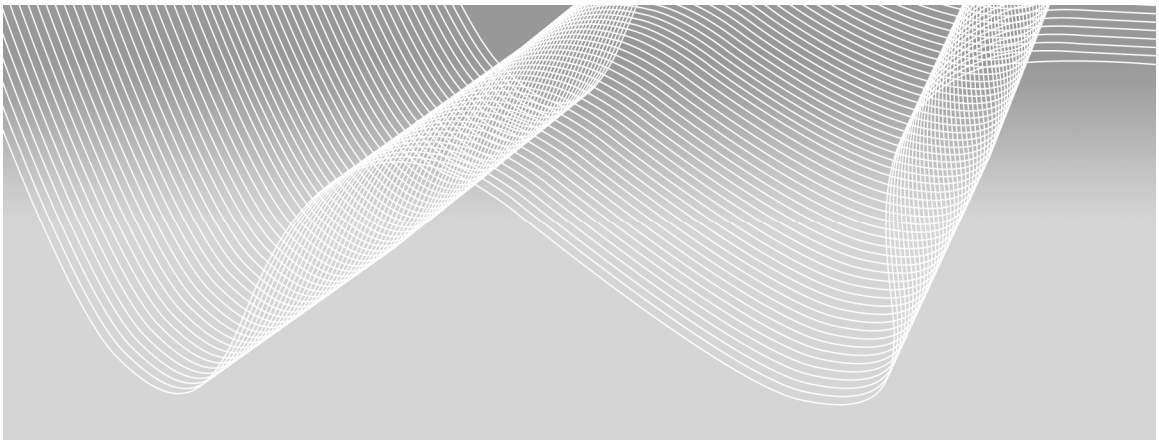


Chapter 1

Gaining Knowledge with Design of Experiments

1.1 Introduction

If you are reading this book, you probably do experiments that you would like to organize well. Ideally, you want to conduct only those experiments whose results yield the best possible information and insights.

This book is written to help you. It describes methods and tools that allow you to consistently design useful experiments, ensuring that you extract the maximum amount of information from each one. You will be able to efficiently solve problems and make decisions while gaining experience using JMP software.

We initially look at how experimental design is integrated into the process of knowledge acquisition. Then, we examine some basic concepts that make it possible to properly define a study, and how to interpret a study's results.

1.2 The Process of Knowledge Acquisition

The process of acquiring knowledge can be thought of as answering carefully posed questions. For example, if a farmer wants to know how a fertilizer influences corn yield, the following (and certainly more) questions are reasonable to ask:

- Can the field produce ten more bushels of corn per acre if I increase the quantity of fertilizer?
- How does the amount of rain affect the effectiveness of fertilizer?
- Will the quality of corn remain good if I use a certain fertilizer?
- How much fertilizer should I apply to get the biggest harvest (the most bushels of corn per acre)?

These questions frame both the problem and the solution. They also specify what work should be carried out. It is therefore important to ask questions that are truly representative of the problem.

Of course, before doing any experiments, it is good practice to verify that the question hasn't already been answered. To this end, you can review current literature from a bibliography, consult subject-matter experts, conduct some theoretical calculations, simulations, or do anything else that may answer your questions without experimentation. If, after conducting this research, you find that the question is answered, then there is no need to conduct experiments. More typically, however, these investigations cause you to modify your questions or generate new ones. This is the point when experiments—well-designed ones—are necessary to completely solve the problem. Unquestionably, this preliminary work is part of the experimenter's job. However, we do not focus on these preliminaries in this book.

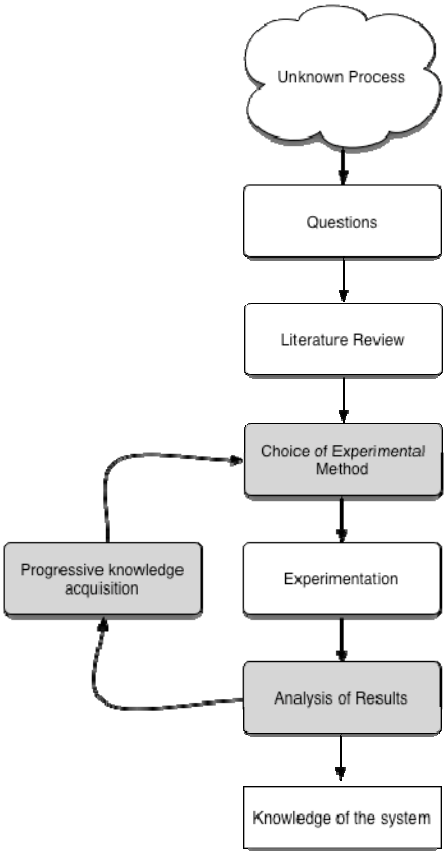
For questions that don't have readily available answers, it is necessary to carry out experiments. How can the experiments be designed so that they accomplish the following goals?

- Quickly arrive at the best possible results
- Omit unnecessary trials
- Give results with the best possible precision
- Progress without failure
- Establish a model for the studied phenomenon
- Discover the optimal solution

The three essential aspects of the knowledge acquisition process are shown in Figure 1.1:

- The choice of the experimental method
- The analysis of the results
- The progressive acquisition of knowledge

Figure 1.1 Design of experiments optimizes the three highlighted parts of the knowledge acquisition process.



Let's look at these three aspects in detail, remembering that experiments are organized to facilitate the analysis of the results and to allow the progressive acquisition of knowledge.

1.2.1 Choosing the Experimental Method

The experimental method must facilitate the interpretation of the results. However, it should also minimize the number of runs without sacrificing quality. Using designed experiments ensures the conditions that give the best possible precision with the smallest number of runs. Designed experiments give maximum efficiency using the smallest number of trials, and therefore the minimum cost.

1.2.2 Analyzing the Results

Analyzing the results of experiments is linked to the initial design choice. If the experiments are well-prepared, the results are easy to interpret, and they are also rich in information.

Thanks to computers and software, the construction of experimental designs and the necessary analysis calculations have become simple. These tools also support graphical representations that illustrate the results spectacularly and increase understanding of the phenomenon.

1.2.3 Progressively Acquiring Knowledge

An experimenter who undertakes a study obviously does not know at the outset the final results or what they reveal. Therefore, it is wise to advance gradually, to be able to adjust the experimental runs based on initial results. Using an initial outline, for example, makes it possible to direct the tests towards only the interesting aspects of the study and to avoid dead ends.

An initial batch of experiments leads to preliminary, tentative conclusions. Using these initial conclusions, the experimenter can carry out a new series of improved tests. Both series of experiments are used to obtain precise results. In this way, the experimenter accumulates only the results that are needed, and has the flexibility to stop when the results are satisfactory.

1.3 Studying a Phenomenon

Studying a phenomenon is often thought of as focusing on a particular measurement: a car's gasoline consumption, the wholesale price of a chemical, or the corn yield per acre. This measurement (consumption, price, or yield) depends on a great number of variables.

For example, gas consumption is related to the speed of the vehicle, the engine horsepower, driving style, the direction and force of the wind, the inflation of the tires, the presence (or not) of a luggage rack, the number of people in the car, the make of car, and so on. The price of a chemical depends on the quality of the raw materials, the yield of each manufacturing unit, external specifications, conditions of manufacture, and many other quantities. Corn yield, too, depends on the quality of the soil, the quantity of incorporated fertilizer, sun exposure, climate, corn variety, and so on.

Mathematically, we can write the measurement of interest as y, (which we will call the *response*) as a function of several variables x_i (which we will call *factors*) as

$$y = f(x_1, x_2, x_3, \cdots, x_k)$$

The study of a phenomenon boils down to determining the function f that relates the response to the factors x_1, x_2, …, x_k.

To look at this approach in more detail, it is necessary to introduce a few special ideas and also some terminology specific to designed experiments.

1.4 Terminology

This section describes terms related to factors and their representation.

1.4.1 Factor Types

The construction of designs and the interpretation of their results depend largely on the types of factors involved in the study. Statisticians distinguish among several types of factors. We discuss four types:

Continuous Factors

Pressure is an example of a continuous factor. For a given interval, any value in the interval can be chosen. Other examples are wavelength, concentration, or temperature. Values taken by continuous factors are therefore represented by continuous numbers.

Discrete Factors

On the other hand, discrete factors can take only particular values. These values are not necessarily numeric. The color of a product (say, blue, red, or yellow) is an example of a discrete factor. A discrete factor can take values that are names,

letters, properties, or numbers. In the latter case, the number is really just a numeric label, and does not represent a numeric quantity. It is merely a name or a reference.

Ordinal Factors

Ordinal factors are discrete factors that can be placed in a logical order. For example, size may be represented as large, medium, or small. Ranks, also, are ordinal: first, second, third, and fourth.

Boolean Factors

Boolean factors are discrete factors which can take only two levels: high and low, open or closed, black and white, −1 and 1, and so on.

The border is sometimes fuzzy among these various types of factors. Color (orange, red, blue, etc.), apparently a discrete factor, can be transformed into an ordinal measurement, and even a continuous measurement if the concept of wavelength is introduced. A continuous factor, like speed, can be transformed into an ordinal or discrete factor: rapid and slow, or speed A and speed B. This possibility is not a disadvantage—it is an additional flexibility that the experimenter can use when interpreting results. In fact, the choice sometimes makes it easier to highlight certain aspects of the study.

Changing the variable type is also a means of adapting the answer to the aim of the study. For example, consider the age of the members of a population. If the study looks at average age, the variable "age" is regarded as continuous. If, however, the study examines the number of people reaching a certain age, the variable "age" is an ordinal variable, since there are several age categories: young people, teenagers, adults, and seniors. Finally, if we were studying only the proportion of young people younger than 18, the variable "age" is Boolean: younger than 18 and older than 18.

1.4.2 Experimental Space

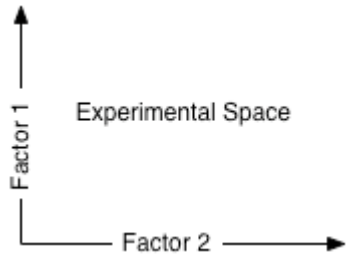
Understanding experimental design also requires grasping the essential concept of *experimental space* of the variables in the study. We now examine this fundamental concept in detail using continuous factors, since they are the most frequently used.

To graphically illustrate an experimental space, we use a two-dimensional area. This representation allows an easy extension to multidimensional spaces.

One continuous factor can be represented by a directed and graduated axis. If there is a second continuous factor, it is represented by a similar axis. This second axis is drawn orthogonally to the first (i.e., they form a 90° angle). Mathematically, this gives a Cartesian plane that defines a Euclidean space in two dimensions. This area is called the

experimental space (Figure 1.2). The experimental space is composed of all the points of the plane factor 1 × factor 2 where each point represents an experimental trial.

Figure 1.2 Each factor is represented by a graduated and oriented axis. The factor axes are orthogonal to each other. The space thus defined is the *experimental space.*

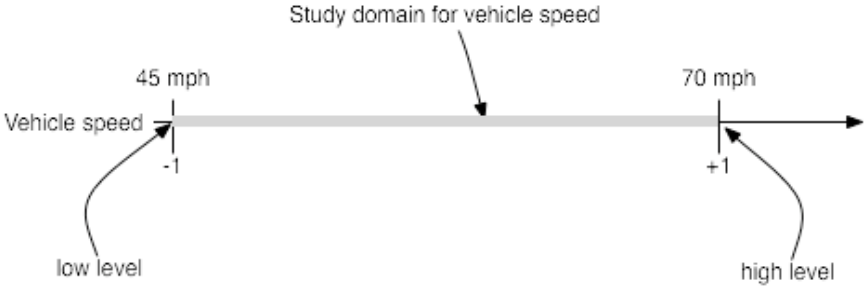


If there is a third factor, it too is represented by a directed and graduated axis, positioned perpendicularly to the first two. With four factors or more, the same construction applies, but it is not possible to represent the space geometrically. A purely mathematical representation (a hypercube of four dimensions) of the experimental space is necessary.

1.4.3 Factor Domain

The value given to a factor while running an experimental trial is called a *level*.

Figure 1.3 The domain of variation for speed contains all the speeds between 45 mph and 80 mph. The low level of the factor is written as −1 and the high level as +1.



When we study the effect of a factor, in general, we restrict its variation to be between two limits. The experimenter defines these two levels according to specifics of the study. The lower limit is the *low level*. The upper limit is the *high level*. For example, to study the effect of a vehicle's speed on its gas usage, its speed is allowed to vary between 45 and 70 miles per hour. The speed of 45 mph is the low level and the speed of 70 mph is the high level. The set containing all the values between the low and the high level that the factor can take is called the factor's *domain of variation* or, more simply, the factor's *domain*.

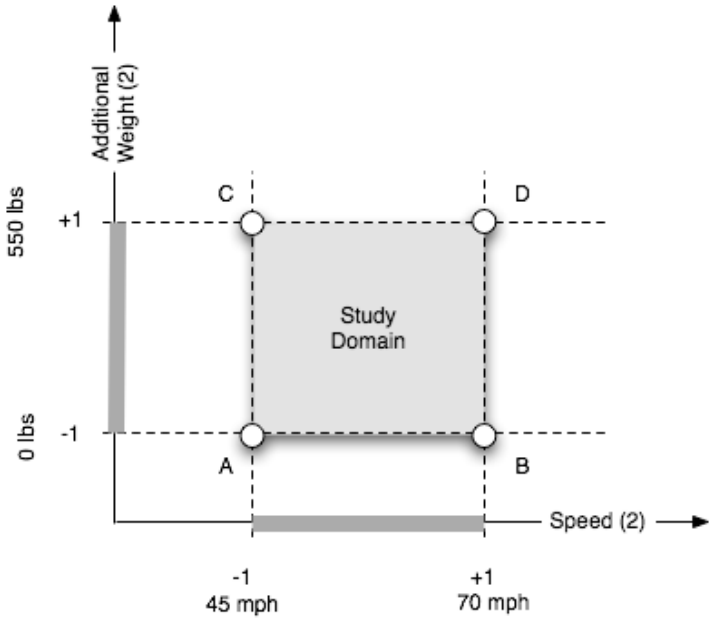
If there are more factors, each has its own domain. Since the different factors may have different units of measurement, it useful to have a common representation for all of them. In design of experiments (DOE), it is common to denote the low levels by –1 and the high levels by +1. Here we designate the speed of 45 mph as the –1 level and 70 mph as the +1 level.

The interior of a factor's domain contains all the values that it can theoretically take. Two, three, or more levels can therefore be chosen according to the needs of the study. For example, if the study uses a second-degree (quadratic) model, three or four levels should be chosen. That is, we should choose three or four different speeds.

1.4.4 Study Domain

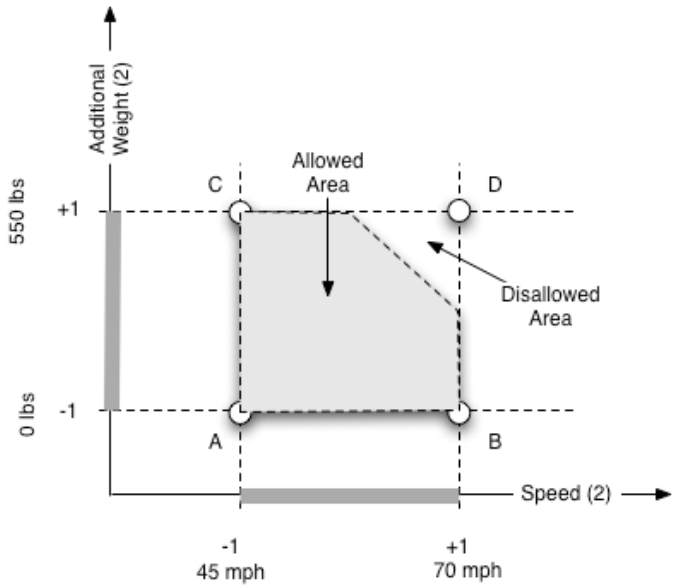
In practice, the experimenter chooses a portion of the experimental space to carry out the study. This special zone of the experimental space is the *study domain* (Figure 1.4). This domain is defined by the high and low levels of all the factors and possibly by constraints among the factors. Let's suppose that the second factor is a vehicle's additional weight, defined as any additional mass aside from that of the vehicle and the driver. The lower level of this additional weight is 0 lbs; the high level may be, say, 550 lbs. If there are no constraints, the study domain is represented by all the points where additional weight lies between 0 and 600 lbs and whose speeds lie between 45 and 70 mph.

Figure 1.4 The study domain is defined by the union of the domains from the different factors. Here, there are no constraints.



There may be constraints on the study domain. For example, it might be impossible to attain a speed of 70 mph with a lot of additional weight. Figure 1.5 illustrates this possible reduction of the initial study domain.

Figure 1.5 The study domain with constraints is represented by the shaded area.



1.5 Centered and Scaled Variables

When the lower level of a factor is represented by −1 and the upper level is represented by +1, two important changes occur:

- The center of the measurements moves.
  In our example, the middle of the interval [−1, 1] is zero and corresponds to the value 57.5 mph. The numerical value of the new zero origin, therefore, differs from the origin when expressed in the original experimental units (sometimes called *engineering units*).
- The measurement units change.
  In our example, the lower level of the speed factor is 45 mph and the high level is 70 mph, so there are 25 mph between these two values, i.e., 25 times the speed unit. Between −1 and 1 there are two new units: the newly defined unit corresponds to 25 mph, and is given the name of *step*.

These two changes involve the introduction of new variables called *centered and scaled variables* (*csv*). Centering refers to the change of origin, and scaling refers to the change of units. These new variables are also commonly called *coded variables* or *coded units*.

The conversion of the original variables A to the coded variables X (and vice versa) is given by the following formula, where A_0 is the central value in engineering units.

$$x = \frac{A - A_0}{\text{Step}} \tag{1.1}$$

The advantage to using coded units lies in their power to present designed experiments in the same way, regardless of the chosen study domains and regardless of the factors. Seen this way, DOE theory is quite generalizable.

The use of coded variables is common in DOE software. For example, finding the best experimental points using the *D*-optimality criterion (see Chapter 11) is possible only when using coded variables.

Coded variables result from the ratio of two same-sized physical units, so they have no dimension. The absence of natural units is due to the fact that all the factors have the same domain of variation (two coded units), allowing direct comparison of the effects of the factors among themselves.

Example 1

An experimenter chooses for the speed factor to be 45 mph at the low level and 70 mph at the high level. In coded units, what is the corresponding speed for 55 mph?

Let's calculate the step for the speed factor. It's equal to half the difference between the high and low levels, so

$$\text{Step} = \frac{A_{+1} - A_{-1}}{2} = \frac{70 - 45}{2} = 12.5$$

A_0 is the center value between the high and low levels; that is, it is half of the sum of the high and low levels:

$$A_0 = \frac{A_{+1} + A_{-1}}{2} = \frac{70 + 45}{2} = 57.5$$

Applying equation (1.1):

$$x = \frac{A - A_0}{\text{Step}} = \frac{50 - 57.5}{12.5} = -0.6$$

A speed of 55 mph is therefore, for this example, equal to –0.6 in coded values.

Example 2

We may also want the value in original units, knowing the coded value. In engineering units, what is the value of the speed factor corresponding to +0.5 in coded units? Write equation (1.1):

$$+0.5 = \frac{A - 57.5}{12.5}$$

So

$$A = 57.5 + 0.5 \times 12.5 = 63.75$$

The coded speed 0.5 corresponds to a speed of 63.75 mph.

1.6 Experimental Points

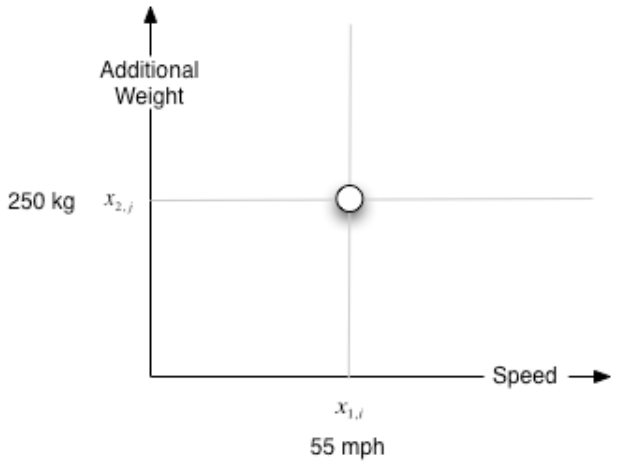
In a two-dimensional space, level i of factor 1, written as $x_{1,i}$, and level j of factor 2, written as $x_{2,j}$, can be considered as the coordinates of a point of the experimental space or of the study domain (Figure 1.6). For example, if the speed level is 55 mph and the additional weight is 250 lbs, the coordinates of the experimental point are

$x_{1,i}$ = 55 mph

$x_{2,j}$ = 250 lbs

One run of the experiment can be represented by a single point in this axis system. This is the reason that a run is often designated by the expression *experimental point*, *experiment point* or, even just *point*. A designed experiment is therefore represented by a collection of experimental points situated in an experimental space. In our current example, the experiment is conducted on a vehicle that is moving at 55 mph with an additional weight of 250 lbs.

Figure 1.6 In the experimental space, the factor levels are specified by experimental points.



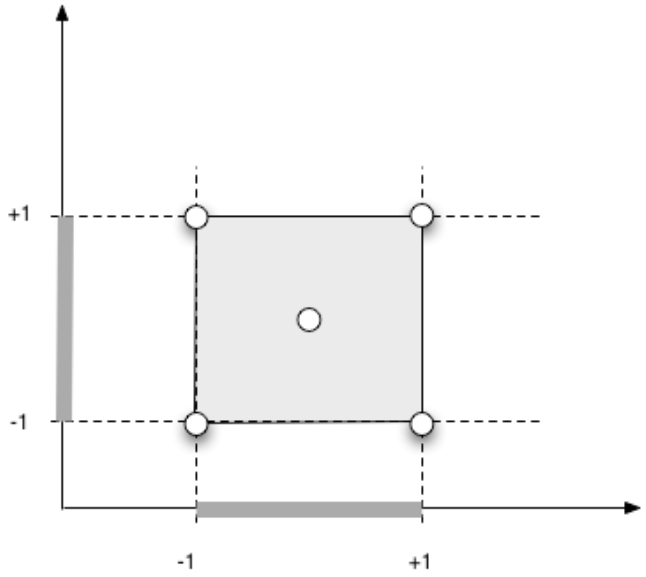
With up to three factors, it is possible to physically draw the study domain. With more than three factors, we use a tabular representation, called a *matrix*, which is more general because it allows representation of the experimental points in a multidimensional space with any number of dimensions.

1.7 Design of Experiments

1.7.1 Methodology of Designs without Constraints

The choice of the number and the placement of experimental points is the fundamental problem of designed experiments. Ideally, we want to carry out the minimum number of trials while reducing the influence of the experimental error on the models that will be used to make decisions. This goal is achieved by considering the mathematical and statistical properties that relate the response to the factors. When there are no constraints on the study domain, there are classical designs that have excellent statistical qualities and which allow the modeling of responses under the best conditions (Figure 1.7). If there are constraints on the design, we must construct custom designs by finding point positions which lead, in the same way, to quality statistics and good response modeling.

Figure 1.7 Example of the arrangement of the experimental points in a domain without constraints



1.7.2 Methodology of Designs with Constraints

The construction procedure of designs where the domain is constrained is as follows:

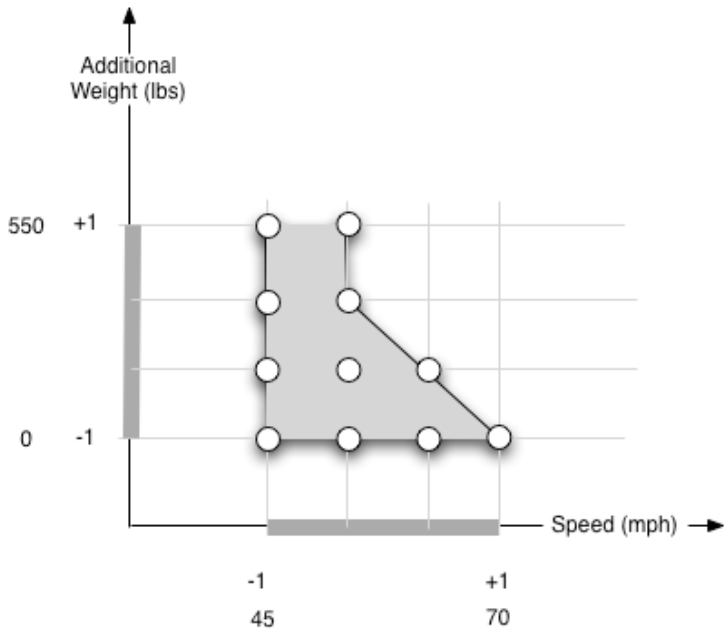
1. Define the domain of each factor (low level and high level).

2. Define any constraints that restrict the factors. These constraints are expressed as inequalities among the factors, and they define allowed areas (where trials can be carried out), and disallowed areas (where trials cannot be carried out).

3. Define any other factor levels that may be interesting for the study (other than the high and low levels, which are already defined). When used, between two and five additional levels for each factor are specified.

4. Construct a grid by taking into account all the combinations of factor levels. This grid should contain only realistic experimental points; that is, it should contain points in the allowed experimental areas. These points form the *candidate set* (Figure 1.8).

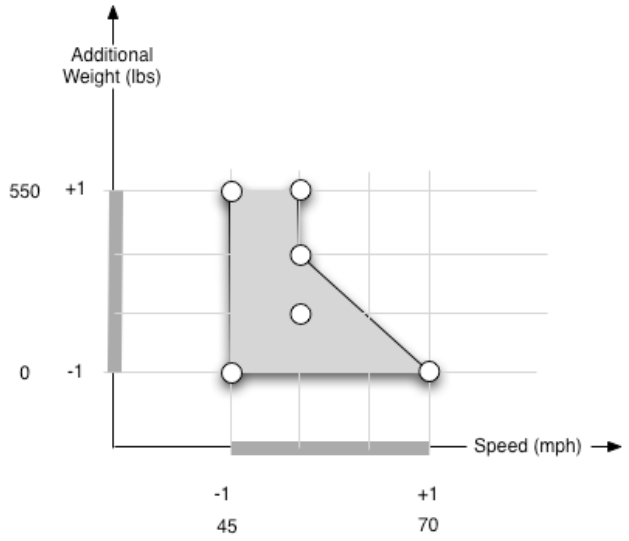
Figure 1.8 The grid of candidate points is made up of possible trials in the study domain.



5. Choose a function a priori that relates the response to the factors.

6. Select, according to the chosen optimality criterion, the number and placement of experimental points most useful for modeling the phenomenon (Figure 1.9). This selection requires many long and tedious calculations, and selection is not possible without the aid of DOE software.

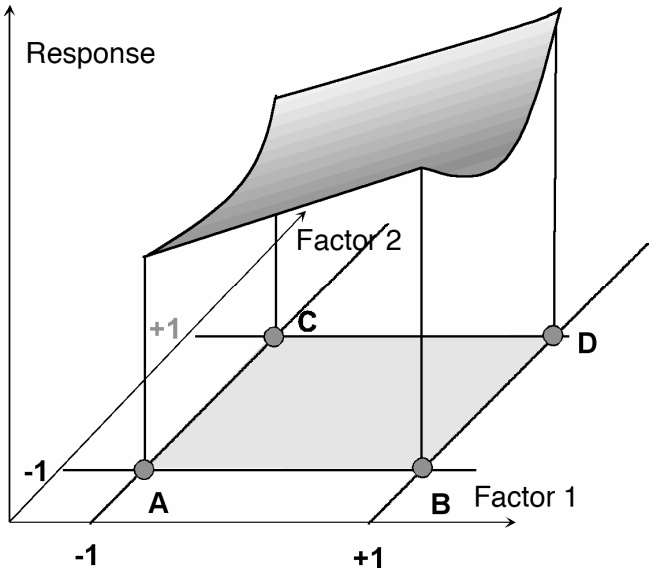
Figure 1.9 Optimal points as selected by software



1.7.3 The Response Surface

Each point in the study domain corresponds to a response. Together, all the points in the study domain correspond to a collection of responses located on a surface. We call this the *response surface* (Figure 1.10).

Figure 1.10 The collection of responses that correspond to all the points in the study domain forms the response surface.



In general, only a few responses are known: those that correspond to the experimental points chosen by the experimenter. To obtain the response surface, it is necessary to interpolate using a mathematical model.

Those points chosen using DOE theory ensure the best possible precision of the form and position of the response surface.

1.7.4 The a priori Mathematical Model of the Response

Mathematical modeling

We use a first-order linear model to approximate the response.

$$y = a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + \sum a_{ii} x_i^2 + \cdots \tag{1.2}$$

where

- y is the response or measurement of interest to the experimenter.
- x_i represents a level of factor i.

- x_j represents a level of factor *j*.
- a_0, a_i, a_{ij}, a_{ii} are the coefficients of the polynomial.

This model is called the *a priori* model, or the *postulated* model.

The predetermined models are valid prediction models inside the study domain, which must always be precisely established. These are not theoretical models based on physiochemical or mechanical laws.

Experimental modeling

Two concepts must be added to the purely mathematical model described above.

The first is the *lack of fit*. This term expresses the fact that the model chosen by the experimenter before the trials is probably a little different from the true model of the studied phenomenon. There is a difference between these two models. This difference is the lack of fit, denoted by the Greek letter delta (Δ).

The second concept is the random nature of the response. In reality, measuring the same response several times at the same experimental point does not give exactly the same result. There is a dispersion of the results. Dispersions like these are called *random error* or *pure error* and are denoted by the Greek letter epsilon (ε).

The general equation (1.2) must be modified as follows:

$$y = f(x_1, x_2, x_3 \cdots, x_n) + \Delta + \varepsilon \tag{1.3}$$

This equation is used in Chapter 5 where we show how to estimate the lack of fit Δ, and the pure error ε.

System of equations

Each experimental point corresponds to a response value. However, this response is modeled by a polynomial whose coefficients are unknowns that must be determined. An experimental design results in a system of *n* equations (for *n* trials) in *p* unknowns (for the *p* coefficients in the *a priori* model). This system can be written in a simple way using matrix notation.

$$\mathbf{y} = \mathbf{X}\ \mathbf{a} + \mathbf{e} \tag{1.4}$$

where

y is the *response vector.*

X is the *model matrix* or the *design matrix* which depends on the experimental points used in the design and on the postulated model.

a is the *coefficient matrix*.

e is the *error matrix*.

This system of equations cannot be, in general, solved simply because there are fewer equations than there are unknowns. There are n equations and $p + n$ unknowns. To find the solution, we must use special matrix methods generally based on the criterion of least squares. The results are estimations of the coefficients, denoted as $\hat{\mathbf{a}}$.

The algebraic result of the least-squares calculations is

$$\hat{\mathbf{a}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \tag{1.5}$$

where $\mathbf{X}'$ is the transpose of X (for details, see Appendix C, "Introduction to Matrix Calculations"). A number of software packages exist (like JMP and SAS/STAT) that will carry out these calculations and that directly give the coefficient values.

Two matrices appear frequently in the theory of experimental design:

- The *information matrix* $\mathbf{X}' \mathbf{X}$.
- The *dispersion matrix* $(\mathbf{X}' \mathbf{X})^{-1}$.