thanks to Elnora Fairbank and Helen Ferris for their dedicated assistance in various stages of manuscript preparation. I'm very grateful to Rick Axelson for his statistical programming support throughout the preparation of the original manuscript and to Harika Basaran for originally preparing the exercise and example data disks. I am indebted to John Kimmel for his confidence and support in the formative stages of this book, and special appreciation is extended to Alex Kugushev for his dedication to excellence. I wish to express my gratitude to Carolyn Crockett, who had the confidence in the viability of the book to promote a second edition. A note of thanks to Kimberly Raburn for her invaluable guidance of all things important during the revision process. Finally, I want to acknowledge the critical reviews and many helpful suggestions by the many reviewers for both editions, including Richard Alldredge, Daniel C. Coster, Shu Geng, Robert Heckard, Hui-Kuang Hsieh, David Jowett, Larry J. Ringer, Oliver Schabenberger, and G. Morris Southward.

Robert O. Kuehl

1 Research Design Principles

All of our activities associated with planning and performing research studies have statistical implications. The principles we encounter in Chapter 1 form a basis for the structure of a research study; the structure, in turn, defines the study's function. If the structure is sound the study will function properly and produce the information it was designed to provide. If the structure is faulty the study will not function properly, and it will produce either incomplete or misleading information. The statistical principles are those associated with the collection of observations to gain maximum information for a research study in an efficient manner. They include treatment design, local control of variability, replication, randomization, and the efficiency of experiments.

1.1 The Legacy of Sir Ronald A. Fisher

No one person had as much impact on the statistical principles of experimental design in his time as Ronald A. Fisher. In October 1919, Fisher joined the staff at Rothamsted Experimental Station near Harpenden, England. He had been asked to come for a period of six months to one year to apply a thorough statistical analysis to agricultural research data that had been accumulated by the staff.

It was during his tenure at Rothamsted, where he remained until 1933, that he developed and consolidated basic principles of design and analysis that we today take as necessary practices for valid research results. From 1919 to 1925, he studied and analyzed wheat experiments on Rothamstead Experimental Station that had been conducted since 1843. From his statistical investigations of these experiments and others, Fisher developed the analysis of variance and unified his ideas on basic principles of experimental design.

In 1926 he published the first full account of his ideas in a paper, "The Arrangement of Field Experiments" (Fisher, 1926). In that seminal paper he outlined

and advanced three fundamental components for experiments in agricultural field trials: **local control** of field conditions for the reduction of experimental error, **replication** as a means to estimate experimental error variance, and **randomization** for valid estimation of experimental error variance. Although replication and local control were practiced at the time, his justifications with regards to experimental error variance were relatively new concepts. Randomization was a radical new concept met with skepticism and resistance by his contemporaries who, for the most part, did not understand its statistical implications. Each of these concepts is discussed in more detail in succeeding sections of this chapter.

Two books that grew out of Fisher's experiences at Rothamsted became standard references for researchers on the design and analysis experimental studies. *Statistical Principles for Research Workers* was first published in 1925 (Fisher, 1925) with 13 subsequent editions, and *The Design of Experiments* was published ten years later (Fisher, 1935) with 7 subsequent editions. His contributions to experimental design were but a few of his contributions to the science of statistics. A biography on the life and times of Fisher, written by one of his daughters (Box, 1978), pays homage to him as the consummate scientist.

1.2 Planning for Research

A **research program** is an organized effort on the part of a scientist to acquire knowledge about a natural or manufactured process. The total program may require many individual studies, each with specific objectives. The individual studies usually answer related questions and provide related pieces of information, which in concert meet the goals of the program. The design and analysis of the individual research study are the object of our attention in this book.

Good planning helps the scientist to organize the required tasks for a research study. The individual study requires the scientist to make a number of critical decisions. Consider the nutrition scientist who wanted to improve the standard method for evaluating nutritional quality of different protein sources. Although the acceptable standard procedure for evaluation was fairly rigidly defined by peer scientists, he hypothesized that the substitution of mice for rats as the standard test animal was more time- and cost-efficient. A test of his research hypothesis required a study to determine whether the mice were more efficient. Among the critical decisions he had to make were the number of mice and rats to use; the amount of protein to use in the diets; the various sources of protein necessary to validate the new protocol; the length of time to run the study; and the number of replications for the full experiment.

Documented Plans Prevent Oversights

The importance of developing a written plan cannot be overemphasized. Frequent reference to an existing document prevents serious oversights. The document also will be useful for subsequent insertions of notes and any alterations relating to the specific items in the original plan.

The astute investigator develops a checklist of specific considerations at the beginning of a study. Some typical items that a checklist can address are

- the specific objectives of the experiment
- identification of influential factors and which of those factors to vary and which to hold constant
- the characteristics to be measured
- the specific procedures for conducting tests or measuring the characteristics
- the number of repetitions of the basic experiment to conduct
- available resources and materials

Bicking (1954) presents a detailed checklist for planning a research study that can be consulted as a guide to develop a written plan.

Simple Questions to Focus Activities

Simple, but challenging, questions aid the design process, even though we may have a well-defined research hypothesis as an impetus for the research study.

Questions that focus our attention throughout the design process include "What is my objective?" "What do I want to know?" and "Why do I want to know it?" Productive follow-up questions for each activity in the process—such as "How am I going to perform this task?" and "Why am I doing this task?"—direct our attention to define the role of each activity in the research study.

Components of the research study are discussed separately in the following sections, but they are interconnected and an investigator must integrate those separate parts into an effective research study. We begin by establishing a small vocabulary to communicate our ideas.

1.3 Experiments, Treatments, and Experimental Units

Accurate communication requires that both parties respond to a common vocabulary with a common meaning. This section establishes the interpretation of some common terms and concepts as they are applied to scientific research studies.

For our purposes, an **experiment** shall be confined to investigations that establish a particular set of circumstances under a specified protocol to observe and evaluate implications of the resulting observations. The investigator establishes and controls the protocols in an experiment to evaluate and test something that for the most part is unknown up to that time.

The **comparative experiment** is the type of experiment familiar to investigators in the fields of biology, medicine, agriculture, engineering, psychology, and

other experimental sciences. The adjective *comparative* implies the establishment of more than one set of circumstances in the experiment, and that responses resulting from the differing circumstances will be compared with one another.

Treatments are the set of circumstances created for the experiment in response to research hypotheses, and they are the focus of the investigation. Examples of treatments are animal diets, cultivars of a crop species, temperatures, soil types, and amounts of a nutrient. Two or more treatments are used in a comparative study, and they are compared with one another for their effects on the subjects of the study.

The **experimental unit** is the physical entity or subject exposed to the treatment independently of other units. The experimental unit, upon exposure to the treatment, constitutes a single replication of the treatment.

Experimental error describes the variation among identically and independently treated experimental units. The various origins of experimental error include (1) the natural variation among experimental units; (2) variability in measurement of the response; (3) inability to reproduce the treatment conditions exactly from one unit to another; (4) interaction of treatments and experimental units; and (5) any other extraneous factors that influence the measured characteristics.

A beef cattle feeding trial provides an example of natural variation among experimental units. Two cattle of the same breed and herd receive the same amount of a diet; yet one steer gains 2.0 pounds per day, and the other gains 2.3 pounds per day over a one-month period.

The inability to reproduce treatment conditions exactly occurs when replicate test tubes are prepared independently, containing the same mixture of compounds, and the resultant chemical product is weighed and found to differ in each of the tubes by 0.1 μg. Weighing or pipetting processes are not exact; therefore, a small amount of variation is introduced at the treatment preparation stage.

A major objective in statistical calculation is the attainment of an estimate for the *variance of experimental error*. In its simplest form, the experimental error variance is the variance of observations on experimental units for which the differences among the observations can be attributed only to experimental error. Many of the statistical procedures we use require an estimate of this variance. Examples include confidence interval estimates for a mean and the two-sample Student t tests for the hypothesis of no differences between the means of two treatment populations.

Comparative observational studies are those studies for which we would like to conduct an experiment but cannot do so for practical or ethical reasons. The investigator has in mind conditions or treatments that have causal effects on subjects for which experiments cannot be conducted to elicit responses.

Investigators in the social sciences, ecology, wildlife, fisheries, and other natural resource sciences often must conduct observational studies in lieu of direct experimentation. The basic unit of study in the investigation may be human subjects, individual animals, habitats, or other microcosms; they have the same role as the experimental unit in the designed experimental study.

The subjects are either self-selected into identifiable groups or they simply exist in their particular circumstances. The groups or circumstances are used as treatment classifications in the observational study. By contrast, the investigator

assigns treatments to the experimental units in a designed experiment. For example, to study nitrification in soils from pure and mixed stands of mature pine and oak trees, a soil scientist selects existing pure and mixed stands of the two species and collects the necessary observations from the selected sites. A true experiment is quite impractical in this case because it requires an extraordinary expenditure of time to establish mature stands of the trees.

Ethical considerations sometimes prevent the use of experiments in lieu of observational studies. Consider a study to compare the severity of automobile accident injuries with and without the use of seat belts. It would be clearly unethical to assign anyone randomly to a "seat belt" or a "no seat belt" treatment and then collide the automobile into a concrete wall nor would anyone agree to this. Rather, investigators would rely on injury data collected from accidents and compare the "seat belt" data with the "no seat belt" data.

The nature of the scientific inference is the primary difference between the designed experiment and the observational study. With the designed experiment it is often possible to assign causal relationships between the responses and the treatments. Observational studies are limited to association relationships between the responses and treatment conditions.

1.4 Research Hypotheses Generate Treatment Designs

The research hypothesis establishes a set of circumstances and the consequences that follow from those circumstances. The treatments are a creation of the circumstances for the experiment. Thus, it is important to identify the treatments relative to the role they each have in the evaluation of the research hypothesis. A failure to clearly delineate the research hypothesis and objective of the study can lead to difficulties in the choice of treatments and to unsuccessful experiments.

The Relationship Between Treatments and Hypotheses

When treatments are chosen properly in response to a research hypothesis the underlying mechanisms may be better understood, whether they be physical, chemical, biological, or social. In some cases the objective may be to "pick the winner" to find one treatment that provides the desired response. In other cases, the experiment is used to elucidate underlying mechanisms associated with the treatments as they affect measured response variables. In the latter case sound research hypotheses motivate the selection of treatments.

It is incumbent on the investigator to ensure that the choice of treatments is consistent with the research hypothesis. It may be sufficient and less difficult to design a study solely to discover the best treatment. However, with a little extra effort even more fundamental information may be derived from the experiment in response to research hypotheses.

The following illustrate treatments used in actual research settings generated by research hypotheses:

- The drinking kinetics of honeybees was studied under different ambient temperatures to address the hypothesis that the energy required by honeybees to retrieve food for the colony was dependent on the ambient air temperature.
- The survival of Euphorbia seedlings under attack by a soil pathogen was determined for different types of fungicide treatment to address the hypothesis that not all fungicides were equally effective in controlling the soil pathogen.
- In traffic engineering, several methods of measuring traffic delay at an intersection were evaluated under different types of traffic signal configurations to address the hypothesis that the method of measuring delay was dependent on the type of configuration used for signaling traffic at an intersection.
- The development of social competence in young children was measured for its relationship to (1) parent education, (2) parent income, (3) family structure, and (4) age of child to address a complex research hypothesis that certain family demographics favorably affect the development of a child.

Note that in some research settings the treatments are conditions imposed by the investigator, such as those involving the honeybees and the survival of seedlings. On the other hand, in the traffic engineering and child development studies the treatments were those of existing conditions. Whether the investigator is performing a designed experiment or an observational study, he or she has the task of selecting the proper treatments to address the research hypotheses.

Frequently, additional treatments are required to fully evaluate the consequences of the hypotheses. An important component of many treatment designs is the *control treatment* discussed in the following section.

Control Treatments Are Benchmarks

The **control treatment** is a necessary benchmark treatment to evaluate the effectiveness of experimental treatments. There are several circumstances in which a control treatment is useful and necessary.

Conditions of the experiment may disallow the effectiveness of the experimental treatments that are known generally to be effective. A control of *no treatment* will reveal the conditions under which the experiment was conducted. For example, nitrogenous fertilizers are generally effective but will fail to produce responses in fields with high fertility. A control of *no nitrogenous fertilizer* will reveal the base fertility conditions for the experiment.

Sometimes treatments require manipulating the experimental units or subjects where the manipulation alone can produce a response. *Placebo* controls establish a basis for treatment effectiveness in these cases. The placebo unit or subject is processed just as the treatment units, but the active treatment is not included in their protocol. One of the most famous health experiments, the 1954 field trial of the Salk poliomyelitis vaccine, utilized placebo controls in approximately one-half of the test areas in the United States. The placebo was prepared to look just like the vaccine, but without the antigenic activity of the poliomyelitis vaccine. The placebo subjects were inoculated in the same fashion as were the subjects receiving the vaccine (Tanur et al., 1978).

Finally, the control may represent a *standard* practice to which the experimental method may be compared. In some situations it is necessary to include two distinct types of controls. For example, the no treatment and the placebo treatment can reveal the effect of manipulating the experimental unit in the absence of any treatment.

Multiple-Factor Treatment Designs Expand Inferences

In "The Arrangement of Field Experiments," Fisher (1926) noted that no proverb in connection with field experiments at that time was more repeated than that which said,

> We must ask Nature few questions, and ideally, one at a time.

He was convinced this was a mistaken view. In this regard, he wrote, "Nature ... will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed."

Fisher understood that in natural systems we don't know whether one treatment's influence is independent of others or its influence is related to variation in the other treatments. Consequently, the conditions under which treatments are compared can be an important aspect of the design.

For example, in a study on nitrogen production by *Rhizobium* bacteria in soil a comparison of interest was the amount of nitrogen produced in normal, saline, and sodic soils. However, the nitrogen production was known to be affected by the temperature and moisture conditions in the soil. In fact, the optimum conditions of temperature and moisture for nitrogen production may well have been different for each soil type. Consequently, the experiment was set up to test the nitrogen production at several temperatures in combination with several moisture conditions for each of the soils. Treatment designs of this type are known as **factorial treatment designs**, wherein one set of treatments, say soils, is tested over one or more other sets of treatments, such as moisture and temperature.

A **factor** is a particular group of treatments. Temperature, moisture, and soil type are each considered a factor. The several categories of each factor are termed *levels* of the factor. Temperature levels are 20°C, 30°C, and 40°C; while soil type

levels are normal, saline, and sodic. A *quantitative factor* has levels associated with ordered points on some metrical scale such as temperature. The levels of a *qualitative factor* represent distinct categories or classifications, such as soil type, that cannot be arranged in order of magnitude.

The factorial arrangement consists of all possible combinations of the levels of the treatment factors. For example, the factorial arrangement for three levels of temperature with three levels of soil type consists of $3 \times 3 = 9$ factorial treatment combinations. The nine combinations are

(20°C, normal)	(30°C, normal)	(40°C, normal)
(20°C, saline)	(30°C, saline)	(40°C, saline)
(20°C, sodic)	(30°C, sodic)	(40°C, sodic)

With this arrangement the investigator was able to evaluate the nitrogen production in each of the soil types and also to determine the influence of temperature conditions on the relative production by soil type. In addition, it was also possible to evaluate separately the influence of temperature on nitrogen production. The statistical analysis of factorial arrangements is discussed in Chapters 6 and 7.

1.5 Local Control of Experimental Errors

Precise and accurate comparisons among treatments over an appropriate range of conditions are the primary objectives of most experiments. These objectives require precise estimates of means and powerful statistical tests. Reduced experimental error variance increases the possibility of achieving these objectives. **Local control** describes the actions an investigator employs to reduce or control experimental error, increase accuracy of observations, and establish the inference base of a study.

The investigator controls (1) technique, (2) selection of the experimental units, (3) blocking or ensuring parity of information on all treatments, (4) choice of experiment design, and (5) measurement of covariates. Each of these is discussed in this section in more detail.

Technique Affects Variation and Bias

If experimental tasks are performed in a slipshod manner the observations will exhibit an increase in variation. *Technique* includes simple tasks such as accurate measurement, preparation of media, pipetting of solutions, or calibration of instruments. On a more complex level the investigator may have several alternative laboratory methods or instruments for measuring chemical or physical properties. The methods may vary in their accuracy, precision, and range of application. The researcher must choose the method or instrument that provides the most accurate and precise set of observations within the budgeted resources.

When technique adversely affects precision the estimated experimental error variances are unnecessarily inflated. Outlying observations caused by recording errors or extraordinary environmental conditions can increase variation. Whatever the

cause the investigator has to decide whether to include these observations in the analysis.

On the other hand, no discernible pattern may exist in the observations other than a general increase in their variability. This type of increased variation could point to faulty technique somewhere in the course of the experiment. The investigator would need to check out the selection of experimental units for the study, treatment protocol, measurement techniques, and personnel for sources of increased error and then attempt to make adjustments wherever necessary.

Poor techniques can affect the accuracy of observations, introducing bias into the results. Further, the variation introduced into the observations by poor technique is not necessarily random and, therefore, not subject to the same probability laws we associate with statistical inference.

Uniform application of treatments throughout the experiment increases the likelihood of unbiased measurement of the treatment effects. For instance, uniform amounts of food intake are required for accurate measures of differences between diets for animals. Uniform amounts of fertilizers applied to the replicate plots are required for accurate measures of crop yield differences in fertilizer trials.

Selection of Uniform Experimental Units

Heterogeneous experimental units produce large values for the experimental error variance. Precise comparisons among the treatments require the selection of uniform experimental units to reduce experimental error. However, an excessively stringent selection of experimental units can produce artificially uniform conditions. The narrow set of conditions restricts the inference base of the study. Therefore, to have reasonable confidence in the conclusions secured from the experiment, it is desirable to have units represent a sufficient range of conditions without unnecessarily increasing the heterogeneity of the experimental units.

The nature of the experiment dictates the balance between range of conditions and uniformity of units. For example, plant selections in a plant-breeding study should be tested over the range of conditions in which future varieties are expected to be cultivated; hence the range of conditions may be quite wide. If the selections are tested in several widely separated locations, then uniformity within a location becomes important. Selecting a uniform set of plots controls the variation within a location. Uniformity of units in an experiment with dairy cows requires selection of cows from the same breed, at the same stage of lactation, and with similar numbers of previous lactations.

Blocking to Reduce Experimental Error Variation

Fisher (1926) argued that no advantage had to be given up to have a valid estimate of error but two things were necessary:

> (a) that a sharp distinction should be drawn between those components of error which are to be eliminated in the field, and those which are not to be eliminated... (b) that

> the statistical process of the estimation of error shall be modified so as to take account of the field arrangement, and so that the components of error actually eliminated in the field shall equally be eliminated in the statistical laboratory.

Blocking provides local control of the environment to reduce experimental error. The experimental units are grouped such that the variability of units within the groups is less than that among all units prior to grouping. The practice of blocking or grouping experimental units into homogeneous sets goes hand in hand with experimental unit selection for uniformity. Treatments are compared with one another within the groups of units in a more uniform environment, and the differences between the treatments are not confused with large differences between experimental units. The variability associated with environmental differences among groups of units can be separated from experimental error in the statistical analysis.

Experimental units are blocked into groups of similar units on the basis of a factor or factors that are expected or known to have some relationship with the response variable or the measurement that is hypothesized to respond differentially to the several treatments.

Four Major Criteria for Blocking

Four major criteria frequently used to block experimental units are (1) proximity (neighboring field plots), (2) physical characteristics (age or weight), (3) time, and (4) management of tasks in the experiment.

The classical blocking practice had its origins in agricultural field experiments when contiguous plots were placed into one group, and each of the treatments was assigned to a plot in that group. Then a second set of contiguous plots was used in the same fashion, and so forth, leading to a *complete block* design.

The rationale for this type of grouping is that plots near one another are more alike than plots separated by greater distance in cultivated fields. Blocking patterns in agricultural trials may not necessarily be nice rectangular configurations. Existing variability patterns can require rather different blocking arrangements to reduce the experimental error.

Another natural blocking unit is the animal litter. The size of litters in some animal species allows several treatments to be accommodated by one litter. Body weight is used to advantage as a blocking factor in animal experiments if variation in body weight amplifies variation in the response variable.

Industrial experiments require homogeneous batches of raw materials. A replicated experiment may require more raw material than that provided by a single batch, and variation from batch to batch may increase experimental error. A single batch sufficiently large for one replication of all treatments can serve as a blocking unit.

Blocking is used to divide the experiment into reasonably sized units for uniform management of time or tasks. Days serve as convenient blocking units if only one replication of treatments can be harvested in the field or processed in the laboratory in a single day.

Individual technicians can serve as blocking units to avoid confusion of observer or technician variability with that of the treatments. Each person can be assigned to one replicate of each treatment when several people are available to record data or perform the laboratory analyses.

A Demonstration of Variance Reduction by Blocking

A **uniformity trial** best illustrates the potential effectiveness blocking may have for variance reduction in a research study. The uniformity trial is essentially an experiment in which the experimental units are measured but have not been subjected to any treatment. For example, the classic uniformity trial in agriculture may have been a field of wheat all of the same variety divided into plots all of the same dimension. The wheat yields were then measured on each plot. Because variation in agricultural fields can generally occur on gradients, it was possible to determine which groupings of adjacent field plots led to groups of plots with the smallest variance within the groups. In following years' experiments the treatments could be allocated within the groups of similar plots based on the results from the uniformity trial.

Similarly, baseline observation prior to treatment application on the measurement of interest, or some variable that is known to have a strong relationship to the measurement of interest, is equivalent to a uniformity trial for blocking purposes. For example, measurements such as weights, age, chemical composition, or cholesterol levels prior to treatment administration may suitably be used as blocking criteria if they have strong relationships to the measurement of interest or are themselves the measurement of interest.

Suppose we have observations on ten units from either a uniformity trial or measurements prior to treatment administration: 43, 72, 46, 66, 49, 68, 50, 76, 42, and 69. The mean and variance for the observations on these ten units are $\bar{x} = 58$ and $s^2 = 175$. Grouping the units into two blocks based on the size of those measurements, we have

Block 1: 43, 46, 49, 50, 42 $\quad \bar{x} = 46 \quad s^2 = 12.5$
Block 2: 72, 66, 68, 76, 69 $\quad \bar{x} = 70 \quad s^2 = 15.2$

Whereas the total variance among the ten units was equal to 175, the variances within the separate blocks have been reduced to 12.5 and 15.2, respectively. The component of error eliminated by blocking will be reflected in the variance between the two block means, 46 and 70. The variability within each of the blocks is now assumed to represent the natural variation that exists among the relatively uniform experimental units unencumbered by controllable environmental differences. Likewise, the comparisons among treatments within those blocks will not be influenced by those same controllable environmental differences.

Matching Strategies to Group Units

Grouping of units often utilizes strategies to match similar units. Subjects or units are chosen for treatment groups on the basis of equality with respect to all influential factors. Any variables thought to have an influence on the value of the observed characteristics of the subjects are candidates for controlling variables. Matched subjects have common values for the controlling variables with the exception of the treatment.

The matching strategies at the study design stage attempt to achieve comparability of subjects on all factors that could unduly bias the comparison of the treatments. *Pair matching* and *nonpair matching* are two general strategies employed in the selection of units or subjects.

With pair matching, a subject from each treatment is paired with subjects having the same controlling variable values in each of the other treatment groups. The values of the controlling variables may be chosen to be (1) *exact value* matches or (2) *caliper* value matches. Exact matches for human subjects are possible with variables such as gender, profession, use of seat belt, and education level. Exact matching of roadway segments in a study of traffic performance is possible with variables such as number of lanes, lane width, presence or absence of a median, and speed limit.

The caliper match allows a certain tolerance in the value of the matching variables. A study on manipulation of a forest ecosystem may require study sites to be matched by tree species composition, slope, and aspect. Exact site matches on composition, slope, and aspect are quite difficult to attain. However, it may be possible to obtain sites with similar values for any or all of the matching variables to the extent that the variation will not have a serious effect on treatment comparisons.

Nonpair matching may be accomplished through (1) a *frequency* or (2) a *mean* strategy. The frequency method stratifies the units into groups on the basis of the controlling variables. Suppose age is a potential influential variable in a study involving human subjects. The subjects can be stratified from a frequency distribution of their ages such that there is a sufficient number of subjects in each age stratum to accommodate all treatment groups.

The mean-based nonpair strategy groups subjects or units in such a way that the average values of the controlling variables are the same in each of the treatment groups. Experimental animals can be grouped such that their average weight is the same for all treatment groups.

The nature of the research study dictates the most effective matching strategy and whether matching is a desirable protocol. Details on matching methods, including advantages and disadvantages, for comparative observational studies can be found in Cochran (1983), Kish (1987), and Fleiss (1981).

Experiment Designs Accommodate Treatment Designs

The **experiment design** is the arrangement of experimental units used to control experimental error and, at the same time, accommodate the treatment design in an experiment. A wide variety of experiment design arrangements to control experimental error exists in the published literature, and there is a natural tendency to design an experiment that conforms to the existing designs. But, a more appropriate attitude is to develop an experiment design that satisfies the demands of the current experiment.

The attainment of maximum information, precision, and accuracy in the results, along with the most efficient use of existing resources, is a guiding principle in choosing an appropriate experiment design.

The association between a treatment design and an experiment design is illustrated here under two different settings. The first illustration shows how three treatments are associated with six experimental units when each treatment is assigned to two experimental units. The second illustration shows how the three treatments are assigned to the six experimental units after the units are blocked into two sets of three homogeneous units.

An Experiment Design Without Blocking

Consider an experiment to compare three gasoline additives for their effect on carbon monoxide emission. Automobile engines are used as experimental units. Each of the three gasoline additives is to be used with two engines. A schematic representation of the design is shown in Figure 1.1. The boxes represent automobile engines as experimental units. One of the three gasoline additives is administered to an engine independently of the other engines in the experiment. The three gasoline additive treatments are randomly allocated to the six engines, two units for each treatment.

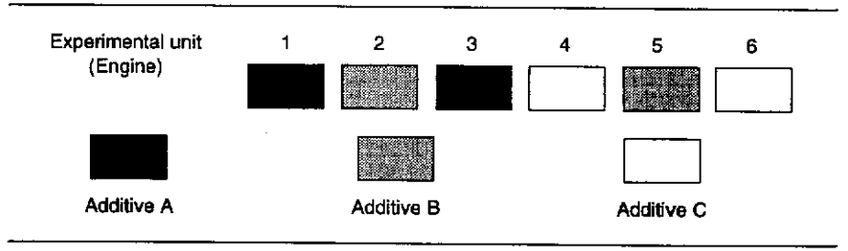

Figure 1.1 Illustration of a completely randomized design, with three treatments, each on two experimental units

This design, known as the **completely randomized design**, is the simplest of the experiment designs. The treatments are allocated to the experimental units at random. Each experimental unit has the same chance of receiving any treatment. The function of randomization in experiment designs is discussed in Section 1.8.

An Experiment with One Blocking Criterion

The completely randomized design provides little control over environmental variation. Several general classes of designs use blocking and grouping of experimental units to control environmental variation. The simplest blocking design is the **randomized complete block design** with one blocking criterion. This design employs one restriction on the random assignment of treatments to experimental units; all treatments must occur an equal number of times in each block.

Consider the arrangement of units in a randomized complete block design for an experiment to study the effects of three diets on the growth of laboratory mice. Three mice of the same gender were selected from each of two litters. The litters were used as blocks for the experiment as shown in Figure 1.2. The three diets were randomly assigned to the three mice in each litter.

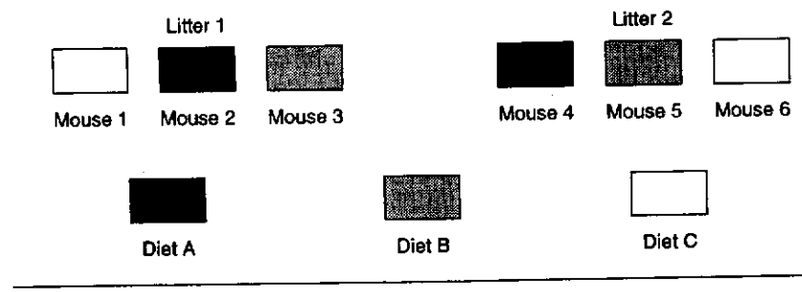

Figure 1.2 Illustration of a randomized complete block design, with three experimental diets tested in two litters of mice

The use of a litter as a block reduces experimental error variation because it isolates the variation among litters from the variation among diets. The diets can be compared under uniform conditions when each of them is tested in the same litter. A complete discussion of the randomized complete block design and its analysis is given in Chapter 8.

Covariates for Statistical Control of Variation

Covariates are variables that are related to the response variable of interest. The information on covariates is used to enact a statistical control on experimental error variance by a procedure known as the *analysis of covariance.*

It was suggested earlier that body weight can be used as a criterion for blocking. After animals are grouped according to weight no further use is made of those values. But, the actual body weights can be used effectively in the statistical model to reduce the estimates of experimental error in lieu of or in addition to using them for blocking. Body weight would be a covariate of weight gain in the experiment. Other examples of covariates might be pretest scores, field plot fertility, previous

year's yield in perennial crops, or purity of raw material in a chemical process; all may vary from unit to unit. Any attribute that is measurable and thought to have a statistical relationship to the variable of primary interest is a candidate for covariate adjustment.

The prime requirement is that the covariate be unaffected by the particular treatments used in the experiment. In practice the covariates are measured before the treatments are administered or before the treatments have had time to develop a response by the unit, or we can assume the treatment never has an effect on the covariate.

The observations in the experiment consist of pairs of observations (x, y) on each experimental unit, where y is the variable of interest in the experiment and x is the covariate. Suppose there are six such pairs of observations for each of two treatment groups, and the data are plotted as shown in Figure 1.3.

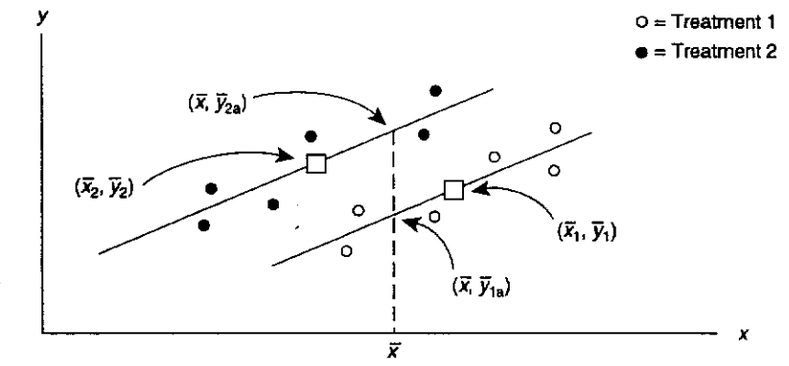

Figure 1.3 An illustration of covariate adjustment for two treatment groups

A portion of the variation in y is associated with x as well as with the treatment effects if there is a statistical relationship between the variable of interest, y, and covariate x. The analysis of covariance estimates the regression relationship between y and the covariate x to statistically reduce the experimental error variance. The average responses to each treatment, $\bar{y}_1$ and $\bar{y}_2$, are adjusted to the same value of the covariate, usually the grand mean $\bar{x}$ as shown in Figure 1.3. A comparison between the adjusted treatment means, $\bar{y}_{1a}$ and $\bar{y}_{2a}$, eliminates the influence of the covariate x on the comparison. For example, if weight gain has the covariate of initial body weight, the average weight gain response to a treatment is adjusted to remove the variation associated with initial body weight. The adjusted treatment means for weight gain represent the weight gains that would be obtained if all animals had the same initial body weight. The analysis of covariance is discussed in Chapter 17.

1.6 Replication for Valid Experiments

The scientific community regards replication of experiments to be a prime requisite for valid experimental results. **Replication** implies an independent repetition of the basic experiment. More specifically, each treatment is applied independently to each of two or more experimental units.

There are several reasons for replicating an experiment, most notably:

- *Replication* demonstrates the results to be reproducible, at least under the current experimental conditions.
- *Replication* provides a degree of insurance against aberrant results in the experiment due to unforeseen accidents.
- *Replication* provides the means to estimate experimental error variance. Even if prior experimentation provided estimates of variance, the estimate from the present experiment may be more accurate because it reflects the current behavior of observations.
- *Replication* provides the capacity to increase the precision for estimates of treatment means. Increasing replication, r, decreases $s^2_{\bar{y}} = s^2/r$, thereby increasing the precision of $\bar{y}$.

Observational Units and Experimental Units Can Be Distinctly Different

The *observational unit* may not be equivalent to the experimental unit. The observational unit can be a sample from the experimental unit, such as individual plant samples from a field plot or serum samples from a subject.

The variance of the observations on the experimental units is the experimental error variance. It is a valid measure of the variation among the experimental units that have *independently* had the treatments administered to them.

The variance among multiple observations from the same experimental unit often is used mistakenly as a measure of experimental error for comparisons among treatments. The following examples may help to clarify the distinction between replicated experimental units and multiple observations from the same experimental unit.

Example 1.1 A simple animal diet ration study has one cage or pen of six animals assigned to ration A and a second cage or pen of six animals assigned to ration B. Weight gain or some other appropriate data are collected to test the efficacy of the rations. The necessary measurements are made on each of the animals in the pens at the end of the study. The schematic in Figure 1.4 illustrates the design.

Typically, the Student t test using the difference between the means of the two pens would be used to test a hypothesis of no difference between the rations.

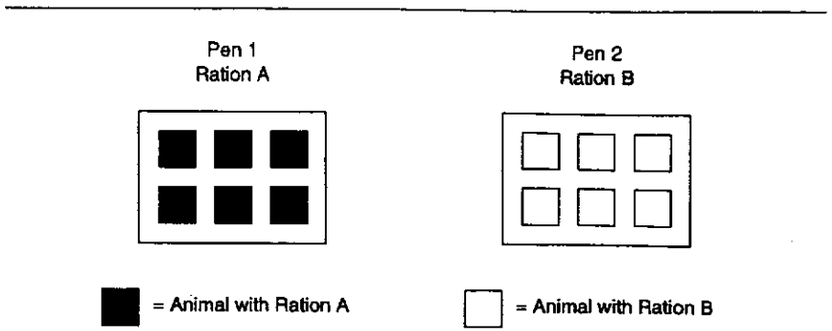

Figure 1.4 Illustration of an unreplicated experiment

However, the differences between the two pens of animals can be caused by the effect of other factors in addition to the treatments. The response to a given ration may vary with different pens of animals. This variation in response could be due to one or more of the factors contributing to experimental error.

The natural variation in the response from one pen to another, or the *pen effect*, will contribute to experimental error. Also, variation in the preparation and presentation of the treatment to the separate pens can cause a treatment–pen interaction. Therefore, any differences between the two rations in Example 1.1 cannot be attributed clearly to the rations alone. The differences may be attributable to combinations of the treatment effects, pen effects, and treatment–pen interactions.

The experiment will not have resolved clearly the question of whether the two rations differed in their effect on weight gain. The experiment has only one true replication. The *pen* is the *experimental unit* because that is the unit to which the treatment was administered independently. The *animals* in the pen are the *observational units*. The variance calculated among the observations on the animals within the pens is only an estimate of observation error within a pen of animals and not an estimate of variance among the experimental units.

Example 1.2 Suppose the experiment in Example 1.1 is restructured such that the animals are randomly divided into four pens of three animals each. Further, each of the two rations is randomly assigned to two pens of animals as shown in Figure 1.5. The rations are administered independently to each of the pens.

Each ration has been replicated properly in Example 1.2. The *experimental units* are the *pens* to which the rations have been administered independently (two pens per ration), and the *animals* within the pens are still the *observational units*. Therefore, the response from the experimental unit is the average pen response.

The estimate of experimental error variance, s^2, calculated as the variance among the pen means within each of the rations, is the appropriate variance for the

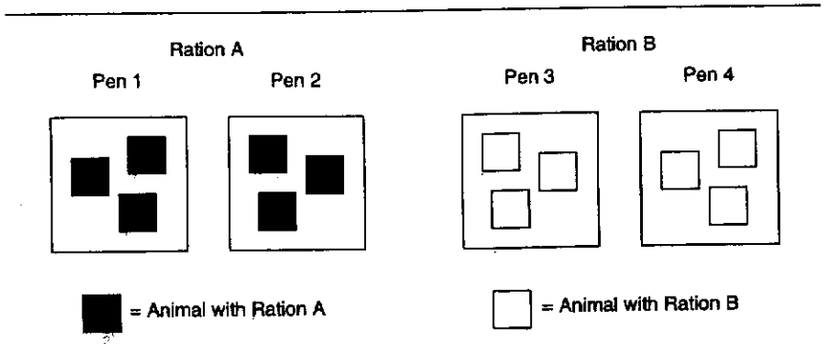

Figure 1.5 Illustration of a replicated experiment

Student t test. The variance among the animals within each pen, say s_w^2, is a measure of the variability of the observational units within the replicate pens.

Thus, two levels of variation are present in this type of study: (1) the variation among observational units *within* experimental units (s_w^2), and (2) variation *among* the experimental units (s^2).

It is important to distinguish which units of the study constitute the experimental unit and consequently which units constitute true replication of the treatment.

More details on the statistical models and analysis for the pseudoreplication and true replication exhibited by Examples 1.1 and 1.2 may be found in Addelman (1970). Examples of pseudoreplication in agronomic studies may be found in Nelson and Rawlings (1983). Hurlbert (1984) provided numerous examples of pseudoreplication in ecological field experiments.

1.7 How Many Replications?

The number of replications in a research study affects the precision of estimates for treatment means and the power of statistical tests to detect differences among the means of treatment groups. However, the cost of conducting research studies constrains the number of replications for a reasonably sized study. Thus, replication numbers are determined on the basis of practical constraints that we can assign to the problem.

Replication Numbers for Testing Hypotheses

The method for determining the number of replications is often based on a test of a hypothesis about differences among treatment group means. An elementary method for experiments with two independent samples is used here to illustrate a few attributes of the replication number problem.

The method is based on a test of a hypothesis about the difference between two treatment group means $d = m_i - m_j$, with known experimental error variance s^2, using the standard normal distribution test statistic. The method determines the number of replications required to test the difference between two independent sample means with specified Type I and Type II errors.

The required number of replications is affected primarily by four factors that are required for calculations:

- the variance (σ^2)
- the size of the difference (that has physical significance) between two means (δ)
- the significance level of the test (α), or the probability of Type I error
- the power of test $1 - \beta$, or the probability of detecting δ, where β is the probability of a Type II error

The required replication number for each treatment group, r, for two-sided alternatives is estimated with

$$r \geq 2[z_{\alpha/2} + z_\beta]^2 \left(\frac{\sigma}{\delta}\right)^2 \tag{1.1}$$

where $z_{\alpha/2}$ is the standard normal variate exceeded with probability $\alpha/2$ and z_β is exceeded with probability β. Probabilities for the standard normal variable are found in Appendix Table I.

The replication number can be estimated with knowledge of the percent coefficient of variation, $\%CV$. The $\%CV$ is substituted for σ in Equation (1.1), where $\%CV = 100(\sigma/\mu)$. The difference (δ) must be expressed as a percentage of the overall expected mean of the experiment, $\%\delta = 100(\delta/\mu)$, in Equation (1.1).

The influence of the coefficient of variation ($\%CV$), the percent difference ($\%\delta$), the power ($1 - \beta$), and the significance level (α) on required numbers of replications is shown in Table 1.1. Although the values in Table 1.1 only apply to two independent samples, the influences are similar in more complex experiments.

Required replication numbers generally increase if

- the variance, $\%CV$ or σ^2, increases
- the size of the difference between two means, $\%\delta$ or δ, decreases
- the significance level of the test, α, decreases
- the power of test, $1 - \beta$, increases

Calculated values for required replication numbers are estimates and approximations. Often they are determined on the basis of variance estimates associated with previous studies and not the variances from the actual study that will be used

Table 1.1 Number of replications required for a given coefficient of variation (%CV) and probability ($1-\beta$) of obtaining a significant difference of %δ between two treatment means with a two-sided test at the α significance level

		$\alpha = .05$		$\alpha = .01$	
		%δ		%δ	
%CV	$1-\beta$	10	20	10	20
5	.80	4	1	6	2
	.95	7	2	9	3
10	.80	16	4	24	6
	.95	26	7	36	9

to compute confidence intervals and tests of hypotheses. The determination of replication numbers for analysis of variance applications will be considered in Chapters 2, 5, 6, and 7. Commercial software programs are available to determine replication numbers for many different types of experimental studies.

1.8 Randomization for Valid Inferences

> In reconciling thus the two desiderata of the reduction of error and of the valid estimation of error,... no principle is in the smallest degree compromised. An experiment either admits of a valid estimate of error, or it does not; whether it does so, or not, depends not on the actual arrangement of plots, but only on the way in which that arrangement was arrived at. (Fisher, 1926)

Replication of the experiment provides the data to estimate the experimental error variance. Blocking provides a means to reduce experimental error. However, replication and blocking alone do not guarantee *valid* estimates of experimental error variance or valid estimates of treatment comparisons.

Fisher (1926) was making the point that **randomization**, alone, in the experiment provides a valid estimate of error variance for justifiable statistical inference methods of estimation and tests of hypotheses. Randomization is the random assignment of treatments to experimental units.

A Rationale for Randomization

Our analysis of data from experiments assumes the observations constitute a random sample from a normally distributed population. This assumption is plausible for comparative observational studies that use random samples of the available observation units from different treatment populations. However, whether experimental units can be considered a random sample is questionable when they are carefully selected, controlled, and monitored in experiments.

Independent observations are critical for estimation and tests of hypotheses because they provide valid estimates of experimental error variance. But, the assumption of independence among the experimental units cannot be justified when relationships exist among them. For example, it is well known that field plots tend to respond more similarly when they are adjacent. Any type of proximity can produce correlated responses whether it be physical location of units or temporal performance of tasks on the units.

Fisher (1926) recognized these potential difficulties with field plot experiments and justified random assignment of treatments to experimental units as the means to obtain valid estimates of experimental error variance. In a more detailed discussion, Fisher (1935) showed that randomization provided appropriate reference populations for statistical inferences free of any assumptions about the distribution of the observations. He showed that significance tests could be based on the distribution created by randomization and that the normal theory tests provided reasonable approximations to these test results. Thus, the random allocation of treatments to the experimental units *simulates* the effect of independence and permits us to proceed as if the observations are independent and normally distributed. These **randomization tests**, illustrated in this section, form the basis for valid statistical inferences in properly randomized experiments.

Further justification for randomization (Cochran & Cox 1957; Greenberg 1951; and Ostle & Mensing 1975) is based on the need to eliminate biases in the comparison of treatments that arise through systematic assignment of treatments to experimental units. If, for example, procedure A is always performed before procedure B, any systematic variation over time will bias the resulting comparisons between A and B. Randomization over these potential systematic sources of variation ensures estimates of treatment means differ from true values only by random variation.

Randomization Tests Show Utility of Randomization

The utility of randomization can be demonstrated with a randomization test that makes no assumptions about the form of the probability distribution for the observations. Randomization creates a population of experiments that could have been performed, although only one arrangement has been chosen at random for the actual experiment.

The randomization test evaluates the test statistic for all possible arrangements of treatments on the experimental units. The *randomization distribution* is the distribution of those values that would be obtained under the null hypothesis of no treatment effects.

Example 1.3 Illustration of a Randomization Test

Consider an experiment in which two treatments are randomly assigned to seven experimental units. Four experimental units receive treatment A and

three experimental units receive treatment B, with the following set of responses:

Unit:	1	2	3	4	5	6	7
Treatment:	B	B	A	A	B	A	A
Response:	14	16	19	17	15	13	17

If there is no difference between the effects of treatments A and B, then they are merely labels on the experimental units and do not affect the results. For example, if the null hypothesis is true, the response for unit 1 would be 14 regardless of the treatment applied. The same is true for each of the other units under the null hypothesis.

The A and B labels (four A's and three B's) may be allocated to the seven experimental units in $7!/4!\,3! = 35$ possible arrangements. These are the 35 experiments possible if the treatments are randomly allocated to the units.

All 35 of the possible arrangements are shown in Table 1.2, along with the difference between the group means $(\overline{y}_A - \overline{y}_B)$ based on the labels assigned to the units in each arrangement. The 35 differences $(\overline{y}_A - \overline{y}_B)$ are 35 possible differences that could occur for the 35 possible randomizations if the null hypothesis is true. They make up the randomization distribution under the null hypothesis.

Table 1.2 Thirty-five possible arrangements of four A's and three B's to seven experimental units with the mean difference $\overline{y}_A - \overline{y}_B$

Arrangement	*Unit:*	1	2	3	4	5	6	7	
	Response:	14	16	19	17	15	13	17	$\overline{y}_A - \overline{y}_B$
1		A	A	B	B	A	A	B	−3.17
2		A	B	B	A	A	A	B	−2.58
3		A	B	B	B	A	A	A	−2.58
4		A	A	B	A	B	A	B	−2.00
5		A	A	B	B	B	A	A	−2.00
6		A	B	A	B	A	A	B	−1.42
7		A	B	B	A	B	A	A	−1.42
8		B	A	B	A	A	A	B	−1.42
9		B	A	B	B	A	A	A	−1.42
10		A	A	A	B	B	A	B	−0.83
11		A	A	B	A	A	B	B	−0.83
12		A	A	B	B	A	B	A	−0.83
13		B	B	B	A	A	A	A	−0.83
14		A	B	A	A	B	A	B	−0.25
15		A	B	A	B	B	A	A	−0.25

(*Continued on next page*)

Table 1.2 (Continued)

Arrangement	*Unit:*	1	2	3	4	5	6	7	
	Response:	14	16	19	17	15	13	17	$\overline{y}_A - \overline{y}_B$
16		A	B	B	A	A	B	A	−0.25
17		B	A	A	B	A	A	B	−0.25
18		B	A	B	A	B	A	A	−0.25
19		A	A	A	B	A	B	B	0.33
20		A	A	B	A	B	B	A	0.33
21		B	B	A	A	A	A	B	0.33
22		B	B	A	B	A	A	A	0.33
23		A	B	A	A	A	B	B	0.92
24		A	B	A	B	A	B	A	0.92
25		B	A	A	A	B	A	B	0.92
26		B	A	A	B	B	A	A	0.92
27		B	A	B	A	A	B	A	0.92
28		A	A	A	A	B	B	B	1.50
29		A	A	A	B	B	B	A	1.50
30		B	B	A	A	B	A	A	1.50
31		A	B	A	A	B	B	A	2.08
32		B	A	A	A	A	B	B	2.08
33		B	A	A	B	A	B	A	2.08
34		B	B	A	A	A	B	A	2.67
35		B	A	A	A	B	B	A	3.25

Consider an alternative hypothesis H_a: $\mu_A - \mu_B \neq 0$ to the null hypothesis H_0: $\mu_A - \mu_B = 0$. The arrangement of the actual experiment is arrangement 30 with a mean difference $(\overline{y}_A - \overline{y}_B) = 1.50$. An absolute difference of 1.50 or larger occurs with 13 arrangements. Under the null hypothesis an absolute mean difference of 1.50 or larger occurs with a frequency of $\frac{13}{35}$, yielding a significance level of 0.37. On the basis of the observed results of the experiment, arrangement 30, there is no reason to reject the null hypothesis.

Under the assumption of a true null hypothesis, H_0: $\mu_A - \mu_B = 0$, the randomization test enabled an evaluation of the test statistic from the actual experiment, $(\overline{y}_A - \overline{y}_B) = 1.50$, against the values for $\overline{y}_A - \overline{y}_B$ from all other members of the population of 35 possible experiments.

Normal Theory Tests Approximate Randomization Tests

Fisher (1935) first demonstrated that normal theory tests are good approximations to the randomization tests provided there has been a random allocation of the treatments to the experimental units and sample sizes are reasonably large.

Approximations to the randomization test by normal theory tests improve as sample size increases. Practical guides to randomization tests for various experimental situations may be found in Edgington (1987) and Manly (1991).

Rigorous treatments of the randomization models and tests of significance in experiment designs are provided by Kempthorne (1952), Scheffé (1959), Mead (1988); and Hinkelmann and Kempthorne (1994). More detailed discussions of randomization in connection with statistical inference may be found in Kempthorne (1966, 1975).

It has become common practice to describe the statistical models for experimental studies in terms of the normal theory models. The formalities of the normal theory models are more straightforward than those for the randomization models, however, the use of normal theory models for experiments can be justified only under the randomization umbrella.

Restricted Randomization for Difficult Circumstances

Randomization may result in an arrangement of treatments that is unsatisfactory from the point of view of scientific validity. Sequential arrangements, such as AAA BBB CCC or ABC ABC ABC, are possible with random assignment. However, the sequential arrays of treatments can lead to problems of bias.

Hurlbert (1984) questions the blind use of randomization in small-scale ecological studies. In small studies, there is a very high possibility of randomized layouts resulting with treatments markedly segregated from each other in space or time. Segregation could lead to spurious treatment effects in which treatment and space effects are confounded. Hurlbert (1984) argues that there must be some physical interspersion of the treatments to avoid the systematic segregation of treatments in small experiments.

Yates (1948) and Youden (1956) independently introduced restricted randomization as a solution to the problem of bad patterns from complete randomization. Restricted randomization omits certain arrangements of treatments that in the opinion of the experimenter are unacceptable for the particular study under consideration.

Industrial experiments may require elaborate dismantling of experimental apparatuses between certain types of treatments. Youden (1972) gave examples of industrial and laboratory experiments in which the cost of switching from one treatment to another might outweigh the advantages of complete randomization.

Bailey (1986, 1987) discussed some history and recent research on the topics of restricted randomization. Schemes for restricted or constrained allocation of treatments have been developed that permit the usual analysis of variance procedures (Bailey, 1986; Youden, 1972).

1.9 Relative Efficiency of Experiment Designs

Relative efficiency measures the effectiveness of blocking in experiment designs to reduce experimental error variance. In practice, relative efficiency is measured to determine the efficiency of the design *actually* used relative to another simpler design that *could have been* used but was not. For example, the efficiency of a randomized complete block design is determined relative to a completely randomized design.

The variance of a treatment mean $\sigma^2_{\bar{y}} = \sigma^2/r$ is a measure of the precision of the estimated treatment means in an experiment. The precision of an estimate for the treatment mean is controlled by the magnitude of σ^2 and the number of replications, r, both of which are to some extent under the control of the investigator. The investigator may increase the number of replications to decrease $\sigma^2_{\bar{y}}$ and increase the precision of the mean estimate. The investigator may also attempt to reduce σ^2 through various local control activities (such as blocking), thereby increasing the precision of the experiment. A method for measuring the effectiveness of blocking is discussed in this section.

The use of $\sigma^2_{\bar{y}}$ as a measure of precision provides a means for comparing the relative precision of two experiment designs. Suppose one design has a true experimental error variance of $\sigma_1^2 = 1$, and a second design has a true experimental error variance of $\sigma_2^2 = 2$. The value of $\sigma^2_{\bar{y}} = \sigma^2/r$ will be the same for the two designs if the second design has twice as many replications as the first design. That is, the variance of a treatment mean in each of the designs is

$$\text{Design 1:}\quad \sigma^2_{\bar{y}_1} = \frac{\sigma_1^2}{r_1} = \frac{1}{r_1}$$

and

$$\text{Design 2:}\quad \sigma^2_{\bar{y}_2} = \frac{\sigma_2^2}{r_2} = \frac{2}{r_2}$$

The variances $\sigma^2_{\bar{y}_1}$ and $\sigma^2_{\bar{y}_2}$ will be the same only if $r_2 = 2r_1$, or Design 2 has twice the replication as that of Design 1. Therefore, Design 1 is more efficient than Design 2 with respect to the number of replications required to have the same precision for an estimate of the treatment mean.

In practice, σ^2 is unknown for each of the designs and must be estimated from the data. Also, the degrees of freedom for the estimate of the variance changes with the designs. Under these circumstances the relative precision of the two designs is determined on the basis of the concept of *information* (Fisher, 1960). Fisher calculated the amount of information that the estimated difference between two means provides about the true difference between the population means. Information calculated from this concept is

$$I = \frac{(f+1)}{(f+3)}\frac{1}{s^2} \tag{1.2}$$

where s^2 is the estimated experimental error variance with f degrees of freedom. If σ^2 is known, then $I = 1/\sigma^2$ and the coefficient $(f+1)/(f+3)$ is replaced by unity. For any reduction in variability, s^2, there is a concurrent increase in the information one has about the mean difference of the populations. Precision and information both increase as the variability decreases.

The *relative efficiency* of two experiment designs is defined as the ratio of information in the two designs. Suppose

$$I_1 = \frac{(f_1+1)}{(f_1+3)}\frac{1}{s_1^2} \quad \text{and} \quad I_2 = \frac{(f_2+1)}{(f_2+3)}\frac{1}{s_2^2}$$

are the estimated information measures of Design 1 and Design 2, respectively. The relative efficiency of Design 1 to Design 2 is estimated as

$$RE = \frac{I_1}{I_2} = \frac{(f_1+1)(f_2+3)}{(f_1+3)(f_2+1)}\frac{s_2^2}{s_1^2} \tag{1.3}$$

If $RE = 1$, then the information in the two designs is equal, and the designs each require the same number of replications to have the same variance of treatment means $\sigma_{\bar{y}}^2$. If $RE > 1$, then Design 1 is more efficient than Design 2. For example, if $RE = 1.5$, then Design 2 requires 1.5 times as many replications as Design 1 to have the same variance of a treatment mean.

1.10 From Principles to Practice: A Case Study

The design of an experimental medical study reported by Moon et al. (1995) illustrates the process of putting principles into practice. The components of the study design provide examples of how the principles of research design, covered in this chapter, helped prepare the investigators to address their research hypothesis. The following is a brief description of the main research design elements in that study. Other details on the study can be found in the publication.

The Problem

Non-melanoma skin cancer, which includes basal cell (BCC) and squamous cell (SCC) carcinomas, is the most common type of cancer. Residents of Arizona experience a three to seven times greater incidence of these cancers than the general U.S. population. Although non-melanoma cancers are usually not life threatening, they result in substantial morbidity and treatment expense.

A history of actinic keratoses (AK), a type of skin lesion, has generally been accepted as a marker for identifying individuals at increased risk of skin cancer. In

many cases AK lesions progress to non-melanoma skin cancer and are thus premalignant lesions. Those individuals who have a history of actinic keratoses, but few if any non-melanoma skin cancers, are considered to be at moderate risk. Also, fair-skinned, older people with a long history of sun exposure are more at risk for non-melanoma skin cancer.

Results from recent clinical and laboratory studies suggested that vitamin A and other retinoids have a preventative effect against cancer in epithelial tissues such as skin. However, these studies included a small number of subjects and did not produce reliable estimates of the vitamin A effect in the primary prevention of human skin cancer.

A five-year clinical trial was to be conducted in southern Arizona to evaluate the effectiveness of vitamin A or retinol supplementation to reduce the risk of non-melanoma skin cancer for individuals already at moderate risk.

Research Hypothesis

Retinol (vitamin A) supplementation reduces the incidence of skin cancer in moderate-risk individuals with a history of at least ten actinic keratoses.

Treatment Design

Important considerations for the dosage level of retinol (vitamin A) included the need to elevate the daily intake of retinol above the common intake of most adults and to avoid a dosage that could potentially induce adverse side effects known to be associated with excessive retinol intake. A placebo treatment was necessary as a comparison group for the treated group; these were subjects with the same characteristics as the treated subjects and on the same follow-up protocols during the course of the study. The subjects could not be told to which treatment group they were assigned in order to keep the subjects from both groups on the same regimen.

Treatment:	Daily, self-administered, dietary supplement of 25,000 IU retinol in capsule form.
Placebo:	Daily, self-administered, placebo capsule

Measurements of Interest

The hypothesis addressed the risk of skin cancer in its relationship to levels of retinol. Therefore, the measurement of interest was whether a subject developed skin cancer during the course of the study. The analysis could consider several approaches to test the hypothesis. The approaches included whether a cancer developed, a binary outcome; how many cancers developed, a count variable; or how long it took for a cancer to develop, which is a time-to-event measurement used in survival analysis. Each of these approaches could be extracted by recording

the time it took to develop a cancer, if one developed. Thus, the measured variables were the

- time to the first and each subsequent, if any, basal cell carcinoma (BCC)
- time to the first and each subsequent, if any, squamous cell carcinoma (SCC)

Selection of Subjects with Common Characteristics

Subjects for this study were required to be representatives of a healthy adult population with a moderate risk to non-melanoma skin cancer, willing to participate in the study, and not currently ingesting an excessive amount of vitamin A in their diet. More than 11,000 subjects were screened for the study; approximately 25% were deemed eligible. Following are some of the criteria used by the investigators to screen the subjects.

The subjects were recruited through referrals by dermatologists and media announcements. The subjects could be men or women with a history of at least ten actinic keratoses diagnosed clinically, with the most recent diagnosed in the past year. They could have had no more than two prior occurrences of SCC or BCC and no cancer diagnosis other than SCC or BCC in the previous year.

The eligible individuals had to be between 21 and 84 years of age, be ambulatory, be capable of self-care, have no diagnosis of a life-threatening disease, intend to be a resident of Arizona for the next five years, and be willing to commit to semiannual follow-up clinic visits for that period. They also had to be willing to limit non-study vitamin A supplementation to no more than 10,000 IU per day. Further, they had to be within the 95% normal range for total cholesterol, liver function, white blood cell count, hemoglobin, and platelet count.

Some Techniques to Reduce Nonrandom Error and Bias

Several precautions had to be taken to ensure that nonrandom error and bias in response was minimized in the study. This included assurance that the subjects would take the medication regularly, retain their willingness to remain in the study, return for their follow-up clinical visits for evaluation, and remain unaware of the treatment group to which they were assigned.

Conceivably, if subjects knew they were in the placebo group they could self-administer extra vitamin A on their own accord with the hope of reducing their personal risk and unknowingly bias the comparison of treatment to placebo. The investigators provided an insurance against this sort of event by not revealing to the subject or the clinician dispersing the capsules whether the subject was receiving the treatment or placebo. This type of trial is known as a *double-blind* trial.

A three-month run-in period was established to evaluate the subjects' ability and willingness to adhere to the study protocol. They received a bottle containing 100 placebo capsules, of which they were to take one per day. Subjects who had

taken at least 75% of the capsules during the run-in period and were willing to continue the study were assigned to a treatment or placebo regimen, given a six-month supply of the appropriate capsules, and scheduled for a follow-up visit.

Subjects returned to the clinic every six months to receive an examination for any symptoms of BCC or SCC. As a safety measure, they were also examined for any potential side effects from elevated retinol intake. The subject's remaining medication was weighed to evaluate adherence to dosage. During the interview, subjects' questions were answered and they were motivated to adhere to the medication schedules. They were then given the next six months' supply of capsules and scheduled for the next follow-up visit. Subjects were reminded of their upcoming follow-up visit by postcard and telephone contact.

Once a year a blood specimen was collected from a random sample of subjects for analyses of retinyl palmitate levels to obtain supplementary information on group adherence to the retinol supplement. The levels of retinyl palmitate should have been greater for the treatment group than the placebo group if the treatment group had adhered to the capsule regimen.

Replication

The trial was conducted from two separate clinics, one each in Tucson and Phoenix, Arizona. The number of subjects required for the study was based on assumptions about the average annual incidence of skin cancers in the placebo and retinol treatment groups and incidences of anticipated deviation from the prescribed study protocols by the subjects. The required sample size was determined to be 1118 subjects in each of the treatment groups and was based on a power of .80 and a .05 two-sided Type I error rate, using techniques specific to studies on time-to-event measurements.

Blocking to Reduce Experimental Error

A subject's risk to non-melanoma skin cancer was anticipated to be related to the amount of time spent in the sun and whether the person had fair skin. Fair-skinned people are anticipated to be at greater risk for skin cancer and will sunburn more readily in a 30-minute period in southern Arizona. Anyone who spends a greater amount of time in the sun is anticipated to be at greater risk for skin cancer. Information on weekly sun exposure and anticipated skin-burning reaction after 30 minutes of sun exposure was collected from each subject during the first contact.

These were the most likely factors to interfere with risk comparisons between retinol treatment and placebo, so they were used as blocking factors prior to assignment of treatment to subjects. Subjects were categorized according to two levels of sun exposure: < 10 hours versus ≥ 10 hours per week. They were also grouped according to levels of skin reaction after 30 minutes: always or usually burns versus burns moderately, rarely, or never.

Thus, subjects were placed into one of the four block types constructed when the two factors were placed in all four combinations of their levels. For example,

subjects exposed to the sun < 10 hours per week and whose skin moderately, rarely, or never burned after 30 minutes exposure to the sun would be placed in the same block. Equal numbers of subjects would be assigned to the retinol treatment and placebo in each of the blocks; thus, any potential differences between retinol treatment and placebo would not be interfered with by differential subject risks based on sun exposure and skin sensitivity to the sun.

Randomization

Subjects enter into clinical trials over a period of time as they are identified by their physicians as candidates for the study and as they respond to calls for subject volunteers. Therefore, the assignment of subjects to treatments sometimes takes place over a period of one or more years until the required number of subjects (replications) has been achieved for the study. For this study, the enrollment period required more than four years to acquire a sufficient number of subjects.

The subjects were assigned to their respective blocking type (described above) as they entered the study. They were randomly assigned to the placebo or retinol treatment by groups of four subjects with the same blocking criteria. For example, if the first two subjects entering in a block were assigned at random to the placebo the next two would automatically receive the retinol treatment. Then the randomization would start over with the next four subjects entering into any one of the blocking groups. This method of assignment in blocks of four subjects ensured an equal distribution of subjects on placebo and retinol treatment over the entry timeline of the trial, effectively removing the chance of having a greater number of subjects on one treatment for a longer period of time than on another treatment.

The randomizations were done separately for the clinics in Tucson and Phoenix. Thus, the clinics became a de facto blocking factor in the study.

Measured Covariates for Statistical Control of Experimental Error

Numerous other factors could conceivably have some relationship to risk for skin cancer. Those factors thought to have the greatest potential for affecting the comparison of risk between retinol treatment and placebo were used as blocking factors: sun exposure and skin sensitivity to sun exposure. To have included more factors in the blocking scheme might have made the study unnecessarily cumbersome, particularly if no hard evidence suggested these factors had major impacts on skin cancer risk.

Other factors the investigators considered to be potentially influential and for which measurements were recorded included age, gender, prior skin cancers (0, 1, or 2), number of moles and freckles, vitamin use, dietary vitamin A at the start of the study determined from a diet interview, and serum retinyl palmitate at the start of the study.

These factors could have been used as covariates in the data analysis to reduce experimental error for comparisons of retinol treatment to placebo. A post-randomization check after the four-year enrollment period was made to determine

Table 1.3 Cross tabulations of subjects by treatment groups and covariates in the skin cancer prevention trial

Characteristic	*Placebo* $n = 1140$	*Retinol* $n = 1157$
Age		
< 63	558	584
≥ 63	582	573
Gender		
Female	345	334
Male	795	823
Prior skin cancers		
0	932	920
1	152	177
2	45	45
> 2	11	15
Moles and freckles		
0–7	550	541
> 7	301	326
Unknown	289	290
Weekly sun exposure		
0–10 hours	452	493
> 10 hours	688	664
Skin burns		
Always/usually	490	517
Moderately/rarely/never	650	640
Clinical center		
Phoenix	429	420
Tucson	711	737
Vitamin use		
No	309	312
Sometimes	322	331
Yes	509	514
Dietary vitamin A (IU)		
1,194–6,979	337	355
6,980–10,627	324	369
10,628–41,404	365	326
Unknown	114	107
Serum retinyl palmitate (mg/ml)		
0.0–6.0	397	400
6.1–20.0	345	346
> 20.0	350	378
Unknown	48	33

whether the subjects in placebo and retinol treatment groups were equally distributed with respect to each of these factors. The cross tabulations shown in Table 1.3 indicate a relatively equal distribution of subjects on placebo and retinol treatment for each of the major covariates considered in the study. Notice 26 subjects with more than two previous skin cancers were enrolled in the study even though they did not meet the selection criterion of no more than two previous skin cancers. Of course, these could be either recording errors or an oversight by the clinicians who could have mistakenly allowed these subjects to enroll. The reason for these types of mistakes often remains unknown.

EXERCISES FOR CHAPTER 1

1. a. Find the definitions of *research* and *hypothesis* in a dictionary, and formulate a definition of the term *research hypothesis*.
   b. How does a research hypothesis differ from the *statistical hypothesis* formulated for statistical tests, such as H_0: $\mu = 0$ versus H_a: $\mu \neq 0$?

2. Choose a journal article in your field of study that reports the results of a comparative experiment or observational study. Identify and briefly (one or two sentences) describe the following:
   a. Research hypothesis
   b. Treatments
   c. Experimental (observational) units
   d. Type of experiment design
   e. Criteria for any grouping, blocking, or matching done in the study
   f. Provide the article citation

3. a. Choose a practical situation from your own field of study and describe a problem whose solution must be determined experimentally.
   b. Indicate the following for the problem described in part (a):
      (i) a research hypothesis
      (ii) the treatments necessary to evaluate the hypothesis
      (iii) what constitutes an experimental unit

4. Choose a journal article from your special field of interest and review it for the purpose of evaluating the application of good research design. Many aspects of research design have been discussed separately in this chapter relative to their effect on statistical and scientific inference. The case study in Section 1.10 illustrates the elements of a reported research study that one must identify for a validation of the reported findings from a research project.

Choose an article that reports on either an experiment conducted to compare two or more treatments or a comparative observational study conducted to compare two or more "treatment" conditions that existed a priori.

Some good questions to ask yourself for the critique are "Did they include all important elements of good research design in the study?" "If so, did they implement them properly?" "Are the elements of the work described in such a way that I could understand or duplicate what they did?"

Pay particular attention to the following items in your review:

a. Review of the literature
b. Statement of the problem
c. Research hypothesis and study objectives: discuss whether they were present and reasonable
d. Treatment design: describe how it did or did not address the hypothesis and objectives
e Experiment design or observational study design
f. Use of randomization (experiments) or random sampling (observational study) and replication
g. Statistical hypotheses and statistical analysis procedures
h. Conclusions and statistical reliability of conclusions
i. Self-evaluation of the study by the author(s) and potential for future investigations
j. Provide the article citation

Your critique should describe and evaluate the author's approach in the research itself and in the article relative to each of the important elements of the research based on the preceding list. Include a page reference from the article for your comments on each of the items.

5. A study is planned on the physiology of exercises with human subject volunteers. The two treatments in the study are two methods of aerobic exercise training (call the methods A and B). At the end of a ten-week exercise period, each subject will undergo a treadmill test for standard respiratory and cardiovascular measurements.

Individual	*Sex*	*Age*	*Individual*	*Sex*	*Age*
1	M	54	10	M	18
2	M	38	11	M	31
3	F	41	12	F	18
4	F	18	13	M	58
5	F	19	14	M	74
6	F	39	15	F	58
7	M	51	16	F	21
8	F	44	17	M	35
9	M	62	18	M	34
			19	F	38

Nineteen volunteers are listed in the table by sex and age. All volunteers are in good health and in the normal weight range for their age, sex, and height. Eight individuals will be tested in each of the methods (A or B), so that only 16 of the 19 volunteers will be used; a subject will participate only in one of the methods.

a. Explain how you would group the individuals prior to the assignment of treatments so that experimental error variance could be kept at a minimum.
b. Explain why you grouped as you did.
c. Show your final assignment of individuals to the treatment groups.

6. An experiment is planned to compare three treatments applied to shirts in a test of durable press fabric treatments to produce wrinkle-free fabrics. In the past formaldehyde had been used to produce wrinkle-free fabric, but it was considered an undesirable chemical treatment. This study is to consider three alternative chemicals: (a) PCA (1-2-3 propane tricarbolic acid), (b) BTCA (butane tetracarboxilic acid), and (c) CA (citric acid).

Four shirts will be used for each of the treatments. First, the treatments are applied to the shirts, which are then subjected to simulated wear and washing in a simulation machine. The chemical treatments will not contaminate one another if they are all placed in the same washing machine during the test. The machine can hold one to four shirts in a single simulation run. At the end of the simulation run each of the shirts is measured for tear and breaking strength of the fabric and how wrinkle-free they are after being subjected to the simulated wear and washing. The comparisons among the treatments can be affected by (a) the natural variation from shirt to shirt; (b) measurement errors; (c) variation in the application of the durable press treatment; and (d) variation in the run of the simulation of wear and washing by the simulation machine. Following is a brief description of three proposed methods of conducting this simple experiment.

Method I. The shirts are divided randomly into three groups of four shirts. Each group receives a durable press treatment as one batch and then each batch is processed in one run of the simulation machine. Each run of the simulation machine has four shirts that have received the same treatment. There are three runs of the simulation machine.

Method II. The shirts are divided randomly into three treatment groups of four shirts each, and the durable press treatments are applied independently to single shirts. The shirts are grouped into four sets of three, one shirt from each durable press treatment in each of the four sets, and each set of three so constructed is used in one run of the simulation machine. There are four runs of the simulation machine.

Method III. The shirts are divided randomly into three groups of four shirts. The durable press treatments are applied independently to single shirts. The simulation of wear and washing is done as in Method I.

a. Which method do you favor?
b. Why do you favor the method you have chosen?
c. Briefly, what are the disadvantages of the other two methods?

7. Explain what is meant by the term *replication* in the context of (a) an experiment in which the effectiveness of several antibiotics is tested on animals in a laboratory and (b) an observational study to determine the differences in grass species present in pure stands of mesquite and pure stands of oak in southern Arizona.

8. An experiment is planned to compare three methods of instruction. Each is tested with a single classroom of 25 students. A different instructor is to be used for each classroom and consequently each instruction method.

a. Write a short critique of the proposed experiment.
b. How could the experiment be improved?

9. An experiment is planned to compare the strengths of three different asphalt mixtures for road surfaces. A single batch is to be manufactured for each experimental mixture. Several asphalt specimens will be made from each of the mixtures and tested for tensile strength.

a. Write a short critique of the proposed experiment.
b. How could the experiment be improved?

10. Suppose you want to randomize the allocation of two treatments to 16 experimental units. How many randomizations are possible if 8 units are to be assigned to each of the treatments? How many randomizations are possible if you want to assign 6 units to one treatment and 10 units to a second treatment?

11. An experiment with four treatments and five replications of each treatment will require 20 experimental units. How many randomizations are possible for this experiment?

12. An experiment with two treatments and three replications per treatment had the following randomization of treatments to the experimental units shown along with the measured response on each unit:

Unit:	1	2	3	4	5	6
Treatment:	A	B	B	A	A	B
Response:	7	10	9	5	10	12

Conduct a randomization test of the null hypothesis, H_0: no difference in the treatment effects of A and B, versus the alternative, H_a : the effect of treatment B is greater than the effect of treatment A. Use the test statistic $\overline{y}_B - \overline{y}_A$. (*Hint*: It is only necessary to identify one-half of the randomizations directly. Each randomization has a "mirror" randomization in which the letters A and B are interchanged. For example, the "mirror" for the actual experiment randomization is the randomization B, A, A, B, B, A.)

13. In the table, the coefficients of variation and relative efficiencies (randomized complete block to completely randomized) of the same experiment conducted at four locations are given. Each trial used a randomized complete block design.

Location	*Coefficient of Variation (%)*	*Relative Efficiency (%)*
Tucson	10	100
Phoenix	10	150
Los Angeles	20	200
San Francisco	20	125

a. How many more replications of a completely randomized design would be necessary at Los Angeles to obtain the same precision as the randomized complete block design for estimating the treatment means? Explain your answer.
b. If you were asked the question in part (a) relative to San Francisco, would you require more or fewer replications in San Francisco than for Los Angeles? Explain your answer.
c. Suppose four replications are required with the randomized complete block design at Tucson to detect differences of $\delta = 20\%$ with a test at the .05 level of significance and a probability (power) of .90. How many replications would be required in Phoenix with the same criteria for a randomized complete block design? Explain your answer.
d. Would you require more or fewer replications in Los Angeles than in Tucson with the same criteria for a randomized complete block design? Explain your answer.

2 Getting Started with Completely Randomized Designs

Chapter 1 presented the principles of design in relation to the stated goals of research hypotheses—accuracy and precision of the observations and validity of the resulting analysis. Some of those principles are illustrated in this chapter, using an experiment with a completely randomized design. A statistical model is developed with parameters to describe the experiment according to the research hypothesis. The parameters are then estimated, using the least squares method. Experimental error variance is estimated and used to estimate standard errors and confidence intervals for the parameters and to test statistical hypotheses about them. The fundamental partition for the sum of squares of the observations is derived and summarized in the traditional analysis of variance table.

2.1 Assembling the Research Design

The *research hypothesis, treatment design,* and *experiment* or *observational study design* constitute the **research design** for the study. Treatments are developed to address specific research questions and hypotheses that arise in the research program. For example, a microbiologist hypothesizes that the activity of soil microbes depends upon soil moisture conditions. Treatments with different amounts of soil moisture are set up to measure the microbe activity at different levels of soil moisture to evaluate the hypothesis. A traffic engineer hypothesizes that traffic speed is related to the width of street lanes. Lanes of different width are selected by the engineer, and traffic speed is measured at each lane width to evaluate the hypothesis.

The treatment design has to be integrated into an experiment design. The investigator must decide what constitutes an experimental unit, how many