CHAPTER 1

Introduction to Response Surface Methodology

In this chapter we discuss, in a preliminary way, some general philosophy necessary to the understanding of the theory and practice of response surface methodology (RSM). Many of the points made here receive more detailed treatment in later chapters.

1.1. RESPONSE SURFACE METHODOLOGY (RSM)

The mechanisms of some scientific phenomena are understood sufficiently well that useful mathematical models that flow directly from the physical mechanism can be written down. Although a number of important statistical problems arise in the building and study of such models they are not considered in this book. Instead, the methods we discuss will be appropriate to the study of phenomena that are presently not sufficiently well understood to permit the mechanistic approach.

Response surface methodology comprises a group of statistical techniques for empirical model building and model exploitation. By careful design and analysis of experiments, it seeks to relate a *response*, or *output* variable to the levels of a number of *predictors*, or *input* variables, that affect it.

The variables studied will depend on the specific field of application. For example, response in a chemical investigation might be *yield* of sulfuric acid and the input variables affecting this yield might be the *pressure* and *temperature* of the reaction. In a psychological experiment, an investigator might want to find out how a test *score* (output) achieved by certain subjects depended upon the *duration* (input 1) of the period during which they studied the relevant material, and the *delay* (input 2) between study and test. In mathematical language, in this latter case, we can say that the

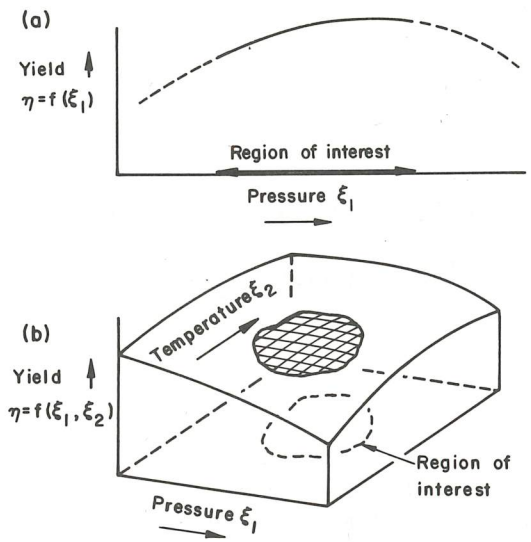

FIGURE 1.1. (*a*) A response curve. (*b*) A response surface.

investigator is interested in a presumed *functional relationship*

$$\eta = f(\xi_1, \xi_2), \tag{1.1.1}$$

that expresses the response or output score, η (Greek letter eta), as a function of two input variables *duration*, ξ_1, and *delay*, ξ_2 (Greek letters xi). Both our examples involve two inputs but, in general, we shall have not two, but k input variables $\xi_1, \xi_2, \ldots, \xi_k$, and the functional relationship between the mean response and the levels of the k inputs can then be written

$$\eta = f(\xi_1, \xi_2, \ldots, \xi_k). \tag{1.1.2}$$

More compactly, if $\boldsymbol{\xi}$ denotes a column vector* with elements $\xi_1, \xi_2, \ldots, \xi_k$, the mean response function may be written as

$$\eta = f(\boldsymbol{\xi}). \tag{1.1.3}$$

If there is only one input variable ξ_1, we can relate an output η to a single input ξ_1 by a *response curve* such as that in Figure 1.1*a*. If we have two inputs ξ_1 and ξ_2 and we draw, in three-dimensional space, a graph of η

*The word vector is used here only as a shorthand word referring to all k elements.

against ξ_1 and ξ_2, we obtain a *response surface* such as that shown in Figure 1.1*b*. When k ξ's are involved and k is greater than 2, we shall still talk of a response surface in the $k + 1$ dimensional space of the variables even though only sectional representation of the surface is possible in the three-dimensional space actually available to us.

The operation of the system with the k inputs adjusted to some definite set of levels $\boldsymbol{\xi}$ is referred to as an *experimental run*. If repeated runs are made at the same conditions $\boldsymbol{\xi}$, the measured response will vary because of measurement errors, observational errors, and basic variability in the experimental material. We regard η therefore as the mean response at particular conditions $(\xi_1, \ldots, \xi_k)$. An actual observed response result, which we call y, would fall in some statistical distribution about its mean value η. We say that the expected value of y equals η, that is, $E(y) = \eta$. In any particular run, we shall refer to the discrepancy $y - \eta$ between the observed value y and the hypothetical mean value η as the *error* and denote it by ε (Greek epsilon). In general then, our object is to investigate certain aspects of a functional relationship affected by error and expressed as

$$y = f(\boldsymbol{\xi}) + \varepsilon. \tag{1.1.4}$$

Extending the Scope of Graphical Procedures

The underlying relationship between η and a single-input variable ξ can often be represented by a smooth curve like that in Figure 1.1*a*. Even though the exact nature of the function f was unknown, a smooth curve such as a polynomial acting as a kind of mathematical French curve might be used to represent the function *locally*. Similarly, the relationship between η and two inputs ξ_1 and ξ_2 can often be represented by a smooth surface like that in Figure 1.1*b* and a suitable polynomial, acting as a two-dimensional French curve could be used to approximate it locally. Mathematical French curves, like other French curves, are, of course, not expected to provide a global representation of the function but only a local approximation over some limited region of current interest.

The empirical use of graphical techniques is an important tool in the solution of many scientific problems. The basic objective of the use of response surface methods is to extend this type of empirical procedure to cases where there are more than just one or two inputs. In these higher dimensional cases, simple graphical methods are inadequate, but we can still fit graduating polynomials and, by studying the features of these fitted empirical functions, greatly extend the range of application of empirical procedures.

So that the reader can better appreciate where response surface methodology fits in the overall picture of scientific experimentation, we now briefly survey some aspects of scientific method. A more extensive review of the role of statistics in the scientific method is given by Box (1976).

1.2. INDETERMINANCY OF EXPERIMENTATION

At first sight, the conduct of an experimental investigation seems to be a highly arbitrary and uncertain process. For example, suppose that we collected 10 teams* of experimenters competent in a particular field of science or technology, locked each team in a separate room, presented them all with the same general scientific problem, and asked each team to submit a plan that could lead to the problem's solution. It is virtually certain that no two teams would present the same plan. Consider, in particular, some of the questions which would arise in planning an initial set of experiments, but on which the teams would be unlikely to agree.

1. Which input variables $\xi_1, \xi_2, \ldots,$ should be studied?

If it were a chemical reaction that was being studied, for example, it might be that most investigators would regard, say, temperature and pressure as being important but that there might be a diversity of opinion about which should be included among other input variables such as the initial rate of addition of the reactant, the ratio of certain catalysts, the agitation rate, and so on. Similar and perhaps even stronger disagreement might occur in a psychological experiment.

2. Should the input variables ξ be examined in their original form, or should transformed input variables be employed?

An input variable ξ, such as energy, may produce a linear increase in a response η, such as perceived loudness of noise, when ξ is varied on a *geometrical* scale—or, equivalently, when $\ln \xi$ is varied on a linear scale. It is simpler to express such a relationship therefore by first transforming the input ξ to its logarithm. Another input might be related to the response η by an inverse square law, suggesting an inverse square transformation ξ^{-2}. Other examples can be found leading to transformations such as the square root $\xi^{1/2}$, the inverse square root $\xi^{-1/2}$, and the reciprocal ξ^{-1}. A choice of a transformation for a single variable is often called a *choice of metric* for that variable. More generally, a transformation on the input variables can involve two or more of the original inputs. Suppose, for example, that the

*Or 10 individual experimenters.

amounts ζ_1 and ζ_2 (Greek zeta) of two nitrogenous fertilizers were being investigated. Rather than employing ζ_1 and ζ_2 themselves as the input variables, their sum $\xi_1 = \zeta_1 + \zeta_2$, the total amount of nitrogenous fertilizer, and their ratio $\xi_2 = \zeta_1/\zeta_2$ might be used if it were likely that the response relationship in terms of ξ_1 and ξ_2 could be more simply expressed. In some instances, the theory of dimensionless groups can be used to indicate appropriate transformations of this kind, but usually the best choice of metrics and transformations is not clearcut and, initially at least, will be the subject of conflicting opinion.

3. How should the response be measured?

It is often far from clear how the response should be defined. For example, in a study to improve a water treatment plant, the biochemical oxygen demand (BOD) of the effluent would often be regarded by experts as a natural measure of purity. However, it is possible to have an effluent with zero BOD which is lethal to both bacteria and humans. Thus the appropriate response depends on the end use of the response and will often be the subject of debate.

In (2), we discussed transformation of the input variables. Transformation of the response variable is also possible, and the appropriate metric for the response is not always clear. In subsequent chapters we show how transformation in the response can lead both to simpler representation and simpler assumptions.

4. At which levels of a given input variable ξ should experiments be run?

Suppose temperature is an important input. One experimenter might feel that, for the particular system under study, experiments covering temperatures over the range 100–140°C should be made. Another experimenter, believing the system to be very sensitive to temperature, might choose a range from 115 to 125°C. A third experimenter, believing that considerably higher temperatures would be needed might pick a range of from 140 to 180°C. In practice, of course, not one but several inputs must be considered simultaneously. If the temperature is to be changed over a 20° range, for example, what is a suitable* "commensurate" change for concentration?

In brief, the investigator must choose not only a location for his experiments but also an appropriate scaling for each of the variables.

*A "rule of thumb" sometimes used by chemists is that a 10°C change in temperature roughly doubles the rate of a chemical reaction. Thus, a commensurate change for concentration in this example would consist of quadrupling the concentration. However, here again, the effect of varying concentration differs for different chemical systems and, in particular, for the reaction order.

5. *How complex a model is necessary in a particular situation?*

This question is, of course, related to questions (2), (3), and (4). By definition, the more appropriately transformations and metrics are chosen, the simpler is the model that may be employed. Also, the more extensive the region of interest is in the ξ space, the more complex will be the model needed. Thus again we see that this question is a hazy one and is intimately bound up with questions considered under (2), (3), and (4).

6. *How shall we choose qualitative variables?*

The foregoing discussion is posed entirely in terms of *quantitative* inputs such as temperature and concentration. Similar indeterminacies occur in experimentation with *qualitative* inputs such as type of raw material, type of catalyst, identity of operator, and variety of seed. For example, if we are to compare three kinds of seed, which varieties should be employed? Should the tested varieties include the cheapest, that believed to give the greatest yield, that which is most popular? The answer depends on the objectives of the investigation, and that, partially at least, depends on opinion.

7. *What experimental arrangement (experimental design) should be used?*

This question is intimately bound up with all of the questions already studied. For instance, consider the hypothetical experiment of question (2). If the amounts ζ_1 and ζ_2 of the two nitrogen fertilizers are each tested at three levels in all nine possible combinations, the experimental arrangement is that of Figure 1.2*a*. However, if the same experimental plan is set out in terms of total nitrogen $\xi_1 = \zeta_1 + \zeta_2$, and nitrogen ratio $\xi_2 = \zeta_1/\zeta_2$, we

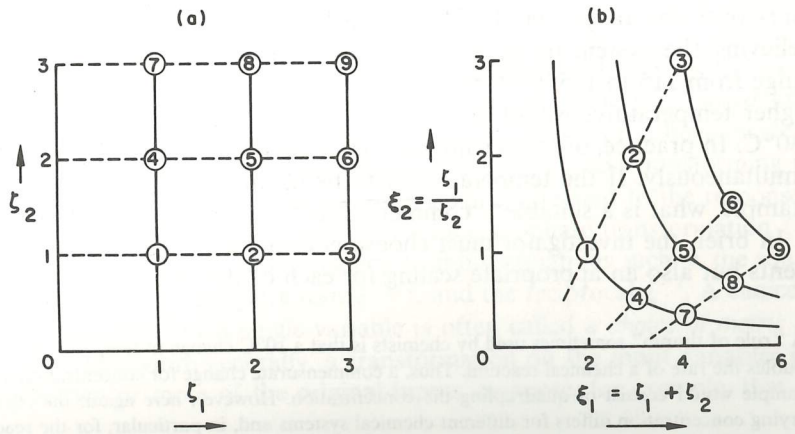

FIGURE 1.2. The effect of a transformation on the shape of an experimental design: (*a*) A balanced design in the (ζ_1, ζ_2) space. (*b*) The same design in the $\{(\zeta_1 + \zeta_2), \zeta_1/\zeta_2\}$ space.

obtain the very unbalanced arrangement of Figure 1.2*b*. In general, the question of "design optimality" is intimately bound up with arbitrary choices of metric, transformation, and the size and shape of the region to be studied.

In brief then, the investigator deals with a number of entities whose natures are necessarily matters of opinion. Among these are (a) the identity of the space of the inputs and outputs in which the experiments should be conducted; (b) the scales, metrics, and transformations in which the variables should be measured, and (c) the location of the region of interest, the specification of the model over it, and the experimental arrangement that should be used to explore the region of interest.

1.3. ITERATIVE NATURE OF THE EXPERIMENTAL LEARNING PROCESS

Faced with so many indeterminancies and uncertainties, the reader might easily despair of a successful outcome of any kind. His spirits may be sustained, however, by the thought that these difficulties have nothing to do with statistical methods per se, they are hurdles that confront and have always confronted all experimenters. In spite of them, practical experimentation is frequently successful. How does this come about? The situation appears more hopeful when we remember that any group of experimental runs is usually only *part of an iterative sequence*, and that an investigational strategy should be aimed at the overall furthering of knowledge rather than just the success of any individual group of trials. Our problem is to so organize matters that we are likely in due course to be led to the right conclusions even though our initial choices of the region of interest, the metrics, the transformations, and levels of the input variables may not all be good. Our strategy must be such as will allow any poor initial choices to be rectified as we proceed. Obviously, the path to success is not unique (although it may seem so to the first investigator in a field). It is not *uniqueness* of the path that we should try to achieve, therefore, but rather the probable and rapid convergence of an iterative sequence to the right conclusions.

This iterative process of learning by experience can be roughly formalized. It consists essentially of the successive and repeated use of the sequence

CONJECTURE—DESIGN—EXPERIMENT—ANALYSIS

as illustrated in Figure 1.3. Here, the words "design" and "analysis" do not

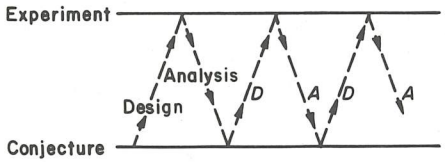

FIGURE 1.3. The iterative nature of experimentation.

refer only to *statistical* design and analysis. By "design," we mean the synthesis of a suitable experiment to test, estimate, and develop a current conjectured model. By "analysis" we mean the treatment of the experimental results, leading either to verification of a postulated model and the working out of its consequences, or to the forming of a new or modified conjecture. These complementary processes are used many times during an investigation and, by their intelligent alternation, the experimenter's knowledge of the system becomes steadily greater.

The indeterminancy of the learning process, but the ultimate possibility of its convergence, is familiarly exemplified by the playing of the game of 20 questions. Ten players engaged independently in playing 20 questions all might succeed eventually in discovering that the object in question was the left ear of the Statue of Liberty but they would almost certainly reach that conclusion by different routes. Let us consider how they do it.

A player may be supplied initially with the information that the object is "mineral." This limits consideration to a certain class of objects which he has to reduce to one as rapidly as possible. When the game is played so that each question concerns a choice between two specific alternatives, the best design of experiment (each "experiment" here is a question asked) is one which will classify the possible objects into two approximately equi-probable groups. A good (but not uniquely good) experiment would thus consist of asking "Is it metal or stone?;" a poor experiment would consist of asking "Is it the stone in my ring?." The experiment having been performed and the data (in the form of an answer) having become available, the player would now analyze the reply in the light of any relevant prior knowledge he possessed and would then form some new or modified conjecture as to the nature of the object in question. To the question, "Is it metal or stone?" the answer "stone" might conjure up in his mind, on the one hand, buildings, monuments, mountains, and so on, and, on the other hand, small stones, both precious (such as diamonds, rubies) and nonprecious (such as pebbles). The player might feel that a good way to discriminate between these two types of objects in the light of information then available would be the question "Is it larger or smaller than this room?." The answer to this question would now result in new analysis which would raise new conjectures and give rise to a new question and so the game would continue.

As another example of iterative investigation, consider a detective employed in solving a mystery. Some data in the form of a body, and the facts that the door was locked on the inside and that Mr. X would benefit from the will, are already known, or have been discovered. These data lead to conjectures on the part of the investigator which in turn lead to certain data-producing actions, or "experiments." As readers of detective novels will realize, the sequence of events which finally leads to the detection of the culprit is by no means unique. Alternative good (but by no means unique) experiments might be a visit to Baker Street subway station to question the ticket collector about a one-armed man with a scar under his left eye, or a visit to the Manor House to find out if the flower bed beneath Lady Cynthia's window shows a tell-tale footprint. The skill of the detective is partially measured by his ability to conceive, at each stage, appropriate conjectures and to select those experiments which, in the light of these conjectures, will best illuminate current aspects of the case. He pursues a strategy which (he hopes) will cause him to follow *one* of the paths leading to the unmasking of the assassin (preferably before the latter has had time to flee the country).

It is through the iterative process of "learning as we go" that the problem of indeterminancy which we mention may be resolved. We shall find that, in the application of the investigative processes described in this book, a multidimensional iteration occurs in which modification in the *location* of the experimental runs, in *scaling*, in *transformation*, and in the *complexity* of the contemplated model, all can occur. While we cannot, and should not, attempt to ensure uniqueness for the investigation, we can ensure that the crucial components of good design and analysis on which convergence depends are organized so as to illuminate and stimulate the ideas of the experimenter as effectively as possible, and so lead him quickly along one (nonunique) path to the truth. Figure 1.4 illustrates this general idea.

The truly iterative nature of some investigations may sometimes be obscured by the length of time taken by each iterative cycle. In this case it may be possible to see the wider iteration only by "stepping back" and examining what occurs over months or years. In this wider context, iteration may skip from one investigator to another, even from one country to another, and its phases may be very long. Even in this situation, however, it is important to bear in mind that the important consideration is the *overall* acquisition of knowledge.

What is being said is not, of course, peculiar to response surface methodology. "Traditional" designs such as randomized blocks, latin squares, and factorial designs have, since their inception, been used by statisticians as building blocks in iterative learning sequences. The possibility of rapid convergence of such a sequence depends, to an important

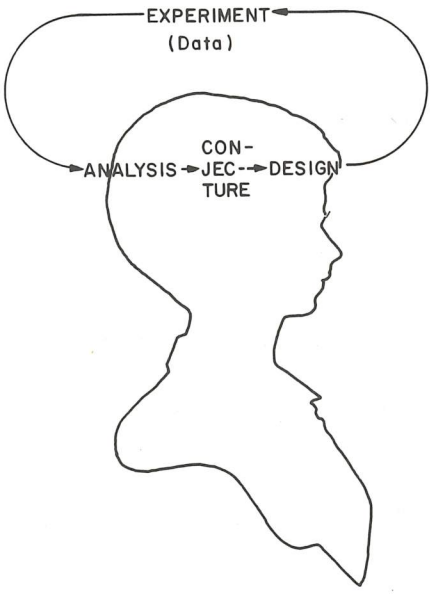

FIGURE 1.4. The iterative experimental process in relation to the experimenter.

extent, on the efficiency of these designs and their associated analyses, and their ability to illuminate and stimulate the ideas of the investigator. This notion was inherent in R. A. Fisher's* general attitude toward the use of these designs in scientific investigation.

1.4. SOME CLASSES OF PROBLEMS (WHICH, HOW, WHY)

Within the foregoing iterative context, we now consider some specific classes of scientific problems and the characteristics of models used in solving these problems.

Mechanistic and Empirical Models

It is helpful in this discussion to distinguish between empirical models and mechanistic models. We consider first what we might mean by a purely mechanistic model. Suppose that, in the study of some physical phenomenon, we know enough of its physical mechanism to *deduce* the form of the

*Sir Ronald Fisher (1890–1962), the famous British statistician, was responsible for many of the basic ideas of experimental design. His books *Statistical Methods for Research Workers* and *The Design of Experiments* are classic.

functional relationship linking the mean value η of the output to the levels $\boldsymbol{\xi}$ of the inputs via an expression

$$E(y) = \eta = f(\boldsymbol{\xi}, \boldsymbol{\theta}), \tag{1.4.1}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)'$ is a set of input variables, measuring, for example, initial concentrations of reactants, temperatures, and pressures, and where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ represents a set of physical parameters measuring such things as activation energies, diffusion coefficients, and thermal conductivities. Then we should say that Eq. (1.4.1) represented a mechanistic model.

In practice some or all of the θ's would need to be estimated from the data. Also, mechanistic knowledge might most naturally be expressed by a set of differential equations or integral equations for which (1.4.1) was the solution.

Often, however, the necessary physical knowledge of the system is absent or incomplete and consequently no mechanistic model is available. In these circumstances, it could often be realistically assumed that the relationship between η and $\boldsymbol{\xi}$ would be smooth and, consequently, that $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ could be locally *approximated* (over limited ranges of the experimental variables $\boldsymbol{\xi}$) by an interpolation function $g(\boldsymbol{\xi}, \boldsymbol{\beta})$, such as a polynomial. In this latter expression, the β's, which are the elements of $\boldsymbol{\beta}$, would be coefficients in the interpolation function. They would be related to, but be distinct from, the parameters (the θ's) of the physical system. The interpolation function $g(\boldsymbol{\xi}, \boldsymbol{\beta})$ could provide a local empirical model for the system and, as we have said, would act simply as a mathematical French curve.

Now the theoretical mechanistic model $\eta = f(\boldsymbol{\xi}, \boldsymbol{\theta})$, and the purely empirical model $\eta \simeq g(\boldsymbol{\xi}, \boldsymbol{\beta})$, as defined above, represent extremes. The former would be appropriate in the extreme case where a great deal was accurately known about the system, and the latter would be appropriate in the other extreme case, where nothing could be assumed except that the response surface was locally smooth. The situation existing in most real investigations is somewhere in between and, as experimentation proceeds, and we gain information, the situation can change. Because real problems occur at almost all points between the extremes mentioned above, a variety of statistical tools is needed to cope with them.

Both the state of ignorance in which we begin our experimental work, and the state of comparative knowledge to which we wish to be brought, will determine our approach. It must be realized, of course, that no real problem ever *quite* fits any prearranged category. With this proviso in mind, it is nevertheless helpful to distinguish the basic *types* of problems shown in Table 1.1. These are categorized in terms of what is unknown about the

TABLE 1.1. Some scientific problems

Supposed unknown	Objective	Descriptive name	Stage
f, $\boldsymbol{\xi}$, $\boldsymbol{\theta}$	Determine the subset $\boldsymbol{\xi}$ of important variables from a given larger set Ξ of potentially important variables	Screening variables	Which
f, $\boldsymbol{\theta}$	Determine empirically the effects of the known input variables $\boldsymbol{\xi}$	Empirical model building	How
f, $\boldsymbol{\theta}$	Determine a local interpolation approximation $g(\boldsymbol{\xi}, \boldsymbol{\beta})$ to $f(\boldsymbol{\xi}, \boldsymbol{\theta})$	(Response surface methodology)	How
f, $\boldsymbol{\theta}$	Determine f	Mechanistic model building	Why
$\boldsymbol{\theta}$	Determine θ	Mechanistic model fitting	Why

"true mechanistic model." For reference purposes we have given each type of problem a descriptive name, as well as a briefer but more pointed "stage" name which makes it clear what stage of investigation is involved.

At the WHICH stage, our object is to determine *which* ones of all the suggested input variables have a significant and important effect upon the response variable. At the HOW stage we would like to learn more about the pattern of the response behavior as changes are made in these important variables. At the WHY stage we attempt to find the mechanistic reasons for the behavior we have observed in the previous stage. (In practice, the stages would usually overlap.)

We now briefly sketch some of the procedures that may be employed to deal with these problems. See also Box, Hunter, and Hunter (1978, Chapters 13–16). Our book will then discuss more fully the use of response surface methods for empirical model building.

Screening (WHICH Stage)

It often happens at the beginning of an investigation that there is rather a long list of variables $\xi_1, \xi_2, \ldots$ which could be of importance in affecting η. One way to reduce the list to manageable size is to sit down with the investigator (the biologist, chemist, psychologist, etc.) and to ask him to pick

out the variables he believes to be most important. To press this too far, however, is dangerous because, not infrequently, a variable initially believed unimportant turns out to have a major effect. A good compromise is to employ a preliminary screening design such as a two-level fractional factorial (see Chapter 5) to pick out variables worthy of further study.

In one investigation, for instance, the original list of variables $\xi_1, \xi_2, \ldots$ that might have affected the response η contained 11 candidates. Three of these were, after careful thought, eliminated as being sufficiently unimportant to be safely ignored. A 16-run two-level fractional factorial design was run on the remaining eight variables and four of the eight were designated as probably influential over the ranges studied. Three of these four had already been selected by the investigator as likely to be critical, confirming his judgement. The fourth was unexpected and turned out to be of great importance. Screening designs are often carried out sequentially in small blocks and are very effective when performed in this way. For additional details see the papers by Box and Hunter (1961a, b, especially pp. 334–337) or see Box, Hunter, and Hunter (1978, Chapter 13).

Empirical Model Building (HOW Stage)

When input variables are quantitative, and the experimental error is not too large compared with the range covered by the observed responses (see Section 8.2), it may be profitable to attempt to *estimate* the response function within some area of immediate interest. In many problems, the form of the true response function $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ is unknown and cannot economically be obtained but may be capable of being locally approximated by a polynomial or some other type of graduating function, $g(\boldsymbol{\xi}, \boldsymbol{\beta})$, say. Suitable experimental designs for this purpose have been developed. The essentially iterative nature of *response surface methodology* (RSM) would ensure that, as the investigation proceeded, it would be possible to learn about (a) the amount of replication needed to achieve sufficient precision, (b) the location of the experimental region of most interest, (c) appropriate scalings and transformations for the input and output variables, and (d) the degree of complexity of an approximating function, and hence of the designs, needed at various stages. These matters are discussed in their Chapters 14 and 15 by Box, Hunter, and Hunter (1978), and, more fully, in this book.

Mechanistic Model Building (WHY Stage)

If it were possible, we might like to use the *true* functional form $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ to represent the response, rather than to approximate it by a graduating

function. In some problems, we can hope to achieve useful working mechanistic models which, at least, take account of the *principal* features of the mechanism. These models often are most naturally expressed in terms of differential equations or other nonexplicit forms, but modern developments in computing facilities and in the theory of nonlinear design and estimation have made it possible to cope with the resulting problems. A mechanistic model has the following advantages:

1. It contributes to our scientific understanding of the phenomenon under study.
2. It usually provides a better basis for extrapolation (at least to conditions worthy of further experimental investigation if not through the entire ranges of all input variables).
3. It tends to be parsimonious (i.e., frugal) in the use of parameters and to provide better estimates of the response.

Results from fitting mechanistic models have sometimes been disappointing because not enough attention has been given to discovering what *is* an appropriate model form. It is easy to collect data that never "place the postulated model in jeopardy" and so it is common (e.g., in chemical engineering) to find different research groups each advocating a different model for the same phenomenon and each proffering data that "prove" their claim. In such cases, methods that discriminate between the various candidate models must be applied. See Box, Hunter, and Hunter (1978, Chapter 16) for additional discussion.

Comment

As we have said, the lines between the WHICH, the HOW, and the WHY stages of investigation are often ill-defined in practice. Thus, toward the end of a (WHICH) screening investigation, for example, we may employ a fractional factorial design that is then used as the first of a series of building blocks in a response surface study (HOW). In other circumstances (see, e.g., Box and Youle, 1955) careful study of an empirical fitted response surface at the HOW stage can generate ideas about the possible underlying mechanism (WHY). The present text is concerned mostly with empirical model building, that is, the HOW stage of experimentation, and specifically with what has become to be called *response surface methodology* or RSM. The various aspects of RSM will be dealt with in the chapters that follow.

1.5. NEED FOR EXPERIMENTAL DESIGN

It sometimes happens (e.g., in investigations of industrial plant processes) that a large amount of past operating data is available. It may then be urged that no experimentation is actually needed because it ought to be possible to extract information relating changes in a response of interest to changes that have occurred *naturally* in the input variables. Such investigations are often valuable as preliminary studies, but the existence of such data rarely eliminates the need for further planned experimentation. There are several reasons for this. For example:

1. Important input variables affecting the response are often the very ones that are not varied.
2. Relations between the response variable and various input variables may be induced by unrecorded "lurking" variables that affect both the response and the input variables. These can give rise to "nonsense correlations."
3. Historical operating data often contain gaps and omit important ancillary information.

Matters such as these are more fully discussed in Box, Hunter, and Hunter (1978, pp. 487–498).

1.6. GEOMETRIC REPRESENTATION OF RESPONSE RELATIONSHIPS

Our ability to think about systems with multiple inputs and outputs is greatly enhanced by geometrical representation and, in particular, by the use of contour diagrams. Some examples are given in Figure 1.5. The curves of Figure 1.5*a*, for example, represent a relationship between yield of product η, and reaction ξ_1 at three different levels of catalyst concentration ξ_2.

In Figure 1.5*b*, a three-dimensional representation is given, and the three curves in Figure 1.5*a* are now sections of the response surface in Figure 1.5*b*.

If we were to slice off portions of the response surface at various horizontal levels of the mean yield, η, and then project the outlines of the slices onto the (ξ_1, ξ_2) base of Figure 1.5*b*, we would obtain the *contours* (or *contour lines*) in Figure 1.5*c*. We call Figure 1.5*c* a *contour diagram*. A

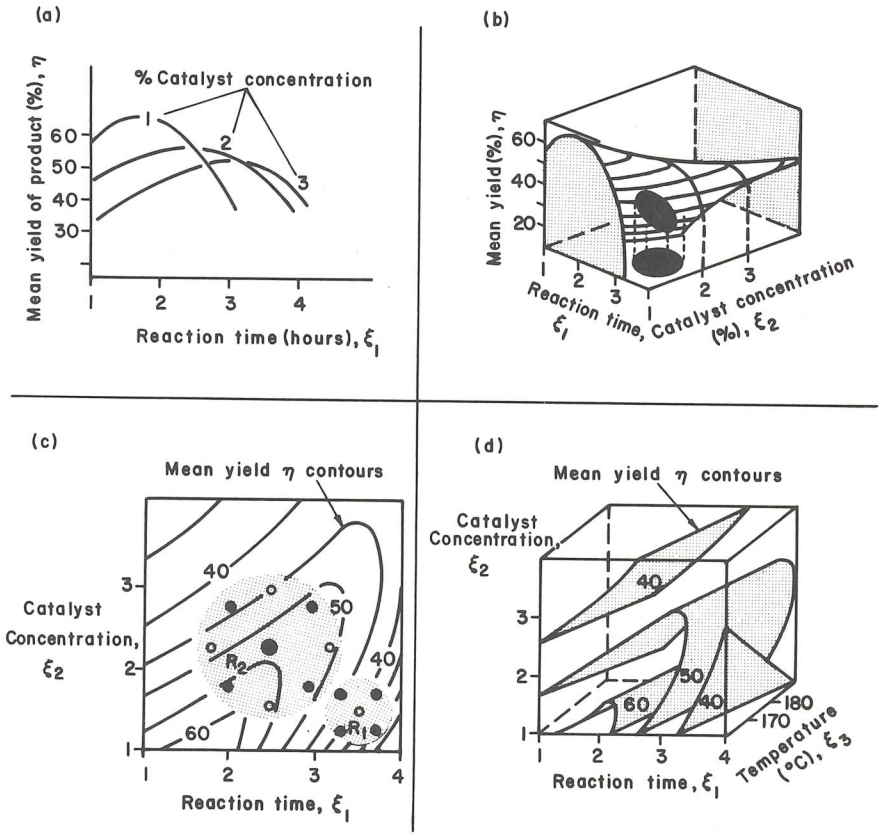

FIGURE 1.5. Geometrical representations of a relationship connecting the mean yield of product (%), η, with reaction time, ξ_1, catalyst concentration (%), ξ_2, and temperature (°C), ξ_3.

contour diagram has the advantage that a relationship between a response η and two predictors ξ_1 and ξ_2 (i.e., a relationship that involves three variables) can be represented in only two-dimensional space. Figure 1.5*d* is also a contour diagram but one in which a third dimension, representing a third predictor variable, temperature, has been added, and three-dimensional *contour surfaces* drawn. If we took a section, at a selected, fixed temperature, of the three-dimensional representation in Figure 1.5*d*, we would obtain a two-dimensional contour diagram of the form of Figure 1.5*c*. We see that, with only *three* dimensions at our disposal, we can actually think about the relationships between *four* variables η, ξ_1, ξ_2, ξ_3. Because the contour diagrams are drawn in the space of the input variables,

the ξ's, each potential experimental run will appear simply as a point in this input space. A special selected pattern of points chosen to investigate a response function relationship is called an *experimental design*. Two such experimental design patterns are shown in the regions R_1 and R_2 in Figure 1.5*c*.

1.7. THREE KINDS OF APPLICATIONS

Response surface methods have proved valuable in the solution of a large variety of problems. We now outline three common classes of such problems using an industrial example for illustration.

A particular type of extruded plastic film should possess a number of properties, one of which is high transparency. This property, regarded as a response η_1, will be affected by such input variables as the screw speed of extrusion ξ_1, and the temperature of the extruder barrel, ξ_2.

1. Approximate Mapping of a Surface Within a Limited Region

Suppose we were exploring the capability of an extruding machine used in actual manufacture. Such a machine would usually be designed to run at a particular set of conditions, but moderate adjustments in the speed ξ_1, and temperature ξ_2, might be possible. If the unknown response function, $f(\xi_1, \xi_2)$ could be graduated by, say, a polynomial $g(\xi_1, \xi_2)$ in the limited region $R(\xi_1, \xi_2)$ over which the machine could be run, we could approximately predict the transparency of the product for any specified adjustment of the machine.

2. Choice of Operating Conditions to Achieve Desired Specifications

It frequently happens that we are interested in more than one property of the product. For example, in the case of the extruded film, we might be interested in tear strength η_2 in addition to transparency η_1.

It is convenient to visualize the situation by thinking of the contours of η_1 and η_2 superimposed in the ξ_1 and ξ_2 space. (See Figure 1.6) For clarity, we can think of the contours of η_1 as being drawn in red ink and those of η_2 drawn in green ink. Actual diagrams and models in which two or more sets of contours are superimposed are in fact often used in elucidating possibilities for industrial systems.

Suppose we desired a high transparency $\eta_1 > \eta_{10}$ and a high tear strength $\eta_2 > \eta_{20}$ where η_{10}, η_{20} are specified values. Then, if the responses

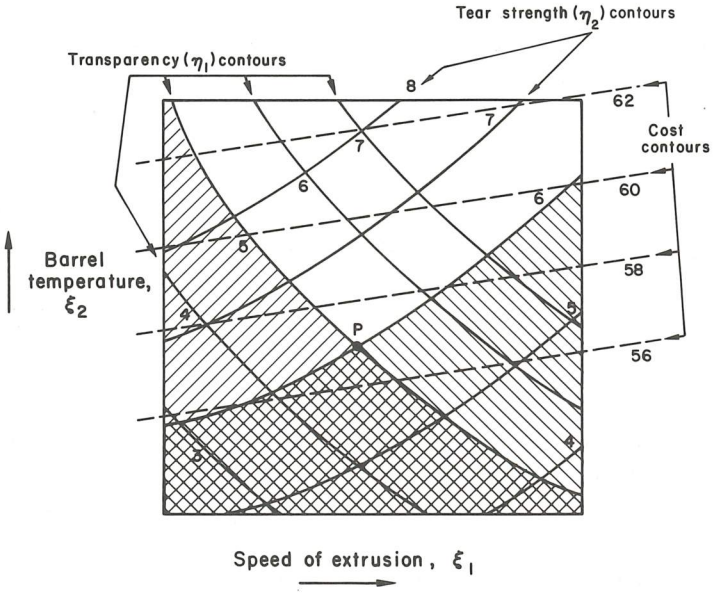

FIGURE 1.6. The unshaded region contains values (ξ_1, ξ_2) in which the desired product with transparency $\eta_1 > 5$ and tear strength $\eta_2 > 6$ is obtained.

transparency and tear strength were known to be graduated, respectively, by

$$\eta_1 = g_1(\xi_1, \xi_2), \qquad \eta_2 = g_2(\xi_1, \xi_2),$$

we could draw the approximate critical contours $\eta_1 = \eta_{10}, \eta_2 = \eta_{20}$ on the same figure and hence determine the region in the (ξ_1, ξ_2) space in which *both* inequalities were satisfied and so *both* responses attained satisfactory values. Figure 1.6 illustrates how this might be done. The shading indicates the regions in the space of the input variables in which an undesirable product would be obtained. Transparencies greater than 5 and tear strengths greater than 6 can be obtained within the unshaded region.

In practice, of course, there may be more than two predictors and/or more than two response variables. One response that would often need to be examined is the overall average cost associated with any particular set of operating conditions. Thus we might seek to obtain a product which satisfied the specification $\xi_1 > \xi_{10}$, $\xi_2 > \xi_{20}$ *and* had the smallest possible cost η_3. We see from Figure 1.6 that manufacture at conditions corresponding to the point P would just satisfy the specifications and at the same time, lead to a minimum cost.

The formal description for the problem above has many aspects closely related to linear and nonlinear programming problems. The situation we consider here however, is more complicated, because the relationships between the η's and ξ's are not given but must be estimated.

3. Search for Optimal Conditions

In the examples above, we imply that reasonably satisfactory manufacturing conditions have already been determined. However, at an earlier stage of process development, work would have been carried out in the laboratory and pilot plant to determine suitable conditions. Frequently, there would be a large region in the space of the variables ξ_1 and ξ_2 where the process was operable. An important problem is that of finding, within this operability region, the *best* operating conditions. Let us suppose that the merit of any particular set (ξ_1, ξ_2) of input conditions can be assessed in terms of a response η representing, say, an overall measure of profitability. The problem would be that of locating, and determining the characteristics of, the optimum conditions.

In the chapters that follow, we shall discuss further the details of response surface methodology and how the various applications of it can be made.