11 Response Surface Methods and Other Approaches to Process Optimization

11-1 INTRODUCTION TO RESPONSE SURFACE METHODOLOGY

Response surface methodology, or **RSM**, is a collection of mathematical and statistical techniques that are useful for the modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response. For example, suppose that a chemical engineer wishes to find the levels of temperature (x_1) and pressure (x_2) that maximize the yield (y) of a process. The process yield is a function of the levels of temperature and pressure, say

$$y = f(x_1, x_2) + \epsilon$$

where ϵ represents the noise or error observed in the response y. If we denote the expected response by $E(y) = f(x_1, x_2) = \eta$, then the surface represented by

$$\eta = f(x_1, x_2)$$

is called a **response surface.**

We usually represent the response surface graphically, such as in Figure 11-1 (on the next page), where η is plotted versus the levels of x_1 and x_2. We have seen response surface plots such as this before, particularly in the chapters on factorial designs. To help visualize the shape of a response surface, we often plot the contours of the response surface as shown in Figure 11-2 (on the next page). In the contour plot, lines of constant response are drawn in the x_1, x_2 plane. Each contour corresponds to a particular height of the response surface. We have also previously seen the utility of contour plots.

In most RSM problems, the form of the relationship between the response and the independent variables is unknown. Thus, the first step in RSM is to find a suitable approximation for the true functional relationship between y and the set of independent variables. Usually, a low-order polynomial in some region of the independent variables is employed. If the response is well modeled by a linear function of the independent variables, then the approximating function is the **first-order model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (11\text{-}1)$$

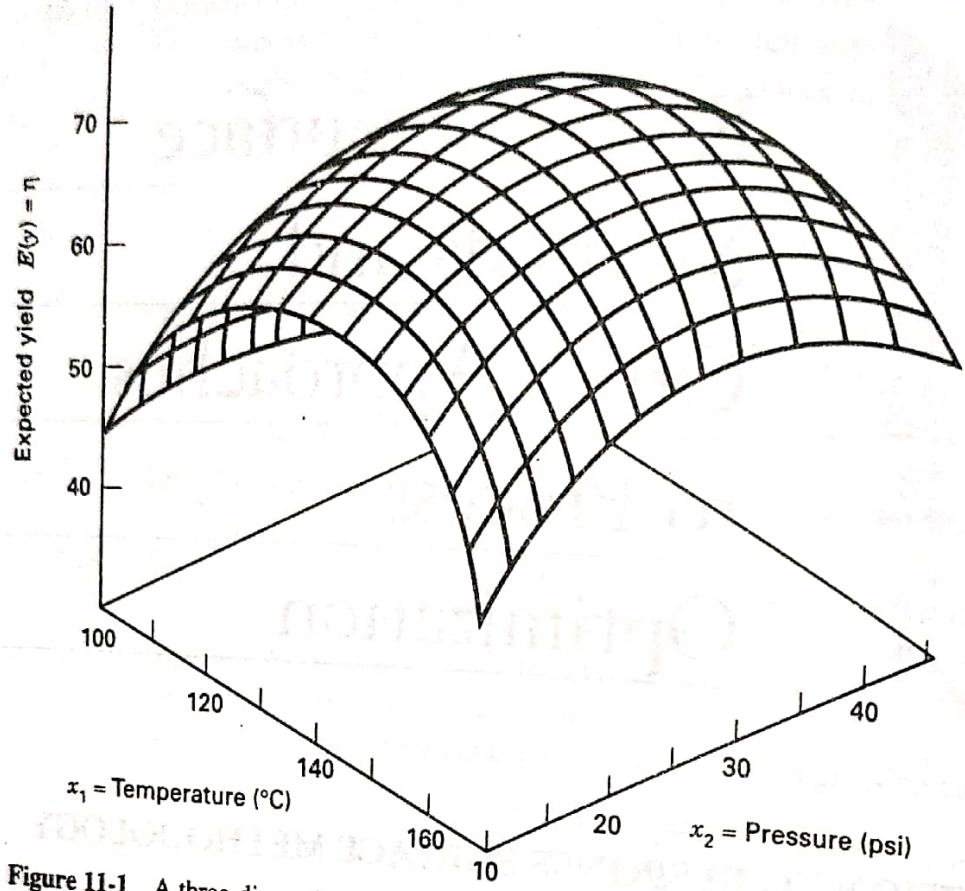

Figure 11-1 A three-dimensional response surface showing the expected yield (η) as a function of temperature (x_1) and pressure (x_2).

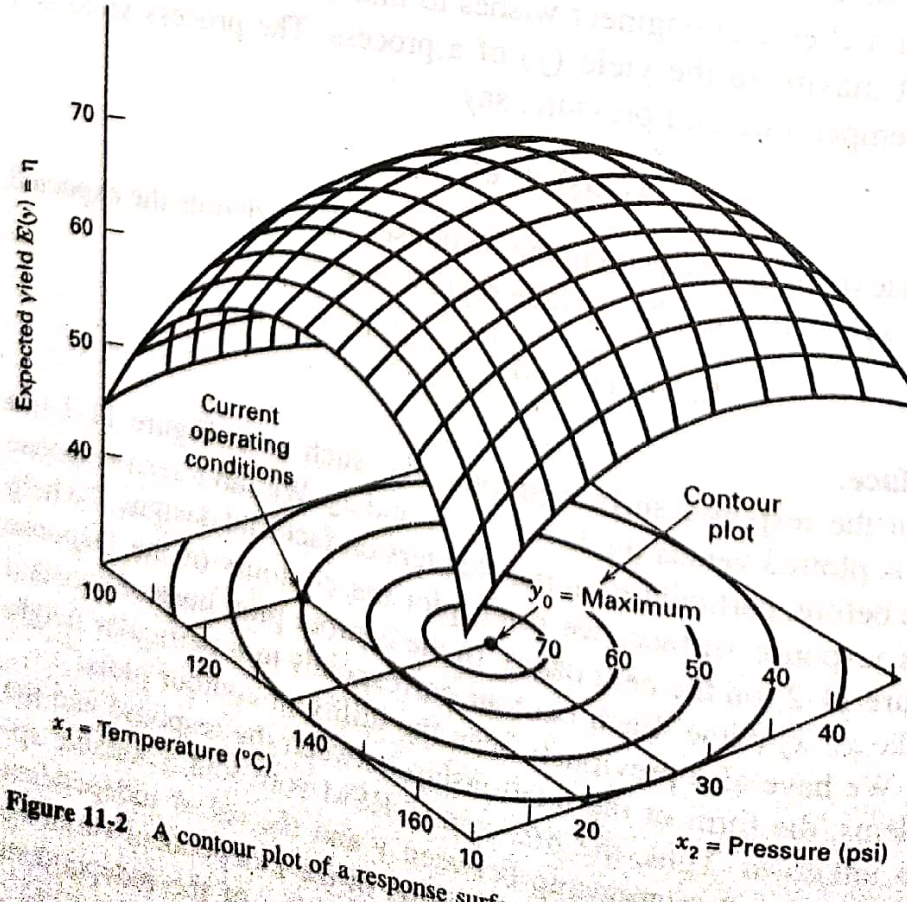

Figure 11-2 A contour plot of a response surface.

If there is curvature in the system, then a polynomial of higher degree must be used, such as the **second-order model**

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum\sum_{i<j} \beta_{ij} x_i x_j + \epsilon \qquad (11\text{-}2)$$

Almost all RSM problems use one or both of these models. Of course, it is unlikely that a polynomial model will be a reasonable approximation of the true functional relationship over the entire space of the independent variables, but for a relatively small region they usually work quite well.

The method of least squares, discussed in Chapter 10, is used to estimate the parameters in the approximating polynomials. The response surface analysis is then performed using the fitted surface. If the fitted surface is an adequate approximation of the true response function, then analysis of the fitted surface will be approximately equivalent to analysis of the actual system. The model parameters can be estimated most effectively if proper experimental designs are used to collect the data. Designs for fitting response surfaces are called **response surface designs**. These designs are discussed in Section 11-4.

RSM is a **sequential procedure**. Often, when we are at a point on the response surface that is remote from the optimum, such as the current operating conditions in Figure 11-3, there is little curvature in the system and the first-order model will be appropriate. Our objective here is to lead the experimenter rapidly and efficiently along a path of improvement toward the general vicinity of the **optimum**. Once the region of the optimum has been found, a more elaborate model, such as the second-order model, may be employed, and an analysis may be performed to locate the optimum. From Figure 11-3, we see that the analysis of a response surface can be thought of as "climbing a hill," where the top of the hill represents the point of maximum response. If the true optimum is a point of minimum response, then we may think of "descending into a valley."

The eventual objective of RSM is to determine the optimum operating conditions for the system or to determine a region of the factor space in which operating require-

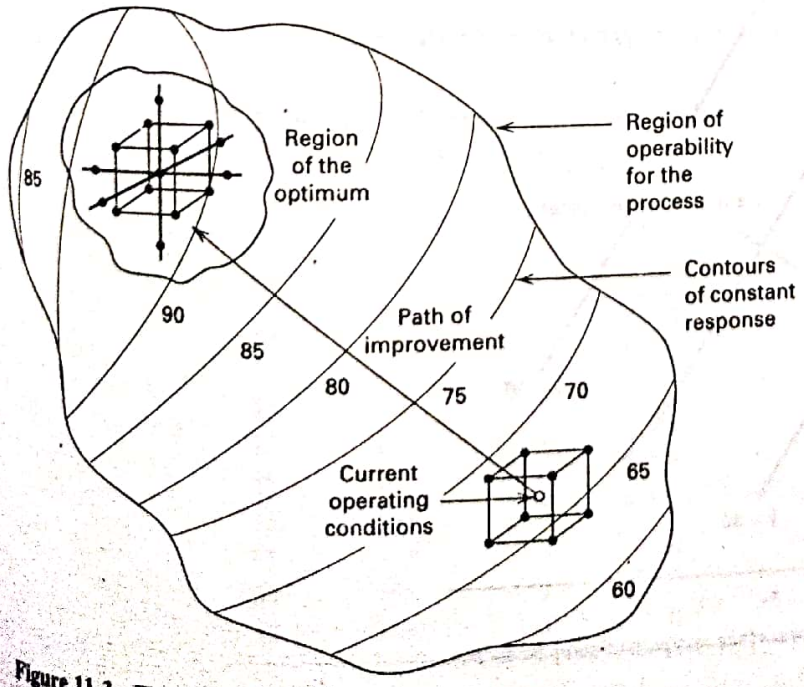

Figure 11-3 The sequential nature of RSM.

ments are satisfied. More extensive presentations of RSM are in Myers and Montgomery (1995), Khuri and Cornell (1996), and Box and Draper (1987).

11-2 THE METHOD OF STEEPEST ASCENT

Frequently, the initial estimate of the optimum operating conditions for the system will be far from the actual optimum. In such circumstances, the objective of the experimenter is to move rapidly to the general vicinity of the optimum. We wish to use a simple and economically efficient experimental procedure. When we are remote from the optimum, we usually assume that a first-order model is an adequate approximation to the true surface in a small region of the x's.

The **method of steepest ascent** is a procedure for moving sequentially along the path of steepest ascent, that is, in the direction of the maximum increase in the response. Of course, if minimization is desired, then we call this technique the **method of steepest descent**. The fitted first-order model is

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i \tag{11-3}$$

and the first-order response surface, that is, the contours of $\hat{y}$, is a series of parallel lines such as that shown in Figure 11-4. The direction of steepest ascent is the direction in which $\hat{y}$ increases most rapidly. This direction is parallel to the normal to the fitted response surface. We usually take as the **path of steepest ascent** the line through the center of the region of interest and normal to the fitted surface. Thus, the steps along the

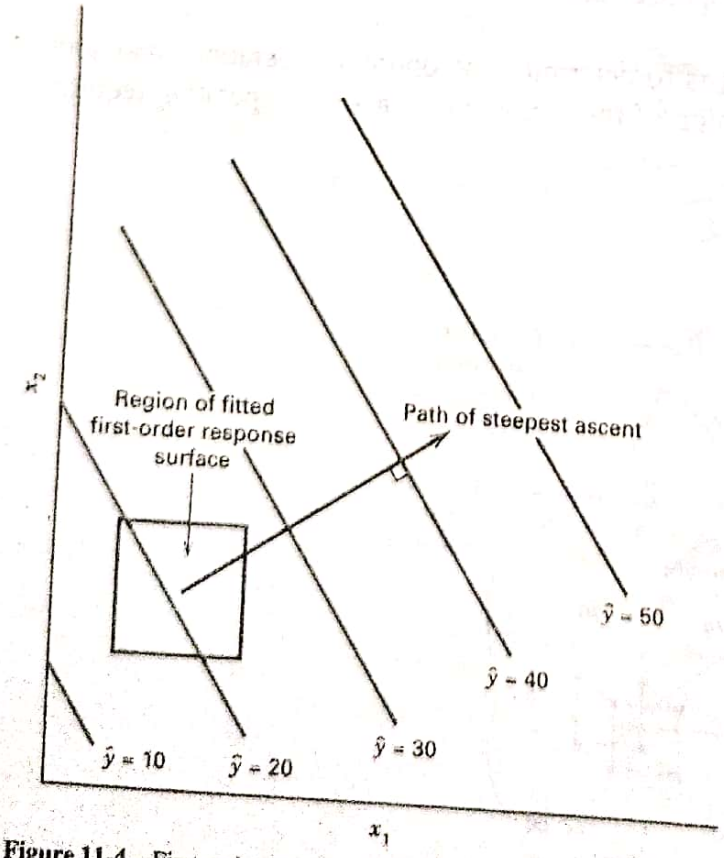

Figure 11-4 First-order response surface and path of steepest ascent.

path are proportional to the regression coefficients $\{\hat{\beta}_i\}$. The actual step size is determined by the experimenter based on process knowledge or other practical considerations.

Experiments are conducted along the path of steepest ascent until no further increase in response is observed. Then a new first-order model may be fit, a new path of steepest ascent determined, and the procedure continued. Eventually, the experimenter will arrive in the vicinity of the optimum. This is usually indicated by lack of fit of a first-order model. At that time additional experiments are conducted to obtain a more precise estimate of the optimum.

EXAMPLE 11-1

A chemical engineer is interested in determining the operating conditions that maximize the yield of a process. Two controllable variables influence process yield: reaction time and reaction temperature. The engineer is currently operating the process with a reaction time of 35 minutes and a temperature of 155°F, which result in yields of around 40 percent. Because it is unlikely that this region contains the optimum, she fits a first-order model and applies the method of steepest ascent.

The engineer decides that the region of exploration for fitting the first-order model should be (30, 40) minutes of reaction time and (150, 160)°F. To simplify the calculations, the independent variables will be coded to the usual $(-1, 1)$ interval. Thus, if ξ_1 denotes the **natural variable** time and ξ_2 denotes the **natural variable** temperature, then the coded variables are

$$x_1 = \frac{\xi_1 - 35}{5} \qquad \text{and} \qquad x_2 = \frac{\xi_2 - 155}{5}$$

The experimental design is shown in Table 11-1. Note that the design used to collect this data is a 2^2 factorial augmented by five center points. Replicates at the center are used to estimate the experimental error and to allow for checking the adequacy of the first-order model. Also, the design is centered about the current operating conditions for the process.

A first-order model may be fit to these data by least squares. Following the methods for two-level designs, we obtain the following model in the coded variables:

$$\hat{y} = 40.44 + 0.775x_1 + 0.325x_2$$

Table 11-1 Process Data for Fitting the First-Order Model

Natural Variables		Coded Variables		Response
ξ_1	ξ_2	x_1	x_2	y
30	150	−1	−1	39.3
30	160	−1	1	40.0
40	150	1	−1	40.9
40	160	1	1	41.5
35	155	0	0	40.3
35	155	0	0	40.5
35	155	0	0	40.7
35	155	0	0	40.2
35	155	0	0	40.6

Before exploring along the path of steepest ascent, the adequacy of the first-order model should be investigated. The 2^2 design with center points allows the experimenter to

1. Obtain an estimate of error
2. Check for interactions (cross-product terms) in the model
3. Check for quadratic effects (curvature)

The replicates at the center can be used to calculate an estimate of error as follows:

$$\hat{\sigma}^2 = \frac{(40.3)^2 + (40.5)^2 + (40.7)^2 + (40.2)^2 + (40.6)^2 - (202.3)^2/5}{4}$$

$$= 0.0430$$

The first-order model assumes that the variables x_1 and x_2 have an **additive effect** on the response. Interaction between the variables would be represented by the coefficient β_{12} of a cross-product term x_1x_2 added to the model. The least squares estimate of this coefficient is just one-half the interaction effect calculated as in an ordinary 2^2 factorial design, or

$$\hat{\beta}_{12} = \tfrac{1}{4}[(1 \times 39.3) + (1 \times 41.5) + (-1 \times 40.0) + (-1 \times 40.9)]$$
$$= \tfrac{1}{4}(-0.1)$$
$$= -0.025$$

The single degree of freedom sum of squares for interaction is

$$SS_{\text{Interaction}} = \frac{(-0.1)^2}{4}$$
$$= 0.0025$$

Comparing $SS_{\text{Interaction}}$ to $\hat{\sigma}^2$ gives a lack-of-fit statistic

$$F = \frac{SS_{\text{Interaction}}}{\hat{\sigma}^2}$$
$$= \frac{0.0025}{0.0430}$$
$$= 0.058$$

which is small, indicating that interaction is negligible.

Another check of the adequacy of the straight-line model is obtained by applying the check for pure quadratic curvature effect described in Section 6-6. Recall that this consists of comparing the average response at the four points in the factorial portion of the design, say $\bar{y}_F = 40.425$, with the average response at the design center, say $\bar{y}_C = 40.46$. If there is quadratic curvature in the true response function, then $\bar{y}_F - \bar{y}_C$ is a measure of this curvature. If β_{11} and β_{22} are the coefficients of the "pure quadratic" terms x_1^2 and x_2^2, then $\bar{y}_F - \bar{y}_C$ is an estimate of $\beta_{11} + \beta_{22}$. In our example, an estimate of the pure quadratic term is

$$\hat{\beta}_{11} + \hat{\beta}_{22} = \bar{y}_F - \bar{y}_C$$
$$= 40.425 - 40.46$$
$$= -0.035$$

Table 11-2 Analysis of Variance for the First-Order Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Model (β_1, β_2)	2.8250	2	1.4125	47.83	0.0002
Residual	0.1772	6			
(Interaction)	(0.0025)	1	0.0025	0.058	0.8215
(Pure quadratic)	(0.0027)	1	0.0027	0.063	0.8142
(Pure error)	(0.1720)	4	0.0430		
Total	3.0022	8			

The single-degree-of-freedom sum of squares associated with the null hypothesis, $H_0: \beta_{11} + \beta_{22} = 0$, is

$$SS_{\text{Pure Quadratic}} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C} = \frac{(4)(5)(-0.035)^2}{4 + 5} = 0.0027$$

where n_F and n_C are the number of points in the factorial portion and the number of center points, respectively. Because

$$F = \frac{SS_{\text{Pure Quadratic}}}{\hat{\sigma}^2} = \frac{0.0027}{0.0430} = 0.063$$

is small, there is no indication of a pure quadratic effect.

The analysis of variance for this model is summarized in Table 11-2. Both the interaction and curvature checks are not significant, whereas the F test for the overall regression is significant. Furthermore, the standard error of $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$se(\hat{\beta}_i) = \sqrt{\frac{MS_E}{4}} = \sqrt{\frac{\hat{\sigma}^2}{4}} = \sqrt{\frac{0.0430}{4}} = 0.10 \qquad i = 1, 2$$

Both regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are large relative to their standard errors. At this point, we have no reason to question the adequacy of the first-order model.

To move away from the design center—the point ($x_1 = 0, x_2 = 0$)—along the path of steepest ascent, we would move 0.775 units in the x_1 direction for every 0.325 units in the x_2 direction. Thus, the path of steepest ascent passes through the point ($x_1 = 0$, $x_2 = 0$) and has a slope 0.325/0.775. The engineer decides to use 5 minutes of reaction time as the basic step size. Using the relationship between ξ_1 and x_1, we see that 5 minutes of reaction time is equivalent to a step in the *coded* variable x_1 of $\Delta x_1 = 1$. Therefore, the steps along the path of steepest ascent are $\Delta x_1 = 1.0000$ and $\Delta x_2 = (0.325/0.775)\Delta x_1 = 0.42$.

The engineer computes points along this path and observes the yields at these points until a decrease in response is noted. The results are shown in Table 11-3 (on the next page) in both coded and natural variables. Although the coded variables are easier to manipulate mathematically, the natural variables must be used in running the process.

Table 11-3 Steepest Ascent Experiment for Example 11-1

Steps	Coded Variables		Natural Variables		Response
	x_1	x_2	ξ_1	ξ_2	y
Origin	0	0	35	155	
Δ	1.00	0.42	5	2	
Origin + Δ	1.00	0.42	40	157	
Origin + 2Δ	2.00	0.84	45	159	41.0
Origin + 3Δ	3.00	1.26	50	161	42.9
Origin + 4Δ	4.00	1.68	55	163	47.1
Origin + 5Δ	5.00	2.10	60	165	49.7
Origin + 6Δ	6.00	2.52	65	167	53.8
Origin + 7Δ	7.00	2.94	70	169	59.9
Origin + 8Δ	8.00	3.36	75	171	65.0
Origin + 9Δ	9.00	3.78	80	173	70.4
Origin + 10Δ	10.00	4.20	85	175	77.6
Origin + 11Δ	11.00	4.62	90	179	80.3
Origin + 12Δ	12.00	5.04	95	181	76.2
					75.1

Figure 11-5 plots the yield at each step along the path of steepest ascent. Increases in response are observed through the tenth step; however, all steps beyond this point result in a decrease in yield. Therefore, another first-order model should be fit in the general vicinity of the point ($\xi_1 = 85$, $\xi_2 = 175$).

A new first-order model is fit around the point ($\xi_1 = 85$, $\xi_2 = 175$). The region of exploration for ξ_1 is [80, 90], and for ξ_2 it is [170, 180]. Thus, the coded variables are

$$x_1 = \frac{\xi_1 - 85}{5} \quad \text{and} \quad x_2 = \frac{\xi_2 - 175}{5}$$

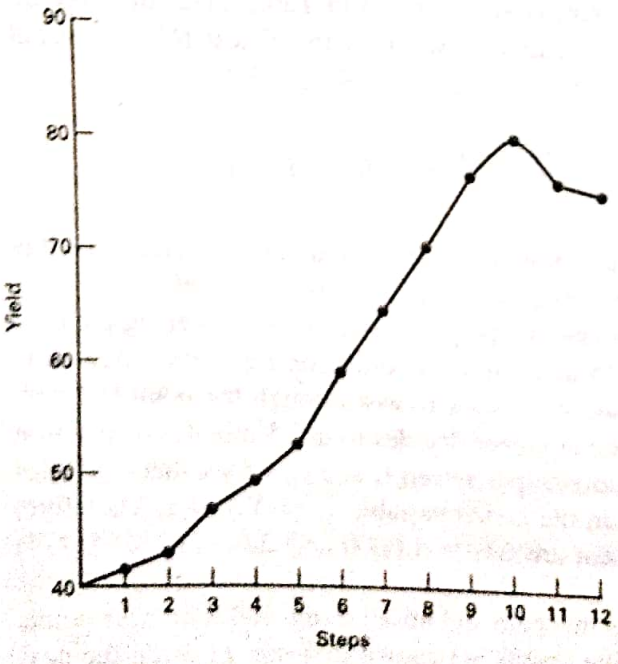

Figure 11-5 Yield versus steps along the path of steepest ascent for Example 11-1.

Table 11-4 Data for Second First-Order Model

Natural Variables		Coded Variables		Response
ξ_1	ξ_2	x_1	x_2	y
80	170	−1	−1	76.5
80	180	−1	1	77.0
90	170	1	−1	78.0
90	180	1	1	79.5
85	175	0	0	79.9
85	175	0	0	80.3
85	175	0	0	80.0
85	175	0	0	79.7
85	175	0	0	79.8

Once again, a 2^2 design with five center points is used. The experimental design is shown in Table 11-4.

The first-order model fit to the coded variables in Table 11-4 is

$$\hat{y} = 78.97 + 1.00x_1 + 0.50x_2$$

The analysis of variance for this model, including the interaction and pure quadratic term checks, is shown in Table 11-5. The interaction and pure quadratic checks imply that the first-order model is not an adequate approximation. This curvature in the true surface may indicate that we are near the optimum. At this point, additional analysis must be done to locate the optimum more precisely.

From Example 11-1 we notice that the *path of steepest ascent is proportional to the signs and magnitudes of the regression coefficients* in the fitted first-order model

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i$$

It is easy to give a general algorithm for determining the coordinates of a point on the path of steepest ascent. Assume that the point $x_1 = x_2 = \ldots = x_k = 0$ is the base or origin point. Then

1. Choose a step size in one of the process variables, say Δx_j. Usually, we would select the variable we know the most about, or we would select the variable that has the largest absolute regression coefficient $|\hat{\beta}_j|$.

Table 11-5 Analysis of Variance for the Second First-Order Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Regression	5.00	2			
Residual	11.1200	6			
(Interaction)	(0.2500)	1	0.2500	4.72	0.0955
(Pure quadratic)	(10.6580)	1	10.6580	201.09	0.0001
(Pure error)	(0.2120)	4	0.0530		
Total	16.1200	8			

11-4 EXPERIMENTAL DESIGNS FOR FITTING RESPONSE SURFACES

Fitting and analyzing response surfaces is greatly facilitated by the proper choice of an experimental design. In this section, we discuss some aspects of selecting appropriate designs for fitting response surfaces.

When selecting a response surface design, some of the features of a desirable design are as follows:

1. Provides a reasonable distribution of data points (and hence information) throughout the region of interest
2. Allows model adequacy, including lack of fit, to be investigated
3. Allows experiments to be performed in blocks
4. Allows designs of higher order to be built up sequentially
5. Provides an internal estimate of error
6. Provides precise estimates of the model coefficients
7. Provides a good profile of the prediction variance throughout the experimental region
8. Provides reasonable robustness against outliers or missing values
9. Does not require a large number of runs
10. Does not require too many levels of the independent variables
11. Ensures simplicity of calculation of the model parameters

These features are sometimes conflicting, so judgment must often be applied in design selection. For more information on the choice of a response surface design, refer to Myers and Montgomery (1995), Box and Draper (1987), and Khuri and Cornell (1996).

11-4.1 Designs for Fitting the First-Order Model

Suppose we wish to fit the first-order model in k variables

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon \tag{11-14}$$

There is a unique class of designs that minimize the variance of the regression coefficients $\{\hat{\beta}_i\}$. These are the **orthogonal first-order designs**. A first-order design is orthogonal if the off-diagonal elements of the $(\mathbf{X}'\mathbf{X})$ matrix are all zero. This implies that the cross-products of the columns of the $\mathbf{X}$ matrix sum to zero.

The class of orthogonal first-order designs includes the 2^k factorial and fractions of the 2^k series in which main effects are not aliased with each other. In using these designs, we assume that the low and high levels of the k factors are coded to the usual ± 1 levels.

The 2^k design does not afford an estimate of the experimental error unless some runs are replicated. A common method of including replication in the 2^k design is to augment the design with several observations at the center (the point $x_i = 0, i = 1, 2, \ldots, k$). The addition of center points to the 2^k design does not influence the $\{\hat{\beta}_i\}$ for $i \geq 1$, but the estimate of β_0 becomes the grand average of all observations. Furthermore, the addition of center points does not alter the orthogonality property of the design. Example 11-1

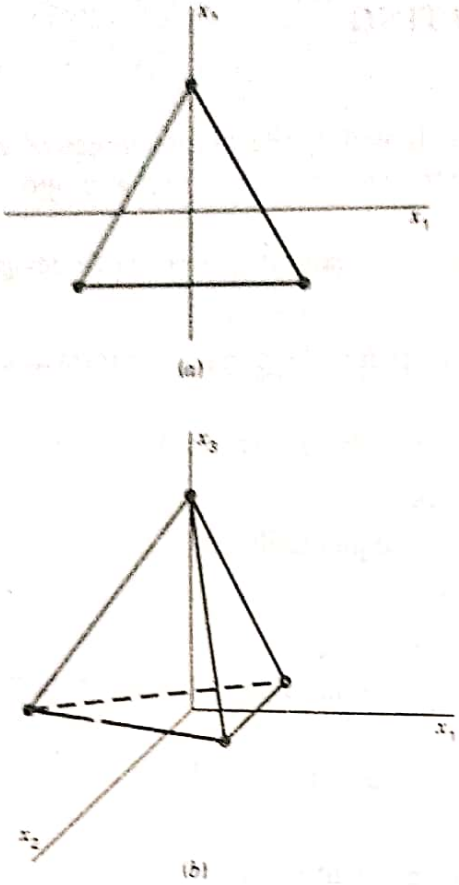

Figure 11-19 The simplex design for (a) $k = 2$ variables and (b) $k = 3$ variables.

illustrates the use of a 2^2 design augmented with five center points to fit a first-order model.

Another orthogonal first-order design is the **simplex**. The simplex is a regularly sided figure with $k + 1$ vertices in k dimensions. Thus, for $k = 2$ the simplex design is an equilateral triangle and for $k = 3$ it is a regular tetrahedron. Simplex designs in two and three dimensions are shown in Figure 11-19.

11-4.2 Designs for Fitting the Second-Order Model

We have informally introduced in Example 11-2 (and even earlier, in Example 6-6) the **central composite design** or CCD for fitting a second-order model. This is the most popular class of designs used for fitting these models. Generally, the CCD consists of a 2^k factorial (or fractional factorial of resolution V) with n_F runs, $2k$ axial or star runs, and n_C center runs. Figure 11-20 (on the facing page) shows the CCD for $k = 2$ and $k = 3$ factors.

The practical deployment of a CCD often arises through **sequential experimentation**, as in Examples 11-1 and 11-2. That is, a 2^k has been used to fit a first-order model, this model has exhibited lack of fit, and the axial runs are then added to allow the quadratic terms to be incorporated into the model. The CCD is a very efficient design for fitting the second-order model. There are two parameters in the design that must be specified: the distance α of the axial runs from the design center and the number of center points n_C. We now discuss the choice of these two parameters.

CHAPTER 7

Designs for Bio-assays and Response Surfaces

7.1 BIO-ASSAYS

In certain investigations it is necessary to compare the efficacy of two or more substances in respect of some of their common effects. Such comparisons are not possible by comparing the effects of individual doses of the substances. The techniques used in bio-assays are designed for such comparisons. Bio-assays are thus a type of experiments with the object of comparing the efficacy of two or more substances, or preparations, like drugs, by using responses produced by them on suitable living organisms. This technique is used more in pharmacological investigations for comparing the potency of two or more preparations of individual drugs.

Normally, two preparations having a common effect are taken for assaying. One of the preparations is of known strength and is called the standard preparation and the other is of unknown strength and is called test preparation. The objective of the assay is to estimate the potency of the test preparation relative to that of the standard preparation. The potency of the test preparation is defined to be the ratio of two doses, one from the standard preparation and the other from the test preparation such that each of them produces the same response. This response is in form of some of their common effects when these are applied to suitable living animals or other organisms.

Let z_s and z_t denote the doses of the standard and the test preparations respectively such that each of them produces a pre-assigned response in some living organism. Then the ratio

$$\rho = \frac{z_s}{z_t}$$

is called the *relative potency* of the test preparation. If ρ is greater than unity, it shows that a smaller dose of the test preparation produces as much response as a relatively larger dose of the standard preparation and hence the potency of the test preparation is greater than that of the standard preparation. Similarly when ρ is less than 1 the potency of the test preparation is smaller than that of the standard preparation.

An assay with two preparations containing the same effective ingredient which is responsible for the response, is called *analytical dillusion assay.*

An assay with two preparations which have a common effect but do not contain the same effective ingredient, is called a comparative dillusion *assay*. The results obtained from analytical dillusion assays hold in general and are not necessarily limited only to experimental conditions, while the results obtained from comparative dillusion assays are limited by the experimental conditions. For example, two preparations which do not contain the same effective ingredient may produce a common effect on a certain organism, say A, while they may not produce the same effect on a certain other organism say, B. Then the potency estimated from such an assay holds only for A organism and not for B. This is, however, not the case with analytical dillusion assays. We shall discuss here only analytical dillusion assays.

The purpose of bio-assays is ultimately to conduct an experiment for estimating the doses of the standard and test preparations, such that each of them produces the same response. Depending on the nature of the preparations, the experimental subjects, that is, the type of living being or organism being used as experimental units and the type of response, such two doses can be estimated by various methods. One of these methods attempts to estimate such doses directly and is called direct bio-assay. In many situations this technique is not applicable and indirect methods are adopted for the estimation of such doses. A detailed discussion of the application of statistical methods to bio-assays is available in the works of Finney (1964). We have discussed here only some of the topics which are necessary for statistical designs for bio-assays.

7.2 DIRECT ASSAYS

In direct assays the response is in general preassigned as death of experimental subjects as a result of application of each preparation. Other types of responses like the appearance of a symptom which can be recognized as soon as it happens, can also be used as preassigned response. It is further assumed that when the dose of a preparation of a drug or a poison under assay, is administered to a subject, it produces the effect immediately without any time lag, and the dose corresponding to the preassigned response can be measured as soon as the response occurs.

If a subject can tolerate up to a dose "d" of a preparation, such that any dose greater than d will always kill it and any dose less than d will not kill it, then d is called the tolerance of the subject relative to the preparation. The direct assay technique requires that the tolerance dose of each subject should be measurable as soon as death occurs.

From a population of subjects as experimental units, we can think of a population of tolerance doses relative to a preparation. It will be assumed that for a given type of subject and preparation the distribution of such tolerance is normal. In some situations the logarithm of tolerance is also assumed to be normally distributed.

Design for the Assay

The assay technique consists of taking a number of subjects and dividing them into two groups at random. One of the groups is allotted to the standard preparation and the other to the test preparation. To each animal the corresponding drug is administered by a suitable device such that the administration can be stopped as soon as the preassigned response occurs. This dose is recorded as an observation. Likewise, for each preparation, observations on all the subjects allotted to it are obtained.

The averages of these observations for the preparations are taken as the estimates of their equipotent doses. Denoting these two average doses by $\bar{y}_s$, and $\bar{y}_t$, we get an estimate of the relative potency (R) as below:

$$R = \frac{\bar{y}_s}{\bar{y}_t}. \tag{7.1}$$

A measure of precision of R cannot be obtained in any straight-forward way. It is obtained by using a theorem due to Fieller (see Finney. 1964). The theorem is reproduced below.

Fieller's Theorem

Let a and b be two random variables distributed normally such that

$$E(a) = \alpha$$
$$E(b) = \beta$$
$$\mu = \frac{\alpha}{\beta}$$
$$m = \frac{a}{b}$$
$$v(a) = v_{11}s^2$$
$$v(b) = v_{22}s^2$$
$$\text{cov}(a,b) = v_{12}s^2$$

The fiducial limits of μ are obtained from the following:

$$m_l, m_u = \frac{m - \frac{v_{12}}{v_{22}}g \pm \frac{t_0 s}{b}\left(v_{11} - 2mv_{12} + m^2_{v_{22}} + \frac{v_{12}^2}{v_{22}}g - v_{12}g\right)^{1/2}}{1-g}$$

where $g = t_0^2 s^2 v_{22}/b^2$ and s_2 is the error variance.

The precision of R depends on the precisions of $\bar{y}_s$ and $\bar{y}_t$. One way of increasing the precisions of $\bar{y}_s$ and $\bar{y}_t$ is to allot a larger number of subjects to each preparation so that $v(\bar{y}_s)$ and $v(\bar{y}_t)$ are smaller. But this technique increases the cost of the experiment and sometimes a large number of suitable subjects may not be available.

The other method is to take a homogeneous group of subjects, such that they are similar in respect of those characters which influence the response under consideration. These animals are then divided into two random groups for allocation to the two preparations. It is, however, risky to choose two homogeneous groups of subjects without regards to the between group variation and then allot them to the two preparations, This procedure may increase the precision of the estimates $\bar{y}_s$ and $\bar{y}_t$ but it introduces bias in the estimation of relative potency when the animals in the two groups differ materially.

Fieller's theorem gives a measure of precision of R by obtaining fiducial limits of R. This method makes the assumption that tolerance is normally distributed. The expression for the fiducial limits of R are some what complicated. If, however, the logarithm of tolerance is normally distributed, estimation of precision of R gets simplified as shown below.

When the logarithm of tolerance is normally distributed, the individual tolerance values collected from the assay are transformed to the logarithmic scale. If $\bar{x}_s$ and $\bar{x}_t$ denote the averages of the logarithm of the tolerance values for the standard and test preparations respectively, then

$$\log R = \bar{x}_s - \bar{x}_t .$$

The variance of $\bar{x}_s - \bar{x}_t$ can be obtained easily and provides measure of precision of $\log R$. The estimate of R is obtained by taking the antilog of $\bar{x}_s - \bar{x}_t$.

In many situations there may be a time lag between the administration of the dose and the appearance of response. Alternative assay techniques are necessary in such situations. Some of these techniques along with their appropriate designs are described in Section 7.3.

7.3 INDIRECT BIO-ASSAYS

In indirect bio-assays the relationship between the dose and response of each preparation is first ascertained. Then the dose corresponding to a given response is obtained from the relation for each preparation separately.

In order to obtain such a relation two or more doses of the standard preparation are taken. Then their responses are obtained through an appropriate experiment.

Let there be k doses of the standard preparation and each dose be administered to n subjects. Some suitable effect of the doses on the subjects is then observed from each of the subjects. Here the response is not preassigned but takes its own value according to the strength of the dose and nature of the subject. With the help of these responses which are assumed to be normally and independently distributed with a constant variance, the dose response relationship can be investigated as given here. First the following table is made with the response observations.

Table 7.1: ***Response from k Doses of the Standard Preparation***

Response	Doses				
	d_1	d_2	d_3	...	d
	y_{11}	y_{21}	y_{31}	...	y_{k1}
	y_{12}	y_{22}	y_{32}	...	y_{k2}
	.	.	.	...	.
	.	.	.	...	.
	.	.	.	...	.
	y_{1n}	y_{2n}	y_{3n}	...	y_{kn}
Total	S_1	S_2	S_3	...	s_k

Let $\Sigma S_i = G.$

First, it is tested if the dose response relationship is linear. For this purpose the following table is made.

Table 7.2: ***Analysis of Variance for Testing the Linearity of Dose Response Relation***

Sources of variation	d.f	s.s	m.s = s.s / d.f.	F
Between doses	$k-1$	$\sum_i \frac{s_i^2}{n} - \frac{G^2}{kn} = D$		
Regression of response on does	1	$\sum_i d_i s_i - \frac{(\sum d_i)G}{k} = R$		
Deviation from regression	$k-2$	$D-R$	S_d^2	S_d^2/S^2
Within doses (Error)	$k(n-1)$	$T-D$	S^2	
Total	$kn-1$	$\sum_{ij} Y_i^2 - \frac{G^2}{kn} = T$		

The linearity of regression is tested by applying F-test on the ratio of the mean squares due to *deviations from regression* and the *within dose mean squares*. If the deviation mean squares is not significant, the relationship is linear and further investigation is made accordingly. If this mean squares is significant, then the doses are transformed to get another variate, say, x. Usually, the following transformations are used

(*i*) $x = \log$ (dose)

(*ii*) $x = (\text{dose})^{\lambda}$

where λ is a suitable constant.

Using any of these transformations we get a value of x for each dose as recorded in Table 7.1. These x values are called the dose metameters. Now replacing the dose values by the dose metameters in the above analysis of variance, we can form a separate analysis of variance table for each metameter and then test the linearity of relation of the response on the dose metameter. It will be seen that due to such transformation the regression s.s. and the deviation from regression s.s. change. If for a certain dose metameter the mean squares due to the deviations from regression is not significant, then the relationship between response and the metameter is taken to be linear. Normally, simple values like 1, 2, ½, etc. are chosen for λ. Experience shows that suitable dose metameters are available in usual situations, though it is not necessary that in all situations such transformations are available.

Sometimes, the response variable is also transformed for linearizing the relationship between the dose and response metameters. But in the present case we shall restrict the investigation only to find out suitable transformation of the dose variable.

The design and analytical techniques of the assay depends on such linearizing transformations on the dose metameters. Depending on these metameters, the indirect assays are divided into two broad categories, called parallel lines assays and slope ratio assays. These have been discussed below.

***Case* 1:** When x – log (dose) is the linearizing transformation.

Let
$$y = a_s + bx_s \tag{7.2}$$
denote the relation between the response y and x_s where $x_s = \log z_s$ and z_s denotes the dose of the standard preparation.

Denoting by z_t a dose equipotent to z_s, we have $\rho = z_s/z_t$ that is,

$$\log \rho = \log z_s - \log z_t = x_s - x_t$$

That is,
$$x_s = \log \rho + x_t$$

Substituting for x_s in the relation of the standard preparation,

$$y = a_s + bx_s$$

we get the relation for the test preparation as

$$y = a_s + b(\log \rho + x_t)$$

that is
$$y = a_t + bx_t \tag{7.3}$$

where
$$a_t = a_s + b \log \rho$$

Hence, the relationship for the test preparation is also linear like that of the standard preparation for the same transformation.

An examination of the two equations for the two preparations shows that the lines have the same slope and are, therefore, parallel.

As $$a_t = a_s + b \log \rho$$

we get $$\log \rho = \frac{a_t - a_s}{b}.$$

where a_s and a_t are the intercepts of the lines on the y axis and b is the common slope.

If in an assay k doses are taken for each of the two preparations and $\bar{x}_s$ and $\bar{x}_t$ denote the averages of the dose metameters and $\bar{y}_s$ and $\bar{y}_t$ are the average responses for the preparations, then it is known that

$$a_s = \bar{y}_s - b\bar{x}_s.$$

and $$a_t = \bar{y}_t - b\bar{x}_t.$$

Subsituting these values in

$$\log \rho = \frac{a_t - a_s}{b}$$

we get an estimate R of ρ from

$$\log R = \bar{x}_s - \bar{x}_t - \frac{\bar{y}_s - \bar{y}_t}{b} \qquad (7.4)$$

From equations (7.2) and (7.3) it is seen that the two lines for the two preparations should be parallel when the dose metameter is log (dose). The assays corresponding to this transformation are, therefore, called parallel line assays. Before going to estimate the relative potency, it is desirable to test with the help of the data collected from the assay if the two lines are parallel. Such a test is called validity test. It is also desirable to test if the relation between response and dose metameter is linear. This gives rise to another validity test. While planning the assay it is necessary that suitable data are available for conducting these validity tests before relative potency is estimated.

Case 2: If the linearizing transformation is

$$x = z^{\lambda}$$

where z denotes the dose, then it is seen by following a similar procedure as in parallel line assays that the equations of the two, lines for the two preparations are

(1) $y = a + b_s x_s$

(2) $y = a + b_t x_t$

where $b_t = b_s \rho^{\lambda}$.

Hence, $$\rho^{\lambda} = \frac{b_t}{b_s}. \qquad (7.5)$$

Since the relative potency is estimated from the ratio of the slopes of the two preparations, the assays corresponding to the transformation z^{λ} are called slope ratio assays.

An examination of the equations of the two lines shows that they intersect on the response axis. A test of this fact gives rise to one validity test. The other validity

test is to see if the relation between the response and dose metameter is linear. It is seen that the first validity test for slope ratio assays is different from the first validity test for the parallel line assays, though the second test is the same for both the types. So the plans for the two types of assays have to be different. The plans for parallel line assays have been discussed in Section 7.4.

7.4 PARALLEL LINE ASSAYS

A parallel line assay in which each of the preparations has an equal number of doses and an equal number of subjects is allotted to each of the doses, is called a symmetrical parallel line assay. We shall discuss here only symmetrical parallel line assays.

Let the number of doses of each of the preparations be k. As there are in all $2k$ doses in this assay, it is called a $2k$-point symmetrical parallel line assay or simply $2k$-point assay.

Let n subjects be allotted to each of the doses and a suitable response be measured from each subject. Suppose further that. $s_1, s_2 \ldots \rho_k$ denote the doses of the stander preparation and $t_1, t_2, \ldots, t_k$ the same for the test preparation. Denoting the response of the rth subject allotted to the pth dose of the standard preparation by y_{spr} and the rth response from the qth subject of the test preparation by y_{tqr}, the response data are first arranged as in Table 7.3.

Table 7.3: ***Response Data from 2k-Point Assay***

Response	Standard preparation				Test preparation			
	Doses				Doses			
	s_1	s_2	...	s_k	t_1	t_2	...	t_k
	y_{s11}	y_{s21}	...	$y_s k_1$	y_{t11}	y_{t21}	...	y_{tk1}
	y_{s12}	y_{s22}	...	$y_s k_2$	y_{t12}	y_{t22}	...	y_{tk2}
	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...
	y_{s1n}	y_{s2n}	...	y_{skn}	y_{t1n}	y_{t2n}	...	y_{tkn}
Total	S_1	S_2	...	S_k	T_1	T_2	...	T_k

The analysis of the assay for conducting validity tests and for estimating relative potency becomes very much simplified when the doses of each of the preparations

are taken in geometric progression as shown below:

$$s,\ cs,\ c^2s \ldots,\ c^{k-1}s \text{ and } t,\ ct,\ c^2t,\ldots,\ c^{k-1}t$$

where s and t are suitable starling doses of the standard and test preparation respectively and c is a constant which is the same for both the preparations. A further precaution necessary while choosing the doses is that the doses should be evenly distributed in the range of response in which the dose response relationship was investigated for obtaining the linearing transformation.

Let $xs_i = \log c^i s = \log s + i \log c \ (i = 0, 1, 2,\ldots, k-1)$

and $xt_i = \log c^i t = \log t + i \log c \ (i = 0, 1, 2,\ldots, k-1).$

Denoting by $\bar{x}_s$ and $\bar{x}_t$ the averages of the doses of the two preparations, we get

$$\bar{x}_s = \log s + \frac{k-1}{2} \log c \tag{7.6}$$

and

$$\bar{x}_t = \log t + \frac{k-1}{2} \log c \tag{7.7}$$

So, $$xs_i - \bar{x}_s = \left(i - \frac{k-1}{2}\right) \log c$$

and $$xt_i - \bar{x}_t = \left(i - \frac{k-1}{2}\right) \log c.$$

Case 1: When k is odd we choose the base of the logarithm as c so that log c is 1. The log dose as deviates from their mean can now be written as below:

Standard preparation

$$-\frac{k-1}{2}, -\frac{k-3}{2}, \ldots, -1, 0, 1, .., \frac{k-1}{2}.$$

Test preparation

$$-\frac{k-1}{2}, -\frac{k-3}{2}, \ldots, -1, 0, 1, .., \frac{k-1}{2}.$$

By choosing the base of the logarithm as above, these deviate values could be made integers.

Case 2: When k is even, the base of the logarithm is taken as c so that log c becomes 2 and hence all the dose deviates $(i - (k-1)/2)$ log c become odd integers as shown below.

Standard preparation

$$-(k-1), -(k-3), \ldots, -1, 1, 3, \ldots, (k-1)$$

Test preparation

$$-(k-1), -(k-3), \ldots, 1, 1, 3, \ldots, (k-1).$$

The regression contrast for each preparation can now be obtained by multiplying these deviate values by the corresponding dose totals and adding them.

Analysis

As stated earlier the purpose of analysis of indirect bio-assays is two fold. First, it is tested through the analysis of variance technique if, (*i*) the dose metameter and response relationship is linear; and (*ii*) the two lines for the two preparations are parallel. If the tests reveal that the relationship is linear and the lines are parallel, then the relative potency of the test preparation is estimated from the relation

$$\log R = \bar{x}_s - \bar{x}_t - \frac{\bar{y}_s - \bar{y}_t}{b}$$

We have already obtained $\bar{x}_s$ and $\bar{x}_t$ at (7.6) and (7.7), $\bar{y}_s - \bar{y}_t$ is given by

$$\bar{y}_s - \bar{y}_t = \frac{\sum_i S_i - \sum_i T_i}{kn}$$

The combined regression coefficient of the two preparations as obtained at (7.8) below gives the value of b.

For the first part of the analysis the following contrasts among the dose totals are obtained.

$$\text{Preparation contrast } (L_p) = -\sum_i S_i + \sum_i T_i \qquad (7.8)$$

Combined regression contrast

$$(L_1) = -\frac{k-1}{2}(S_1 + T_1) - \frac{k-3}{2}(S_2 + T_2) - \ldots + \frac{k-1}{2}(S_k + T_k)$$

when k is odd.

Combined regression contrast

$$(L_1) = -(k-1)(S_1 + T_1) - (k-3)(S_2 + T_2) - \ldots + (k-1)(S_k + T_k)$$

when k is even.

The difference between the two regression contrasts of the two preparations is the parallelism contrast.

Parallelism contrast

$$(L_1') = -\frac{k-1}{2}(S_1 - T_1) - \frac{k-3}{2}(S_2 - T_2) - \ldots + \frac{k-1}{2}(S_k - T_k)$$

when k is odd

$$= -(k-1)(S_1 - T_1) - (k-3)(S_2 - T_2) - \ldots + (k-1)(S_k - T_k)$$

when k is even

It is seen that $\bar{y}_s - \bar{y}_t = -\frac{L_p}{kn}$.

Again, when k is odd

$$b = \frac{L_1}{4\left\{\left(\frac{k-1}{2}\right)^2 + \left(\frac{k-3}{2}\right)^2 + \ldots + 1^2\right\}}$$

$$= \frac{6L_1}{kn(k^2-1)}.$$

when k is even

$$b = \frac{L_1}{4\left\{(k-1)^2 + (k-3)^2 + \ldots + 1^2\right\}}$$

$$= \frac{6L_1}{kn(k^2-1)}.$$

The following analysis of variance table is then written for the validity tests and estimation of error variance.

Table 7.4 : ***Analysis of Variance in 2k-Point Assays for Validity Tests***

Sources of variation	d.f.	s.s.	m.s.	F
Preparation (L_P)	1	$L_P^2/2kn$		
Regression (combined) (L_1)	1	L_1^2/D		
Parallelism (L'_1)	1	$L_1'^2/D$	s_b^2	s_b^2/s^2
Deviation from regression	$2k-4$	By subtraction	s_d^2	s_d^2/s^2
Doses	$2k-1$	$\frac{\sum_i S_i^2 + \sum_i T_i^2}{n} - \frac{\{\Sigma(S_i+T_i)\}^2}{2kn}$		
Within doses (error)	$2k(n-1)$	By subtraction	s^2	
Total	$2kn-1$	$\sum_{pr} y_{spr}^2 + \sum_{qr} y_{tqr}^2 - \frac{\{\Sigma(S_i+T_i)\}^2}{2kn}$		

The value of D, the divisor for the regression and the parallelism sums of squares in the above table is $(kn\,(k^2-1))/6$ when k is odd and $(2kn\,(k^2-1)/3$ when k is even.

For testing the linearity of regression, the mean squares for the deviations from regression is tested by the F-test using the within mean squares as error. For testing parallelism, the "parallelism" component is tested.

If both these tests are not significant, then the relative potency can be estimated as below,

$$\log R = \bar{x}_s - \bar{x}_t - \frac{\bar{y}_s - \bar{y}_t}{b}$$

$$= \log s - \log t + \frac{L_p}{kn} \cdot \frac{kn\left(k^2-1\right)}{6L_1}$$

when k is odd

$$= \log \frac{s}{t} + \frac{\left(k^2-1\right)}{6} \cdot \frac{L_p}{L_1}$$

or

$$R = \frac{s}{t} \text{antilog} \left\{ \frac{d\left(k^2-1\right)}{6} \frac{L_p}{L_1} \right\}$$

Where $d = \log_{10} c$.

When k is even

$$\log R = \log \frac{s}{t} + \frac{L_p}{kn} \cdot \frac{2kn(k^2-1)}{3L_1}.$$

That is,

$$R = \frac{s}{t} \text{antilog} \left\{ \frac{d(k^2-1)}{3} \frac{L_p}{L_1} \right\}.$$

Precision of R can be estimated through Fieller's theorem.

Some Particular Cases

It is seen that in 4-point assay a test of linearity of regression is not available, as the deviation from regression component does not exist in the analysis of variance. In 6-point assay there are 2 degrees of freedom for the deviation from regression. The following two contrasts of the dose totals are the two deviation contrasts.

Quadratic (L_2) combined $= (S_1 - 2S_2 + S_3) + (T_1 - 2T_2 + T_3)$

Difference between quadratics $(L_2') = (S_1 - 2S_2 + S_3) - (T_1 - 2T_2 + T_3)$.

When k is 4, there are 4 degrees of freedom for deviation from regression. In general, the contrast L_n is represented by the nth degree orthogonal polynomial (see Fisher and Yates table, 1973). The. difference between two such polynomials is denoted by L'_n.

Example 1: 6-Point Symmetrical Parallel Line Assay

Let s and t denote two initial doses of the standard and the test preparations respectively. The other doses are shown in Table 7.5.

Designs for Symmetrical Parallel Line Assays

The precision of the estimate of relative potency depends on the precision of L_p and L_1 and the error variance. In the previous planning of the assays we assumed completely randomized designs. But taking all the doses as treatments, randomized block designs can also be adopeted, if

Table 7.5: *Contrasts of Validity Tests. Regression and Preparation in 6-point Assay*

Dose (z)	Standard preparation			Test preparation			
	s	cs	c^2s	t	ct	c^2t	
$x_{si} - \bar{x}_s$	–1	0	1	–1	0	1	
Dose total	S_1	S_2	S_3	T_1	T_2	T_3	Divisors
Preparation contrast	– 1	– 1	– 1	+ 1	+ 1	+ 1	$6n$
Regression	–1	0	1	–1	0	1	$4n$
Parallelism	–1	0	1	1	0	– 1	$4n$
Quadratic	1	–2	1	1	–2	1	$12n$
Difference between quadratics	1	– 2	1	– 1	2	–1	$12n$

Table 7.6: *Analysis of Variance*

Some of variation	d.f.	s.s.	m.s.	F
Preparation	1	$(-S_1 - S_2 - S_2 + T_1 + T_2 + T_3)^2/6n$		
Regression	1	$(-S_1 + S_3 - T_1 + T_3)^2/4n$		
Parallelism	1	$(-S_1 + S_3 + T_1 - T_3)^2/4n$	S_p^2	S_p^2/S^2
Quadratic	1	$(S_1 - 2S_2 + S_3 + T_1 - 2T_2 + T_3)^2/12n$	S^2_q	S^2_q/S^2
Difference between quadratics	1	$(S_1 - 2S_2 + S_2 - T_1 + 2T_2 - T_3)^2/_{12n}$	S^2_{DQ}	S^2_{DQ}/S^2
Within dose (error)	$(n-1)$	By subtractics	S^2	
Total		$6n - 1 \sum_{ij} y_{ij}^2 - \frac{\left(\sum y_{ij}^2\right)}{n}$		

the experimental animals can be made available in homogeneous groups. This design is, however, not applicable when k is (say) greater than 3 or 4. Usually small animals like rats, cats, guinea pigs, etc. are used as experimental units in bio-assays. These can be formed into homogeneous groups of required size by equalizing them in respect of age, breed, housing and other management, etc., to form blocks. Animals belonging to the same litter are expected to be very much homogeneous and hence whenever available, litters can be used as blocks. For a $2k$-point assay each block consists of $2k$ units to which the $2k$ doses are allotted at random when a randomized block design is adopted. The analysis of the data is the same as indicated earlier excepting that a component due to a blocks is also obtained. If the number of doses is small a latin square design can also be adopted to increase further the precision of the estimate of the relative potency. Availability of suitable animals, however, stands in the way of adopting latin square designs.

7.5 INCOMPLETE BLOCK DESIGNS FOR BIO-ASSAYS

When the number of doses is large it may not always be possible to get suitable homogeneous groups of experimental units for adopting randomized block designs. If litters are used as blocks a sufficient number of litters of required size may not be available, even if the number of doses is not very large. If, again, twin calves are used as blocks, it is not possible to use a randomized block design when the total number of doses is more than two. All these point to the necessity of incomplete block designs for bio-assays. We have discussed below some incomplete block designs which are suitable for parallel line assays. This work is mainly due to Das and Kulkarni (1966).

The use of incomplete block designs for bio-assays is limited mainly due to the inflexibility of the existing incomplete block designs. The main purpose of these designs is the estimation of the difference between all pairs of treatment effects with equal or nearly equal variances. In bio-assays all contrasts are not of equal importance. The preparation and the combined regression contrasts in parallel line assays are more important because these two are used to estimate the relative potency. The other contrasts are used for testing the validity of the underlying assumptions which are normally likely to be satisfied. They are, therefore, not as important as the preparation and regression contrasts.

When an incomplete block design is used for an assay the block effects are not orthogonal to the dose effects. The dose effects are, therefore, estimated first by the method appropriate for the incomplete block design and these effects are then used in place of the unadjusted dose totals in the various contrasts indicated earlier. The sums of squares of these contrasts are then obtained by squaring each contrast value and then dividing it by a suitable divisor. The method of finding the

Source :

1. M.N. Das and N.P. Giri : Design and Analysis of Experiments, New Age International, 2nd Edtion, 2008.

2. Montgomery : Design and Analysis of Experiments, John Wiley & Sons (p)

Ltd.. 5th Edition 2009.