Experiments

Planning, Analysis, and Parameter Design Optimization

C. F. Jeff Wu

Michael Hamada

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Contents

CHAPTER 1

Basic Principles and Experiments with a Single Factor

Some basic concepts and principles in experimental design are introduced in this chapter. The simplest class of experiments, that with a single factor, is then considered. Analysis techniques like regression analysis, variable selection, analysis of variance (ANOVA), multiple comparisons, and residual analysis are presented. These techniques are applicable to more complex experiments to be considered in later chapters.

1.1 INTRODUCTION AND HISTORICAL PERSPECTIVE

Experimentation is one of the most common activities that people engage in. It covers a wide range of applications from household work like food preparation to technological innovation in material science, semiconductors, robotics, life science, etc. It allows an investigator to find out what happens to the output or response when the settings of the input variables in a system are purposely changed. Statistical or common-sense analysis can then be used to study the relationship between the input and output values. A better understanding of how the input variables affect the performance of a system can thereby be achieved. This provides a basis for selecting optimum input settings. Experimental design is a body of knowledge and techniques that enables an investigator to conduct better experiments, analyze data efficiently, and make the connections between the conclusions from the analysis and the original objectives of the investigation.

Experimentation is used to understand and/or improve a system. A system can be a product or process. A product can be one developed in engineering, biology, or the physical sciences. A process can be a manufacturing process, a process that describes a physical phenomenon, or a non-physical process such as those found in service or administration. Although most examples in the book are from engineering or the physical and biological

sciences, the methods can also be applied to other disciplines, such as business, medicine, and psychology. For example, in studying the efficiency and cost of a payroll operation, the entire payroll operation can be viewed as a process with key input variables such as the number of supervisors, the number of clerks, method of bank deposit, level of automation, administrative structure, etc. A computer simulation model can then be used to study the effects of changing these input variables on cost and efficiency.

Modern experimental design dates back to the pioneering work of the great statistician R. A. Fisher in the 1930s at the Rothamsted Agricultural Experimental Station in the United Kingdom. Fisher's work and the notable contributions by F. Yates and D. J. Finney were motivated by problems in agriculture and biology. Because of the nature of agricultural experiments, they tend to be large in scale, take a long time to complete, and must cope with variations in the field. Such considerations led to the development of blocking, randomization, replication, orthogonality, and the use of analysis of variance and fractional factorial designs. The theory of combinatorial designs, to which R. C. Bose has made fundamental contributions, was also stimulated by problems in block designs and fractional factorial designs. The work in this era also found applications in social science research and in the textile and woolen industries.

The next era of rapid development came soon after World War II. In attempting to apply previous techniques to solve problems in the chemical industries, G. E. P. Box and co-workers at Imperial Chemical Industries discovered that new techniques and concepts had to be developed to cope with the unique features in process industries. The new techniques focused on process modeling and optimization rather than on treatment comparisons, which was the primary objective in agricultural experiments. The experiments in process industries tend to take less time but put a premium on run size economy because of the cost of experimentation. These time and cost factors naturally favor sequential experimentation. The same considerations led to the development of new techniques for experimental planning, notably central composite designs and optimal designs. Their analysis relies more heavily on regression modeling and graphical analysis. Process optimization based on the fitted model is also emphasized. Because the choice of design is often linked to a particular model (e.g., a second-order central composite design for a second-order regression model) and the experimental region may be irregularly shaped, a flexible strategy for finding designs to suit a particular model and/or experimental region is called for. With the availability of fast computational algorithms, optimal designs (which was pioneered by J. Kiefer) have become an important part of this strategy.

The relatively recent emphasis on variation reduction has provided a new source of inspiration and techniques in experimental design. In manufacturing the ability to make many parts with few defects is a competitive advantage. Therefore variation reduction in the quality characteristics of these parts has become a major focus of quality and productivity improvement. G. Taguchi advocated the use of robust parameter design to improve a system

(i.e., a product or process) by making it less sensitive to variation, which is hard to control during normal operating or use conditions of the product or process. The input variables of a system can be divided into two broad types: control factors, whose values remain fixed once they are chosen, and noise factors, which are hard to control during normal conditions. By exploiting the interactions between the control and noise factors, one can achieve robustness by choosing control factor settings that make the system less sensitive to noise variation. The new paradigm is variation modeling and reduction. Traditionally, when the mean and variance are both considered, variance is used to assess the variability of the sample mean as in the t test or of the treatment comparisons as in the analysis of variance. The focus on variation and the division of factors into two types led to the development of new concepts and techniques in the planning and analysis of robust parameter design experiments. The original problem formulation and some basic concepts were developed by G. Taguchi. Other basic concepts and many sound statistical techniques have been developed by statisticians since the mid-1980s.

Given this historical background, we now classify experimental problems into five broad categories according to their objectives.

1. *Treatment Comparisons.* The main purpose is to compare several treatments and select the best ones. For example, in the comparison of six barley varieties, are they different in terms of yield and resistance to drought? If they are indeed different, how are they different and which are the best? Examples of treatments include varieties (rice, barley, corn, etc.) in agricultural trials, sitting positions in ergonomic studies, instructional methods, machine types, suppliers, etc.

2. *Variable Screening.* If there is a large number of variables in a system but only a relatively small number of them is important, a screening experiment can be conducted to identify the important variables. Such an experiment tends to be economical in that it has few degrees of freedom left for estimating error variance and higher order terms like quadratic effects or interactions. Once the important variables are identified, a follow-up experiment can be conducted to study their effects more thoroughly. This latter phase of the study falls into the category discussed next.

3. *Response Surface Exploration.* Once a smaller number of variables is identified as important, their effects on the response need to be explored. The relationship between the response and these variables is sometimes referred to as a response surface. Usually the experiment is based on a design that allows the linear and quadratic effects of the variables and some of the interactions between the variables to be estimated. This experiment tends to be larger (relative to the number of variables under study) than the screening experiment. Both parametric and semi-parametric models may be considered. The latter is more computer intensive but also more flexible in model fitting.

4. *System Optimization.* In many investigations, interest lies in the optimization of the system. For example, the throughput of an assembly plant or

the yield of a chemical process is to be maximized; the amount of scrap or number of reworked pieces in a stamping operation is to be minimized; the time required to process a travel claim reimbursement is to be reduced. If a response surface has been identified, it can be used for optimization. For the purpose of finding an optimum, it is, however, not necessary to map out the whole surface as in a response surface exploration. An intelligent sequential strategy can quickly move the experiment to a region containing the optimum settings of the variables. Only within this region is a thorough exploration of the response surface warranted.

5. *System Robustness.* Besides optimizing the response, it is important in quality improvement to make the system robust against noise (i.e., hard-to-control) variation. This is often achieved by choosing control factor settings at which the system is less sensitive to noise variation. Even though the noise variation is hard to control in normal conditions, it needs to be systematically varied during experimentation. The response in the statistical analysis is often the variance (or its transformation) among the noise replicates for a given control factor setting.

1.2 A SYSTEMATIC APPROACH TO THE PLANNING AND IMPLEMENTATION OF EXPERIMENTS

In this section, we provide some guidelines on the planning and implementation of experiments. The following seven-step procedure summarizes the important steps that the experimenter must address.

1. *State Objective.* The objective of the experiment needs to be clearly stated. All stakeholders should provide input. For example, for a manufactured product, the stakeholders may include design engineers who design the product, process engineers who design the manufacturing process, line engineers who run the manufacturing process, suppliers, lineworkers, customers, marketers, and managers.

2. *Choose Response.* The response is the experimental outcome or observation. There may be multiple responses in an experiment. Several issues arise in choosing a response. Responses may be *discrete* or *continuous.* Discrete responses can be counts or categories, e.g., binary (good, bad) or ordinal (easy, normal, hard). Continuous responses are generally preferable. For example, a continuous force measurement for opening a door is better than an ordinal (easy, normal, hard to open) judgment; the recording of a continuous characteristic is preferred to the recording of the percent that the characteristic is within its specifications. Trade-offs may need to be made. For example, an ordinal measurement of force to open a door may be preferable to delaying the experiment until a device to take continuous measurements can be developed. Most importantly, there should be a good

measurement system for measuring the response. In fact, an experiment called a *gauge repeatability and reproducibility* (R&R) *study* can be performed to assess a continuous measurement system (AIAG, 1990). When there is a single measuring device, the variation due to the measurement system can be divided into two types: variation between the operators and variation within the operators. Ideally, there should be no between-operator variation and small within-operator variation. The gauge R&R study provides estimates for these two components of measurement system variation. Finally, the response should be chosen to increase understanding of mechanisms and physical laws involved in the problem. For example, in a process that is producing underweight soap bars, soap bar weight is the obvious choice for the response in an experiment to improve the underweight problem. By examining the process more closely, there are two subprocesses that have a direct bearing on soap bar weight: the mixing process that affects the soap bar density and the forming process that impacts the dimensions of the soap bars. In order to better understand the mechanism that causes the underweight problem, soap bar density and soap bar dimensions are chosen as the responses. Even though soap bar weight is not used as a response, it can be easily determined from its density and dimensions. Therefore, no information is lost in studying the density and dimensions. Such a study may reveal new information about the mixing and forming subprocesses, which can in turn lead to a better understanding of the underweight problem. Further discussions on and other examples of the choice of responses can be found in Phadke (1989) and León, Shoemaker and Tsui (1993).

The chosen responses can be classified according to the stated objective. Three broad categories will be considered in this book: **nominal-the-best**, **larger-the-better**, and **smaller-the-better**. The first one will be addressed in Section 3.2 and the last two in Section 5.2.

3. *Choose Factors and Levels.* A **factor** is a variable that is studied in the experiment. In order to study the effect of a factor on the response, two or more values of the factor are used. These values are referred to as **levels** or settings. A **treatment** is a combination of factor levels. When there is a single factor, its levels are the treatments. For the success of the experiment, it is crucial that potentially important factors be identified at the planning stage. There are two graphical methods for identifying potential factors. First, a **flow chart** of the process or system is helpful to see where the factors arise in a multi-stage process. In Figure 1.1, a rough sketch of a paper pulp manufacturing process is given which involves raw materials from suppliers, a chemical process to make a slurry which is passed through a mechanical process to produce the pulp. Involving all the stakeholders is invaluable in capturing an accurate description of the process or system. Second, a **cause-and-effect diagram** can be used to list and organize the potential factors that may impact the response. In Figure 1.2, a cause-and-effect diagram is given which lists the factors thought to affect the product quality of an injection molding process. Traditionally, the factors are organized under the headings:

Man, Machine, Measurement, Material, Method, and Environment (Mother Nature for those who like M's). Because of their appearance, cause-and-effect diagrams are also called *fishbone diagrams*. Different characteristics of the factors need to be recognized because they can affect the choice of the experimental design. For example, a factor such as furnace temperature is *hard to change*. That is, after changing the temperature setting, it may take a considerable amount of time before the temperature stabilizes at the new setting. A factor may also be *hard to set* so that the actual level used in the experiment may be different than the intended level. For example, the actual impact force of a pellet projected at an automobile windshield can only be set within 3 psi of the intended impact force. Other factors that may be hard or impossible to control are referred to as *noise* factors. Examples of noise factors include environmental and customer use conditions. (An in-depth discussion of noise factors will be given in Section 10.3.)

Factors may be *quantitative* and *qualitative*. Quantitative factors like temperature, time, and pressure take values over a continuous range. Qualitative factors take on a discrete number of values. Examples of qualitative factors include operation mode, supplier, position, line, etc. Of the two types of factors, there is more freedom in choosing the levels of quantitative factors. For example, if temperature (in °C) is in the range 100–200°C, one could choose 130 and 160°C for two levels or 125, 150, and 175°C for three levels. If only a linear effect is expected, two levels should suffice. If curvature is expected, then three or more levels are required. In general, the levels of quantitative factors must be chosen far enough apart so that an effect can be detected but not too far so that different physical mechanisms are involved (which would make it difficult to do statistical modeling and prediction). There is less flexibility in choosing the levels of qualitative factors. Suppose there are three testing methods under comparison. All three must be included as three levels of the factor "testing method," unless the investigator is willing to postpone the study of one method so that only two methods are compared in a two-level experiment.

When there is flexibility in choosing the number of levels, the choice may depend on the availability of experimental plans for the given combination of factor levels. In choosing factors and levels, *cost* and *practical constraints* must be considered. If two levels of the factor "material" represent expensive and cheap materials, a negligible effect of material on the response will be welcomed because the cost can be drastically reduced by replacing the expensive material by the cheap alternative. Factor levels must be chosen to meet practical constraints. If a factor combination (e.g., high temperature and long time in an oven) can potentially lead to disastrous results (e.g., burned or overbaked), it should be avoided and a different plan should be chosen.

4. *Choose Experimental Plan.* Use the fundamental principles discussed in Section 1.3 as well as other principles presented throughout the book. The choice of the experimental plan is crucial. A poor design may capture little

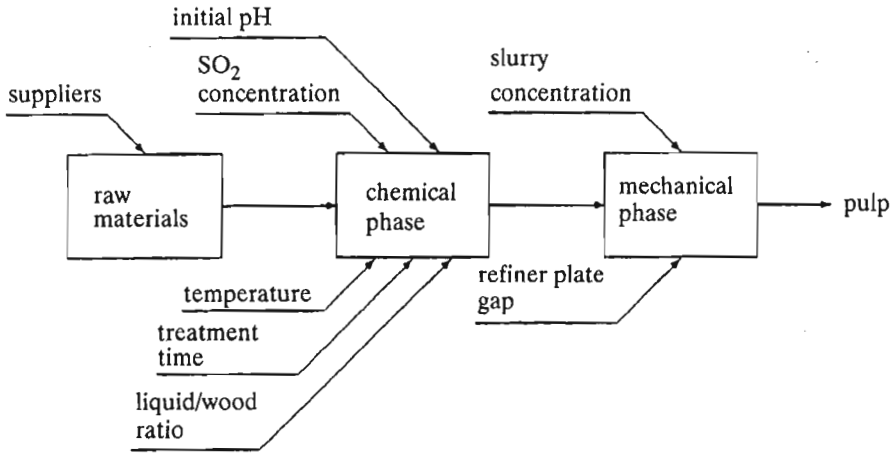


Figure 1.1. Flow Chart, Pulp Manufacturing Process.

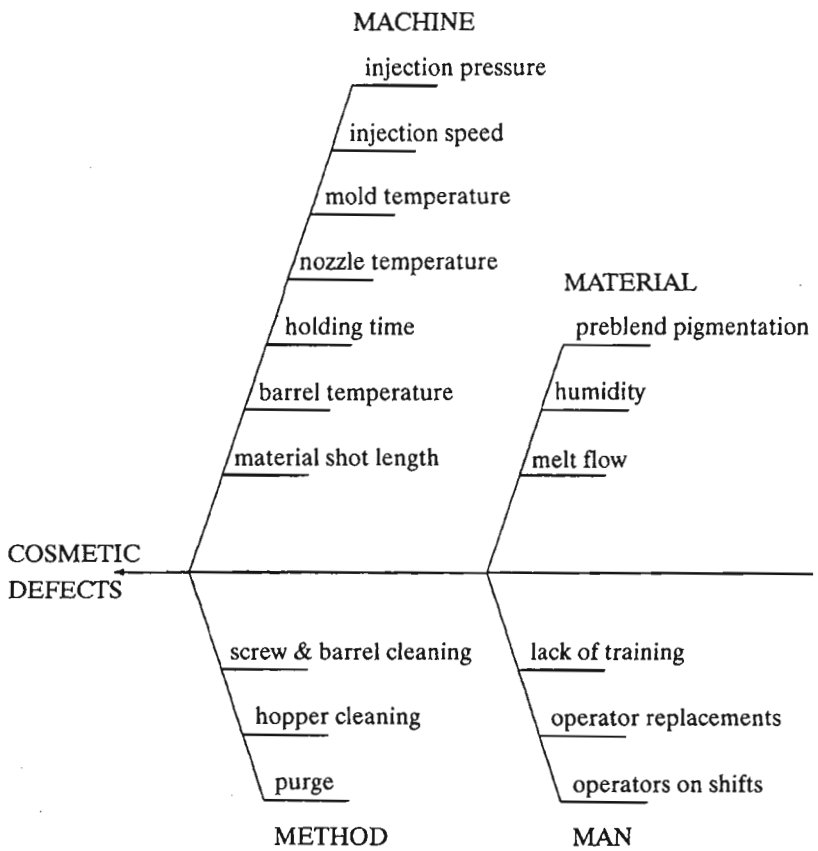


Figure 1.2. Cause-and-Effect Diagram, Injection Molding Experiment.

Man, Machine, Measurement, Material, Method, and Environment (Mother Nature for those who like M's). Because of their appearance, cause-and-effect diagrams are also called *fishbone diagrams*. Different characteristics of the factors need to be recognized because they can affect the choice of the experimental design. For example, a factor such as furnace temperature is *hard to change*. That is, after changing the temperature setting, it may take a considerable amount of time before the temperature stabilizes at the new setting. A factor may also be *hard to set* so that the actual level used in the experiment may be different than the intended level. For example, the actual impact force of a pellet projected at an automobile windshield can only be set within 3 psi of the intended impact force. Other factors that may be hard or impossible to control are referred to as *noise* factors. Examples of noise factors include environmental and customer use conditions. (An in-depth discussion of noise factors will be given in Section 10.3.)

Factors may be *quantitative* and *qualitative*. Quantitative factors like temperature, time, and pressure take values over a continuous range. Qualitative factors take on a discrete number of values. Examples of qualitative factors include operation mode, supplier, position, line, etc. Of the two types of factors, there is more freedom in choosing the levels of quantitative factors. For example, if temperature (in °C) is in the range 100–200°C, one could choose 130 and 160°C for two levels or 125, 150, and 175°C for three levels. If only a linear effect is expected, two levels should suffice. If curvature is expected, then three or more levels are required. In general, the levels of quantitative factors must be chosen far enough apart so that an effect can be detected but not too far so that different physical mechanisms are involved (which would make it difficult to do statistical modeling and prediction). There is less flexibility in choosing the levels of qualitative factors. Suppose there are three testing methods under comparison. All three must be included as three levels of the factor "testing method," unless the investigator is willing to postpone the study of one method so that only two methods are compared in a two-level experiment.

When there is flexibility in choosing the number of levels, the choice may depend on the availability of experimental plans for the given combination of factor levels. In choosing factors and levels, *cost* and *practical constraints* must be considered. If two levels of the factor "material" represent expensive and cheap materials, a negligible effect of material on the response will be welcomed because the cost can be drastically reduced by replacing the expensive material by the cheap alternative. Factor levels must be chosen to meet practical constraints. If a factor combination (e.g., high temperature and long time in an oven) can potentially lead to disastrous results (e.g., burned or overbaked), it should be avoided and a different plan should be chosen.

4. *Choose Experimental Plan.* Use the fundamental principles discussed in Section 1.3 as well as other principles presented throughout the book. The choice of the experimental plan is crucial. A poor design may capture little

information which no analysis can rescue. On the other hand, if the experiment is well planned, the results may be obvious so that no sophisticated analysis is needed.

5. *Perform the Experiment.* The use of a *planning matrix* is recommended. This matrix describes the experimental plan in terms of the actual values or settings of the factors. For example, it lists the actual levels such as 50 or 70 psi if the factor is pressure. To avoid confusion and eliminate potential problems of running the wrong combination of factor levels in a multifactor experiment, each of the treatments, such as temperature at 30°C and pressure at 70 psi, should be put on a separate piece of paper and given to the personnel performing the experiment. It is also worthwhile to perform a *trial run* to see if there will be difficulties in running the experiment, namely, if there are problems with setting the factors and measuring the responses. Any deviations from the planned experiment need to be recorded. For example, for hard-to-set factors, the actual values should be recorded.

6. *Analyze the Data.* An analysis appropriate for the design used to collect the data needs to be carried out. This includes model fitting and assessment of the model assumptions through an analysis of residuals. Many analysis methods will be presented throughout the book.

7. *Draw Conclusions and Make Recommendations.* Based on the data analysis, conclusions are presented which include the important factors and a model for the response in terms of the important factors. Recommended settings or levels for the important factors may also be given. The conclusions should refer back to the stated objectives of the experiment. A *confirmation experiment* is worthwhile, for example, to confirm the recommended settings. Recommendations for further experimentation in a *follow-up experiment* may also be given. For example, a follow-up experiment is needed if two models explain the experimental data equally well and one must be chosen for optimization.

For further discussion on the planning of experiments, see Coleman and Montgomery (1993), Knowlton and Keppinger (1993), and Barton (1997).

1.3 FUNDAMENTAL PRINCIPLES: REPLICATION, RANDOMIZATION, AND BLOCKING

There are three fundamental principles that need to be considered in the design of an experiment: **replication**, **randomization**, and **blocking**. Other principles will be introduced later in the book as they arise.

An *experimental unit* is a generic term that refers to a basic unit such as material, animal, person, machine, or time period, to which a treatment is applied. By *replication*, we mean that each treatment is applied to experimental units that are representative of the population of units to which the conclusions of the experiment will apply. It enables the estimation of the

magnitude of experimental error (i.e., the error variance) against which the differences among treatments are judged. Increasing the number of replications, or *replicates*, decreases the variance of the treatment effect estimates and provides more power for detecting differences in treatments. A distinction needs to be made between replicates and *repetitions*. For example, three readings from the same experimental unit are repetitions while the readings from three separate experimental units are replicates. The error variance from the former is less than that from the latter because repeated readings only measure the variation due to errors in reading while the latter also measures the unit-to-unit variation. Underestimation of the true error variance can result in the false detection of effect significance.

The second principle is that of *randomization*. It should be applied to the allocation of units to treatments, the order in which the treatments are applied in performing the experiment, and the order in which the responses are measured. It provides protection against variables that are unknown to the experimenter but may impact the response. It reduces the unwanted influence of subjective judgment in treatment allocation such as in a physician's assignment of medical treatments to patients. Moreover, randomization ensures validity of the estimate of experimental error and provides a basis for inference in analyzing the experiments. For an in-depth discussion on randomization, see Hinkelmann and Kempthorne (1994).

A group of homogeneous units is referred to as a *block*. Examples of blocks include days, weeks, morning vs. afternoon, batches, lots, sets of twins, and pairs of kidneys. For *blocking* to be effective, the units should be arranged so that the within-block variation is much smaller than the between-block variation. By comparing the treatments within the same block, the block effects are eliminated in the comparison of the treatment effects, thereby making the experiment more efficient. For example, there may be a known day effect on the response so that if all the treatments can be applied within the same day, the day-to-day variation is eliminated.

If blocking is effective, it should be applied to remove the block-to-block variation. Randomization can then be applied to the assignments of treatments to units within the blocks to further reduce the influence of unknown variables. This strategy of **block what you can and randomize what you cannot** is used in randomized block designs, to be discussed in Section 2.2.

These three principles are generally applicable to physical experiments but not to computer experiments because the same input in a computer experiment gives rise to the same output. Computer experiments are not considered in the book.

A simple example will be used to explain these principles. Suppose two keyboards denoted by A and B are being compared in terms of typing efficiency. Six different manuscripts denoted by 1–6 are given to the same typist. First the test is arranged in the following sequence:

$$1.\ A, B,\quad 2.\ A, B,\quad 3.\ A, B,\quad 4.\ A, B,\quad 5.\ A, B,\quad 6.\ A, B.$$

Even though the experiment is replicated six times (with six manuscripts) and blocking is used to compare two keyboards on the same manuscript, the design has a serious flaw. After typing the manuscript on keyboard A, the typist will get more familiar with the content of the manuscript when he or she is typing the same manuscript on keyboard B. This "learning effect" will unfairly help the performance of keyboard B. The observed difference between A and B is the combination of the treatment effects (which measures the intrinsic difference between A and B) and the learning effect. For the given test sequence, it is impossible to disentangle the learning effect from the treatment effect. Randomization would help reduce the unwanted influence of the learning effect, which might not have been known to the investigator who planned the study. By randomizing the typing order for each manuscript, the test sequence may appear as follows:

$$1.\ A,B,\quad 2.\ B,A,\quad 3.\ A,B,\quad 4.\ B,A,\quad 5.\ A,B,\quad 6.\ A,B.$$

With four AB's and two BA's in the sequence, it is a better design than the first one. A further improvement can be made. The design is not balanced because B benefits from the learning effect in four trials while A only benefits from two trials. There is still a residual learning effect not completely eliminated by the second design. The learning effect can be completely eliminated by requiring that half of the trials have the order AB and the other half the order BA. The actual assignment of AB and BA to the six manuscripts should be done by randomization. This method is referred to as *balanced randomization.* Balance is a desirable design property, which will be discussed later.

For simplicity of discussion, we have assumed that only one typist was involved in the experiment. In a practical situation, such an experiment should involve several typists that are representative of the population of typists so that the conclusions made from the study would apply more generally. This and other aspects of the typing experiment will be addressed in the exercises.

With these principles in mind, a useful addition to the cause-and-effect diagram is to indicate how the proposed experimental design addresses each listed factor. The following designations are suggested: **E** for an experimental factor, **B** for a factor handled by blocking, **O** for a factor held constant at one value, and **R** for a factor handled by randomization. This designation clearly indicates how the proposed design deals with each of the potentially important factors. The designation **O**, for "one value," serves to remind the experimenter that the factor is held constant during the current experiment but may be varied in a future experiment. An illustration is given in Figure 1.3 from the injection molding experiment discussed in Section 1.2.

Other designations of factors can be considered. For example, experimental factors can be further divided into two types (control factors and noise

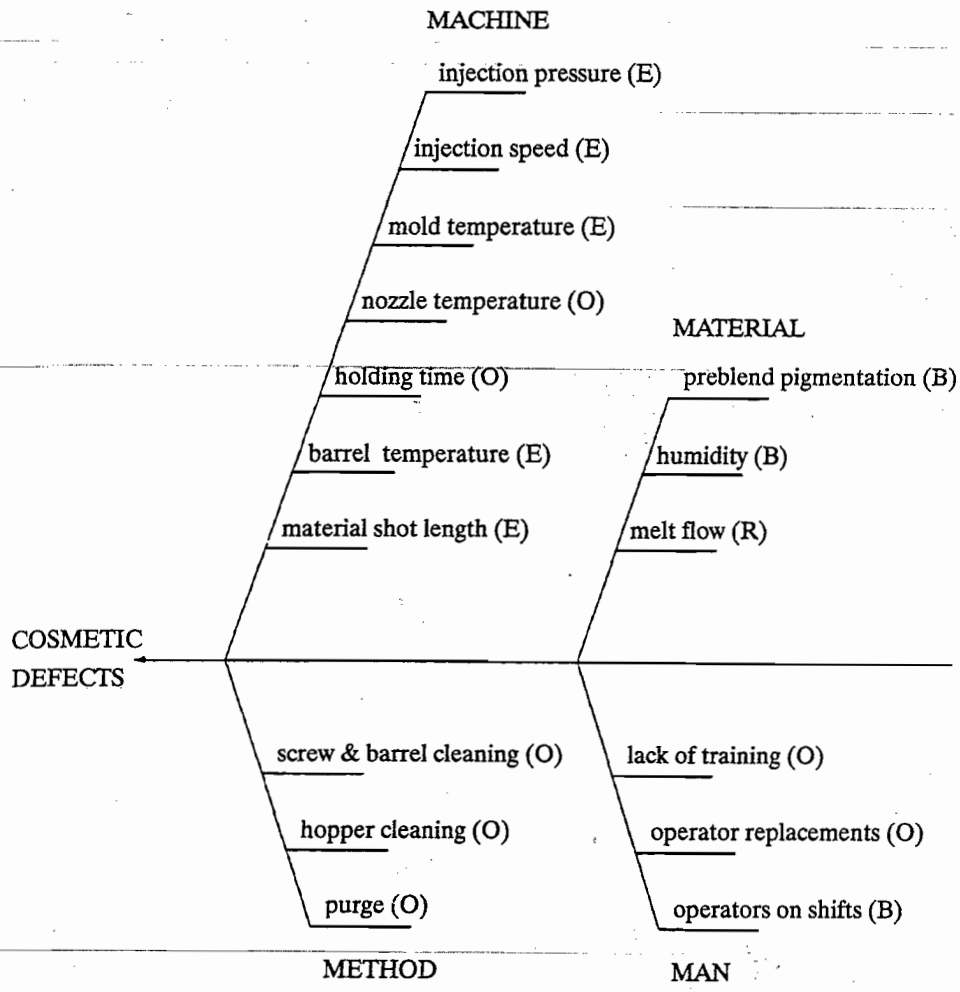


Figure 1.3. Revised Cause-and-Effect Diagram, Injection Molding Experiment.

implementation of experiments, we may also designate an experimental factor as "hard-to-change" or "easy-to-change." These designations will be considered later as they arise.

1.4 THE GENERAL LINEAR MODEL

Experimental data can often be modeled by the general linear model (also called the multiple regression model). Suppose that the response y is related to p covariates (also called explanatory variables, regressors, predictors) $x_1, x_2, \ldots, x_p$ as follows:

where ϵ is the random part of the model which is assumed to be normally distributed with mean 0 and variance σ^2, i.e., $\epsilon \sim N(0, \sigma^2)$; because ϵ is normally distributed, so is y and $Var(y) = \sigma^2$. The structural part of the model is

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + E(\epsilon)$$
$$= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Here, $E(y)$ is linear in the β's, the regression coefficients, which explains the term **linear model.**

If N observations are collected in an experiment, the model for them takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, N, \tag{1.2}$$

where y_i is the ith value of the response and $x_{i1}, \ldots, x_{ip}$ are the corresponding values of the p covariates.

These N equations can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$ is the $N \times 1$ vector of responses, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the $(p+1) \times 1$ vector of regression coefficients, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ is the $N \times 1$ vector of errors, and $\mathbf{X}$, the $N \times (p+1)$ **model matrix**, is given as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix}. \tag{1.4}$$

The unknown parameters in the model are the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2. Thus, the purpose for collecting the data is to estimate and make inferences about these parameters. For estimating $\boldsymbol{\beta}$, the least squares criterion is used; i.e., the least squares estimators (LSEs), denoted by $\hat{\boldsymbol{\beta}}$, minimize the following quantity:

$$\sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 \tag{1.5}$$

which in matrix notation is

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.6}$$

In other words, the squared distance between the response vector $\mathbf{y}$ and the

squared residuals, the vector of **residuals**

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{1.7}$$

needs to be perpendicular to the vector of *fitted values*

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{1.8}$$

that is, the cross product between these two vectors should be zero:

$$\mathbf{r}^T\hat{\mathbf{y}} = \mathbf{r}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

An equivalent way of stating this is that the columns of the model matrix $\mathbf{X}$ need to be perpendicular to $\mathbf{r}$, the vector of residuals, and thus satisfy

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}. \tag{1.9}$$

The solution to this equation is the **least squares estimate** which is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.10}$$

In fitting the model, one wants to know if any of the variables (regressors, predictors, covariates) has explanatory power. None of them has explanatory power if the null hypothesis

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \tag{1.11}$$

holds. In order to test this null hypothesis, one needs to assess how much of the total variation in the response data can be explained by the model relative to the remaining variation after fitting the model, which is contained in the residuals.

Recall how the model was fitted: the residuals are perpendicular to the fitted values so that we have a right triangle. This brings to mind the Pythagorean theorem: the squared length of the hypotenuse is equal to the sum of the squared lengths of its opposite sides. In vector notation, the squared distance of a vector $\mathbf{a}$ is simply $\mathbf{a}^T\mathbf{a} = \sum a_i^2$. Thus, from the least squares fit, we obtain

$$\mathbf{y}^T\mathbf{y} = (\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$
$$= \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbf{y}^T\mathbf{y}$ is the *total sum of squares (uncorrected)*, $\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$ is the *regression sum of squares (uncorrected)*, and

Table 1.1 ANOVA Table for General Linear Model

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	p	$\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2$	$(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2)/p$
residual	$N-p-1$	$(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$	$(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/(N-p-1)$
total (corrected)	$N-1$	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

is the *residual* (or *error*) *sum of squares*. In order to test the null hypothesis (1.11), the contribution from estimating the intercept β_0 needs to be removed. Subtracting off its contribution $N\bar{y}^2$, where $\bar{y}$ is the average of the N observations, yields

$$\begin{aligned} CTSS &= \mathbf{y}^T\mathbf{y} - N\bar{y}^2 = RegrSS + RSS \\ &= \left(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2\right) + \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right), \end{aligned} \tag{1.12}$$

where *CTSS* is called the *corrected total sum of squares* and is equal to $\sum_{i=1}^{N}(y_i-\bar{y})^2$, which measures the variation in the data, and *RegrSS* is called the *corrected regression sum of squares*. In the remainder of this book, "corrected" will be dropped in reference to various sums of squares but will be implied. Thus, the variation in the data is split into the variation explained by the regression model plus the residual variation. This relationship is given in a table called the ANalysis Of VAriance or ANOVA table displayed in Table 1.1.

Based on (1.12), we can define

$$R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}. \tag{1.13}$$

Because the R^2 value measures the "proportion of total variation explained by the fitted regression model $\mathbf{X}\hat{\boldsymbol{\beta}}$," a higher R^2 value indicates a better fit of the regression model. It can be shown that R is the correlation between $\mathbf{y} = (y_i)_{i=1}^N$ and $\hat{\mathbf{y}} = (\hat{y}_i)_{i=1}^N$ and thus is called the *multiple correlation coefficient*.

The degrees of freedom are those associated with each sum of squares. The mean square is the sum of squares divided by the corresponding degrees of freedom. The residual mean square is commonly referred to as the *mean-squared error* (MSE) and is an estimate $\hat{\sigma}^2$ for σ^2, i.e.,

If the null hypothesis (1.11) holds, the F statistic

$$\frac{\left(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2\right)/p}{\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)/(N-p-1)} \tag{1.15}$$

(the regression mean square divided by the residual mean square) has an F distribution with parameters p and $N-p-1$, which are the degrees of freedom of its numerator and denominator, respectively. The p value is calculated by evaluating

$$Prob\left(F_{p,N-p-1} > F_{obs}\right), \tag{1.16}$$

where $Prob(\cdot)$ denotes the probability of an event, $F_{p,N-p-1}$ has an F distribution with parameters p and $N-p-1$, and F_{obs} is the observed value of the F statistic. The F critical values can be found in Appendix D. The p value in (1.16) can be obtained from certain pocket calculators or by interpolating the values given in Appendix D. An example of an F distribution is given in Figure 1.4 along with its critical values.

Note that the **p value** gives the probability under the null hypothesis that the F statistic value for an experiment conducted in comparable conditions will exceed the observed value F_{obs}. The smaller the p value, the stronger is the evidence that the null hypothesis does not hold. Therefore it provides a quantitative measure of the significance of effects in the experiment under study. The same interpretation can be applied when other test statistics and null hypotheses are considered.

It can be shown that the least squares estimate $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and variance-covariance matrix

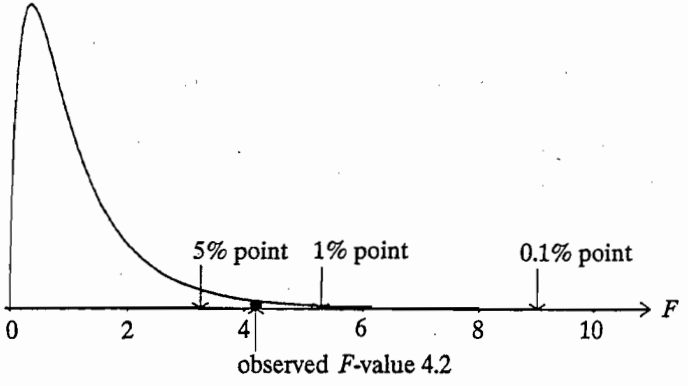


Figure 1.4. Observed F Value of 4.20 in Relation to an F Distribution With 3 and 16

$\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, i.e.,

$$\hat{\boldsymbol{\beta}} \sim MN\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right), \tag{1.17}$$

where MN stands for multivariate normal. The (i, j)th entry of the variance-covariance matrix is $Cov(\hat{\beta}_i, \hat{\beta}_j)$ and the jth diagonal element is $Cov(\hat{\beta}_j, \hat{\beta}_j) = Var(\hat{\beta}_j)$. Therefore, the distribution for the individual $\hat{\beta}_j$ is $N(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj})$, which suggests that for testing the null hypothesis

$$H_0 : \beta_j = 0, \tag{1.18}$$

the following t statistic be used:

$$\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}. \tag{1.19}$$

Under H_0, it has a t distribution with $N-p-1$ degrees of freedom. This can also be used to construct confidence intervals since the denominator of the t statistic is the standard error of its numerator $\hat{\beta}_j$:

$$\hat{\beta}_j \pm t_{N-p-1,\,\alpha/2}\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}, \tag{1.20}$$

where $t_{N-p-1,\,\alpha/2}$ is the upper $\alpha/2$ quantile of the t distribution with $N-p-1$ degrees of freedom. See Appendix C for t critical values.

Besides testing the individual β_j's, testing linear combinations of the β_j's can be useful. For testing $\mathbf{a}^T\boldsymbol{\beta} = \sum_{j=0}^{p} a_j\beta_j$, where $\mathbf{a}$ is a $(p+1)\times 1$ vector, it can be shown that

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \sim N\left(\mathbf{a}^T\boldsymbol{\beta}, \sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\right). \tag{1.21}$$

This suggests using the test statistic

$$\frac{\mathbf{a}^T\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}}, \tag{1.22}$$

which has a t distribution with $N-p-1$ degrees of freedom.

Extra Sum of Squares Principle

The *extra sum of squares principle* will be useful later for developing test statistics in a number of situations. Suppose that there are two models, say Model I and Model II. Model I is a special case of Model II, denoted by Model I $\subset$ Model II. Let

and

$$\text{Model II: } y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \beta_{q+1}x_{i,q+1} + \cdots + \beta_p x_{ip} + \epsilon_i'. \tag{1.24}$$

Model I $\subset$ Model II since $\beta_{q+1} = \cdots = \beta_p = 0$ in Model I. Then, for testing the null hypothesis that Model I is adequate, i.e.,

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0 \tag{1.25}$$

holds, the extra sum of squares principle employs the F statistic:

$$\frac{(RSS(\text{Model I}) - RSS(\text{Model II}))/(p-q)}{RSS(\text{Model II})/(N-p-1)}, \tag{1.26}$$

where RSS stands for the residual sum of squares. It follows that

$$\begin{aligned} &RSS(\text{Model I}) - RSS(\text{Model II}) \\ &\quad = RegrSS(\text{Model II}) - RegrSS(\text{Model I}), \end{aligned} \tag{1.27}$$

where $RegrSS$ denotes the regression sum of squares; thus, the numerator of the F statistic in (1.26) is the gain in the regression sum of squares for fitting the more general Model II relative to Model I, i.e., the *extra sum of squares*. When (1.25) holds, the F statistic has an F distribution with parameters $p-q$ (the difference in the number of estimated parameters between Models I and II) and $N-p-1$. The extra sum of squares technique can be implemented by fitting Models I and II separately, obtaining their respective residual sums of squares, calculating the F statistic above, and then computing its p value.

1.5 VARIABLE SELECTION IN REGRESSION ANALYSIS

(The material in this section will not be used until Chapter 5.)

In the regression fitting of the linear model (1.2), those covariates whose regression coefficients are not significant may be removed from the full model. A more parsimonious model (i.e., one with fewer covariates) is preferred as long as it can explain the data well. It is also known that a model that fits the data too well may give poor predictions. The goal of variable selection in regression analysis is to identify the smallest subset of the covariates that explains the data well; one hopes to capture the true model or at least the covariates of the true model with the largest regression coeffi-

to a model) with the best value of the criterion. This is referred to as **best subset regression**. To maintain a balance between data fitting and prediction, a good model selection criterion should reward good model fitting as well as penalize model complexity. The R^2 in (1.13) is not a suitable criterion because it increases as the number of covariates increases. That is, it does not penalize excessively large models.

A commonly used criterion is the C_p **statistic** (Mallows, 1973). Suppose there are a total of q covariates. For a model that contains p regression coefficients corresponding to $p-1$ covariates and an intercept term β_0, define its C_p value as

$$C_p = \frac{RSS}{s^2} - (N-2p), \tag{1.28}$$

where RSS is the residual sum of squares for the model, s^2 is the mean-squared error (see (1.14)) for the model containing all q covariates and β_0, and N is the total number of observations. As the model gets more complicated, the RSS term in (1.28) decreases while the p value in the second term increases. The counteracting effect of these two terms prevents the selection of extremely large or small models. If the model is true, $E(RSS) = (N-p)\sigma^2$. Assuming that $E(s^2) = \sigma^2$, it is then approximately true that

$$E(C_p) \approx \frac{(N-p)\sigma^2}{\sigma^2} - (N-2p) = p.$$

Thus one should expect the best fitting models to be those with $C_p \approx p$. Further theoretical and empirical studies suggest that models whose C_p values are low and are close to p should be chosen.

For moderate to large q, fitting all subsets is computationally infeasible. An alternative strategy is based on adding or dropping one covariate at a time from a given model, which requires fewer model fittings but can still identify good fitting models. It need not identify the best fitting models as in any optimization that optimizes sequentially (and locally) rather than globally. The main idea is to compare the current model with a new model obtained by adding or deleting a covariate from the current model. Call the smaller and bigger models Model I and Model II, respectively. Based on the extra sum of squares principle in Section 1.4, one can compute the F statistic in (1.26), also known as a *partial F*, to determine if the covariate should be added or deleted. The partial F statistic takes the form

$$\frac{RSS(\text{Model I}) - RSS(\text{Model II})}{RSS(\text{Model II})/\nu}, \tag{1.29}$$

One version is known as **backward elimination**. It starts with the full model containing all q covariates and computes partial F's for all models with $q-1$ covariates. At the kth step, Model II has $q-k+1$ covariates and Model I has $q-k$ covariates, so that $\nu = N-(q-k+1)-1 = N-q+k-2$ in the partial F in (1.29). At each step, compute the partial F value for each covariate being considered for removal. The one with the lowest partial F, provided it is smaller than a preselected value, is dropped. The procedure continues until no more covariates can be dropped. The preselected value is often chosen to be $F_{1,\nu,\alpha}$, the upper α critical value of the F distribution with 1 and ν degrees of freedom. Choice of the α level determines the stringency level for eliminating covariates. Typical α's range from $\alpha = 0.1$ to 0.2. A conservative approach would be to choose a smaller F (i.e., a large α) value so that important covariates are not eliminated. Note that the statistic in (1.29) does not have a proper F distribution so that the F critical values serve only as guidelines. The literature often refers to them as *F-to-remove* values to make this distinction.

Another version is known as **forward selection**, which starts with the model containing an intercept and then adds one covariate at a time. The covariate with the largest partial F [as computed by (1.29)] is added, provided it is larger than a preselected F critical value, which is referred to as an *F-to-enter* value. The forward selection procedure is not recommended as it often misses important covariates. It is combined with backward elimination to form the following stepwise selection procedure.

The **stepwise selection** procedure starts with two steps of the forward selection and then alternates between one step of backward elimination and one step of forward selection. The F-to-remove and F-to-enter values should be chosen to be the same. A typical choice is $F_{1,\nu,\alpha}$ with $\alpha = 0.05,\ 0.1, 0.15$. The choice varies from data to data and can be changed as experience dictates. Among the three selection procedures, stepwise selection is known to be the most effective and is therefore recommended for general use.

For a comprehensive discussion on variable selection, see Draper and Smith (1998).

1.6 ONE-WAY LAYOUT

Consider the following experiment, reported by Sheldon (1960), which was performed at a pulp mill. Plant performance is based on pulp brightness as measured by a reflectance meter. Each of the four shift operators (denoted by A, B, C, and D) made five pulp handsheets from unbleached pulp. Reflectance was read for each of the handsheets using a brightness tester, as reported in Table 1.2. A goal of the experiment is to determine whether there are differences between the operators in making the handsheets and reading their brightness.

Table 1.2 Reflectance Data, Pulp Experiment

Operator			
A	*B*	*C*	*D*
59.8	59.8	60.7	61.0
60.0	60.2	60.7	60.8
60.8	60.4	60.5	60.6
60.8	59.9	60.9	60.5
59.8	60.0	60.3	60.5

treatment is a factor level combination applied to the experimental units. Since there is a single factor, the k treatments correspond to the k levels of the factor. Replication is used here with n_i observations taken for treatment i. For the pulp experiment, $k=4$, $n_1=n_2=n_3=n_4=5$, and $N=20$.

Although Sheldon (1960) did not provide further details, randomization could have been applied in several ways in this experiment. First, 20 containers holding enough pulp to make a handsheet could have been set up and the 20 containers randomly distributed among the four shift operators. For randomization, there are

$$\binom{20}{5\,5\,5\,5} = \frac{20!}{5!5!5!5!} \approx 11.7\times 10^9$$

different allocations of the units and one such allocation can be randomly chosen by taking a random permutation of the numbers 1–20, then assigning the first five numbers to operator A, and so on. Second, the order of brightness measurements for the 20 handsheets could have been randomized.

The linear model for the one-way layout is

$$y_{ij} = \eta + \tau_i + \epsilon_{ij}, \quad i=1,\ldots,k; \quad j=1,\ldots,n_i, \tag{1.30}$$

where y_{ij} is the jth observation with treatment i, τ_i is the ith treatment effect, the errors ϵ_{ij} are independent $N(0,\sigma^2)$ with mean 0 and variance σ^2, k is the number of treatments, and n_i is the number of observations with treatment i.

In terms of the general linear model (1.3), for the pulp experiment $\boldsymbol{\beta}=(\eta,\tau_1,\tau_2,\tau_3,\tau_4)^T$, $\mathbf{y}$ is the column vector (59.8, 59.8, 60.7, 61.0, 60.0, 60.2, 60.7, 60.8, 60.8, 60.4, 60.5, 60.6, 60.8, 59.9, 60.9, 60.5, 59.8, 60.0, 60.3, 60.5)^T.

corresponding model matrix $\mathbf{X}$ is the 20×5 matrix

$$\mathbf{X} = \begin{pmatrix}
1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\
1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 1&1&0&0&0\\
1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\
1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\
1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1
\end{pmatrix}. \tag{1.31}$$

The ANOVA table for the general linear model in Table 1.1 can be shown to reduce to Table 1.3 for the one-way layout, where $N=\sum_{i=1}^k n_k$, $p=k-1$, and $\bar{y}_{\cdot\cdot}$ denotes the average of all N observations. Generally, the dot subscript indicates the summation over the particular index, e.g., $\bar{y}_{i\cdot}$ is the mean of observations for the ith treatment (i.e., averaged over the second index, $j=1,\ldots,n_i$).

Instead of using the matrix algebra of Section 1.4, the ANOVA for the one-way layout can be derived directly as follows. Using the decomposition

$$\begin{aligned} y_{ij} &= \hat{\eta} + \hat{\tau}_i + r_{ij} \\ &= \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot}), \end{aligned} \tag{1.32}$$

Table 1.3 ANOVA Table for One-Way Layout

Source	Degrees of Freedom	Sum of Squares
treatment	$k-1$	$\sum_{i=1}^k n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$
residual	$N-k$	$\sum_{i=1}^k\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$

where

$$\hat{\eta} = \bar{y}_{..},\ \hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..},\ r_{ij} = y_{ij} - \bar{y}_{i.}, \tag{1.33}$$

then using

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}), \tag{1.34}$$

and squaring both sides and summing over i and j yield

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2. \tag{1.35}$$

For each i,

$$(\bar{y}_{i.} - \bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) = 0, \tag{1.36}$$

so that these cross-product terms do not appear in (1.35). The corrected total sum of squares on the left equals the treatment sum of squares plus the residual sum of squares. These three terms are given in the ANOVA table in Table 1.3. The treatment sum of squares is also called the *between-treatment sum of squares* and the residual sum of squares the *within-treatment sum of squares*.

Thus, the F statistic for the null hypothesis that there is no difference between the treatments, i.e.,

$$H_0 : \tau_1 = \cdots = \tau_k, \tag{1.37}$$

is

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/(N-k)}, \tag{1.38}$$

which has an F distribution with parameters $k-1$ and $N-k$.

For the pulp experiment,

$$\bar{y}_{1.} = 60.24, \quad \bar{y}_{2.} = 60.06, \quad \bar{y}_{3.} = 60.62, \quad \bar{y}_{4.} = 60.68, \quad \bar{y}_{..} = 60.40,$$
$$n_1 = n_2 = n_3 = n_4 = 5.$$

Therefore the treatment (i.e., operator) sum of squares in Table 1.3 is

$$5(60.24 - 60.40)^2 + 5(60.06 - 60.40)^2$$

Table 1.4 ANOVA Table, Pulp Experiment

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
operator	3	1.34	0.447	4.20
residual	16	1.70	0.106	
total	19	3.04		

Because there are four operators, $k = 4$ and the degrees of freedom for the treatment sum of squares is 3 $(= 4 - 1)$. Its mean square is then

$$\frac{1.34}{3} = 0.447.$$

Both 1.34 and 0.447 are given in the "operator" row of the ANOVA table in Table 1.4. Similarly, the residual sum of squares is

$$(59.8 - 60.24)^2 + (60 - 60.24)^2 + \cdots + (60.5 - 60.68)^2 = 1.70,$$

which has 16 $(= N - k = 20 - 4)$ degrees of freedom. Then the mean-squared error is

$$\hat{\sigma}^2 = \frac{1.70}{16} = 0.106.$$

Both 1.70 and 0.106 are given in the "residual" row of the ANOVA table in Table 1.4. The F statistic in (1.38) has the value

$$\frac{0.447}{0.106} = 4.20,$$

which is given in Table 1.4 under the column F. Under the null hypothesis H_0, the F statistic has an F distribution with 3 and 16 degrees of freedom. The area under the curve (in Figure 1.4) to the right of the observed F value of 4.20 is the p value

$$Prob(F_{3,16} > 4.20) = 0.02.$$

Recalling the interpretation of p values given after (1.16), the small value of 0.02 provides some evidence that there is an operator-to-operator difference. Another way to interpret the value 0.02 is that for the pulp experiment the F test rejects the null hypothesis H_0 at the 0.02 level.

Once H_0 is rejected, an immediate question is: what pairs of treatments are different? This question will be addressed by the method of multiple

So far we have not considered the estimation of the treatment effects τ_i in (1.30). Because there are k types of observations but $k+1$ regression parameters in (1.30), the model (1.30) is *over-parameterized*. If one attempts to fit the model, i.e., to calculate the least squares estimate $\boldsymbol{\beta}$ in (1.10), the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ based on (1.31) does not exist. The matrix $\mathbf{X}$ is not of full rank since the sum of columns 2–5 equals column 1; that is, the five columns are not linearly independent so that $\mathbf{X}^T\mathbf{X}$ is singular. In order to make $\mathbf{X}^T\mathbf{X}$ a nonsingular matrix, one constraint needs to be put on the parameters. Two types of constraints will be considered in the remaining part of the section.

Constraint on the Parameters

The more commonly used constraint is

$$\sum_{i=1}^{k} \tau_i = 0, \tag{1.39}$$

which is called the **zero-sum constraint**. An example is the $\hat{\tau}_i$ in (1.33) for the ANOVA decomposition. It is readily verified that, for $n_i = n$,

$$\sum_{i=1}^{k} \hat{\tau}_i = \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = 0.$$

Given τ_i, $i = 1, \ldots, k-1$, $\tau_k = -\sum_{i=1}^{k-1}\tau_i$. In substituting τ_k by $-\sum_{i=1}^{k-1}\tau_i$ in the model (1.30), the number of parameters is reduced by 1, i.e., $\boldsymbol{\beta} = (\eta, \tau_1, \tau_2, \ldots, \tau_{k-1})^T$. The remaining parameters have a different meaning. For example,

$$\frac{1}{k}\sum_{i=1}^{k} E(y_{ij}) = \frac{1}{k}\sum_{i=1}^{k} (\eta + \tau_i) = \eta + 0 = \eta,$$

i.e., η represents the grand mean. Also,

$$E(y_{ij}) - \eta = \eta + \tau_i - \eta = \tau_i,$$

which is the *offset between the expected treatment i response and the average response*. Since treatment k is a linear combination of the remaining treat-

With $\boldsymbol{\beta} = (\eta, \tau_1, \tau_2, \tau_3, \tau_4)^T$, (1.30) leads to

$$\mathbf{X}\boldsymbol{\beta} = \eta\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + \tau_1\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + \tau_2\begin{pmatrix}0\\1\\0\\0\end{pmatrix} + \tau_3\begin{pmatrix}0\\0\\1\\0\end{pmatrix} + \tau_4\begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{1.40}$$

for the first four rows of the $\mathbf{X}$ matrix. Substituting $\tau_4 = -\tau_1 - \tau_2 - \tau_3$ in (1.40) leads to

$$\begin{aligned}\mathbf{X}\boldsymbol{\beta} &= \eta\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + \tau_1\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + \tau_2\begin{pmatrix}0\\1\\0\\0\end{pmatrix} + \tau_3\begin{pmatrix}0\\0\\1\\0\end{pmatrix} + (-\tau_1 - \tau_2 - \tau_3)\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\\ &= \eta\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + \tau_1\begin{pmatrix}1\\0\\0\\-1\end{pmatrix} + \tau_2\begin{pmatrix}0\\1\\0\\-1\end{pmatrix} + \tau_3\begin{pmatrix}0\\0\\1\\-1\end{pmatrix},\end{aligned}$$

which leads to the following model matrix $\mathbf{X}$ with $\boldsymbol{\beta} = (\eta, \tau_1, \tau_2, \tau_3)^T$:

$$\mathbf{X} = \begin{pmatrix}
1 & 1 & 0 & 0\\
1 & 0 & 1 & 0\\
1 & 0 & 0 & 1\\
1 & -1 & -1 & -1\\
1 & 1 & 0 & 0\\
1 & 0 & 1 & 0\\
1 & 0 & 0 & 1\\
1 & -1 & -1 & -1\\
1 & 1 & 0 & 0\\
1 & 0 & 1 & 0\\
1 & 0 & 0 & 1\\
1 & -1 & -1 & -1\\
1 & 1 & 0 & 0\\
1 & 0 & 1 & 0\\
1 & 0 & 0 & 1\\
1 & -1 & -1 & -1
\end{pmatrix}. \tag{1.41}$$

For the pulp experiment,

$$\hat{\boldsymbol{\beta}} = \left(\hat{\eta}, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3\right)^T = (60.40, -0.16, -0.34, 0.22)^T. \tag{1.42}$$

Although $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ depend on the choice of constraints, $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$ do not. That is, $\mathbf{X}\boldsymbol{\beta}$ in (1.31) and $\boldsymbol{\beta} = (\eta, \tau_1, \tau_2, \tau_3, \tau_4)^T$ is equal to $\mathbf{X}\boldsymbol{\beta}$ for $\mathbf{X}$ in (1.41) and $\boldsymbol{\beta} = (\eta, \tau_1, \tau_2, \tau_3)^T$.

The second type of constraint is

$$\tau_1 = 0\,, \tag{1.43}$$

which amounts to dropping τ_1 from the vector of parameters $\boldsymbol{\beta}$ and dropping the corresponding column in the model matrix $\mathbf{X}$. Now $(\mathbf{X}^T\mathbf{X})^{-1}$ exists, so that the least squares estimate $\hat{\boldsymbol{\beta}}$ can be obtained. How should the remaining parameters be interpreted? It can be shown that $\eta = E(y_{1j})$, i.e., the expected response value from treatment 1;

$$E(y_{2j}) - E(y_{1j}) = \eta + \tau_2 - \eta = \tau_2, \tag{1.44}$$

i.e., τ_2 represents the comparison of treatments 1 and 2 in terms of their expected responses; and the remaining τ parameters have a similar interpretation, e.g., $\tau_i = E(y_{ij}) - E(y_{1j})$. Thus, τ_i, $i \geq 2$, represent the comparisons between the first and the rest of the k treatments. Other treatment comparisons can be expressed as linear combinations of the τ_i, e.g., the comparison of treatments 2 and 3 is given by $\tau_3 - \tau_2$, which is a linear combination, $\mathbf{a}^T\boldsymbol{\beta}$, of $\boldsymbol{\beta}$, where $\mathbf{a}^T = (0, -1, 1, 0, \ldots, 0)$. Any linear combination $\mathbf{a}^T\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ with $\Sigma a_j = 0$ is called a *contrast*.

The constraint in (1.43) is natural when treatment 1 is a standard or existing treatment and the other $k-1$ treatments are new treatments. The performance of the new treatments is measured by their comparisons with the standard treatment. Treatment 1 can also be interpreted as the baseline for studies in medicine and social sciences. It is referred to as the **baseline constraint** in the book.

For the pulp experiment, $\boldsymbol{\beta} = (\eta, \tau_2, \tau_3, \tau_4)^T$. Then, $\mathbf{X}$ is obtained by dropping the second column of (1.31) and

$$\hat{\boldsymbol{\beta}} = (60.24, -0.18, 0.38, 0.44)^T. \tag{1.45}$$

Again $\mathbf{X}\boldsymbol{\beta}$ (and respectively, $\mathbf{X}\hat{\boldsymbol{\beta}}$) under $\tau_1 = 0$ is the same as $\mathbf{X}\boldsymbol{\beta}$ (and respectively, $\mathbf{X}\hat{\boldsymbol{\beta}}$) under $\Sigma_{i=1}^k \tau_i = 0$.

1.7 MULTIPLE COMPARISONS

Because different pairs (and sets) of treatments are being compared, this

between the ith and jth treatments in the one-way layout, it is common to use the t statistic:

$$t_{ij} = \frac{\bar{y}_{j\cdot} - \bar{y}_{i\cdot}}{\hat{\sigma}\sqrt{1/n_j + 1/n_i}}, \tag{1.46}$$

where $\bar{y}_{i\cdot}$ denotes the average of the n_i observations for treatment i and $\hat{\sigma}$ is the square root of the MSE from the ANOVA table (which is also called the root-mean-square error, or RMSE). Note that the denominator in (1.46) is the standard error of the numerator.

Suppose that the null hypothesis $H_0 : \tau_1 = \cdots = \tau_k$ is rejected. An immediate question is to determine which pairs of treatments are significantly different. Using the *two-sample t test*, treatments i and j are declared significantly different at level α if

$$|t_{ij}| > t_{N-k,\alpha/2}, \tag{1.47}$$

where $t_{N-k,\alpha/2}$ is the upper $\alpha/2$ quantile of a t distribution with $N-k$ degrees of freedom. This test is valid for testing one pair of treatments.

Suppose that k' such tests are performed. It is easy to show that, under H_0, the probability of declaring at least one pair of treatments significantly different (which is called the *experimentwise error rate*) exceeds α for $k' > 1$. For larger k', the experimentwise error rate is higher. (Its proof is left as an exercise.) Therefore, the standard t test cannot be applied in the multiple comparisons of treatments.

To control the experimentwise error rate, two methods are available: the Bonferroni and the Tukey methods. They are convenient to use and have known theoretical properties.

The Bonferroni Method

The Bonferroni method for testing $\tau_i = \tau_j$ versus $\tau_i \neq \tau_j$ declares "τ_i different from τ_j at level α/k'" if

$$|t_{ij}| > t_{N-k,\alpha/2k'}, \tag{1.48}$$

where k' denotes the number of pairs being tested. In the case of the one-way layout with k treatments,

$$k' = \binom{k}{2} = \frac{1}{2}k(k-1).$$

Denote the set of observed data that satisfy (1.48) by A_{ij}. From the distribution of the t statistic,

$$\text{Prob}(A_{ij} \mid \tau_i = \tau_j) = \alpha \tag{1.49}$$

For any pair (i, j), declare "τ_i different from τ_j" if (1.48) for (i, j) is satisfied. Under $H_0: \tau_1 = \cdots = \tau_k$,

$$\begin{aligned} &Prob(\text{at least one pair is declared significantly different} | H_0) \\ &= Prob\left(\bigcup_{i<j} A_{ij} | H_0 \right) \le \sum_{i<j} Prob\{A_{ij} | \tau_i = \tau_j\} \\ &= \sum_{i<j} \alpha/k' = \frac{k'\alpha}{k'} = \alpha. \end{aligned} \tag{1.50}$$

Therefore, the probability of mistakenly declaring any pair of treatments significantly different when they are not (i.e., the experimentwise error rate) is at most α. This method is very easy to use. It is conservative but works for very general problems because the so-called *Bonferroni inequality* in (1.50) is general. For the one-way layout, the Tukey method (to be introduced next) is recommended. For multiple comparison problems for which the Tukey method is not applicable, the Bonferroni method is recommended.

The method can also be adapted to construct conservative *simultaneous confidence intervals* for the k' pairs of differences $\{\tau_i - \tau_j\}_{i<j}$. Solving for

$$\left| \bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\tau_i - \tau_j) \right| \le t_{N-k,\alpha/2k'} \hat{\sigma} \sqrt{(1/n_i + 1/n_j)} \tag{1.51}$$

leads to the confidence interval for $\tau_i - \tau_j$ as

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{N-k,\alpha/2k'} \hat{\sigma} \sqrt{1/n_i + 1/n_j}\,. \tag{1.52}$$

That is, after identifying which pairs are different, the confidence interval in (1.52) quantifies how different the two treatment effects are.

For the pulp experiment, the means for operators A–D are 60.24, 60.06, 60.62, and 60.68, respectively. The t statistics (1.46) are given in Table 1.5 for the six pairs of treatments. For example, the A-vs.-B t statistic is calculated as

$$\frac{60.06 - 60.24}{\sqrt{0.106}\sqrt{1/5 + 1/5}} = -0.87, \tag{1.53}$$

Table 1.5 Multiple Comparison t Statistics, Pulp Experiment

A vs. B	A vs. C	A vs. D	B vs. C	B vs. D	C vs. D

where $\hat{\sigma}^2 = 0.106$ is from Table 1.4 and $n_1 = n_2 = 5$. Notice that for the A-vs.-B comparison, the mean of A is subtracted from the mean of B in (1.53). *This convention is followed throughout the book when the t statistic for a pairwise comparison is presented in a table.* To apply the Bonferroni method in (1.48) at level $\alpha = 0.05$, first compute the $t_{N-k,\alpha/2k'}$ value, which is

$$t_{16,0.05/2\times 6} = 3.008, \tag{1.54}$$

because $N = 20$, $k = 4$, and $k' = 6$. By comparing the t values in Table 1.5 with 3.008, only the B-vs.-D comparison has a t value that exceeds 3.008. Therefore, only operators B and D are significantly different at the 0.05 level.

The Tukey Method

The only difference between the Tukey and Bonferroni methods is in the choice of the critical value. The Tukey method is described as follows: for any pair (i, j) with $1 \le i < j \le k$, declare "τ_i different from τ_j" if

$$|t_{ij}| > \frac{1}{\sqrt{2}} q_{k,N-k,\alpha}, \tag{1.55}$$

where t_{ij} is defined in (1.46) and $q_{k,N-k,\alpha}$ is the upper α quantile of the **Studentized range** distribution with parameter k and $N-k$ degrees of freedom. Recall that k is the number of treatments. See Appendix E for the Studentized range critical values. This method for equal sample sizes has been widely used for many years. A proof that the procedure (1.55) works for general n_i and n_j, i.e., the experimentwise error rate is at most α, can be found in Hochberg and Tamhane (1987). Details on the Studentized range distribution can be found in the same book.

This method is known to be generally the most effective among conservative methods for the one-way ANOVA, that is, its Type II error is generally the smallest (or equivalently, its confidence bound is the tightest). It is recommended unless the critical value $q_{k,N-k,\alpha}$ is not tabled.

For the balanced one-way layout (i.e., $n_i = n$), the experimentwise error rate for the Tukey method is exactly α. To prove this, note that

$$\begin{aligned} &Prob(\text{at least one pair is declared significantly different} | H_0) \\ &= Prob\left(\max_{i,j} \frac{\left| \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \right|}{\hat{\sigma}\sqrt{(1/n + 1/n)}} > \frac{1}{\sqrt{2}} q_{k,N-k,\alpha} | H_0 \right) \\ &= Prob\left(\frac{\max \bar{y}_{i\cdot} - \min \bar{y}_{i\cdot}}{\hat{\sigma}\sqrt{1/n}} > q_{k,N-k,\alpha} | H_0 \right) = \alpha. \end{aligned} \tag{1.56}$$

The last equality in (1.56) holds because under H_0

$$\sqrt{n}\,(\max \bar{y}_{i\cdot} - \min \bar{y}_{i\cdot})/\hat{\sigma}$$

is the *Studentized range* statistic with parameters k and $N-k$.

By solving

$$\frac{\left|(\bar{y}_{j\cdot} - \bar{y}_{i\cdot}) - (\tau_j - \tau_i)\right|}{\hat{\sigma}\sqrt{1/n_j + 1/n_i}} \le \frac{1}{\sqrt{2}} q_{k,N-k,\alpha}$$

for $\tau_j - \tau_i$, the simultaneous confidence intervals for $\tau_j - \tau_i$ are

$$\bar{y}_{j\cdot} - \bar{y}_{i\cdot} \pm \frac{1}{\sqrt{2}} q_{k,N-k,\alpha}\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_i}}, \tag{1.57}$$

for all i and j pairs. Since the Bonferroni method is conservative, the simultaneous confidence intervals based on the Tukey method are shorter.

For the pulp experiment, according to (1.55) at $\alpha = 0.05$,

$$\frac{1}{\sqrt{2}} q_{k,N-k,0.05} = \frac{1}{\sqrt{2}} q_{4,16,0.05} = \frac{4.05}{\sqrt{2}} = 2.86.$$

By comparing 2.86 with the t values in Table 1.5, the Tukey method also identifies that operators B and D are different. The 2.86 used here is smaller than the 3.008 of the Bonferroni method because the Bonferroni method is more conservative. For multiple comparisons at the 0.05 level, the two methods reach the same conclusion, but the simultaneous confidence intervals for the Tukey method are shorter.

1.8 QUANTITATIVE FACTORS AND ORTHOGONAL POLYNOMIALS

Mazumdar and Hoa (1995) performed an experiment which dealt with the laser-assisted manufacturing of a thermoplastic composite. The experimental factor is laser power at 40, 50, and 60 watts. The response is interply bond strength of the composite as measured by a short-beam-shear test. The strength data for the composite experiment are displayed in Table 1.6.

By treating the experimental design as a one-way layout, the ANOVA table for the experiment is computed and given in Table 1.7. The p value for the test of significance for the laser factor $Prob(F_{2,6} > 11.32)$ is 0.009, where $11.32 = 112.09/9.90$ is the observed F statistic value F_{obs}. Thus, the experiment provides strong evidence that laser power affects the strength of the

Table 1.6 Strength Data, Composite Experiment

Laser Power		
40 W	50 W	60 W
25.66	29.15	35.73
28.00	35.09	39.56
20.65	29.79	35.66

Table 1.7 ANOVA Table, Composite Experiment

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
laser	2	224.184	112.092	11.32
residual	6	59.422	9.904	
total	8	283.606		

its significance can be further studied by decomposing the sum of squares for the laser factor (with two degrees of freedom) into a linear component and a quadratic component.

Suppose that a factor is quantitative and has three levels with evenly spaced values. For example, laser power in the composite experiment has evenly spaced levels 40, 50, and 60. Then, one can investigate whether the relationship between the factor and response is linear or quadratic over the three levels. Denote the response value at the low, medium, and high levels by y_1, y_2, and y_3, respectively. Then the linear relationship can be evaluated using

$$y_3 - y_1 = -1y_1 + 0y_2 + 1y_3,$$

which is called the **linear contrast**. To define a quadratic effect, one can use the following argument. If the relationship is linear, then $y_3 - y_2$ and $y_2 - y_1$ should approximately be the same, i.e., $(y_3 - y_2) - (y_2 - y_1) = 1y_1 - 2y_2 + 1y_3$ should be close to zero. Otherwise, it should be large. Therefore, the **quadratic contrast**

$$y_1 - 2y_2 + y_3$$

can be used to investigate a quadratic relationship. The linear and quadratic contrasts can be written as $(-1, 0, 1)(y_1, y_2, y_3)^T$ and $(1, -2, 1)(y_1, y_2, y_3)^T$. The coefficient vectors $(-1, 0, 1)$ and $(1, -2, 1)$ are called the *linear contrast vector* and *the quadratic contrast vector*. Two vectors $\mathbf{u} = (u_i)_1^l$ and $\mathbf{v} = (v_i)_1^l$ are said to be **orthogonal** if their cross product $\mathbf{uv}^T = \sum_{i=1}^{l} u_i v_i = 0$. It is easy to verify that the linear and quadratic contrast vectors are orthogonal, i.e., their cross product $(-1, 0, 1)(1, -2, 1)^T = (-1)(1) + (0)(-2) + (1)(1) = -1 +$

contrasts and the tests based on them are statistically independent. To provide a consistent comparison of their regression coefficients, these vectors should be scaled by their lengths, i.e., $\sqrt{2}$ $(=[(-1)^2+0^2+1^2]^{1/2})$ and $\sqrt{6}$ $(=[(-1)^2+2^2+(-1)^2]^{1/2})$, respectively. These scaled vectors become the covariates in the model matrix.

To see whether laser power has a linear and/or quadratic effect on strength, the linear model with linear and quadratic contrasts [i.e., $(-1,0,1)/\sqrt{2}$ for linear, $(1,-2,1)/\sqrt{6}$ for quadratic] can be fitted and their effects tested for significance. Thus,

$$\mathbf{y} = (25.66, 29.15, 35.73, 28.00, 35.09, 39.56, 20.65, 29.79, 35.66)^T$$

and the model matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix}, \tag{1.58}$$

whose second and third columns need to be divided by $\sqrt{2}$ and $\sqrt{6}$, respectively. The formulas in (1.10) and (1.17) are used to calculate the estimates and standard errors, which are given in Table 1.8 along with the corresponding t statistics. (The least squares estimate for the intercept is 31.0322.) The results in Table 1.8 indicate that laser power has a strong linear effect but no quadratic effect on composite strength. While the ANOVA in Table 1.7 indicates that laser power has a significant effect on composite strength, the additional analysis in Table 1.8 identifies the linear effect of laser power as the main contributor to the significance.

Suppose that the investigator of the composite experiment would like to predict the composite strength for other settings of the laser power, such as 55 or 62 watts. In order to answer this question, we need to extend the notion of orthogonal contrast vectors to **orthogonal polynomials**. Denote the three

Table 1.8 Tests for Polynomial Effects, Composite Experiment

Effect	Estimate	Standard Error	t	p value
linear	8.636	1.817	4.75	0.003

evenly spaced levels by $m-\Delta$, m, and $m+\Delta$, where m denotes the middle level and Δ the distance between consecutive levels. Then define the first- and second-degree polynomials

$$P_1(x) = \frac{x-m}{\Delta}, \tag{1.59}$$

$$P_2(x) = 3\left[\left(\frac{x-m}{\Delta}\right)^2 - \frac{2}{3}\right]. \tag{1.60}$$

It is easy to verify that $P_1(x) = -1, 0, 1$ and $P_2(x) = 1, -2, 1$ for x equal to $m-\Delta$, m, and $m+\Delta$, respectively. These two polynomials are more general than the linear and quadratic contrast vectors because they are defined for a whole range of the quantitative factor and give the same values as the contrast vectors at the three levels of the factor where the experiment is conducted. Based on P_1 and P_2, we can use the following model for predicting the y value at *any* x value in the range,

$$y = \beta_0 + \beta_1 P_1(x)/\sqrt{2} + \beta_2 P_2(x)/\sqrt{6} + \epsilon, \tag{1.61}$$

where $\sqrt{2}$ and $\sqrt{6}$ are the scaling constants used for the linear and quadratic contrast vectors and ϵ are independent $N(0, \sigma^2)$. Because the y values are observed in the experiment only at $x = m-\Delta$, m and $m+\Delta$, the least squares estimates of β_0, β_1, and β_2 in (1.61) are the same as those using regression analysis with the $\mathbf{X}$ matrix in (1.58).

For the composite experiment, fitting model (1.61) including the quadratic effect (that was not significant) leads to the prediction model:

$$\begin{aligned}\text{Predicted strength} = {} & 31.0322 + 8.636 P_1(\text{laser power})/\sqrt{2} \\ & - 0.381 P_2(\text{laser power})/\sqrt{6},\end{aligned} \tag{1.62}$$

where the estimated coefficients 8.636 and -0.381 are the same as in Table 1.8. The model in (1.62) can be used to predict the strength for laser power settings in [40, 60] and its immediate vicinity. For example, at a laser power of 55 watts,

$$P_1(55) = \frac{55-50}{10} = \frac{1}{2}, \qquad \frac{1}{\sqrt{2}} P_1(55) = \frac{1}{2\sqrt{2}} = 0.3536,$$

$$P_2(55) = 3\left[\left(\frac{55-50}{10}\right)^2 - \frac{2}{3}\right] = 3\left(\frac{1}{2}\right)^2 - 2 = -\frac{5}{4},$$

and

$$\frac{1}{\sqrt{6}} P_2(55) = \frac{-5}{4\sqrt{6}} = -0.5103$$

= 50 and $\Delta = 10$. Therefore at 55 watts

icted strength = $31.0322 + 8.636(0.3536) - 0.381(-0.5103)$
$= 34.2803.$

:) to extrapolate far outside the experimental region [40, 60], where may no longer hold, should be avoided.

venly spaced levels, orthogonal polynomials of degree $1, \ldots, k-1$ nstructed. Orthogonal polynomials of degree 1–4 are given as

$$\lambda_1\left(\frac{x-m}{\Delta}\right),$$

$$\lambda_2\left[\left(\frac{x-m}{\Delta}\right)^2 - \left(\frac{k^2-1}{12}\right)\right],$$

$$\lambda_3\left[\left(\frac{x-m}{\Delta}\right)^3 - \left(\frac{x-m}{\Delta}\right)\left(\frac{3k^2-7}{20}\right)\right],$$

$$\lambda_4\left[\left(\frac{x-m}{\Delta}\right)^4 - \left(\frac{x-m}{\Delta}\right)^2\left(\frac{3k^2-13}{14}\right) + \frac{3(k^2-1)(k^2-9)}{560}\right],$$

the distance between the levels of x, k is the total number of $\{\lambda_i\}$ are constants such that the polynomials have integer values. of the orthogonal polynomials and values of the λ_i for $k \le 7$ are pendix G. The value C in each column of the table in Appendix s the sum of squares of the coefficients. The contrast vector es are the coefficients in the column) can be scaled (i.e., divided) esponding $\sqrt{C}$ value so that the regression coefficients of the effects can be consistently compared.

el in (1.61) can be extended to a k-level factor by using higher nomials. Polynomials with fourth and higher degrees, however, e used unless they can be justified by a physical model. Data can d by using a high-degree polynomial model but the resulting l will lack predictive power. In regression analysis this phe-

models is to fit a low-degree polynomial over a small interval or region and patch these polynomials together over the entire region to make it into a smooth function or surface.

1.9 RESIDUAL ANALYSIS: ASSESSMENT OF MODEL ASSUMPTIONS

Before making inferences using hypothesis testing and confidence intervals, it is important to assess the model assumptions:

(i) Have all important effects been captured?
(ii) Are the errors independent and normally distributed?
(iii) Do the errors have constant (the same) variance?

We can assess these assumptions graphically by looking at the **residuals**

$$r_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, N,$$

where $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ is the fitted (or predicted) response at $\mathbf{x}_i$ and $\mathbf{x}_i$ is the ith row of the matrix $\mathbf{X}$ in (1.4). Writing $\mathbf{r} = (r_i)_{i=1}^N$, $\mathbf{y} = (y_i)_{i=1}^N$, $\hat{\mathbf{y}} = (\hat{y}_i)_{i=1}^N = \mathbf{X}\hat{\boldsymbol{\beta}}$, we have

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{1.63}$$

In the decomposition, $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{r}$, $\hat{\mathbf{y}}$ represents information about the assumed model, and $\mathbf{r}$ can *reveal* information about possible model violations.

Based on the model assumptions it can be shown (the proofs are left as two exercises) that the residuals have the following properties:

(a) $E(\mathbf{r}) = \mathbf{0}$,
(b) $\mathbf{r}$ and $\hat{\mathbf{y}}$ are independent, and
(c) $\mathbf{r} \sim MN(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$, where $\mathbf{I}$ is the $N \times N$ identity matrix and

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

is the so-called *hat matrix* since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, i.e., $\mathbf{H}$ puts the hat $\hat{}$ on $\mathbf{y}$.

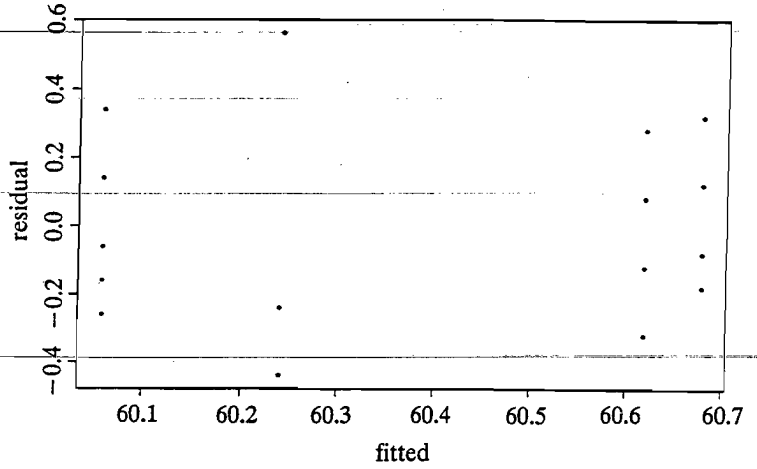
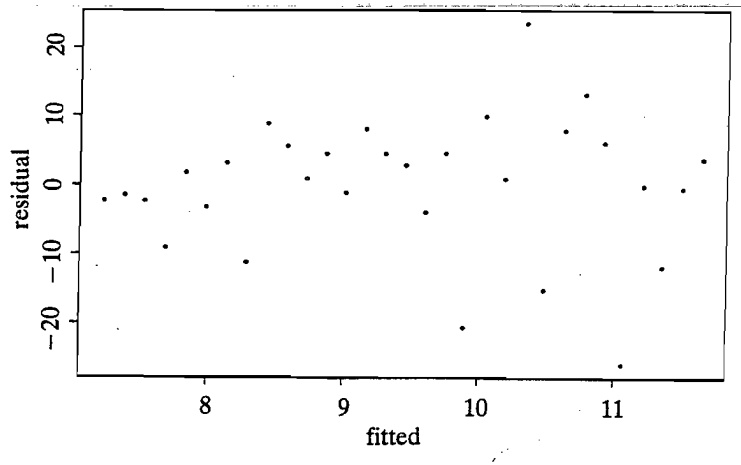


Figure 1.5. r_i vs. $\hat{y}_i$, Pulp Experiment.

Violation of the properties in (a)–(c) would suggest where the model assumptions may be going wrong and how to use the following plots to detect them:

1. **Plot r_i versus $\hat{y}_i$**—The plot should appear as a parallel band [from property (b)] centered about zero [from property (a)]. An example is given in Figure 1.5 for the pulp experiment. If the spread of r_i increases as $\hat{y}_i$ increases, it suggests that the error variance increases with the mean. An example is given in Figure 1.6, which is not related to the pulp experiment. Such a pattern would suggest that the response y needs to be transformed. Transformation of y will be considered in Section 2.5.



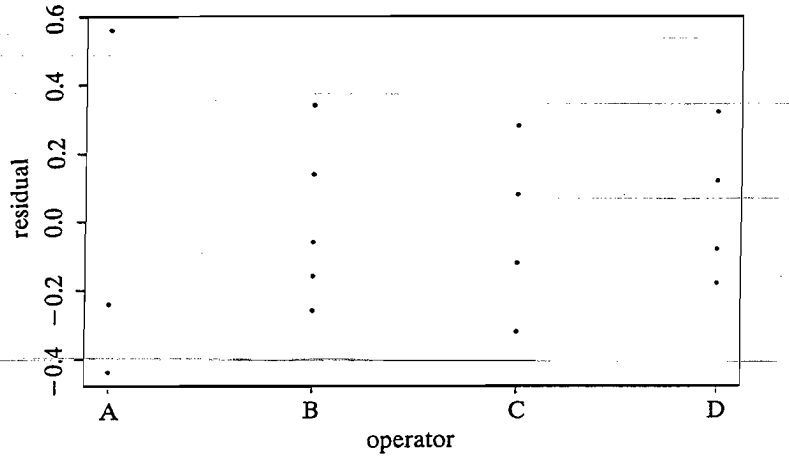


Figure 1.7. r_i vs. x_i, Pulp Experiment.

2. **Plot r_i versus x_i**—Property (c) suggests that the plot should appear as a parallel band centered about zero. An example is given in Figure 1.7 for the pulp experiment. If there is a systematic pattern, it would suggest that the relationship between x_i and the response has not been captured fully in the structural part of the model.
3. **Plot r_i versus time sequence i**, where i is the time sequence in which the observations were taken—The plot should be a parallel band centered about zero. If there is a systematic pattern, it would suggest that the observations are not independent and there is possibly correlation over time.
4. **Plot r_i from replicates grouped by treatment**—The spread of the residuals should be the same for all treatments. Unequal spreads would suggest that the error variance σ^2 also depends on the experimental factor(s). An example is displayed in Figure 1.7 for the pulp experiment; in this case, because only a single experimental factor was studied, this plot is the same as the r_i versus x_i plot.

If there is a large number of replicates per treatment, a **box-whisker** plot of the residuals for each treatment is recommended. A box-whisker plot given in Figure 1.8 displays the minimum, 25th percentile, median, 75th percentile, and maximum, where the box ends correspond to the 25th and 75th percentiles and the line inside the box is the median. Denote the 25th percentile by Q_1, the 75th percentile by Q_3, and the interquartile range $Q_3 - Q_1$ by IQR. Then the two whiskers denote the minimum and maximum values within the range $[Q_1 - 1.5\ IQR,\ Q_3 + 1.5\ IQR]$. Any values outside the

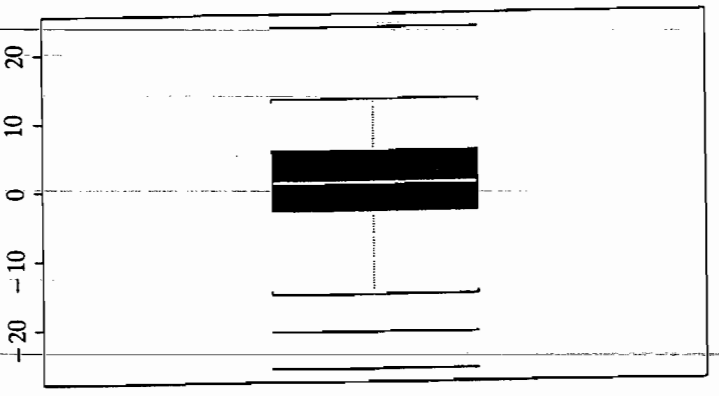
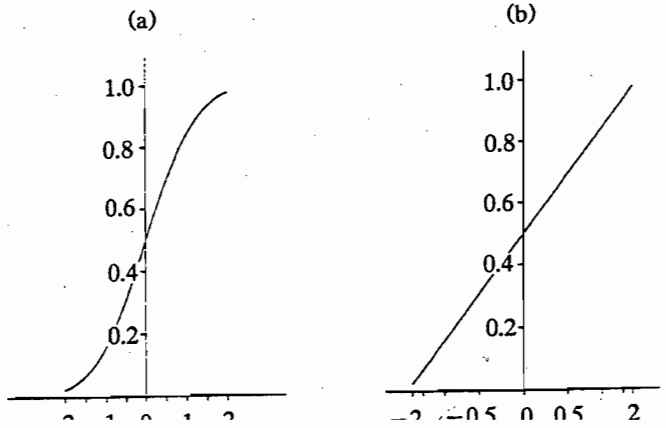


Figure 1.8. Box-Whisker Plot.

indicates one positive outlier and two negative outliers. If there are no outliers, the minimum and maximum are used as the whiskers instead of $Q_1 - 1.5\ IQR$ and $Q_3 + 1.5\ IQR$. The box-whisker plot enables the location, dispersion, skewness, and extreme values of the replicated observations to be displayed. Its use will be demonstrated later for the bolt experiment discussed in Section 2.3.

The normality assumption of the errors can be assessed by the following method. Let $r_{(1)} \leq \cdots \leq r_{(N)}$ denote the ordered residuals. If the errors were normally distributed, then the plot of the cumulative probabilities $p_i = (i - 0.5)/N$ versus the ordered residuals $r_{(i)}$ should ideally be S-shaped, which is the shape of the normal cumulative distribution function as depicted in Figure 1.9(a). The human eye has trouble recognizing departures from a curved line but can easily detect departures from a straight line. By stretching the scale at both ends, the ideal curve becomes a straight line on the



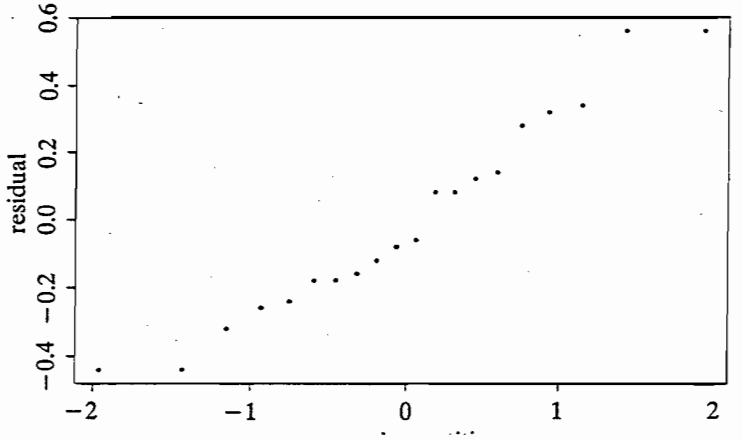
transformed scale, as shown in Figure 1.9(b). By plotting the ordered residuals on the transformed scale, any deviation of the plot from a straight line is an indication of the violation of normality. This method is developed and justified as follows. Suppose that the residuals r_i are normally distributed with the same variance (which is true for most balanced designs considered in the book.) Then, $\Phi(r_i)$ has a uniform distribution over $[0, 1]$. This implies that the expected values of $\Phi(r_{(i)})$, $i = 1, \ldots, N$, are spaced uniformly over $[0, 1]$, i.e., the N points $(p_i, \Phi(r_{(i)}))$, $p_i = (i - 0.5)/N$, should fall on a straight line. By applying the Φ^{-1} transformation to the horizontal and vertical scales, the N points

$$\left(\Phi^{-1}(p_i), r_{(i)}\right), \qquad i = 1, \ldots, N, \tag{1.64}$$

which form the **normal probability plot** of residuals, should plot roughly as a straight line. (Its rigorous justification can be found in Meeker and Escobar, 1998.) A marked deviation of the plot from a straight line would indicate that the normality or constant variance assumptions for the errors do not hold. The normal probability plot can also be used for quantities other than the residuals. A major application is in factorial designs, where the $r_{(i)}$ in (1.64) are replaced by ordered factorial effect estimates. (See Section 3.9.)

For the pulp experiment, the r_i vs. $\hat{y}_i$ and r_i vs. x_i plots are displayed in Figures 1.5 and 1.7. No patterns are evident in these plots. Moreover, the normal probability plot in Figure 1.10 appears close to a straight line. Thus, the residuals are consistent with the model assumptions. An unusually large residual would suggest that the associated observation may be an *outlier*. An outlier is an indication of model inadequacy or suggests that something peculiar happened to the experimental run for the associated observation.

For more details on residual analysis, see Draper and Smith (1998).



1.10 PRACTICAL SUMMARY

1. Experimental problems can be divided into five broad categories:
   (i) treatment comparisons,
   (ii) variable screening,
   (iii) response surface exploration,
   (iv) system optimization,
   (v) system robustness.
2. Statistical process control tools such as control charts are often used to monitor and improve a process. If a process is stable but needs to be further improved, more active intervention like experimentation should be employed.
3. There are seven steps in the planning and implementation of experiments:
   (i) state objective,
   (ii) choose response,
   (iii) choose factors and levels,
   (iv) choose experimental plan,
   (v) perform the experiment,
   (vi) analyze the data,
   (vii) draw conclusions and make recommendations.
4. Guidelines for choosing the response:
   (i) It should help understand the mechanisms and physical laws involved in the problem.
   (ii) A continuous response is preferred to a discrete response.
   (iii) A good measurement system should be in place to measure the response.
5. For response optimization, there are three types of responses: nominal-the-best, larger-the-better, and smaller-the-better.
6. A cause-and-effect diagram or a flowchart should be used to facilitate the identification of potentially important factors and to provide a system view of the problem.
7. Three fundamental principles need to be considered in experimental design: replication, randomization, and blocking. Blocking is effective if the within-block variation is much smaller than the between-block variation.
8. Factors can be designated as **E** (experimental), **B** (blocking), **O** (constant level), and **R** (randomization).
9. A summary of linear model theory is given in Section 1.4 as the foundation for regression analysis used in the book. Variable selection

10. One-way layout (comparison of treatments with no blocking):
    (i) Use model (1.30) with either of the constraints $\sum_1^k \tau_i = 0$ (zero sum) or $\tau_1 = 0$ (baseline). Interpretation of the τ_i for each constraint can be found in Section 1.6.
    (ii) Use the ANOVA table in Table 1.3 and the F test in (1.38) for testing the null hypothesis: $\tau_1 = \tau_2 = \cdots = \tau_k$.
    (iii) If the null hypothesis is rejected, multiple comparisons of the τ_i's should be considered. The Tukey method in (1.55) is recommended. The Bonferroni method in (1.48) can be used in very general situations. It is recommended in situations where the critical values for the Tukey method are not available.
11. For a quantitative factor, use orthogonal polynomials to further model the main effect of the factor. First- and second-degree polynomials are commonly used. Fourth- and higher degree polynomials are rarely used because of the problems associated with overfitting and interpretation.
12. For checking the model assumptions, use the following residual plots:
    (i) plot r_i versus $\hat{y}_i$,
    (ii) plot r_i versus x_i,
    (iii) plot r_i versus time sequence i,
    (iv) plot r_i from replicates grouped by treatment.

    If any of these plots shows a systematic pattern, one or more of the model assumptions are violated. Countermeasures as described in Section 1.9 should be taken. If there is a large number of replicates per treatment, a box-whisker plot is recommended. It enables the location, dispersion, skewness, and extreme values of the replicated observations to be visually compared.

EXERCISES

1. Use a real example to illustrate the seven-step procedure in Section 1.2.

2. Use two examples, one from manufacturing and another from service, to illustrate the construction of the cause-and-effect diagram. Designate each factor on the diagram as **E**, **B**, **O**, or **R**.

3. Give examples of hard-to-change factors. How do you reconcile the hard-to-change nature of the factor with the need for randomization?

4. (a) For the typing experiment considered in Section 1.3, use a statistical model to quantify the gains from using randomization (as illustrated in the second sequence) and from using balance in addition to

(b) Suppose that the following sequence is obtained from using balanced randomization:

$$1\ A,B,\quad 2.\ A,B,\quad 3.\ A,B,\quad 4.\ B,A,\quad 5.\ B,A,\quad 6.\ B,A.$$

Would you use it for the study? If not, what would you do? What aspect of the sequence makes you uneasy? Can you relate it to the possibility that the advantage of the learning effect may diminish over time and express it in more rigorous terms? (Hint: The terms in the model should represent the effects you identified as potentially influencing the comparison.)

5. The typing experiment can be further improved by employing more typists that are representative of the population of typists. Suppose three typists are chosen for the study. Devise an experimental plan and discuss its pros and cons. (Some of the more elaborate plans may involve strategies that will be introduced in the next chapter.)

6. For the pulp experiment obtain the 95% simultaneous confidence intervals for the six pairs of treatment differences using the Bonferroni method and the Tukey method. Which gives shorter intervals?

7. (a) For the pulp experiment show that neither the Bonferroni nor the Tukey method declares any pair of treatments as different at the 0.01 level.

   (b) How do you reconcile the finding in (a) with the result in Section 1.6 that the F test rejects the null hypothesis H_0 at the 0.05 level? After rejecting the null hypothesis, do you expect the multiple comparison method to identify at least one pair of treatments as different? (Hint: One is at the 0.01 level while the other is at the 0.05 level.)

   (c) Recall that the p value for the observed F statistic value 4.20 is 0.02. How can you use this fact to reach the same conclusion in (a) without actually performing the multiple comparisons? (Hint: Use the relationship between the p value and the significance level of the F test.)

8. Make various residual plots for the composite experiment data to support the finding in Table 1.8 that the linear effect is significant while the quadratic effect is not.

9. Use the prediction model in (1.62) to predict the composite strength at 62 watts. If it is suggested to you that the model be used to predict the composite strength at 80 watts, what argument would you use to dissuade

10. Prove that $E(\mathbf{r}) = 0$ and that the covariance matrix between $\mathbf{r}$ and $\hat{\mathbf{y}}$ is zero, i.e., $\mathbf{r}$ and $\hat{\mathbf{y}}$ are independent.

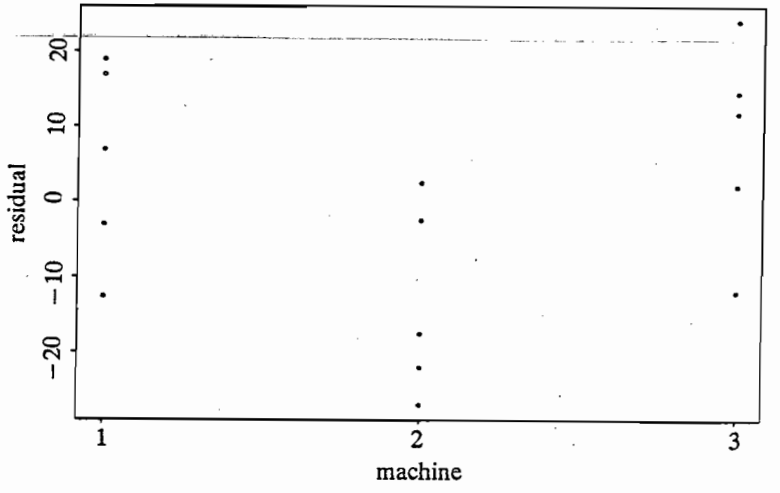
11. Show that $\mathbf{r} \sim MN(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$, where $\mathbf{I}$ is the $N \times N$ identity matrix and $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

12. Prove that under H_0 in (1.37) the probability of declaring at least one pair of treatments significantly different, based on (1.47), exceeds α for $k' > 1$ and increases as k' increases. (Hint: Write the event in (1.47) as C_{ij} and express the rejection region as a union of the C_{ij}'s.)

13. If the plot of residuals against time exhibits a quadratic trend (going up and then going down), what does it suggest to you regarding the model currently entertained and what remedial measures would you take?

14. In order to analyze possible differences between five treatments, a one-way layout experiment was carried out. Each of the treatments was tested on three machines, resulting in a total of 15 experimental runs. After fitting the one-way model (1.30) (which has no block effect) to the data, the residuals were plotted against machine number, as shown in Figure 1.11. What do you learn from the plot? How would you modify your model and analysis?

15. The bioactivity of four different drugs A, B, C, D for treating a particular illness was compared in a study and the following ANOVA table was



given for the data:

Source	Sum of Squares	Degrees of Freedom	Mean Square
between treatments	64.42	3	21.47
within treatments	62.12	26	2.39
total	126.54	29	

(a) Describe a proper design of the experiment to allow valid inferences to be made from the data.

(b) Use an F test to test at the 0.01 level the null hypothesis that the four treatments have the same bioactivity. Compute the p value of the observed F statistic.

(c) The treatment averages are as follows: $\bar{y}_A = 66.10$ (7 samples), $\bar{y}_B = 65.75$ (8 samples), $\bar{y}_C = 62.63$ (9 samples), $\bar{y}_D = 63.85$ (6 samples). Use the Tukey method to perform multiple comparisons of the four treatments at the 0.01 level.

(d) It turns out that A and B are brand-name drugs and C and D are generic drugs. To compare brand-name vs. generic drugs, the contrast $\frac{1}{2}(\bar{y}_A + \bar{y}_B) - \frac{1}{2}(\bar{y}_C + \bar{y}_D)$ is computed. Obtain the p value of the computed contrast and test its significance at the 0.01 level. Comment on the difference between brand-name and generic drugs.

16. In the winter, a plastic rain gauge cannot be used to collect precipitation data because it will freeze and crack. As a way to record snowfall, weather observers were instructed to collect the snow in a metal standard 2.5 can, allow the snow to melt indoors, pour it into a plastic rain gauge, and then record the measurement. An estimate of the snowfall is then obtained by multiplying the measurement by 0.44. (The factor 0.44 was theoretically derived as the ratio of the surface area of the rectangular opening of the rain gauge and of the circular metal can.) One observer questioned the validity of the 0.44 factor for estimating snowfall. Over one summer, the observer recorded the following rainfall data collected in the rain gauge and in the standard 2.5 can, both of which were mounted next to each other at the same height. The data (courtesy of Masaru Hamada) appear in Table 1.9, where the first column is the amount of rain collected in the standard 2.5 can (x) and the second column is the amount of rain collected in the rain gauge (y).

(a) Plot the residuals $y_i - 0.44x_i$ for the data. Do you observe any systematic pattern to question the validity of the formula $y = 0.44x$?

(b) Use regression analysis to analyze the data in Table 1.9 by assuming a general β_0 (i.e., an intercept term) and $\beta_0 = 0$ (i.e., regression line through the origin). How well do the two models fit the data? Is

Table 1.9 Rainfall Data

x	y	x	y	x	y
0.11	0.05	2.15	0.96	1.25	0.62
1.08	0.50	0.53	0.32	0.46	0.23
1.16	0.54	5.20	2.25	0.31	0.17
2.75	1.31	0.00	0.06	0.75	0.33
0.12	0.07	1.17	0.60	2.55	1.17
0.60	0.28	6.67	3.10	1.00	0.43
1.55	0.73	0.04	0.04	3.98	1.77
1.00	0.46	2.22	1.00	1.26	0.58
0.61	0.35	0.05	0.05	5.40	2.34
3.18	1.40	0.15	0.09	1.02	0.50
2.16	0.91	0.41	0.25	3.75	1.62
1.82	0.86	1.45	0.70	3.70	1.70
4.75	2.05	0.22	0.12	0.30	0.14
1.05	0.58	2.22	1.00	0.07	0.06
0.92	0.41	0.70	0.38	0.58	0.31
0.86	0.40	2.73	1.63	0.72	0.35
0.24	0.14	0.02	0.02	0.63	0.29
0.01	0.03	0.18	0.09	1.55	0.73
0.51	0.25	0.27	0.14	2.47	1.23

Note: x = amount of rain collected in metal can, y = amount of rain collected in plastic gauge.

(c) Because of evaporation during the summer and the can being made of metal, the formula $y = 0.44x$ may not fit the rainfall data collected in the summer. An argument can be made that supports the model with an intercept. Is this supported by your analyses in (a) and (b)?

17. Analyze the mandrel portion of the torque data in Table 2.8 by treating it as a one-way layout. Your analysis should include ANOVA, residual analysis, and multiple comparisons of the three plating methods.

18. Data from a one-way layout are given in Table 1.10. The response is the muzzle velocity of mortar-like antipersonnel weapon. The quantitative factor is the discharge hole area (in inches), which has four levels in the experiment. An inverse relationship between muzzle velocity and discharge hole area was expected because a smaller hole would increase the pressure pulse of the propellant gases. Analyze the data in two ways: (i) by treating it as a one-way layout and using an F test and multiple comparisons, (ii) by using orthogonal polynomials to model the linear and quadratic effects. (Note: These data are obtained by collapsing and

Table 1.10 Adapted Muzzle Velocity Data

Discharge Hole Area			
0.016	0.030	0.044	0.058
294.9	295.0	270.5	258.6
294.1	301.1	263.2	255.9
301.7	293.1	278.6	257.1
307.9	300.6	267.9	263.6
285.5	285.0	269.6	262.6
298.6	289.1	269.1	260.3
303.1	277.8	262.2	305.3
305.3	266.4	263.2	304.9
264.9	248.1	224.2	216.0
262.9	255.7	227.9	216.0
256.0	245.7	217.7	210.6
255.3	251.0	219.6	207.4
256.3	254.9	228.5	214.6
258.2	254.5	230.9	214.3
243.6	246.3	227.6	222.1
250.1	246.9	228.6	222.2

19. In tree crop spraying, an airblast sprayer was used with and without an air oscillator on grapefruit and orange trees in an experiment to evaluate the delivery of a solution. Data for the four treatments (grapefruit trees with oscillator, grapefruit trees without oscillator, orange trees with oscillator, orange trees without oscillator) consisted of 72 observations. The corresponding sample means and sample standard deviations of the solution deposited in nanograms per square centimeter (ng/cm^2) appear in Table 1.11.

(a) Analyze the data as a one-way layout by constructing the corresponding ANOVA table. Are there significant differences between the treatments? (Hint: The mean-squared error can be calculated by pooling the sample variances and the treatment sum of squares can be determined from the sample means.)

(b) The analysis in (a) assumes the error variance does not depend on the particular treatment. Are the data consistent with this assumption?

Table 1.11 Summary Data, Airsprayer Experiment

Treatment	Mean	Standard Deviation
grapefruit trees with oscillator	514	330
orange trees with oscillator	430	360

REFERENCES

AIAG (Automotive Industry Action Group) (1990), *Measurement Systems Analysis Reference Manual*, AIAG. Troy, MI.

Barton, R.R. (1997), "Pre-Experiment Planning for Designed Experiments: Graphical Methods," *Journal of Quality Technology*, 29, 307–316.

Coleman, D.E. and Montgomery, D.C. (1993), "A Systematic Approach to Planning for a Designed Industrial Experiment" (with discussion), *Technometrics*, 35, 1–27.

Draper, N.R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd ed., New York: John Wiley & Sons.

Hinkelmann, K. and Kempthorne, O. (1994), *Design and Analysis of Experiments*, Vol. 1, New York: John Wiley & Sons.

Hochberg, Y. and Tamhane, A.C. (1987), *Multiple Comparison Procedures*, New York: John Wiley & Sons.

Knowlton, J. and Keppinger, R. (1993), "The Experimentation Process," *Quality Progress*, February, 43–47.

León, R.V., Shoemaker, A.C., and Tsui, K.L. (1993), Discussion of "A Systematic Approach to Planning for a Designed Industrial Experiment" by Coleman, D.E. and Montgomery, D.C. , *Technometrics*, 35, 21–24.

Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–676.

Mazumdar, S.K. and Hoa, S.V. (1995), "Application of Taguchi Method for Process Enhancement of On-line Consolidation Technique," *Composites*, 26, 669–673.

Meeker, W.Q. and Escobar, L.A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

Phadke, M.S. (1989), *Quality Engineering Using Robust Design*, Englewood Cliffs, NJ: Prentice-Hall.

Sheldon, F.R. (1960), "Statistical Techniques Applied to Production Situations," *Industrial and Engineering Chemistry*, 52, 507–509.

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer.