CHAPTER 1

Introduction

Humans have always been curious about nature. Since prehistoric times, they have tried to understand how the universe around them operates. Their curiosity and ingenuity have led to innumerable scientific discoveries that have fundamentally changed our lives for the better. This progress has been achieved primarily through careful observation and experimentation. Even in cases of serendipity, for example, Alexander Fleming's discovery of penicillin when a petri dish in which he was growing cultures of bacteria had a clear area (because the bacteria were killed) where a bit of mold had accidentally fallen (Roberts, 1989, pp. 160–161) or Charles Goodyear's discovery of vulcanization of rubber when he inadvertently allowed a mixture of rubber and sulfur to touch a hot stove (Roberts, 1989, p. 53), experimental confirmation of a discovery is a must. This book is about how to design experiments and analyze the data obtained from them to draw useful conclusions. In this chapter we introduce the basic terminology and concepts of experimentation.

The outline of the chapter is as follows. Section 1.1 contrasts observational studies with experimental studies. Section 1.2 gives a brief history of the subject. Section 1.3 defines the basic terminology and concepts followed by a discussion of principles in Section 1.4. Section 1.5 gives a summary of the chapter.

1.1 OBSERVATIONAL STUDIES AND EXPERIMENTS

Observational studies and experiments are the two primary methods of scientific inquiry. In an observational study the researcher is a passive observer who records variables of interest (often categorized as independent/explanatory variables or factors and dependent/response variables) and draws conclusions about associations between them. In an experiment the researcher actively manipulates the factors and evaluates their effects on the response variables.
For example, an observational study may find that people who exercise regularly live healthier lives. But is it the exercise that makes people healthy or is it something else that makes people exercise regularly and also makes them healthy? After all, there are many other variables such as diet, sleep, and use of medication that can affect a person’s health. People who exercise regularly are likely to be more disciplined in their dietary and sleep habits and hence may be healthy. These variables are not controlled in an observational study and hence may confound the outcome. Only a controlled experiment in which people are randomly assigned to different exercise regimens can establish the effect of exercise on health.

An observational study can only show association, not causation, between the factors of interest (referred to as treatment factors) and the response variable. This is because of possible confounding caused by all other factors that are not controlled (referred to as noise factors) and are often not even recognized to be important to be observed (hence referred to as lurking variables). Any conclusion about cause–effect relationships is further complicated by the fact that some noise factors may affect not only the response variable but also the treatment factors. For example, lack of good diet or sleep may cause a person to get tired quickly and hence not exercise.

Epidemiological studies are an important class of observational studies. In these studies the suspected risk factors of a disease are the treatment factors, and the objective is to find out whether they are associated with the disease. These studies are of two types. In prospective studies, subjects with and without risk factors are followed forward in time and their disease outcome (yes or no) is recorded. In retrospective studies (also called case–control studies), subjects with and without disease are followed backward in time and their exposure to suspected risk factors (yes or no) is recorded. Retrospective studies are practically easier, but their results are more likely to be invalidated or at least more open to question because of uncontrolled lurking variables. This is also a problem in prospective studies, but to a lesser extent. For instance, if a study establishes association between obesity and hypertension, one could argue that both may be caused by a common gene rather than obesity causing hypertension. This general phenomenon of a lurking variable influencing both the predictor variable and the response variable is depicted diagrammatically in Figure 1.1. An even more perplexing possibility is that the roles of “cause” and “effect” may be reversed. For example, a person may choose not to exercise because of poor health.

On the other hand, an experiment can establish causation, that is, a cause–effect relationship between the treatment factors that are actively changed and the response variable. This is because the treatment factors are controlled by the investigator and so cannot be affected by uncontrolled and possibly unobserved noise factors. Furthermore, selected noise factors may be controlled for experimental purposes to remove their confounding effects, and the effects of the others can be averaged out using randomization; see Section 1.4.

In addition to establishing causation, another advantage of experimentation is that by active intervention in the causal system we can try to improve its
performance rather than wait for serendipity to act. Even if an improvement occurs due to serendipity, we are left to guess as to which input variables actually caused the improvement.

The general goal of any experiment is knowledge and discovery about the phenomenon under study. By knowledge we mean a better understanding of the phenomenon; for example, which are the key factors that affect the outcomes of the phenomenon and how. This knowledge can then be used to discover how to make improvements by tuning the key design factors. This process is often iterative or sequential since, as our knowledge base expands, we can make additional adjustments and improvements. The sequential nature of experimentation is discussed in Section 1.4.2.

Some specific goals of an experiment include the following:

(a) Screen the treatment factors to identify the important ones.
(b) Determine the factor space, that is, the ranges of values of the treatment factors (current or new ones suggested by previous experiments), for follow-up experiments.
(c) Select the best combination of the treatment factor settings to optimize the response.
(d) Select the best combination of treatment factor settings to make the response robust (i.e., insensitive) to variations in noise factors.
(e) Fit a model that can be used to make predictions and/or to adjust the treatment factors so as to keep the response on target.
(f) Determine or expand the scope of applicability of the treatment factors and the predictive model based on them.

Statistics plays a crucial role in the design and analysis of experiments and of observational studies. The design of an experiment involves many practical considerations. Statistics is especially useful in determining the appropriate combinations of factor settings and the necessary sample sizes. This book focuses mainly on the statistical analyses of data collected from designed experiments. Often the same methods of data analysis are used for observational studies and
experiments, but as explained above, stronger conclusions are possible from experiments.

1.2 BRIEF HISTORICAL REMARKS

The field of statistical design and analysis of experiments was founded by Sir Ronald A. Fisher (1890–1962) in the 1920s and 1930s while he was working at the Rothamsted Agricultural Experimental Station in England. Fisher was an intellectual giant who made seminal contributions to statistics and genetics. In design of experiments he invented many important basic ideas (e.g., randomization), experimental designs (e.g., Latin squares), and methods of analysis (e.g., analysis of variance) and wrote the first book on the subject (Fisher 1935). Figure 1.2 shows a picture of Fisher in his younger days taken from his excellent biography by his daughter, Joan Fisher-Box (1978). Fisher was followed in his position at Rothamsted by Sir Frank Yates (1902–1994), who proposed novel block designs and factorial designs and their methods of analysis.

In the 1940s and 1950s, Sir George Box, while working at the Imperial Chemical Industry, developed response surface methodology (Box and Wilson, 1953) as a statistical method for process optimization. There are some crucial differences

![Figure 1.2](image-url)
between agricultural experimentation, the original setting of the subject, and industrial experimentation, the setting in which Box and his co-workers extended the subject in new directions:

(a) Agricultural experiments can be performed only once or twice a year, and data do not become available until the growing and harvesting seasons are over. Industrial experiments, on the other hand, are usually much shorter in duration and data often become immediately available. Therefore agricultural experiments tend to be a few in number but large in size, while several small and sequential experiments are feasible (and preferable) in industrial settings.

(b) Many industrial experiments are performed online and hence are likely to disrupt an ongoing production process. Therefore it is preferable to conduct them sequentially with several small experiments rather than one large experiment.

(c) In agricultural experiments the focus is on comparisons between crop varieties or fertilizers. As a result, analysis of variance techniques with the associated significance tests of equality of means are common. On the other hand, in industrial experiments the focus is on process modeling, optimization, and quality improvement.

In the 1950s, a mathematical theory of construction of experimental designs based on combinatorial analysis and group theory was developed by Raj Chandra Bose (1901–1987) and others. Later a theory of optimal designs was proposed by Jack Kiefer (1923–1980).

Around the same time, A. Bradford Hill (1897–1991) promoted randomized assignments of patients in clinical trials. Psychology, education, marketing, and other disciplines also witnessed applications of designed experiments. A random assignment of human subjects is not always ethical and sometimes not even practical in social and medical experiments. This led to the development of quasi-experiments in the fields of psychology and education by Donald Campbell (1916–1996) and Julian Stanley.

The most recent infusion of new ideas in design of experiments came from engineering applications, in particular designing quality into manufactured products. The person primarily responsible for this renaissance is the Japanese engineer Genichi Taguchi, who proposed that a product or a process should be designed so that its performance is insensitive to factors that are not easily controlled, such as variations in manufacturing conditions or field operating conditions. The resulting methodology of planning and analysis of experiments is called robust design.

1.3 BASIC TERMINOLOGY AND CONCEPTS OF EXPERIMENTATION

In designed experiments the factors whose effects on the response variable are of primary interest are referred to as treatment factors or design factors. The
different settings of a treatment factor are called its **levels**. Because the experi-
menter can set the levels of the treatment factors, they are said to be **controllable**
factors. In the health–exercise example, exercise (yes or no) is the treatment fac-
tor, whose effect on the subjects’ health is evaluated by comparing a group that
follows a prescribed exercise regimen with another group that does not exercise.
The other factors that may also possibly affect the response variable can be
broadly divided into two categories: noise factors and **blocking factors**. These
are discussed in more detail later.

In this book we restrict discussion to a single response variable but possibly
multiple treatment factors. A **qualitative factor** has categorical (nominal or
ordinal) levels, while a **quantitative factor** has numerical levels. For example,
the type of a drug (e.g., three analgesics: aspirin, tylenol, and ibuprofen) is a
qualitative factor, while the dose of a drug is a quantitative factor. A particu-
lar combination of factor levels is called a **treatment combination** or simply a
**treatment**. (If there is a single factor, then its levels are the treatments.)

The treatments are generally applied to physical entities (e.g., subjects, items,
animals, plots of land) whose responses are then observed. An entity receiving
an **independent** application of a treatment is called an **experimental unit**. An
experimental **run** is the process of “applying” a particular treatment combination
to an experimental unit and recording its response. A **replicate** is an independent
run carried out on a different experimental unit under the same conditions. The
importance of independent application of a treatment is worth emphasizing for
estimation of **replication error** (see the next section for a discussion of different
errors). If an experimental unit is further subdivided into smaller units on which
measurements are made, then they do not constitute replicates, and the sample
variance among those measurements does not provide an estimate of replication
error. As an example, if a batch of cookie dough is made according to a
certain recipe (treatment) from which many cookies are made and are scored
by tasters, then the batch would be an experimental unit—not the cookies. To
obtain another replicate, another batch of dough must be prepared following the
same recipe.

A **repeat measurement** is another measurement of the same response of a
given experimental unit; it is not an independent replicate. Taste scores on cookies
made from the same batch and assigned by the same taster can be viewed as repeat
measurements assuming that the cookies are fairly homogeneous and the only
variation is caused by variation in the taster’s perception of the taste. Sample
variance among repeat measurements estimates **measurement error**—not the
replication error that is needed to compare the differences between different
recipes for dough. The measurement error is generally smaller than the replication
error (as can be seen from the cookie example). If it is incorrectly used to compare
recipes, then it may falsely find nonexisting or negligible differences between
recipes as significant.

All experimental units receiving the same treatment form a **treatment group**.
Often, an experiment includes a **standard** or a **control** treatment, which is used as
a benchmark for comparison with other, so-called **test treatments**. For example,
in a clinical trial a new therapy is compared to a standard therapy (called an **active control**) if one exists or a therapy that contains no medically active ingredient, called a **placebo** or a **passive control** (e.g., the proverbial “sugar pill”). All experimental units receiving a control treatment form a **control group**, which forms a basis for comparison for the treatment group.

Let us now turn to noise and blocking factors. These factors differ from the treatment factors in that they represent intrinsic attributes of the experimental units or the conditions of the experiment and are not externally “applied.” For example, in the exercise experiment the age of a subject (young or old) may be an important factor, as well as diet, medications, and amount of sleep that a subject gets. The noise factors are not controlled or generally not even measured in observational studies. On the other hand, blocking factors are controlled in an experiment because their effects and especially their interactions with the treatment factors (e.g., consistency or lack thereof of the effects of the treatment factors across different categories of experimental units) are of interest since they determine the scope and robustness of the applicability of the treatments. For example, different varieties of a crop (treatment factor) may be compared in an agricultural experiment across different fields (blocking factor) having different growing conditions (soils, weather, etc.) to see whether there is a universal winner with the highest yield in all growing conditions. In designed experiments some noise factors may be controlled and used as blocking factors mainly for providing uniform conditions for comparing different treatments. This use of blocking to reduce the variation or bias caused by noise factors is discussed in the next section.

**Example 1.1 (Heat Treatment of Steel: Treatment and Noise Factors)**

A metallurgical engineer designing a heat treatment method wants to study the effects of furnace temperature (high or low) and quench oil bath temperature (high or low) on the surface hardness of steel, which is the response variable. The treatment factors are furnace temperature and quench bath temperature, each with two levels. This gives $2 \times 2 = 4$ treatments. Noise factors include deviations from constant furnace and quench oil bath temperatures, variations between steel samples, and so on.

Suppose that 20 steel samples are available for experimentation. In order to regard them as experimental units, each sample must receive an independent application of furnace heating followed by a quench bath, and the temperatures of each should be independently set in a random order (subject to the condition that all four treatments are replicated five times to have a balanced design). But this may not be feasible in practice. If the engineer can assure us that the furnace and quench bath temperatures are perfectly controllable, then a simpler experiment can be conducted in which 10 samples are heated together in the furnace at one temperature followed by the remaining 10 samples at the other temperature. Each group of 10 samples is then randomly divided into two subgroups of five samples each, which are then quenched at two different temperatures. If replication error is estimated from the samples, it will underestimate the
true replication error if the assumption of perfect controllability of furnace and quench bath temperatures is not correct. In this case different methods of analyses are required.

As mentioned above, often multiple treatment factors are studied in a single experiment. An experiment in which the factors are simultaneously varied (in a random order) is called a factorial experiment. In contrast, in a one-factor-at-a-time experiment only one factor is varied at a time, keeping the levels of the other factors fixed. In a full factorial experiment all factor-level combinations are studied, while in a fractional factorial experiment only a subset of them are studied.

In a factorial experiment each factor can be classified as fixed or random. The levels of a fixed factor are chosen because of specific a priori interest in comparing them. For example, consider a clinical trial to compare three different therapies to treat breast cancer: mastectomy, chemotherapy, and radiation therapy. The therapy is then a fixed factor. The levels of a random factor are chosen at random from the population of all levels of that factor. The purpose generally is not to compare the specific levels chosen but rather (i) to estimate the variability of the responses over the population of all levels and (ii) to assess the generalizability of the results to that population. For example, consider an experiment to compare the mean assembly times using two types of fixtures. Suppose three different operators, Tom, Dick, and Harry, are chosen to participate in the experiment. Clearly, the fixture is a fixed factor. The operator would be a fixed factor if Tom, Dick, and Harry are chosen because the experimenter was specifically interested in comparing them or because they are the only operators in the factory. If there are many operators in the factory from whom these three are chosen at random, then the operator would be a random factor. In this latter case, there would be less interest in comparing Tom, Dick, and Harry with each other since they simply happened to be chosen. However, the variability among these three can be used to estimate the variability that could be expected across all operators in the factory. In practice, however, comparisons will and should be made between the chosen operators if there are large differences between them to determine the causes for the differences.

The parameters that quantify how the mean response depends on the levels of a factor are called its effects. For a fixed factor, the effects are fixed quantities and are called fixed effects; for a random factor they are random quantities and are called random effects. The corresponding mathematical models for the response variable are called the fixed-effects model (also called model I) and the random-effects model (also called model II). In the former case the goal is to estimate the fixed effects while in the latter case the goal is to estimate the variances of the random effects. Some experiments have a combination of fixed and random factors. The corresponding mathematical model is then called the mixed-effects model (also called model III). Chapters 1–10 cover designs which use fixed-effects models. Chapters 11–13 cover designs which use random and mixed-effects models.
Any two factors in a factorial experiment may be crossed with each other or one may be nested within the other. If the same levels of two factors are used in combination, then the factors are said to be crossed. This is possible if the levels of the factors can be set independently of each other. In the fixture–operator example above, the factors are crossed since each operator works with each fixture. Suppose the fixture experiment is carried out in two manufacturing plants and three operators are chosen from each plant. Thus the operators in plant 1 are not the same as the operators in plant 2. In this case the operators are said to be nested within the plants. More generally, if there are two factors, A and B, and different levels of B are observed in combination with each level of A, then B is said to be nested within A. In this case the levels of the factors cannot be set independently of each other because the levels of B are different for each level of A. A schematic of these two factorial designs is shown in Figure 1.3.

1.4 BASIC PRINCIPLES OF EXPERIMENTATION

1.4.1 How to Minimize Biases and Variability?

As noted in the previous section, the effects of treatment factors are of primary interest in an experiment. However, the treatment effects may be distorted by biases or masked by variability in the responses of the experimental units. Therefore biases and variability need to be minimized in order to detect practically important factor effects with high probability. In other words, we need to maximize the signal-to-noise ratio. To understand how to do this, let us look at the main components of the “noise”:

- **Systematic biases/errors** are caused by systematic differences between the experimental units in different treatment groups. For example, suppose a new teaching method is to be compared with a standard method by offering them in two separate sections of a course. If which section a student registers in is voluntary, then students in the two classes are likely to differ systematically; for example, intellectually more adventurous ones may register in the new teaching method section, thus biasing the results. The noise factors on which the experimental units differ are said to **confound** or **bias** the treatment comparisons.
• **Replication or random errors** are caused by the inherent variability in the responses of similar experimental units given the same treatment. They manifest in the variation between replicate observations.

• **Measurement errors** are caused by imprecise measuring instruments or inspectors. They manifest in the variation between repeat measurements.

Replication and measurement errors are together referred to as **experimental errors**. In subsequent chapters, experimental errors are used synonymously with replication errors as measurement errors are often a much smaller part.

We first discuss how to account for the effects of noise factors in order to reduce systematic biases. As seen in Example 1.1, not all noise factors are specified or even known. Noise factors that are specified and measured may or may not be controllable. Strategies to reduce systematic biases differ depending on which type of noise factor we are dealing with. There are two common strategies that are used to design out the effects of noise factors:

• **Blocking**: Use of blocking factors for evaluating the consistency of the effects of treatment factors has been discussed in the previous section. But a noise factor may also be used as a blocking factor if it is controllable or its value for each experimental unit is known before the experiment begins. In that case, its effect can be separated from the treatment effects by dividing the experimental units into homogeneous groups (i.e., having the same or similar values of the noise factor). These groups are called **blocks**. The treatments are assigned at random to the experimental units within each block; thus the effect of the noise factor is the same across all treatments within each block and cancels out when the responses to the treatments are compared to each other.

  In Example 1.1, suppose that the 20 steel samples come from five different batches of steel, four per batch. The differences between the batches are not of primary interest. Since the batches can be identified before the experiment begins, they can be used as a blocking factor; see Example 1.3.

• **Randomization**: What about other noise factors, whether known or unknown? The experimental units may differ on these factors as well, thus biasing the results. To make the treatment groups equal on every single noise factor is impossible. Fisher’s brilliant solution to this problem was to make the treatment groups **probabilistically** equal on **all** noise factors by **randomly assigning** experimental units to the different treatments. Note that randomization does not imply equality of experimental units for all treatments but that no treatment is favored.

  The chance element is introduced intentionally to reduce bias. Another important consequence of random assignment is that it makes statistical inference possible because the statistics calculated from the resulting data have well-defined sampling distributions. If assignment is selective, then such distributions may exist but cannot be derived or simulated in general.
In summary, the basic strategy for dealing with noise factors at the design stage is as follows: block over those noise factors that can be easily controlled; randomize over the rest.

Borrowing from the widely publicized quote by former U.S. Secretary of Defense, Donald Rumsfeld, block over known knowns (known noise factors) and randomize over known and unknown unknowns (unknown noise factors).

If the noise factors are observed, then their effects can also be taken into account at the analysis stage by using regression methods or by forming the blocks postrandomization and data collection. Normally, only a few noise factors can be used for blocking. Some noise factors cannot be used for blocking because they are not observed before assignment of experimental units to the treatments or their levels cannot be fixed. Such noise factors are called covariates. In Example 1.1 percentage of carbon content of steel samples may be regarded as a covariate. A regression analysis that includes effects of both the treatment factors and the covariates is called analysis of covariance (see Section 3.8).

The effect of random errors can be reduced by replicating the experimental runs, that is, by making multiple independent runs under identical treatment conditions. The effect of measurement errors can be reduced by making repeat measurements. In both cases some average is used to smooth out the variability in the individual measurements. Another benefit of replicates and repeat measurements is that they can be used to estimate the corresponding error variances.

To summarize, the following four strategies are useful to improve the precision of an experiment:

(a) Blocking
(b) Randomization
(c) Replication
(d) Repeat measurements.

The first two are smart strategies because they do not require larger samples. The last two are costly strategies, but they are necessary if random and measurement errors must be estimated and controlled.

Example 1.2 (Lady Tasting Tea Experiment)
Fisher (1935) illustrated the principles of experimentation by using the following actual episode that took place at the Rothamsted Experimental Station as described in Fisher-Box (1978, p. 134). In fact, the experiment is so well-known in statistical literature that there is a book with the title *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* by Salsburg (2001) and the experiment is described in detail in Chapter 1 of the book.

Dr. B. Muriel Bristol, an algologist, claimed that she could tell by tasting a cup of tea made with milk whether brewed tea infusion or milk was added to the cup first. It was suggested to test her claim. Fisher (1935) formulated this
as a hypothesis testing problem with the null hypothesis ($H_0$) being that the lady has no discriminating ability and picks one of the two choices at random, each with probability $1/2$. Rejection of $H_0$ would be taken as a proof of the lady’s claim.

One of the key decisions when planning an experiment such as this is the sample size, that is, how many cups to test. The other decisions include: in what order to offer the cups, whether to block over some factors of interest, how to control for noise factors, and the decision rule for rejecting $H_0$. We assume that an equal number of cups are made using each method and the lady is told of this fact. To illustrate statistical considerations used to determine the sample size, first suppose that six cups of tea are tested in a random order, three made by adding milk first (method $A$) and three made by adding tea infusion first (method $B$). The lady classifies them into two groups ($A$ and $B$) of three each. There are $\binom{6}{3} = 20$ different ways of classifying the cups. Under $H_0$, all 20 classifications are equally likely. If correct classification of all six cups is taken as a proof of the lady’s discriminating ability, then the type I error probability, which is the probability of rejecting a true null hypothesis, is $\frac{1}{20} = 0.05$. This is exactly equal to the commonly used standard, called the level of significance, or $\alpha$, that the type I error probability is required not to exceed. Therefore the experimenter may want to use eight cups. A similar calculation shows that the type I error probability for correct classification of all eight cups is $\frac{1}{70} = 0.014$, which may be deemed to be acceptable. In addition to such statistical considerations, practical considerations (e.g., time and other resources available for the experiment, how many cups can be tested in one sitting, etc.) are also used to determine the sample size.

The above probability calculation would be different if a different design is used, for example, if the $A$ and $B$ cups were offered in pairs for tasting with a random order within each pair. This design improves the precision of the experiment because $A$ and $B$ are directly compared with each other, which allows the lady to better discriminate between them. Assuming that $H_0$ is true, the probability of classifying all four pairs correctly is $\left(\frac{1}{2}\right)^4 = 0.063$. This illustrates how the statistical analysis of an experiment depends on its design.

The conditions for making and tasting the tea cups made by the two methods should be as similar as possible, so that any taste difference can be attributed solely to the method of preparation. If the conditions are not the same, then that could introduce bias; for example, if sugar is added to all the $A$ cups but not to the $B$ cups, then the lady’s discriminating ability could be impaired by the sweetness of sugar. Besides introducing bias, this could change the null distribution of the number of correct classifications. In the extreme case, the classification of cups could depend solely on whether sugar is added or not; thus there will not be 20 classifications, but only two (one correct, one wrong) each with probability $\frac{1}{2}$ under $H_0$.

Some factors are relatively easy to control, for example, the type of cup used and the amount of sugar added to each cup. Other factors are not so easy to control, for example, the temperature of tea and the aftertaste effect. It is difficult,
if not impossible, to make the conditions exactly equal for the two methods of tea preparation with respect to all such noise factors. Fisher suggested to resolve this problem by randomizing the order in which the cups are offered to the lady to taste. Note that randomization not only helps to avoid the bias but also makes the null distribution of the test statistic (i.e., all classifications equally likely) valid.

A need for blocking may arise in this experiment if a fully randomized order of tasting is not feasible or desirable for some reason. Offering A and B cups in pairs to improve the precision of the experiment is an example of blocking. Another example would be if it was of interest to find out if the lady’s discriminating ability ranges over different brands of tea, for example, Earl Grey and English Breakfast. In that case, the brand would be a blocking factor with an equal number of A and B cups of each brand.

We now introduce two basic designs that result from the principles of randomization and blocking. The first is called a completely randomized (CR) design, in which all experimental units are assigned at random to the treatments without any restriction. This design is generally used when the treatments are applied to a single group of essentially homogeneous experimental units. Any heterogeneity among these units is averaged out by random assignment of treatments to them.

Heterogeneity among experimental units can be more explicitly accounted for by designating selected measurable noise factors as blocking factors and forming blocks of units which are similar with regard to those factors. Treatments are then randomly assigned to the units within each block. Thus randomization is done subject to a blocking restriction. The resulting design is called a randomized block (RB) design. This design enables more precise estimates of the treatment differences because comparisons between treatments are made among homogeneous experimental units in each block.

Example 1.3 (Heat Treatment of Steel: CR and RB Designs)
Refer to Example 1.1. Suppose that 20 samples come from five different batches, four per batch. There are four treatment combinations: $A = ($Heating Temp.: Low, Quench Temp.: Low$)$, $B = ($Heating Temp.: High, Quench Temp.: Low$)$, $C = ($Heating Temp.: Low, Quench Temp.: High$)$, $D = ($Heating Temp.: High, Quench Temp.: High$)$. A CR design consists of randomly assigning the 20 steel samples to $A, B, C, D$ without regard to the batches that the samples came from; see Figure 1.4a. It is not required to assign an equal number to each treatment, although randomization is often done under that restriction.

The steel samples from the same batch are likely to be more similar to each other than are samples from different batches. Therefore we can form blocks of four samples from each of five batches and randomly assign them to the four treatments. We then have one replicate per treatment from each batch. The resulting RB design is shown in Figure 1.4b.
### Example 1.4  (Chemical Process Yield: CR and RB Designs)

The yield of a chemical process is to be compared under three different operating conditions: \( A, B, \) and \( C \). Suppose that six runs can be made in a day (three in the morning, three in the afternoon) allowing two runs per treatment. The CR design would randomize the order of the runs so as to neutralize any confounding effect of the time of day. One possible random order is \( \{A, C, A, B, B, C\} \). A practical difficulty in using a completely random order is that the operating conditions may not be easy to change. Then it may be necessary to make the two runs for each process condition consecutively; the resulting design is *not* a CR design but a **split-plot design** (discussed in Chapter 13).

Now suppose it is known from past experience that due to temporal trends the morning (AM) runs are different from the afternoon (PM) runs in terms of the yield obtained. Then the CR design given above is not balanced with respect to this noise factor since both \( A \) runs are made in the morning while both \( B \) runs are made in the afternoon. Therefore the \( A \)-versus-\( B \) difference is confounded with the AM-versus-PM difference. An RB design that uses the AM and PM runs as blocks provides the necessary balance. An example is \( \{B, C, A, C, B, A\} \).

---

### 1.4.2 Sequential Experimentation

Knowledge and discovery represent an iterative *exploratory* process in which learning occurs through induction–deduction cycles; see Box, Hunter, and Hunter (2005, Chapter 1) for a thorough discussion. One-shot experiments are used mainly for **confirmatory** purposes. In the exploratory phase the investigator generally begins with a theory or a hypothesis or a model (either suggested by previously collected data or derived from some basic scientific principles). For example, the theory might state that certain factors have positive effects on the process and certain other factors have negative effects. At the beginning of an investigation, the phenomenon may be very poorly understood with little good-quality data; therefore the theory may be rather fuzzy or almost nonexistent. To arrive at a testable theory, first an exploratory experiment must be conducted. At each later stage of the investigation, the data from the experiment must be checked if they conform with the predictions made by the current theory under
test. In rare cases the theory is confirmed and no further experimentation is necessary. More often, the theory is modified or a new theory is proposed because the original theory is refuted. This leads to a new cycle of experiments and a further modification of the theory.

In practice, as one learns from data, not only may the theory be modified but also the experimental conditions may be altered, as the following examples illustrate:

(a) Some of the factors used in earlier experiments may be found unsuitable and other more promising factors may replace them.
(b) The ranges of the factors may be changed because the previously used ranges may not have produced desired improvements or may have produced undesirable side effects.
(c) The response variable may itself be changed or redefined to better capture the outcome that the investigator wants to measure.
(d) The design may itself be changed, for example, from a CR design to an RB design.

There is no general methodology for designing sequential experiments that would deal with all of the above and possibly other changes dictated by data and any external scientific knowledge that becomes available during the course of the study. Methodologies have been developed for some specific applications. Two specific methodologies are sequential assemblies of fractional factorial designs (follow-up designs) to resolve confounding patterns (see Section 8.6) and the other is response surface optimization (see Section 10.1). The rest of this book deals with single experiments which are used as building blocks of sequential experiments.

When conducting experiments to search for the optimum factor-level combination to achieve the best response, the importance of a final confirmation experiment cannot be overemphasized. This experiment consists of a few runs to verify that the optimizing factor-level combination indeed meets the design objectives.

1.5 CHAPTER SUMMARY

(a) In an observational study an investigator passively observes potential predictor variables and the response variable. In an experimental study an investigator actively manipulates potential predictor variables to study their effects on the response variable. An observational study can only establish association between the predictor variables and the response variable. A controlled experiment can also establish a cause–effect relationship.
(b) A basic goal in any experiment is to identify the effects of the treatment factors on the response variable unconfounded from those of any noise factors. Blocking is used to evaluate the consistency of the effectiveness
of the treatments across experimental units having wide-ranging characteristics. It is also used to minimize the biasing effects of the noise factors that are easily controllable and on which data are available before experimentation begins. Randomization is used to minimize the biasing effects of all other noise factors. Completely randomized (CR) and randomized block (RB) designs are two basic designs based on these two techniques. The variation caused by random and measurement errors can be minimized through replication and repeat measurements, respectively.

EXERCISES

Section 1.1 (Observational Studies and Experiments)

1.1 Tell in each of the following instances whether the study is experimental or observational:

(a) Two computing algorithms are compared in terms of the CPU times required to do the same six test problems.

(b) A psychologist measures the response times of subjects under two stimuli; each subject is observed under both stimuli in a random order.

(c) A group of smokers and a group of nonsmokers are followed for 10 years and the numbers of subjects getting cancer over this period are noted.

(d) An advertising agency has come up with two different TV commercials for a household detergent. To compare which one is more effective, a pretest is conducted in which a sample of 50 adults is randomly divided into two groups. Each group is shown a different commercial, and the people in the group are asked to score the commercial.

(e) Two different school districts are compared in terms of the scores of the students on a standardized test.

Section 1.3 (Basic Terminology and Concepts of Experimentation)

1.2 A drilling experiment was conducted to study the effects of the bit size (0.25 and 0.5 in.), rotational speed (1000 and 2000 rpm), and feed rate (0.1 and 0.2 in./sec) on the surface finish of drilled holes. All combinations of these factors were run with three holes drilled in each case by the same operator in pieces of aluminum cut from the same stock. The coolant temperature was not controlled because its effect was not of interest and is difficult to control in practice.

(a) Which are the treatment and noise factors?

(b) How many levels of each treatment factor are studied? How many treatment combinations are studied?
(c) What are the experimental units? How many replicates are observed per treatment?
(d) Are the treatment factors crossed or nested?

1.3 A bottling plant has four beverage fillers, each with eight filling heads. To see what part of the variation in the filled amounts is due to the differences between the fillers and between the filling heads (with the rest being random variation), a random sample of five bottles is taken from two randomly selected filling heads of each filler (for a total of 8 filling heads and 40 bottles) and their contents are measured.

(a) Which are the treatment factors? Are they fixed or random? Are they crossed or nested?
(b) What are the experimental units? How many replicates are observed per treatment?

1.4 For a statistics project a student compared three methods of ripening raw bananas at home. The methods were: keep bananas in (i) open air, (ii) a brown paper bag, (iii) a basket covered with apples. A dozen bananas were ripened using these methods (four bananas per method) for one week, at the end of which the amount of ripening was measured by counting the number of black dots on the banana skins.

(a) Identify the treatments. Is there a control treatment?
(b) What are some of the noise factors in this experiment?
(c) What are the experimental units? How many replicates are observed per treatment?
(d) What is the response variable?

Section 1.4  (Basic Principles of Experimentation)

1.5 Refer to Exercise 1.2. Suppose that the drill operator randomized the sequence of the eight settings of the drill bit size, speed, and feed rate but made all three holes for each setting one after the other because it was too much work to change the settings after drilling each hole. Why is this a violation of the randomization principle? What systematic errors could this introduce? How should the randomization principle be correctly applied?

1.6 Refer to Exercise 1.4. What are the possible sources of systematic and measurement errors? What precautions would you take to minimize them?

1.7 An experiment is to be conducted to study the effects of having sugar (in the form of a donut) and caffeine (in the form of coffee) for breakfast
on the alertness of students in a morning class. Sixteen students are
selected to participate in the experiment. Alertness will be measured by
reaction times to a certain standardized stimulus. Outline how you would
conduct a completely randomized experiment. Specify the treatment
factors, treatments, experimental units, three possible noise factors, and
replicates.

1.8 Refer to the previous exercise. The sixteen students are recruited so that
there are four from each class: freshman, sophomore, junior, and senior.
Also, there are two men and two women from each class. Which of these
factors (class or gender) would be more appropriate for blocking? Outline
a randomized block design using the chosen blocking factor.

1.9 (From Mason, Gunst, and Hess 2003, Chapter 4, Exercise 1. Reprinted by
permission of John Wiley & Sons, Inc.) A test program was conducted
to evaluate how the quality of epoxy–glass–fiber pipes is affected by
operating condition (normal vs. severe) and water temperature (150 vs.
175°F). The test program required 16 pipes, half of which were man-
ufactured at each of two manufacturing plants. Table 1.1 lists the test
conditions that constituted the experimental protocol. Identify which of
the following features are included in this design and what they are:

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Plant</th>
<th>Operating Condition</th>
<th>Water Temperature (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Normal</td>
<td>175</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Normal</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Severe</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Severe</td>
<td>175</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Normal</td>
<td>175</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Normal</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Normal</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>Severe</td>
<td>175</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>Normal</td>
<td>175</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>Severe</td>
<td>150</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>Normal</td>
<td>150</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>Severe</td>
<td>175</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>Severe</td>
<td>175</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>Severe</td>
<td>150</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>Severe</td>
<td>150</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>Normal</td>
<td>175</td>
</tr>
</tbody>
</table>

Source: Mason, Gunst, and Hess (2003, p. 133). Reprinted by
permission of John Wiley & Sons, Inc.
(a) treatment factors, (b) treatments, (c) blocking factors, (d) experimental units, (e) replications, (f) randomization.

1.10 Consider the lady tasting tea experiment.

(a) Calculate the type I error probability if eight cups of tea are tested but one misclassification is allowed in the decision rule, that is, $H_0$ is rejected if there is more than one misclassification.

(b) What is the minimum number of cups that must be tested if the type I error probability for this decision rule must not exceed $\alpha = 0.05$?