.68 Synthesis of the variance analysis for CFE 2^4, example of an esterification reaction.

Origin of the variance	Sums of the differences	Degrees of freedom	Variances	Computed value of the Fischer variable	Theoretical value of the Fischer variable	Decision
Temperature variation A	$S_A = 138.06$	1	$s^2_A = 138.06$	$F_A = s^2_A/s^2_{rz} = 20.38$	$F_{1,6,\alpha} = 5.99$	$F_A > F_{1,6,\alpha}$ H_A refused
Molar ratio variation B	$S_B = 22.56$	1	$s^2_B = 22.05$	$F_B = s^2_B/s^2_{rz} = 2.63$	$F_{1,6,\alpha} = 5.99$	$F_B < F_{1,6,\alpha}$ H_B accepted
Reaction time variation C	$S_C = 115.56$	1	$s^2_C = 115.56$	$F_c = s^2_B/s^2_{rz} = 17.36$	$F_{1,6,\alpha} = 5.99$	$F_C > F_{1,6,\alpha}$ H_C refused
Catalyst conc. variation D	$S_D = 76.56$		$s^2_D = 6.56$	$F_D = s^2_D/s^2_{rz} = 11.29$	$F_{1,6,\alpha} = 5.99$	$F_D > F_{1,6,\alpha}$ H_C refused
Alcohol type M	$S_M = 12.19$	3	$s^2_M = 4.06$	$F_M = s^2_M/s^2_{rz} = 0.59$	$F_{3,6,\alpha} = 4.76$	$F_M < F_{1,6,\alpha}$ H_M accepted
Interaction AC	$S_{AC} = 52.56$	1	$s^2_{AC} = 52.56$	$F_{AC} = s^2_{AB}/s^2_{rz} = 7.62$	$F_{1,6,\alpha} = 5.99$	$F_{AC} > F_{1,6,\alpha}$ H_{AB} refused
Interaction AB	$S_{AB} = 10.56$	1	$s^2_{AB} = 10.56$	$F_{AB} = s^2_{AB}/s^2_{rz} = 1.55$	$F_{1,6,\alpha} = 5.99$	$F_{AB} < F_{1,6,\alpha}$ H_{AB} accepted
Interaction BC	$S_{BC} = 1.56$	1	$s^2_{BC} = 1.56$	$F_{BC} = s^2_{AB}/s^2_{rz} = 0.23$	$F_{1,6,\alpha} = 5.99$	$F_{BC} < F_{1,6,\alpha}$ H_{BC} accepted
Interaction BD	$S_{BD} = 18.06$	1	$s^2_{BD} = 18.06$	$F_{BD} = s^2_{AB}/s^2_{rz} = 2.66$	$F_{1,6,\alpha} = 5.99$	$F_{BD} < F_{1,6,\alpha}$ H_{BD} accepted
Interaction CD	$S_{CD} = 10.56$	1	$s^2_{CD} = 10.56$	$F_{CD} = s^2_{AB}/s^2_{rz} = 1.55$	$F_{1,6,\alpha} = 5.99$	$F_{CD} < F_{1,6,\alpha}$ H_{CD} accepted
Interaction ABD	$S_{ABD} = 0.56$	1	$s^2_{ABD} = 0.56$	$F_{ABD} = s^2_{AB}/s^2_{rz} = 0.07$	$F_{1,6,\alpha} = 5.99$	$F_{ABD} < F_{1,6,\alpha}$ H_{ABD} accepted
Interaction ACD	$S_{ACD} = 6.00$	1	$s^2_{ACD} = 6.00$	$F_{ACD} = s^2_{AC}/s^2_{rz} = 0.88$	$F_{1,6,\alpha} = 5.99$	$F_{ACD} < F_{1,6,\alpha}$ H_{ACD} accepted
Interaction ABCD	$S_{ABCD} = 0.06$	1	$s^2_{ABCD} = 0.06$	$F_{ABCD} = s^2_{BC}/s^2_{rz} = 0.01$	$F_{1,6,\alpha} = 5.99$	$F_{ACBD} < F_{1,6,\alpha}$ H_C accepted
Residual	$S_r = 47.36$	6	$s^2_{rz} = 6.77$	All interactions without AC		

5.7
Use of Neural Net Computing Statistical Modelling

At the beginning of this chapter, we introduced statistical models based on the general principle of the Taylor function decomposition, which can be recognized as non-parametric kinetic model. Indeed, this approximation is acceptable because the parameters of the statistical models do not generally have a direct contact with the reality of a physical process. Consequently, statistical models must be included in the general class of connectionist models (models which directly connect the dependent and independent process variables based only on their numerical values). In this section we will discuss the necessary methodologies to obtain the same type of model but using artificial neural networks (ANN). This type of connectionist model has been inspired by the structure and function of animals' natural neural networks.

Neural nets are computing programs that behave externally as multi-input multi-output computing blocks. Although artificial neural networks were initially devised for parallel processing, they are being used on sequential machines (von Neumann) as well.

They have been used successfully in several diverse engineering fields [5.37–5.39], such as process control engineering [5.40, 5.41] and non-parametric statistics [5.42–5.44]. A neural network is readily programmed for kinetic prediction where many strongly interacting factors do affect the process rate or when data are either incomplete, not defined or even lacking. With reference to the "black box" used by classical statistics to describe the action of internal parameters of processes and their interaction on the process exit, the ANN methodology is strongly different because it explains the mechanism working inside the black box.

5.7.1
Short Review of Artificial Neural Networks

As mentioned in the introduction, ANNs are models inspired by the structure and the functions of the biological neurons, since they can also recognize patterns, disordered structure data and can learn from observation.

A network is composed of units or simple named nodes, which represent the neuron bodies. These units are interconnected by links that act like the axons and dendrites of their biological counterparts. A particular type of interconnected neural net is shown in Fig. 5.12. In this case, it has one input layer of three units (leftmost circles), a central or hidden layer (five circles) and one output (exit) layer (rightmost) unit. This structure is designed for each particular application, so the number of the artificial neurons in each layer and the number of the central layers is not a priori fixed.

The system behaves like synaptic connections where each value of a connection is multiplied by a connecting weight and then the obtained value is transferred to another unit, where all the connecting inputs are added. If the total sum exceeds a certain threshold value (also called offset or bias), the neuron begins to fire [5.45, 5.46].

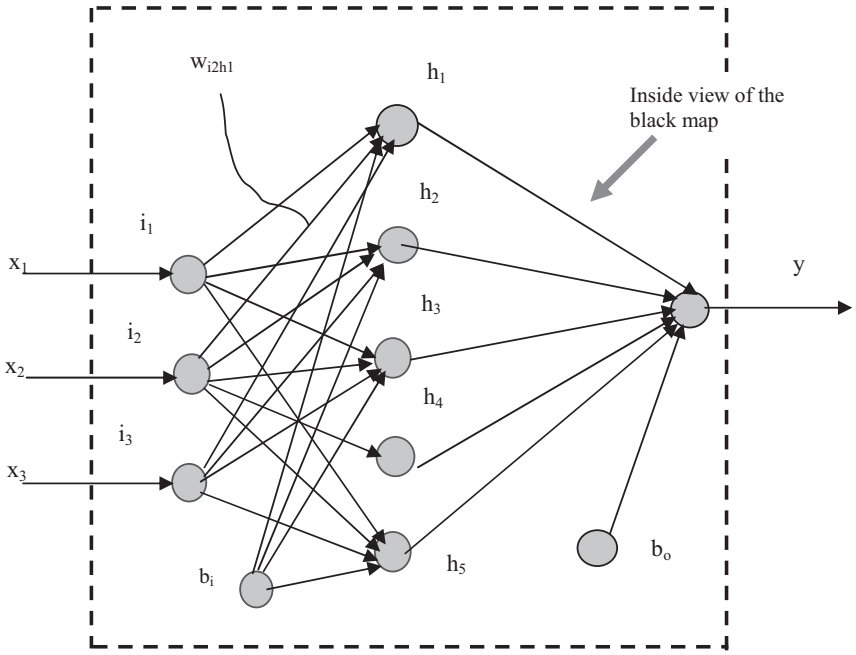


Figure 5.13 Layers of units and connection links in an artificial neuronal network. I_1–i_3: input neurons, h_1–h_5: hidden neurons, b_i, b_o: exit and output bias neurons, w_{i2h1}: weight of transmission, i_2–h_1, x_1–x_3: input process variables and y: output process variable.

The changes brought about in the pattern of neurons constitute the basis for learning.

In biological neurons, learning is carried out by changing the synaptic resistance associated to a change in the activation pattern of neurons.

Neural networks are able to learn because they can change the connection weights between two units which are in direct contact. After learning, the knowledge is somehow stored in the weights.

However, artificial neurons are much simpler than natural ones, the analogy serves to highlight an important feature of ANNs: the ability to learn through training. Just as the brain learns to infer from observations, an ANN learns the key features of a process through repeated training with data and, like in natural learning, its performance improves as it gains experience with a process.

The Latin expression "repetitio est mater studiorum", can be used here to describe the learning process with an ANN. The ANN repeats on and on, gradually adjusting the output to the imposed data output.

5.7.2 Structure and Threshold Functions for Neural Networks

The information flow between two biological neurons is affected by a variable synaptic resistance. In artificial systems, each connecting link has an associated weight. If two units are linked by a connection, the activation value of the emitting units is multiplied by the connecting weight before reaching the receiving unit.

The input value for an arbitrary unit, j, is then the sum of all activations coming from the units of the preceding layer, multiplied by the respective weights, w_{kj}, plus the bias value θ_{ij}. Thus, the total input to unit j, will be written as (5.171) where n represents the number of the neurons preceding neuron j and o_k shows the output.

$$I_j = \sum_{k=1}^{n} w_{kj} o_k + \theta_{ij} \tag{5.171}$$

Even though, most networks use the same type of input, their output generation may differ. In general, the output is computed by means of a transfer function, also called activation function. Concerning the behaviour of the transfer function, a gradual approach is required [5.47]. Therefore, a continuous threshold function is selected, chiefly because its continuity and derivability at all points are required features for the current optimization of the algorithms of learning. This type of function is well suited to the learning procedure that will be described later. A typical continuous threshold function is the following exponential sigmoid:

$$o_j = \frac{1}{1 + e^{-\beta l_j}} \tag{5.172}$$

where o_j is the activation value of neuron j, l_j is the total input to neuron j (as calculated by relation 5.171) and β is a constant which frequently takes a unitary value. The use of β, allows some modifications of the width of the region of the sigmoid, a feature which is useful in setting the learning ability of the net. Table 5.69 shows some other sorts of threshold functions that can be successfully used for developing an application.

Table 5.69 Common threshold functions used in ANN modeling.

Type	Function expression	Symbol signification
Linear	$o_j = \begin{cases} a + bI_j\ ,\ 0 \le I_j \prec 1 \\ 1\ ,\ I_j \ge 1 \end{cases}$	I_j = total input to neuron j
Saturating linear	$o_j = \begin{cases} I_j\ ,\ 0 \prec I_j \prec 1 \\ 1\ ,\ I_j \ge 1 \\ 0\ ,\ I_j \le 0 \end{cases}$	o_j = output from neuron j
Sigmoid classic	$o_j = \frac{1}{1 + e^{-l_j}}$	I^T_j – transpose of I_j
Hyperbolic tangent	$o_j = \tanh(I_j)$	σ_j standard deviation of I_j
Radial basis	$o_j = \exp[-I_j^T/(2\sigma_j^2]$	a, b – numerical constants

The function of the neural net depends not only on the information and processing mode of each isolated unit, but also on its overall topology. The topology considered in Fig. 5.12 must to be considered only as a didactic example. In the case of questions about the necessity of the hidden layer, we can easily give an answer: a hidden layer allows one to increase the network memory and provides some flexibility in the learning process. With the very simple topology considered in Fig. 5.13, the net is able to map linear and nonlinear relationships between inputs and outputs. The number of units in the input layer is determined by the variables that affect the response (x_1, x_2, x_3 in Fig. 5.13). The number of units in the hidden layer will be established during the learning process from a compromise between predicting errors and the number of iterations needed to attain them. In addition to the above units, two bias units are used (in Fig. 5.13, one for the hidden layer and one for the output unit). Their inputs are zero and their outputs or activation values are equal to one. Their use provides the threshold values to the hidden layer and to the output unit.

It is not difficult to observe that the application of an ANN to a problem involves four steps:

1. selection of the network topology (i.e. the layout of the neurons and their inter-connections),
2. specification of the transformation operator for each neuron from the topology,
3. initial assignment of weights w_{kj}, which are updated as the network learns,
4. initial learning, called training, which involves choosing the data and the training method.

As neural network theory has been developed, the empiricism associated with the choices at each step, has given ways to heuristic rules and guidelines [5.48, 5.49]. Nevertheless, experience still plays an important part in designing a network. The network depicted in Fig. 5.13 is the most commonly used and is called the feed-forward network because all signals flow forward.

Even though a number of techniques have been developed for the development of networks, they still remain iterative trial and error procedures. The heuristic approach described here can be used to reduce the trial and error selection process.

A hidden layer, with its appropriate units is capable of mapping any input presentation [5.50] and is thus necessary to restrict the topology to one layer only. So as to determine the optimum hidden units, the learning rate (v_l) and the momentum term (m_τ) will be assigned arbitrarily but with constant values and the gain term will be fixed at a value of one. With all the parameters fixed, various net topologies exhibit the same trends relative to each other, "vis-à-vis" the overall absolute error as a function of the number of iterations in the training mode [5.51]. Thus, it was found possible to determine the optimum net architecture within 50–100 iterations and without using the whole graph which describes the variation of the absolute error as a function of the number of iterations for each topology.

The number of iterations will be used as the criterion whenever on-line predictions are to be made, such as for chemical process control where computation time is important. The selection of v_l and m_τ and the gain term is essentially a trial and error procedure. Contrary to the usual approach, each of these parameters has not been fixed at a constant value for the entire training period. These were initially assigned with arbitrary values ($v_l = 0.8$, $m_\tau = 0.8$, gain = 1 for example, although these values are not a priori imposed). Then, they were updated while the parameters "jolted" the overall absolute error out of the local minima, which is typically encountered in the mechanism of the descendent gradient. Once the net parameters and the net architecture have been fixed, the minimum number of training data sets required for adequate mapping will be determined by trial and error procedures. The net is then ready to learn the data presented using the back-propagation algorithm.

5.7.3
Back-propagation Algorithm

As described below, the required behaviour is taught to the neural net by back propagation. This procedure is carried out by exposing the network to sets consisting of one input vector and its corresponding output vector. By an iterated procedure of trial and error, the convergence to determine the weight values that minimizes a prescribed error value is then achieved.

Back propagation is a kind of rapid descendent method of optimization. However, some authors prefer other optimization algorithms rather than back propagation, for example the Levenberg–Marquardt method [5.52]. The back-propagation algorithm with the delta rule is called a supervised learning method, because weights are adapted to minimize the error between the desired outputs and those calculated by the network. The error is calculated, for convenience, from the following expression in terms of squared deviations:

$$\Phi(w_{kj}^p) = \frac{1}{2}\sum_{p=1}^{r}\sum_{j=1}^{n}(\varsigma_j^p - o_j^p)^2 \tag{5.173}$$

where ς_j^p is the desired value output unit j for the sample pair p, o_j^p is the observed value for the same unit j and sample pair p, and p is the sum index for the total number of pairs r.

The adjustment on weights w_{kj} is done using the sensitivity of the error with respect to that weight, as:

$$\Delta w_{kj} = -\gamma\frac{\partial\Phi}{\partial w_{kj}} \tag{5.174}$$

The expression for the weight change is obtained from Eqs. (5.171) and (5.172) replacing them in the relation (5.174):

$$\Delta w_{hj} = \alpha\gamma(\varsigma_j - o_j)o_j(1 - o_j)o_h \tag{5.175}$$

where the indexes h and j refer to the nodes of the hidden and of the output layer respectively. Equation (5.175) allows one to modify the weights between the hidden and the output layers. On the other hand, the group of equations for the change of the weights between the hidden and output layer is obtained also, as:

$$\Delta w_{in\,k} = \alpha\gamma o_{in} o_k (1 - o_k) \sum_j \delta_j w_{hj} \tag{5.176}$$

where the subscripts "in, h, j" now refer to the input layer, the hidden layer and the output layer, respectively. As for the output layer, δ_j can be expressed as:

$$\delta_j = \alpha(\varsigma_j - o_j) o_j (1 - o_j) \tag{5.177}$$

In this case α and γ represent, respectively, the rate factor in output and the scaling factor of the net. These relations are also related to the sigmoid threshold function.

In order to modify the weights between the input and hidden layer it is necessary to know the weights between the hidden units and the output units. Therefore, during back propagation, first we change the connection weights between the output and hidden layer, and then we change the remaining weights conversely to the direction of information flow during the normal operation of the network: from hidden layer to input layer backwards.

While training is performed, the weights are initialized with values between –0.5 and 0.5 [5.48, 5.53] using a random procedure. The input–output experimental pairs are successively shown to the net and the weights are changed simultaneously. When all pairs have been shown to the network, the error is computed. If it is larger than the value allowed, all the pairs are shown to the network again in order to induce more changes in the weights. This cycle is repeated until convergence is achieved.

5.7.4
Application of ANNs in Chemical Engineering

The ANN techniques can be successfully applied in the field of chemical engineering. The examples presented here give a brief overview of the capacity of ANNs to solve some chemical engineering problems. Readers interested in investigating this topic further can refer to Bulsari's book [5.41] or to other authors referenced in the bibliography [5.39, 5.54]. In addition, to complete the information presented here, an important number of Internet sites can be used as well as some of the current chemical engineering scientific publications, which have given important attention to this subject.

Three major ways can be identified for the use of an ANN in chemical engineering:

1. as a substitute for the complicated models of transport phenomena or stochastic based models;

2. as data support, especially for the equilibrium and kinetic data needed by models based on transport phenomena. This kind of model is recognized as a hybrid neural –regression model;
3. as a model for control and process operation.

In all the above-mentioned cases, once the learning processes have been completed, ANNs have an assistant function which gives one or many answers to an argument of the complex modeled process (parameters, factors or independent variables). When we use an ANN as a substitute for models for stochastic or complicated transport phenomena, the learning process must be as shown in Fig. 5.14 which shows the coupling of an ANN and a complicated mathematical process. The mathematical model gives the input and output vectors for the ANN, which, in normal cases, are represented by the measured data. When the learning process has been completed, the process mathematical model (PMM) and the optimizing algorithm (OA) are decoupled and the ANN is ready to produce the simulation results for the process. This procedure is also used to produce the ANN simulators needed for the control of the processes or their usual automatic operation.

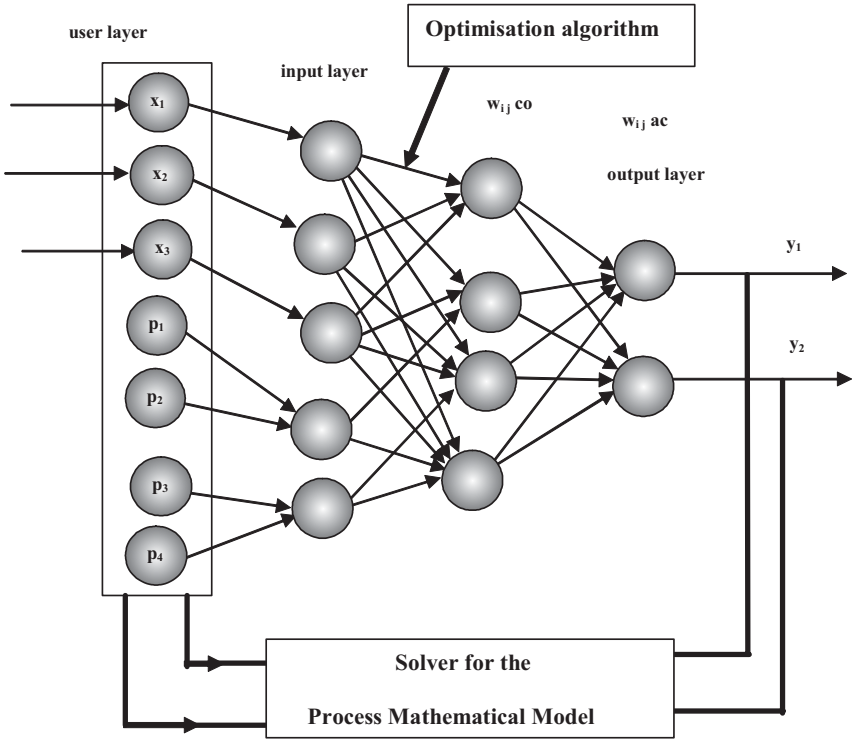


Figure 5.14 The architecture of assembly ANN-PMM-OA for the learning step (ANN – substitute for the complicated PMM).

In Fig. 5.14, it is shown that a previous formal user layer is necessary before using the input layer [5.55, 5.56]. Nevertheless, it is not necessary to have the same number of units in the input layer and in the user layer when each unit introduces a parameter or a process variable for the model.

The design of the input layer, which is a parameter for the ANN topology, will be coupled to the general problem of the topology of the ANN design.

The hybrid neural–regression model, shown in Fig. 5.15, uses one or more ANN(s) as generator for some of the numerical values needed by the base model process. As shown in Eq. (5.172) and in the equations reported in Table 5.69, the obtained answer of the net is uncertain, particularly for the inputs near zero. For those situations, the ANN will be assisted by one or more regression equations. However, why should a regression equation be used instead of neural net computing alone? In fact, the neural system is capable of giving a precise guess in the case of a kinetic yield at given P, T and one or more y_i^{τ} or y_i^{x} variables (see Fig. 5.15) and time or position needed by the complex modeled process. However, this response is not reliable when the value to be found is the time derivative of a neural net-generated curve because in this curve there exist points for which there is a maximum of the rate function.

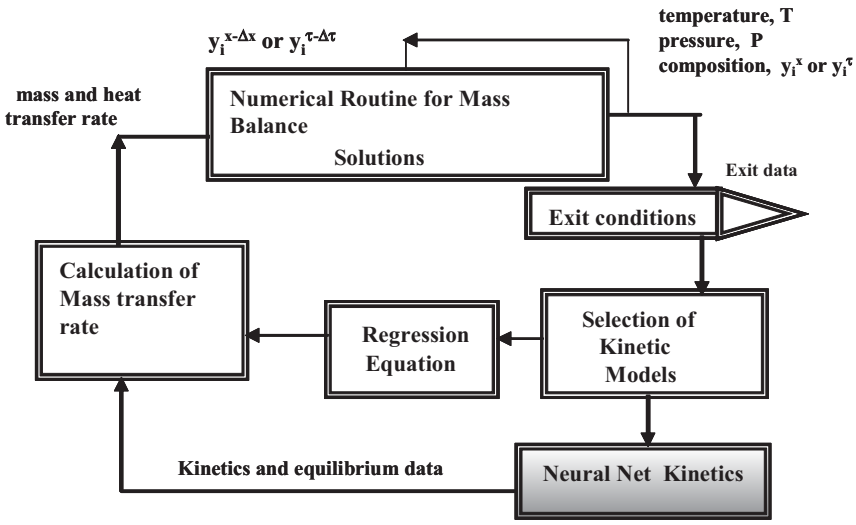


Figure 5.15 Principle of mass transfer model integration with a neural net (hybrid neural-regression model).

Neural networks can be used when traditional computing techniques can also be applied, but they can perform some calculations that would otherwise be very difficult to realize with current computing techniques. In particular, they can design a model from training data. A neural network can also be adapted to perform many different analogue functions such as pattern recognition, image processing, and trend analysis. These tasks are difficult to perform with conventional

digital program computers. An important function of the neural network is its ability to discover the trends in a collection of data. Indeed, trend analysis is very important in dynamic chemical engineering problems, in process control, as well as in chemical formulation, data mining, and decision support. Neural networks are also particularly useful as data sensor analysis and processing of industrial units or integrated chemical plants as well as in commercial activity in industrial chemistry. If the researcher has a good algorithm capable of completely describing the problem, then traditional calculation techniques can, in most cases, give the best solution, but if no algorithm or other digital solution exists to address a complex problem with many variables, then a neural network that learns from examples may provide a more effective solution to the problem.

References

5.1 W. L. Gore, *Statistical Methods for Chemical Experimentation*, Interscience Publishers, New York, 1952.

5.2 A C. Bonnet, L. N. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, John-Wiley, New York, 1954.

5.3 V. Youden, *Statistical Methods for Chemistry*, John Wiley, New York, 1955.

5.4 O. L. Davies, *Statistical Methods in Research and Production*, Oliver and Body, London, 1957.

5.5 V. V. Nalimov, *The Application of Mathematical Statistics by Chemical Analysis*, Pergamon Press, Oxford, 1963.

5.6 K. C. Peng, *The Design of Scientific Experiments*, Addison-Wesley, New York, 1967.

5.7 G. E. P. Box, N. Drapper, *Evolutionary Operation*, John Wiley, New York, 1969.

5.8 A. Gluck, *Mathematical Methods for Chemical Industry*, Technical Book, Bucharest, 1971.

5.9 R. Calcutt, R. Body, *Statistics for Analytical Chemists*, Chapman and Hall, London, 1983.

5.10 I. N. Miller, J. Miller, *Statistics for Analytical Chemistry*, Prentice-Hall, Hemel Hempstead, 1993.

5.11 W. P. Gardiner, *Statistical Analysis Methods for Chemists: A Software-Based Approach*, Royal Society of Chemistry, Cambridge, 1997.

5.12 C. W. Robert, J. A. Melvin (Eds.), *Handbook of Chemistry and Physics*, CRC Press Inc, Boca Raton, FL, 1983.

5.13 V. V. Kafarov, *Cybernetic Methods for Technologic Chemistry*, Mir, 1969.

5.14 W. Cornell, *Experiments with Mixture: Designs, Models and the Analysis of Mixture Data*, John Wiley, New York, 1990.

5.15 T. Dobre, O. Floarea, *Momentum Transfer*, Matrix-Rom, Bucharest, 1997.

5.16 M. Tyron, *Theory of Experimental Errors*, Technical Book, Bucharest, 1974.

5.17 R. Chaqui, *Truth, Possibility and Probability*, Elsevier, Amsterdam, 2001.

5.18 N. R. Drapper, H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1981.

5.19 O. Iordache, *Mathematical Methods for Chemical Engineering-Statistics Methods*; Polytechnic Institute of Bucharest, Bucharest, 1982.

5.20 J. N. Kappar, *Mathematical Modeling*, John Wiley, New York, 1988.

5.21 W. C. Hamilton, *Statistics in Physical Science*, The Ronald Press Co, New York, 1964.

5.22 R. Mihail, *Introduction to Experimental Planning Strategy*, Technical Book, Bucharest, 1983.

5.23 R. Carlson, *Design and Optimization in Organic Synthesis*, Elsevier, Amsterdam, 1997.

5.24 P. G. Maier, R. F. Zund (Eds.), *Statistical Method in Analytical Chemistry*, Wiley-Interscience, New York, 2000.

5.25 G. E. P. Box, K. B. Wilson, *J. R. Soc., Ser. B*, **1951**, *13*(1), 1–15.

5.26 V. Halimov, N. Cernova, *Mathematical Statistics for Chemical Analysis*, Nauka, Moscow, 1965.
5.27 G. E. P. Box, J. S. Hunter, *Ann. Math. Stat.*, **1957**, *28*, 1,195.
5.28 G. W. Lowe, *Trans. Inst. Chem. Eng.*, **1964**, *42* (9), 1334–1342.
5.29 W. E. Deming, *Statistical Adjustment of Data*, Dover, New York, 1964.
5.30 P. R. Bevington, *Data Reduction and Errors Analysis for Physical Science*, McGraw-Hill, New York, 1969.
5.31 C. K. Bayne, T. B. Rubin, *Practical Experimental Design and Optimization Methods for Chemists*, VCH, Deerfield Beach Florida, 1986.
5.32 O. Muntean, G. Bozga, *Chemical Reactors*, Technical Book, Bucharest, 2001.
5.33 M. Giovanni, H. Christoph, W. W. Bernd, R. Darskus, *Anal. Chem.*, **1997**, *69* (4), 601–606.
5.34 J. Tellinghuisen, *J. Phys. Chem. A*, **2001**, *105*, 3917–3924.
5.35 S. Gosh, C. R. Rao (Eds.), *Handbook of Statistics: Design and Analysis of Experiments*, Elsevier, Amsterdam, 1996.
5.36 J. Krauth, *Experimental Design*, Elsevier, Amsterdam, 2000.
5.37 P. G. Lisboa (Ed.), *Neural Network, Current Applications*, Chapman and Hall, London, 1992.
5.38 R. Maus, J. Keyes (Eds.), *Handbook of Expert System in Manufacturing: Neural Nets for Custom Formulation*, McGraw-Hill, New York, 1991.
5.39 L. Fausett, *Fundamentals of Neural Network*, Prentice-Hall, New York, 1994.
5.40 N. V. Bath, I. J. McAvoy, *Comput. Chem. Eng.*, **1990**, *14* (4/5), 573–584.
5.41 A. B. Bulsari (Ed.), *Neural Networks for Chemical Engineering*, Elsevier, Amsterdam, 1995.
5.42 H. S. Stern, *Technometrics*, **1996**, *38*, 205–220.
5.43 F. Blayo, M. Verleysen, *Artificial Neuronal Networks*, PUF, Paris, 1996.
5.44 R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, Amsterdam, 1990.
5.45 K. Gurney, *An Introduction to Neural Network*, UCLA Press, Los Angeles, 1997.
5.46 G. Montague, J. Morris, *Trends. Biotechnol*, **1994**, *6* (12), 312–325.
5.47 (a) D. van Camp, *Come istruire una rete artificiale di neuroni*, Le Scienze n° 291, pp. 134–136, Le Scienze spa, Milano, 1992; (b) T. Masters, *Practical Neural Network in C_{++}*, Academic Press, London, 1993.
5.48 N. V. Bath, T. J. McAvoy, *Comput. Chem. Eng.*, **1992**, *16* (2/3), 271–281.
5.49 M. A. Sartori, J. A. Panos, *IEEE Trans. Neural Networks*, **1991**, *2*, 4–16.
5.50 R. Sharma, D. Singhal, R. Ghosh, A. Dwiedi, *Comput. Chem. Eng.*, **1999**, *23* (5/6), 385–391.
5.51 A. B. Bulsari, S. Palasaori, *Neural Comput. Appl.*, **1993**, *1*,160–165.
5.52 M. L. Mavrovounitios, S. Chang, *Comput. Chem. Eng.*, **1992**, *1* (6), 283–291.
5.53 J. Delleirs (Ed.), *Neural Networks in QSAR and Drug Design*, Academic Press, New York, 1996.
5.54 P. R. Patnaik, *Biotechnol. Adv.*, **1999**, *17*, 477–489.
5.55 S. Zouling, L. K. Marjatta, S. Palosaori, *Chem. Eng. J.*, **2001**, *81*, 101–108.