1 / INTRODUCTION TO BASIC CONCEPTS IN EXPERIMENTAL DESIGN

1.1 INTRODUCTION

The term *experimental design* refers to five interrelated activities required in the investigation of scientific or research hypotheses. These activities, listed in the order performed, are as follows:

1. Formulate statistical hypotheses and make plans for the collection and analysis of data to test the hypotheses. A statistical hypothesis is a statement about one or more parameters of a population. Statistical hypotheses are rarely identical to research or scientific hypotheses but are testable formulations of research hypotheses.
2. State decision rules to be followed in testing the statistical hypotheses.
3. Collect data according to plan.
4. Analyze data according to plan.
5. Make decisions concerning the statistical hypotheses based on decision rules and inductive inferences concerning the probable truth or falsity of the research hypotheses.

The term *experimental design* is also used in a more restricted sense to designate a particular type of plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan. The meaning of the term is generally clear from the context in which it is used.

SUBJECT MATTER AND GENERAL ORGANIZATION OF THIS BOOK

The concepts and procedures involved in carrying out steps 1, 2, 4, and 5 above comprise the subject matter of this book. Experimental design is only one of the many facets of scientific research. A carefully conceived and executed design is of no avail if the scientific hypothesis that originally led to the experiment is without merit.

A detailed examination of the logical and statistical aspects of specific experimental designs begins in Chapter 4. The first three chapters provide an overview of experimental designs. a review of basic statistical

concepts, and a presentation of statistical *tools* used throughout the remainder of the book. The reader is encouraged to review the introductory chapters *after* reading subsequent chapters.

This book emphasizes those experimental designs that are most useful in the behavioral sciences. Many of the chapters conclude with a selected bibliography of contemporary research, which indicates diverse applications of the designs discussed in the chapter. The reader should consult the bibliographies to gain an over-all view of the way experimental designs are used in research.

The validity of inductive inferences that an experimenter draws from research rests on the fulfillment of certain assumptions. These assumptions are explicitly stated for each design as it is presented. Procedures for determining whether or not the assumptions are tenable in the light of sample data are also described.

A list of advantages and disadvantages for each design is provided to aid an experimenter in the selection of an appropriate design. Subsequent sections of this chapter treat general concepts basic to the selection of the best design for a particular research application.

1.2 DEFINITION OF BASIC TERMS

A number of terms must be defined before concepts basic to the selection of an experimental design can be discussed. It is assumed that the reader already has some familiarity with most of the terms that follow. Therefore, the material in this section is intended only to ensure a common vocabulary for the subsequent discussion. The definitions of some terms are oversimplified, but the assumed mathematical background does not permit rigorous definitions for all terms. Additional definitions are listed in the glossary.

Population. A collection of all observations identifiable by a set of rules.

Sample. A subset of observations from a population.

Random Sample. A sample drawn from a population in such a way that all possible samples of size n have the same probability of being selected.

Parameter. A measure computed from all observations in a population. Parameters are designated by Greek letters. For example, the symbols for a population mean and standard deviation are μ and σ, respectively.

Statistic. A measure computed from observations in a sample. Statistics are designated by Latin letters. For example, the symbols for a sample mean and standard deviation are $\bar{X}$ and S, respectively.

Random Variable. A quantity, say X, which may assume a range of possible values, each having an associated probability, say $p(X)$.

Estimator. The particular function of observations in a sample that is chosen to estimate a population parameter. For example, the sample mean is used to estimate the population mean. The numerical value obtained is called an estimate.

Expected Value. The long-run average of a random variable over an indefinite number of samplings. The expected value [E(*X*)] of a discrete random variable *X* is given by $E(X) = \Sigma X p(X)$ = mean of *X*. It should be noted from the above definition that an expected value may be a value that the random variable could not actually have.

Unbiased Estimator. An estimate of a parameter is said to be unbiased if its expected value is equal to the parameter.

Research Hypothesis. A tentative theory or supposition provisionally adopted to account for certain facts and to guide in the investigation of others. The terms research hypothesis and scientific hypothesis may be used interchangeably.

Statistical Hypothesis. A statement about one or more parameters of a population. Null and alternative hypotheses are two forms of a statistical hypothesis.

Null Hypothesis (H_0). A statement concerning one or more parameters that is subjected to statistical test.

Alternative Hypothesis (H_1). The hypothesis that remains tenable when the null hypothesis is rejected.

Statistical Test. A procedure whereby two mutually exclusive statistical hypotheses are evaluated in the light of sample data. The hypothesis that dictates the sampling distribution against which an obtained sample value is compared is said to be the one tested.

Level of Significance (α). Probability of rejecting the null hypothesis when it is true.

Type I Error. Error that occurs when the experimenter rejects the null hypothesis when it is true. The probability of committing a type I error is determined by the level of significance (α) that the experimenter adopts.

Type II Error. Error that occurs when the experimenter fails to reject the null hypothesis when it is false. The probability (β) of committing a type II error is determined by the magnitude of the experimental effect, size of sample, magnitude of random error, and level of significance.

Power of Test. Probability of rejecting the null hypothesis when the alternative hypothesis is true. If β is designated as the probability of committing a type II error, power is equal to $1 - \beta$.

Confidence Interval. A range of values that, considering all possible samples, has some designated probability of including the true population value.

Confidence Limits. Upper and lower boundaries of confidence interval.

Critical Region. A set of outcomes of a statistical test that leads to the rejection of the null hypothesis.

Replication. The collection of two or more observations under a set of identical experimental conditions.

Degrees of Freedom (df). The number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.

Experimental Error. Measure that includes all uncontrolled sources of variation affecting a particular score.

Sampling Distribution. A theoretical probability distribution that describes the functional relation between possible values of a statistic based on N cases drawn at random and the probability associated with each value over all possible samples of size N.

Statistical Model. A mathematical statement concerning the sampling distribution of a random variable that is used in evaluating the outcome of an experiment or in predicting the outcome of future replications of an experiment.

Test Statistic. A statistic whose purpose is to provide a test of some statistical hypothesis. Test statistics such as t and F have known sampling distributions that can be employed in determining the probability of an obtained result under the null hypothesis.

Relative Efficiency of a Statistic. Ratio of experimental error of one statistic to that of another statistic.

Statistical Decision Theory. Branch of mathematics concerned with the problem of decision making and the choice of decision rules under uncertain conditions.

1.3 FORMULATION OF PLANS FOR COLLECTION AND ANALYSIS OF DATA

ACCEPTABLE RESEARCH HYPOTHESES

Some questions cannot currently be subjected to experimental investigation. For example, the questions "Can three or more angels sit on the head of a pin?" and "Does life exist in more than one galaxy in the universe?" cannot be answered because no procedures presently exist for observing either angels or other galaxies. Scientists confine their research hypotheses to questions for which procedures can be devised that offer the possibility of arriving at an answer. This does not mean that the question concerning the existence of life on other galaxies can never be investigated. Indeed, with continuing advances in space science it is probable that this question will eventually be answered.

Questions that provide the impetus for experimental research should be statable in the logical form of the general implication. That is, a question should be reducible to the form, *if A, then B.* For example, *if* albino rats are subjected to microwave radiation, *then* their food consumption will decrease. This research hypothesis can be investigated because procedures both for manipulating radiation level and for measuring food consumption of rats are available.

DISTINCTION BETWEEN DEPENDENT AND INDEPENDENT VARIABLES

In the example just cited, the presence or absence of radiation is designated as the *independent variable*—the variable that is under the

control of the experimenter. The terms independent variable and treatment will be used interchangeably. The *dependent variable* is the amount of food consumed by the rats. The dependent variable reflects any effects associated with manipulation of the independent variable.

SELECTION OF DEPENDENT VARIABLE

The choice of an appropriate dependent variable may be based on theoretical considerations, although in many investigations the choice is determined by practical considerations. In this example, other dependent variables that could also be measured include

1. Activity level of rat in an activity cage.
2. Body temperature of rat.
3. Emotionality of rat as evidenced by amount of defecation and urination.
4. Problem-solving ability.
5. Weight of rat in grams.
6. Speed of running in a straight-alley maze.
7. Visual discrimination capacity.
8. Frequency of mating behavior.

Several independent variables can be employed in an experiment, but the designs described in this book are limited to the assessment of one dependent variable at a time. If it is necessary to evaluate two or more dependent variables simultaneously, a multivariate analysis of variance design can be used.* Some of these multivariate procedures are so complicated or so tedious that they cannot reasonably be carried out without a digital computer. However, the increasing availability of computer facilities makes the use of multivariate procedures more widespread. Univariate procedures can be appropriately applied to most research problems because it is generally impossible to measure more than a limited number of dependent variables, and those that can be measured are often found to be highly correlated.

The selection of the most fruitful variables to measure should be determined by a consideration of the sensitivity, reliability, distribution, and practicality of the possible dependent variables. From previous experience, an experimenter may know that one dependent variable is more sensitive than another to the effects of a treatment or that one dependent variable is more reliable, that is, gives more consistent results, than another variable. Because behavioral research generally involves a sizable investment in time and material resources, the dependent variable should be reliable and maximally sensitive to the phenomenon under investigation. Choosing a dependent variable that possesses these two characteristics

*A discussion of these designs is beyond the scope of this book. The reader is referred to Anderson (1958), Cooley and Lohnes (1962), Fryer (1966), Morrison (1967), and Rao (1952) for a discussion of multivariate procedures.

may minimize the amount of research effort required to investigate a research hypothesis.

Another important consideration in selecting a dependent variable is whether the observations within each treatment level (or combination of treatment levels in the case of multitreatment experiments) would be normally distributed. The assumption of normality, discussed in Chapter 2, is required for the experimental designs described in Chapters 4 through 12. In some cases it may be possible to *transform* nonnormally distributed observations so that the resultant distributions are normal. This procedure is described in Chapter 2. If theoretical considerations do not dictate the selection of a dependent variable and if several alternative variables are equally sensitive and reliable, in addition to being normally distributed, an experimenter should select the variable that is most easily measured.

SELECTION OF INDEPENDENT VARIABLE

The independent variable was defined earlier as the presence or absence of radiation. Such a treatment is described as having two treatment levels. If the experimenter is interested in the effects of different radiation dosages, he can employ three or more levels of radiation. The levels could consist of 0 microwatts, 20,000 microwatts, 40,000 microwatts, and 60,000 microwatts of radiation. This particular treatment is an example of a *quantitative* independent variable in which different treatment levels constitute different amounts of the independent variable.

In general, when the independent variable is quantitative in character there is little interest in the exact values of the treatment levels used in the experiment. In the radiation example, the research hypothesis could also be investigated, using three other levels of radiation, say, 25,000, 50,000, and 75,000 microwatts in addition to the zero-microwatt control level. The treatment levels should be chosen so as to cover a sufficiently wide range to detect effects of the independent variable if real effects exist. In addition, the number and spacing of the levels should be sufficient to define the shape of the function relating the independent and dependent variables. This is necessary if an experimenter is interested in performing a trend analysis as described in Chapter 4.

Selection of appropriate levels of the independent variable may be based on results of previous experiments or on theoretical considerations. In some research areas, it may be helpful to carry out a small pilot experiment to select treatment levels prior to the main experiment.

Under the conditions described in Chapters 2 and 4, the levels of a quantitative independent variable may be selected randomly from a population of treatment levels. If this procedure is followed, an experimenter can extrapolate from the results of his experiment to treatment levels that are not included in the experiment. If the treatment levels are not randomly sampled, the results of an experiment are applicable only to the specific levels included in the experiment.

Preceding paragraphs described a quantitative independent variable. If the treatment levels consisted of unmodulated radiation, amplitude-modulated radiation, and pulse-modulated radiation, the treatment is designated as a *qualitative* independent variable. The different treatment levels represent different *kinds* rather than different *amounts* of the independent variable. The distinction between quantitative and qualitative treatments is important in connection with trend analysis. The specific levels of a qualitative independent variable employed in an experiment are generally of direct interest to an experimenter. The levels chosen are usually dictated by the nature of the research hypothesis.

CONTROL OF NUISANCE VARIABLES

In addition to independent and dependent variables, all experiments include one or more *nuisance* variables. Nuisance variables are undesired sources of variation in an experiment that may affect the dependent variable. As the name implies, the effects of nuisance variables are of no interest per se. In the radiation example, potential nuisance variables include sex of the rats, variation in weight of the rats prior to the experiment, presence of infectious diseases in one or more cages where the rats are housed, temperature variation among the cages, and differences in previous feeding experiences of the rats. Unless controlled, nuisance variables can bias the outcome of an experiment. For example, if rats in the radiated groups suffer from some undetected disease, differences among the groups would reflect the effects of the disease in addition to radiation effects—if the latter effects exist.

Four approaches can be followed in controlling nuisance variables. One approach is to hold the nuisance variable constant for all subjects. For example, use only male rats of the same weight. Although an experimenter may attempt to hold all nuisance variables constant, the probability is high that some variable will escape his attention. A second approach, one that is used in conjunction with the first, is to assign subjects randomly to the experimental conditions. Then known as well as unsuspected sources of variation or bias are distributed over the entire experiment and thus do not affect just one or a limited number of treatment levels. In this case an experimenter increases the magnitude of random variation among observations in order to minimize systematic effects, that is, the effects of nuisance variables that bias all observations in one or more treatment levels in the same manner. Random variation can be taken into account in evaluating the outcome of an experiment, whereas it is difficult or impossible to account for systematic nuisance effects. A third approach to controlling nuisance variables is to include the variable as one of the *treatments* in the experimental design. This approach is illustrated in Section 1.4 in connection with a Latin square design.

The above three approaches for controlling nuisance variables illustrate the application of *experimental control* as opposed to the fourth

approach which is *statistical control.* In some experiments it may be possible —through the use of regression procedures (see Chapter 12)—to remove the effects of a nuisance variable statistically. This use of statistical control is referred to as the analysis of covariance.

CLASSIFICATION OF INDEPENDENT AND NUISANCE VARIABLES

All independent and nuisance variables in behavioral research can be classified in one of three general categories—organismic, environmental, and task variables. In the radiation example, the independent variable of radiation can be classified as an environmental variable. The nuisance variables listed earlier as sex, weight, prior experience, and infectious diseases are examples of organismic variables. The other nuisance variable of temperature variation among the cages is an example of an environmental variable. This radiation experiment does not include a task variable. A task variable could be introduced into the experiment by requiring the rats to perform easy, medium, and difficult visual discriminations before gaining access to food. The effect of the visual discrimination on food consumption represents an additional independent variable that can be classified as a task variable. In the design of experiments, the above classifications may help an experimenter in listing the nuisance variables that should be controlled.

EFFICIENCY AND EXPERIMENTAL DESIGN

An experimenter engaging in research is desirous of arriving at valid conclusions. At the same time he hopes to accomplish this goal as efficiently as possible. Generally several experimental designs can be used in testing a statistical hypothesis. However, alternative designs that are equally valid for testing a hypothesis are rarely equally efficient. Efficiency of alternative research procedures may be defined in different ways. For example, efficiency may be defined in terms of time required to collect data, cost of data collection, ratio of information obtained to cost, and so on. A discussion of relative efficiency by Cochran and Cox (1957, 31) is instructive. A commonly used index for assessing the relative efficiency of two experimental designs is given by the ratio of their respective experimental errors. *Experimental error* refers to all extraneous variation in dependent variable scores that tends to mask the effects of the independent variable. The main sources of experimental error are inherent variability in the behavior of subjects and lack of uniformity in the conduct of the experiment.

A formula that provides insight into factors related to the efficiency of two designs is

$$\text{Efficiency} = \frac{\left(\frac{n_2 C_2}{\hat{\sigma}_1^2}\right)\left(\frac{df_1 + 1}{df_1 + 3}\right)}{\left(\frac{n_1 C_1}{\hat{\sigma}_2^2}\right)\left(\frac{df_2 + 1}{df_2 + 3}\right)},$$

where $\hat{\sigma}^2$ = estimate of experimental error per observation, n = number of subjects, C = cost of collecting data per subject, df = experimental error degrees of freedom, and the subscripts designate the two experimental designs (Federer, 1955, 13). If the ratio is less than one, the second design is more efficient than the first. The converse is true if the ratio is greater than one. The formula calls attention to four factors that are related to the efficiency of experimental designs. Unfortunately, an experimental design that is advantageous with respect to one factor may not be advantageous with respect to the others. For example, if a design has the desirable attribute of a small experimental error, it may have a high cost per subject or a small number of degrees of freedom for experimental error, or it may require a large number of subjects. The problem facing an investigator is to select an experimental design that represents the best compromise obtainable within the constraints of his research situation.

DETERMINATION OF SAMPLE SIZE

Once the independent and dependent variables are specified, the number of subjects required for the experiment must be determined. This is one of the more perplexing problems in experimental design. Five factors must be considered in specifying a sample size that is adequate for testing a statistical hypothesis: (1) minimum treatment effects an experimenter is interested in detecting, (2) number of treatment levels, (3) population error variance, (4) probability of making a type I error, and (5) probability of making a type II error. In general, the population error variance is unknown. It may be possible to make a reasonable estimate of the population error variance on the basis of previous experiments or a pilot study. If the above information can be specified, the size of the sample necessary to achieve a given power can be calculated. The power of a research methodology is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true. Power is equal to 1 − (probability of committing a type II error).

The procedure described here for calculating power was developed by Tang (1938). It assumes that the observations are normally distributed with a common error variance = σ_ε^2. The parameter ϕ is defined as

$$\phi = \frac{\sqrt{\sum_1^k (\mu_j - \mu)^2 / k}}{\sigma_\varepsilon / \sqrt{n}},$$

where $\mu_j - \mu$ = the minimum treatment effect an experimenter is interested in detecting, k = number of treatment levels, σ_ε = square root of population error variance, and n = size of sample. If estimates of $\mu_j - \mu$ and σ_ε can be made, the size of a sample necessary to achieve a designated power can be determined from Table D.14 by a process of trial-and-error. The probability of type I and type II errors, α and β respectively, that the experimenter is willing to accept must also be specified.

Assume that a treatment has three levels and that the smallest treatment effects of interest to the experimenter are -4, -1, and $+5$. That is,

$$\sum_{1}^{k}(\mu_j - \mu)^2 = (-4)^2 + (-1)^2 + (5)^2 = 42.$$

Assume, also, that on the basis of previous research σ_ε is estimated to be six and that the investigator wishes the power of his test $(1 - \beta)$ to equal at least .80 and the probability of a type I error to equal .05. If nine subjects are assigned to each of the three treatment levels, Table D.14 in the appendix can be used to determine if the specified power is achieved. For example,

$$\phi = \frac{\sqrt{\dfrac{(-4)^2 + (-1)^2 + (5)^2}{3}}}{6/\sqrt{9}} = \frac{\sqrt{42/3}}{6/3} = 1.87,$$

with $k - 1 = 2$ and $N - k = 24$ degrees of freedom. The value of $\phi = 1.87$ and $\alpha = .05$ are entered in the table for $k - 1$ degrees of freedom. The curve corresponding to $N - k$ degrees of freedom indicates that the power is equal to .79, which is less than that desired by the experimenter. If the sample size is increased to 30, with ten subjects assigned to each treatment level, the power can be estimated from

$$\phi = \frac{\sqrt{\dfrac{(-4)^2 + (-1)^2 + (5)^2}{3}}}{6/\sqrt{10}} = 1.97,$$

with $k - 1 = 2$ and $N - k = 27$ degrees of freedom. The probability of detecting the specified treatment effects is approximately .83. Thus the required sample size is found to be 30.

If reasonable estimates of the parameters can be made, the required sample size should always be computed before the experiment is begun. If these preliminary calculations indicate that the power of the experimental design is inadequate, the experimenter may choose not to conduct the experiment or may modify it so as to increase its power. The two most common procedures for increasing power are (1) to increase the size of the sample and (2) to employ an experimental design that provides a more precise estimate of treatment effects and a smaller error term. The first procedure was illustrated in this section. The second is described in

Section 1.4. Overall and Dalal (1965) have described a procedure for maximizing the power of a research methodology relative to cost through the optimum allocation of resources to subjects; it requires *a priori* estimates of a number of design parameters. A procedure is described in Section 4.5 for estimating the number of subjects required for an experiment, one that does not necessitate making an estimate of the population error variance.

1.4 OVERVIEW OF TYPES OF EXPERIMENTAL DESIGNS

One of the procedures suggested above for increasing the power of a research methodology was to employ a more *sensitive* experimental design. In this context, the term experimental design refers to the plan by which subjects are assigned to treatment levels and the data analyzed.

An almost bewildering array of kinds of experimental designs exists. Fortunately, most complex experimental designs represent a combination of a relatively small number of basic *building block* designs. For example, most complex designs can be constructed by combining two or more completely randomized, randomized block, or Latin square designs. A simple classification of the experimental designs described in this book is outlined in Table 1.4-1. A more complete classification system appears in Cox (1943); Doxtator, Tolman, Cormany, Bush, and Jensen (1942); and Federer (1955, 6–12).

The category *systematic designs* in the outline is of historical interest only. According to Leonard and Clark (1939), agricultural field research employing systematic designs on a practical scale dates back to 1834. Prior to the work of Fisher, as well as of Neyman and Pearson on the theory of statistical inference, investigators used systematic schemes rather than randomization procedures for assigning treatment levels to plots of land or other suitable experimental units—hence the designation systematic designs for these early field experiments. Impetus for this early experimental research came from a need to improve agricultural techniques. Today the nomenclature of experimental design is replete with terms from agriculture. Systematic designs in which the randomization principle is not followed do not provide a valid estimate of error variance and hence are not subject to powerful tools of statistical analysis, such as analysis of variance.

Modern principles of experimental design, particularly the principle of random assignment of treatment levels to experimental units, received general acceptance as a result of Fisher's work (1922, 1923, 1935). Experimental designs using the randomization principle are called *randomized designs.* Randomized designs can be subdivided into two distinct categories, complete block designs and incomplete block designs, and two pseudo-

TABLE 1.4-1 Outline of Experimental Designs Described in this Book

Experimental Design	*Abbreviated Designation*
I. Systematic Designs	
II. Randomized Designs	
A. Complete Block Designs	
1. Completely randomized design	CR-k*
2. Randomized block design	RB-k
3. Latin square design	LS-k
4. Graeco-Latin square design	GLS-k
5. Hyper-Graeco-Latin square design	HGLS-k
B. Incomplete Block Designs	
1. Balanced incomplete block design	BIB-t
2. Youden square balanced incomplete block design	YBIB-t
3. Partially balanced incomplete block design	PBIB-t
C. Factorial Experiments	
1. Completely randomized factorial design	CRF-pq
2. Randomized block factorial design	RBF-pq
3. Completely randomized hierarchal design	CRH-$p(q)$
4. Completely randomized partial hierarchal design	CRH-$p(q)r$
5. Split-plot design	SPF-$p \cdot q$
6. Randomized block completely confounded factorial design	RBCF-p^k
7. Randomized block partially confounded factorial design	RBPF-p^k
8. Latin square completely confounded factorial design	LSCF-p^k
9. Completely randomized fractional factorial design	CRFF-p^k
10. Randomized block fractional factorial design	RBFF-p^k
11. Latin square fractional factorial design	LSFF-p^k
12. Graeco-Latin square fractional factorial design	GLSFF-p^k
D. Analysis of Covariance Experiments	
1. Completely randomized analysis of covariance design	CRAC-k
2. Randomized block analysis of covariance design	RBAC-k
3. Latin square analysis of covariance design	LSAC-k
4. Completely randomized factorial analysis of covariance design	CRFAC-pq
5. Split-plot factorial analysis of covariance design	SPFAC-$p \cdot q$

*The letter(s) following the dash designates the number and levels of each treatment. Refer to chapters in which the designs are discussed for an explanation of the abbreviated designations.

categories, factorial experiments and analysis of covariance experiments. The former pseudocategory is so designated because a factorial experiment consists of a combination of elementary building block designs. The term factorial experiment refers to the simultaneous evaluation of two or more treatments in one experiment rather than to a distinct kind of experimental design. Analysis of covariance experiments combine building block designs with regression analysis procedures and thus do not represent a distinct type of design. A brief description of some of the simpler designs follows.

COMPLETELY RANDOMIZED DESIGN

The simplest complete block experimental design from the standpoint of assignment of subjects to treatment levels and statistical analysis

is the completely randomized design. This design can be used to compare any number of treatment levels. When two treatment levels are used, the statistical test employed in the analysis is equivalent to a test by means of a t ratio for uncorrelated groups. The general features of the design can be illustrated by the microwave radiation example cited earlier. Let b_1, b_2, and b_3 stand for treatment levels 0, 20,000, and 40,000 microwatts of radiation, respectively. Fifteen albino rats are assigned to the three treatment levels by means of a table of numbers. Food consumption of the rats assigned to each treatment level is indicated by X_{ij}, where i designates the ith rat in treatment level j. Table 1.4-2 shows the layout of a completely randomized design. The *average* food consumption of rats in each treatment level is designated by $\bar{X}_{\cdot j}$. The dot in the subscript indicates the variable over which summation has occurred. In this example, treatment means are obtained by summing the scores over the $i = 1$ through 5 rats. The average food consumption for all 15 rats is designated by $\bar{X}_{\cdot\cdot}$.

TABLE 1.4-2 Completely Randomized Design

	Treatment Levels			
	b_1	b_2	b_3	
	X_{11}	X_{12}	X_{13}	
	X_{21}	X_{22}	X_{23}	
	X_{31}	X_{32}	X_{33}	
	X_{41}	X_{42}	X_{43}	
	X_{51}	X_{52}	X_{53}	
Treatment means =	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$	Grand mean = $\bar{X}_{\cdot\cdot}$

Here conclusions concerning the effects of microwave radiation are restricted to the three treatment levels and to the 15 rats included in the experiment. Edgington (1966) recently emphasized that random assignment of subjects to treatment levels is essential if an experimenter wishes to draw statistical inferences concerning treatment effects from nonrandomly selected subjects.* Because of the importance of the principle of random assignment, an experimenter should always describe his technique for assigning subjects to treatment levels.

Associated with every experimental design is a mathematical model that purports to include all sources of variability affecting individual scores. To the extent that the model accurately represents these sources of variability, the experimenter can evaluate the effects of a treatment. The linear model for a completely randomized design is

*Few experiments in the behavioral sciences are carried out with randomly selected subjects. When a random sample is used, the population sampled is likely to be so specific as to be of little interest.

$$X_{ij} = \mu + \beta_j + \varepsilon_{ij}. \tag{1}$$

According to this model, an individual score is equal to the population mean μ, plus a treatment effect β_j, plus an error effect ε_{ij}, which is unique for each individual subject. In a particular experiment, the parameters μ, β_j, and ε_{ij} are unknown, but sample estimates of these parameters are given by $\hat{\mu}$, $\hat{\beta}_j$, and $\hat{\varepsilon}_{ij}$, respectively. It can be shown by maximum-likelihood methods that unbiased estimates of the required parameters are provided by the statistics

$$\hat{\mu} = \bar{X}.. \qquad \rightarrow \mu$$
$$\hat{\beta}_j = (\bar{X}._j - \bar{X}..) \rightarrow \beta_j$$
$$\hat{\varepsilon}_{ij} = (X_{ij} - \bar{X}._j) \rightarrow \varepsilon_{ij}.$$

The symbol $\rightarrow$ indicates that the term on the left is an estimator of the term on the right. According to the maximum-likelihood method, the best estimate is the one that gives the highest probability of obtaining the observed data. It should be noted that a maximum-likelihood estimator is not necessarily unbiased, although the center of its distribution is generally close to the value of the parameter estimated. Assumptions associated with the mathematical model for a completely randomized design are discussed in Chapter 2 and explicitly stated in connection with the description of each design in subsequent chapters.

The meaning of the term *error effect* is somewhat elusive. An intuitive understanding of this term can be obtained by an examination of Table 1.4-2 and the linear model for the design. It is obvious that the scores for all 5 rats exposed to treatment level b_1 in this table will probably not be identical. Variation among the five scores can be attributed to a variety of sources—experiences of the rats prior to participation in the experiment, unintended variation in administration of the treatment level, lack of reliability in measuring the effect of the treatment level, etc. An error effect is an estimate of all effects *not* attributable to a particular treatment level. This can be seen from the linear model if the terms in equation (1) are rearranged and statistics are substituted for the parameters. The equation can be written

$$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_j - \hat{\mu}.$$

Thus the error effect is that portion of a score remaining after the treatment effect and grand mean are subtracted from it. An experimenter attempts, by using an appropriate design and experimental controls, to minimize the size of the error effect. Designs described in subsequent paragraphs permit an experimenter to accomplish this by isolating additional sources of variation that affect individual scores.

RANDOMIZED BLOCK DESIGN

A randomized block design is based on the principle of assigning subjects to blocks so that the subjects within each block are more homogeneous than subjects in different blocks. Assume that the 15 albino rats

in the previous example were taken from five different litters. Rats from the same litter can be expected to be more homogeneous with respect to genetic characteristics than rats from different litters. In Table 1.4-3 the 3 rats in each row that comprise a block are from the same litter. Differences among the litters can be regarded as a nuisance variable that is experimentally isolated through the use of a randomized block design. The subscripts of X_{ij} designate a particular litter and treatment level, in that order. Differences among the column means reflect treatment effects, whereas differences among the row means reflect litter effects.

TABLE 1.4-3 Randomized Block Design

	Treatment Levels			
	b_1	b_2	b_3	Block means
Block (litter) p_1	X_{11}	X_{12}	X_{13}	$\bar{X}_{1\cdot}$
Block (litter) p_2	X_{21}	X_{22}	X_{23}	$\bar{X}_{2\cdot}$
Block (litter) p_3	X_{31}	X_{32}	X_{33}	$\bar{X}_{3\cdot}$
Block (litter) p_4	X_{41}	X_{42}	X_{43}	$\bar{X}_{4\cdot}$
Block (litter) p_5	X_{51}	X_{52}	X_{53}	$\bar{X}_{5\cdot}$
Treatment means =	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$	Grand mean = $\bar{X}_{\cdot\cdot}$

Assignment of the three treatment levels to the rats is randomized independently for each row. The linear model for this design is

$$X_{ij} = \mu + \beta_j + \pi_i + \varepsilon_{ij}.$$

Unbiased estimates of the parameters are given by the statistics

$$\hat{\mu} = \bar{X}_{\cdot\cdot} \qquad \rightarrow \mu$$

$$\hat{\beta}_j = (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}) \qquad \rightarrow \beta_j$$

$$\hat{\pi}_i = (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \qquad \rightarrow \pi_i$$

$$\hat{\varepsilon}_{ij} = (X_{ij} - \bar{X}_{\cdot j} - \bar{X}_{i\cdot} + \bar{X}_{\cdot\cdot}) \rightarrow \varepsilon_{ij}.$$

The term π_i represents an effect attributable to the ith block of 3 rats. It can be shown, by regrouping terms in the linear model and substituting statistics for parameters, that the error effect in a randomized block design is equal to

$$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_j - \hat{\pi}_i - \hat{\mu}.$$

The error effect for a completely randomized design was given earlier as

$$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_j - \hat{\mu}.$$

Thus the error effect $\hat{\varepsilon}_{ij}$ for a randomized block design is equal to the completely randomized design error effect minus a block effect $\hat{\pi}_i$. It is apparent from this that the error effect for a randomized block design will be smaller than the error effect for a completely randomized design if the block effect $\hat{\pi}_i$ is appreciably greater than zero.

One way of increasing the power of an experimental methodology mentioned in Section 1.3 is to choose an experimental design that provides for a more precise estimate of treatment effects and a smaller error variance. A randomized block design is more powerful than a completely randomized design if the block effects in the former design account for an appreciable portion of the total variance. It should be noted that the increased power of the randomized block design was made possible through the use of matched subjects. In many research situations, the increased experimental effort required to match subjects may not justify the greater power obtainable with a randomized block design.

LATIN SQUARE DESIGN

A Latin square design utilizes the blocking principle to obtain homogeneity with respect to two nuisance variables. The levels of the two nuisance variables are assigned to the rows and columns of a Latin square. Treatment levels are identified within each cell of the Latin square. In the randomized block design example, subjects were equated on the basis of genetic characteristics. It is reasonable to assume that rats in the same litter are also relatively homogeneous in weight. However, because the dependent variable in the radiation example is food consumption, the experimenter might wish to control the extraneous variable of weight. This can be accomplished by assigning the lightest rat in each litter to category b_1, the rat intermediate in weight to category b_2, and the heaviest rat to category b_3. Blocking with respect to both genetic characteristics a_i and weight b_j is shown in Table 1.4-4.

TABLE 1.4-4 Latin Square Design

	Weight Categories of Rats			
	b_1 Lightest	b_2 Intermediate	b_3 Heaviest	Block means
Block (litter) a_1	c_1 X_{111}	c_2 X_{122}	c_3 X_{133}	$\bar{X}_{1\cdot\cdot}$
Block (litter) a_2	c_2 X_{212}	c_3 X_{223}	c_1 X_{231}	$\bar{X}_{2\cdot\cdot}$
Block (litter) a_3	c_3 X_{313}	c_1 X_{321}	c_2 X_{332}	$\bar{X}_{3\cdot\cdot}$
Weight means =	$\bar{X}_{\cdot 1\cdot}$	$\bar{X}_{\cdot 2\cdot}$	$\bar{X}_{\cdot 3\cdot}$	

Grand mean = $\bar{X}_{\cdots}$

Treatment level means: $c_1 = (X_{111} + X_{321} + X_{231})/3 = \bar{X}_{\cdot\cdot 1}$

$c_2 = (X_{212} + X_{122} + X_{332})/3 = \bar{X}_{\cdot\cdot 2}$

$c_3 = (X_{313} + X_{223} + X_{133})/3 = \bar{X}_{\cdot\cdot 3}$

The three subscripts designate a particular block, weight category, and treatment level, in that order. The three treatment levels c_k are randomly assigned to the nine cells with the restriction that each treatment level must occur in any row and any column only once. In order to achieve this balance, a Latin square design must have the same number of rows, columns, and treatment levels. Consequently, only 9 animals can be used in the design shown in Table 1.4-4 instead of the 15 animals used in the two designs described previously.

The linear model for this design is

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}.$$

An individual score is equal to the grand mean μ, plus a block effect α_i, plus a column effect β_j, plus a treatment effect γ_k, plus an error effect ε_{ijk}. If the block and column effects, α_i and β_j, in a Latin square design are appreciably greater than zero, the design may be more powerful than either a completely randomized or a randomized block design. This is apparent if the error effect is examined by means of the procedure used for the two designs described previously. The error effect for a Latin square design is equal to

$$\hat{\varepsilon}_{ijk} = X_{ijk} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_k - \hat{\mu}.$$

INCOMPLETE BLOCK DESIGN

An incomplete block design is particularly applicable to research situations in which the number of subjects available for each block is less than the number of treatment levels. If, for example, only 2 albino rats from each litter are available, and the experimenter wants to use three treatment levels, an incomplete block design is required. This design is shown in Table 1.4-5.

TABLE 1.4-5 Incomplete Block Design

	Treatment Levels			
	b_1	b_2	b_3	Block means
Block (litter)	X_{11}		X_{13}	$\bar{X}_{1\cdot}$
Block (litter)		X_{22}	X_{23}	$\bar{X}_{2\cdot}$
Block (litter)	X_{31}	X_{32}		$\bar{X}_{3\cdot}$
Treatment means =	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$	Grand mean = $\bar{X}_{\cdot\cdot}$

The linear model for this design is

$$X_{ij} = \mu + \beta_j + \pi_i + \varepsilon_{ij}.$$

It should be noted that each block contains the same number of subjects, each treatment level occurs the same number of times, and

subjects are assigned to the treatment levels so that each possible pair of treatment levels occurs together within some block an equal number of times. A design having these characteristics is called a *balanced* incomplete block design. *Partially balanced* designs are those in which some pair of treatment levels occur together within the blocks more often than do other pairs.

FACTORIAL EXPERIMENT

A factorial experiment permits an investigator to evaluate the combined effects of two or more treatments in a single experiment. This is accomplished by combining building block designs so that one level from each of two or more treatments is presented simultaneously. The most commonly used building block designs are the completely randomized design and the randomized block design.

In the microwave radiation example, an investigator can, by using a factorial experiment, evaluate the effects of radiation and also the effects of a second treatment, such as room temperature. Assume that there are two levels of ambient room temperature, $a_1 = 80°$ and $a_2 = 65°$, and three levels of radiation, $b_1 = 0$, $b_2 = 20{,}000$, and $b_3 = 40{,}000$ microwatts. Tables 1.4-6 and 1.4-7 illustrate the use of two frequently used building block designs in a factorial experiment. In Table 1.4-6 the three subscripts designate a particular temperature level, radiation level, and subject, in that order. The three subscripts in Table 1.4-7 designate a particular temperature level, radiation level, and block, in that order.

In the completely randomized factorial design of Table 1.4-6 it is assumed that 18 albino rats are randomly assigned to the six treatment level combinations. In the randomized block factorial design example shown in Table 1.4-7, the treatment level combinations are randomly assigned within each block of litter mates. The models for the completely randomized and randomized block factorial designs are, respectively,

$$X_{ijm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{m(ij)}$$

$$X_{ijm} = \mu + \alpha_i + \beta_j + \pi_m + \alpha\beta_{ij} + \varepsilon_{ijm}.$$

The effect of temperature is designated by α_i, radiation by β_j, interaction of temperature and radiation by $\alpha\beta_{ij}$, experimental error by ε, and litter by π_m. Both designs permit an investigator to determine if radiation dosage has the same effect on food consumption at 80° ambient room temperature as it has at 65° ambient room temperature. It is conceivable that radiation might be more detrimental at a high ambient room temperature than at a low ambient room temperature. If such a result is found, it is called an *interaction effect*.

The error effect for a completely randomized factorial design, using the scheme described previously, can be written

$$\hat{\varepsilon}_{m(ij)} = X_{ijm} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\alpha\beta}_{ij} - \hat{\mu}.$$

If we let τ_{ij} stand for all treatment effects, the error effect can be written

$$\hat{\varepsilon}_{m(ij)} = X_{ijm} - \hat{\tau}_{ij} - \hat{\mu}.$$

In this form the similarity between the error effect for this design and the error effect for a completely randomized design is apparent. This latter error effect was given earlier as

$$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_j - \hat{\mu}.$$

The similarity between the models for a completely randomized design and a completely randomized factorial design is not surprising in view of the fact that the former design is the *building block* for the latter design.

The error effect for a randomized block factorial design is

$$\hat{\varepsilon}_{ijm} = X_{ijm} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\pi}_m - \hat{\alpha\beta}_{ij} - \hat{\mu}.$$

TABLE 1.4-6 Completely Randomized Factorial Design

	Radiation Levels			
Temperature Levels	b_1	b_2	b_3	*A* treatment means
a_1	X_{111} X_{112} X_{113}	X_{121} X_{122} X_{123}	X_{131} X_{132} X_{133}	$\bar{X}_{1\cdot\cdot}$
a_2	X_{211} X_{212} X_{213}	X_{221} X_{222} X_{223}	X_{231} X_{232} X_{233}	$\bar{X}_{2\cdot\cdot}$
B treatment means =	$\bar{X}_{\cdot 1\cdot}$	$\bar{X}_{\cdot 2\cdot}$	$\bar{X}_{\cdot 3\cdot}$	Grand mean = $\bar{X}_{\cdots}$

TABLE 1.4-7 Randomized Block Factorial Design

Temperature Levels	a_1	a_1	a_1	a_2	a_2	a_2	
Radiation Levels	b_1	b_2	b_3	b_1	b_2	b_3	Block mean
Block (litter) p_1	X_{111}	X_{121}	X_{131}	X_{211}	X_{221}	X_{231}	$\bar{X}_{\cdot\cdot 1}$
Block (litter) p_2	X_{112}	X_{122}	X_{132}	X_{212}	X_{222}	X_{232}	$\bar{X}_{\cdot\cdot 2}$
Block (litter) p_3	X_{113}	X_{123}	X_{133}	X_{213}	X_{223}	X_{233}	$\bar{X}_{\cdot\cdot 3}$
Column means	$\bar{X}_{11\cdot}$	$\bar{X}_{12\cdot}$	$\bar{X}_{13\cdot}$	$\bar{X}_{21\cdot}$	$\bar{X}_{22\cdot}$	$\bar{X}_{23\cdot}$	Grand mean = $\bar{X}_{\cdots}$

A_1 treatment mean = $(X_{111} + X_{112} + X_{113} + X_{121} + \cdots + X_{133})/9 = \bar{X}_{1\cdot\cdot}$

A_2 treatment mean = $(X_{211} + X_{212} + X_{213} + X_{221} + \cdots + X_{233})/9 = \bar{X}_{2\cdot\cdot}$

B_1 treatment mean = $(X_{111} + X_{112} + X_{113} + X_{211} + X_{212} + X_{213})/6 = \bar{X}_{\cdot 1\cdot}$

B_2 treatment mean = $(X_{121} + X_{122} + X_{123} + X_{221} + X_{222} + X_{223})/6 = \bar{X}_{\cdot 2\cdot}$

B_3 treatment mean = $(X_{131} + X_{132} + X_{133} + X_{231} + X_{232} + X_{233})/6 = \bar{X}_{\cdot 3\cdot}$

If treatment effects are designated by τ_{ij}, the error effect can be written

$$\hat{\varepsilon}_{ijm} = X_{ijm} - \hat{\tau}_{ij} - \hat{\pi}_m - \hat{\mu}.$$

It is interesting to note the similarity between this error effect and the error effect for a randomized block design, which is the building block for this factorial design. The error effect for the randomized block design was given earlier as

$$\hat{\varepsilon}_{ij} = X_{ij} - \hat{\beta}_j - \hat{\pi}_i - \hat{\mu}.$$

QUESTIONS TO CONSIDER IN SELECTING AN APPROPRIATE DESIGN

Statisticians have provided an experimenter with a vast array of experimental designs. On what basis should an experimenter decide which design to use? Selection of the *best* experimental design for a particular research problem requires (1) a knowledge of the research area and (2) a knowledge of different experimental designs. To arrive at the best experimental design, an experimenter must consider the following questions:

1. What kinds of data are required to test the statistical hypotheses?
   (a) How many treatment levels should be used?
   (b) Should the treatment levels used in the experiment be selected on an *a priori* basis or by random sampling from a population of treatment levels?
   (c) Should a factorial experiment be used so that interaction effects may be evaluated?
   (d) Are all treatments and treatment levels of equal interest to the experimenter? Experimental designs may be used that sacrifice power in evaluating some treatments in order to gain power in evaluating other treatments.
2. Is the proposed sample of subjects large enough to provide adequate precision in testing the statistical hypotheses?
   (a) Do the available subjects represent a random sample from the population of interest to the experimenter?
   (b) Can the subjects be stratified into homogeneous blocks?
   (c) Does the nature of the experiment permit each subject to be observed under more than one treatment level?
   (d) Will the treatment(s) produce physical or psychological injury to the subjects? The use of potentially injurious treatments precludes the employment of human subjects.
3. Is the power of the proposed experimental design adequate to test the statistical hypotheses?
   (a) What is the size of treatment effects that the experimenter considers to be of practical interest?
   (b) What are the consequences of committing type I and type II errors?
4. Does the proposed experimental design provide maximum efficiency in testing the statistical hypotheses?
   (a) Would efficiency be improved more by using a design employing blocks of homogeneous subjects or by using random assignment of a large number of subjects to the treatment levels?

(b) Can efficiency be increased more by the use of a larger sample size or by exercising additional experimental controls during the conduct of the experiment?
(c) Can efficiency be increased by the measurement of one or more characteristics related to the dependent variable in order to use regression techniques?
(d) Can efficiency be increased more by the use of a complex experimental design that requires considerable time to plan and analyze or by using a simple design but a large number of subjects? If subjects are plentiful and time required to obtain the data is sufficient, a simple design utilizing a large number of subjects may be more efficient than a complex design that involves costly planning and statistical analysis.

It should be apparent that the question "What is the best experimental design to use?" is not easily answered. Statistical as well as nonstatistical factors must be considered. The discussion has emphasized economic factors in the selection of a design because rules can be explicitly stated for increasing the precision and power of an experimental methodology; but, when efficiency is considered, such rules are difficult to formulate.

ROLE OF EXPERIMENTER AND STATISTICIAN

It is the conviction of the author that the selection of the best experimental design for a particular research problem can be most expeditiously accomplished when the roles of experimenter and statistician are performed by the same person. This is essentially the same position taken by Finney (1960, 3), who states, ". . . to write of the 'experimenter' and the 'statistician' as though they are separate persons is often convenient; the one is concerned with undertaking a piece of research comprehensively and accurately yet with reasonable economy of time and materials, the other is to provide technical advice and assistance on quantitative aspects both in planning and in interpretation . . . the statistician can produce good designs only if he understands something of the particular field of research, and the experimenter will receive better help if he knows the general principles of design and statistical analysis. Indeed, the two roles can be combined when an experimenter with a little mathematical knowledge is prepared to learn enough of the theory of design to be able to design his own experiments."

CRITERIA FOR EVALUATING AN EXPERIMENTAL DESIGN

Many different sets of criteria could be given for evaluating an experimental design. The criteria presented by Winer (1962, 47) and Lindquist (1953, 6) are most helpful. The following questions, except for number 5, were selected because they touch on the major points presented in this chapter.

1. Does the design permit an experimenter to calculate a valid estimate of the experimental effects and error effects?
2. Does the data-collection procedure produce reliable results?
3. Does the design provide maximum efficiency within the constraints imposed by the experimental situation?
4. Does the design possess sufficient power to permit an adequate test of the statistical hypotheses?
5. Does the experimental procedure conform to accepted practices and procedures used in the research area? Other things being equal, an experimenter should use procedures that offer an opportunity for comparison of his findings with the results of other investigations.

1.5 A REVIEW OF STATISTICAL INFERENCE

In the previous section an overview of experimental designs was presented. This section is written to accomplish the same goal with respect to procedures involved in statistical inference. It is assumed that the reader is already familiar with basic hypothesis-testing concepts. Hays (1963) presents an excellent introduction to this topic.

A distinguishing characteristic of the scientific method is the formulating and testing of hypotheses. The testing of hypotheses requires the *a priori* formulation of decision rules to guide the decision maker. The problem may be stated: Given two mutually exclusive hypotheses about a population, how does one decide on the basis of sample data which hypothesis is supported? It will be apparent that this question lacks a simple answer.

A statistical hypothesis is a statement about one or more parameters of population distributions; and, as such, it refers to a situation that *might* be true. Such a statement is always made with respect to a population and not to a sample. Distinguishing between statistical hypotheses and research or scientific hypotheses is important. Research hypotheses are normally stated in general terms, at least in the initial stages of an inquiry. In this form they are not amenable to evaluation through the use of the procedures and theory of statistical inference. It may be possible, by means of deductive reasoning, to transform a research hypothesis into a statistical hypothesis that can be subjected to test. Statistical hypotheses refer to population parameters, whereas scientific hypotheses refer to the phenomena of nature and man (Clark, 1963).

In logic, the terms *direct statement* and *indirect statement* are analogous in many ways to statistical and scientific hypotheses. A direct statement is made in reference to limited phenomena that are directly observable; for example, "This rat is running." The truth or falsity of such a direct statement can be determined by observing the rat. An indirect statement refers to phenomena that cannot be directly observed or that are so numerous in time that it is impossible to view them all. For example,

"All rats run under condition X." Such hypotheses can be evaluated by inductive inference only and must be reducible to direct statements. The chain of events required in testing an indirect statement is shown in Figure 1.5-1.

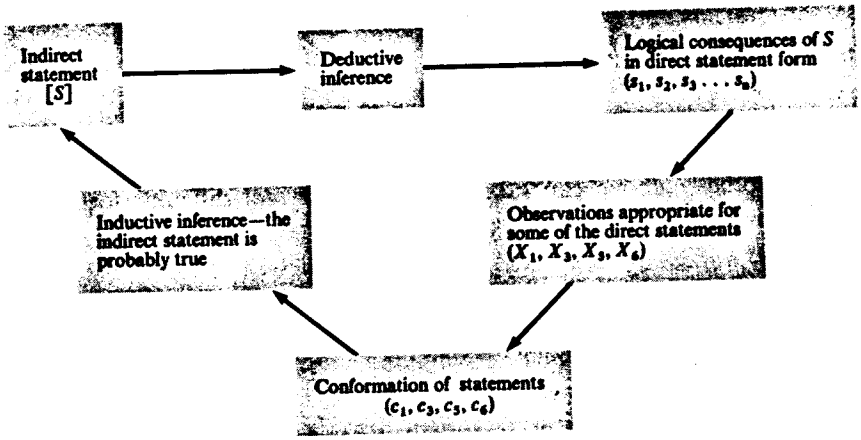

Figure 1.5-1 Test of indirect statement by deductive and inductive inferences.

If, in Figure 1.5-1, the term *scientific hypothesis* is substituted for indirect statement and *statistical hypothesis* for direct statement, the analogy with hypothesis testing is readily apparent. The experimenter has the task of translating his research hypotheses into a dichotomous set of mutually exclusive statistical hypotheses. It should be noted that the chain of deductive reasoning from a *question* concerning nature to a *research hypothesis* to a *statistical hypothesis* and the reverse process of inductive reasoning from the *statistical hypothesis* to the *question* is an exercise in logic rather than statistical inference. If an error occurs in the chain, the statistical hypotheses subjected to test may have no bearing on the original question, or incorrect inferences concerning the question may be made. Grant (1962), Binder (1963), and Edwards (1965) have examined in detail the relation between scientific and statistical hypotheses.

KINDS OF STATISTICAL HYPOTHESES

A null hypothesis (H_0) is the statistical hypothesis that is subjected to a test. The notion that the null hypothesis refers to a parameter value of zero is a simplification; the hypothesis can specify the parameter as having any value, including zero. Less confusion will result if the null hypothesis is considered as the hypothesis that is tested. The hypothesis that remains tenable if the null hypothesis is rejected is called the alternative hypothesis (H_1). Hypothesis testing can be viewed as a procedure whereby an experimenter decides which one of a dichotomous set of mutually exclusive and

exhaustive hypotheses is to be rejected and which one is to be accepted at some specified risk of making an incorrect decision (Clark, 1963).

A statistical hypothesis can be either exact or inexact. The hypothesis that the mean (μ) of population j is equal to 40

$$H_0: \mu_j = 40 \tag{1}$$

is an exact hypothesis. The hypothesis

$$H_0: \mu_j \leq 40 \tag{2}$$

is an example of an inexact hypothesis. The alternative (H_1) to the exact null hypothesis above can take any one of several different forms; for example,

$H_1: \mu_j = 43$ (exact alternative hypothesis)

$H_1: \mu_j \neq 40$ (inexact two-tailed alternative hypothesis).

The alternative to the inexact null hypothesis above can be written

$$H_1: \mu_j > 40.$$

If a comparison of the central tendency of two populations j and j' is of interest, the null and alternative hypotheses can take any of the following forms:

$$H_0: \mu_j - \mu_{j'} = 0$$
$$H_1: \mu_j - \mu_{j'} \neq 0$$

or

$$H_0: \mu_j - \mu_{j'} \leq 0$$
$$H_1: \mu_j - \mu_{j'} > 0$$

or

$$H_0: \mu_j - \mu_{j'} \geq 0$$
$$H_1: \mu_j - \mu_{j'} < 0.$$

It should be noted that hypothesis testing in the behavioral sciences usually involves either two inexact hypotheses or one exact and one inexact hypothesis. The distinction between exact and inexact hypotheses is unimportant from a practical standpoint because the same general test procedures are followed in each case. Although we may speak of testing a single hypothesis, in practice we behave as though we were deciding which one of two mutually exclusive and exhaustive hypotheses is supported by our data. The procedure by which we make this decision is called a statistical test.

STATISTICAL TEST

A statistical test is the comparison of two hypotheses in the light of sample data according to a set of decision rules. The null hypothesis

leads to a prediction, or anticipated value, and to a hypothetical sampling distribution of anticipated values for a sample statistic. If the sample statistic equals the anticipated value, or falls in a region of the sampling distribution designated as a *probable anticipated value*, a decision is made to accept the null hypothesis. On the other hand, if the sample statistic deviates appreciably from the anticipated value, either a rare and improbable event has occurred or the null hypothesis has led to a poor prediction and should be rejected.

HYPOTHESIS TESTING

Hypothesis testing appears to be a straightforward objective procedure until an attempt is made to define such phrases as "probable anticipated value," "deviates appreciably," and "poor prediction." On what basis does one decide which anticipated values are probable, or when the sample statistic deviates appreciably from the anticipated value, or when a null hypothesis leads to a poor prediction? The answer to these questions in the behavioral sciences is that the experimenter falls back on a set of conventions. A branch of mathematics known as *decision theory* deals with the problem of choosing optimum decision rules. Although hypothesis-testing procedures in the behavioral sciences use many notions from decision theory, the application is incomplete and research is frequently conducted according to rules that are less than optimum for the experimenter's purposes.

STEPS FOLLOWED IN TESTING A HYPOTHESIS

What conventions are currently used in testing a hypothesis? These conventions can be summarized in four steps.

Step 1: State a null hypothesis H_0 and an alternative hypothesis H_1.

Step 2: Decide on an appropriate sample statistic and test statistic. The selection of a test statistic is based on (1) H_0, (2) the chosen sample statistic, and (3) tenable assumptions concerning the population distributions. Assumptions underlying the sampling distributions of χ^2, t, and F test statistics are discussed in Section 2.1.

Step 3: Decide on a level of significance α and a sample size N. α and N, together with the sampling distribution of the test statistic under the null hypothesis, determine the *region for rejecting* H_0.
The *location* and *size* of the region for rejection of the null hypothesis are determined by H_1 and α, respectively. An experimenter attempts to select a level of significance so that the region of rejection contains values of the test statistic that have a low probability of occurrence if H_0 is true but a high probability if H_1 is true.

Step 4: Obtain the sample statistic and compute the test statistic. If the value of the test statistic falls in the region of rejection, H_0 is rejected in favor of H_1. If the test statistic falls outside the region of rejection, the experimenter may either accept H_0 or suspend making a decision concerning it.

These four steps and the conventions they summarize require some amplification. First, the selection of an appropriate sample statistic is determined by the experimenter's interest in a particular parameter or characteristic of a population. If only one population is involved, an experimenter in the behavioral sciences is generally interested in testing a hypothesis with respect to the central tendency of the population. If, as is frequently the case, more than one population is involved, hypotheses concerning differences among the populations in terms either of central tendency or of dispersion may be of interest. The measures most often adopted to describe central tendency and dispersion are the mean and standard deviation, respectively.

Test statistics are similar to sample statistics in that both have sampling distributions; however, unlike sample statistics, test statistics are not used to estimate population parameters. Instead, test statistics provide information in the form of a probability statement, which is used by an experimenter in deciding whether or not to reject a null hypothesis. Conventionally, an experimenter specifies a region of the sampling distribution of the test statistic based on α and H_1 that will lead to rejection of the null hypothesis *prior* to computation of the test statistic. This region is specified in such a way as to contain those values of the test statistic that have a small probability of occurring if the null hypothesis is true but a high probability of occurring if the alternative hypothesis is true. If the test statistic falls in the region for rejection, either an improbable event has occurred or the null hypothesis is false and should be rejected.

Two commonly used test statistics are z and t ratios. For testing a hypothesis concerning a single population mean, z and t ratios have the following form:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{and} \quad t = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

where $\bar{X}$ = sample mean used to estimate the population mean μ, μ_0 = value of population mean specified by null hypothesis, σ = population standard deviation, $\hat{\sigma}$ = unbiased estimate of population standard deviation calculated from a sample, and n = number of observations in the sample. If it can be assumed that the population sampled has a normal distribution, z and t are distributed as the normal curve and t distribution, respectively. That is, a z or t ratio can be computed for each conceivable sample of n independent observations drawn from a normal population with mean = μ. The value of z or t will vary over the different samples from the population. If a plot of the probability-density of each z or t value is made, the resulting distributions will be distributed as the normal distribution and t distribution, respectively. The t distribution, unlike the z distribution, actually is a family of distributions. The exact shape of the t distribution varies, depending on the number of observations in the sample. Probabilities associated with obtaining various values of z or t are given in Tables D.3 and D.4, respectively. It should be noted that the denominator

of a z ratio is a constant for any sample of size n because the population parameter σ is a constant. By comparison, the denominator of a t ratio for any sample of size n is a random variable because of sampling variation in estimating the parameter σ. The numerators of both test statistics are subject to sampling variation and hence are random variables.

In practice, the population variance required to compute z is rarely ever known. The ratio

$$\frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

can be treated as a z variable provided that the sample size is large, say around 100, and that the population has a normal distribution. An examination of Appendix Tables D.3 and D.4 reveals that the probabilities associated with values of z and t are quite similar even for samples as small as 30. As the sample size is reduced below 30, the correspondence becomes poorer. Thus, for small samples, the t distribution should be employed if a *sample* standard deviation is used to estimate σ. The t distribution and other useful sampling distributions are discussed in Section 2.1.

EXAMPLE ILLUSTRATING STEPS IN HYPOTHESIS TESTING

An example may help to clarify the concepts and conventions involved in hypothesis testing. Assume that we wish to test the hypothesis that the average performance of some population on a psychological test is greater than 100. An arithmetic mean is chosen as the appropriate measure of central tendency. Two *test statistics* can be suggested for this experiment:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{and} \quad t = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}.$$

Let us assume in our example that the population is normally distributed and that σ is known to equal 15. Under these conditions, the appropriate test statistic is z. The statistical hypotheses can be stated as follows:

$$H_0: \mu \leq 100$$

$$H_1: \mu > 100.$$

We will reject the null hypothesis in favor of the alternative hypothesis only if an observed sample mean is so much larger than 100 that it has a probability of .05 or less of occurring if the population mean really is equal to 100. As written, the null hypothesis is inexact because it states a whole region of possible values for the population mean. However, one exact value is specified, $\mu = 100$. Actually, the hypothesis tested is $\mu = 100$

versus some unspecified alternative greater than 100. If the experimenter can reject the hypothesis that $\mu = 100$ at $\alpha = .05$ level of significance, then he can reject any other hypothesis that $\mu < 100$ at a level of significance $\leq .05$. The decision rule for this example can be stated as follows: If the test statistic falls among the highest 5 percent of z's in a normal distribution under H_0, reject H_0; otherwise do not reject H_0. If H_0 is rejected, the experimenter, in this example, decides in favor of H_1.

We have stated the null and alternative hypotheses, decided on an appropriate sample statistic and test statistic, and specified the level of significance that will be used in the decision process. The final steps in the hypothesis-testing procedure are to specify the size of the sample that will be observed, obtain the sample, compute the sample statistic and test statistic, and make a decision. Suppose that a random sample of $n = 100$ observations has been obtained from the population and that the mean of this sample is equal to $\bar{X} = 102$. Is the deviation of this sample mean from the predicted mean of 100 large enough to lead the experimenter to reject the null hypothesis? The probability associated with obtaining a sample mean as deviant as 102 if the true mean is 100 can be determined from

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{102 - 100}{15/\sqrt{100}} = \frac{2.0}{1.5} = 1.33$$

and the cumulative normal probability table in Appendix D.3. According to Appendix D.3, the probability associated with obtaining a sample mean of 102 if the true mean is 100 is approximately .09. According to the decision rules outlined above, the null hypothesis is not rejected, because $.09 > .05$, and therefore z does not fall in the region for rejection of H_0. The regions for rejection or nonrejection of H_0 are illustrated in Figure 1.5-2. If the

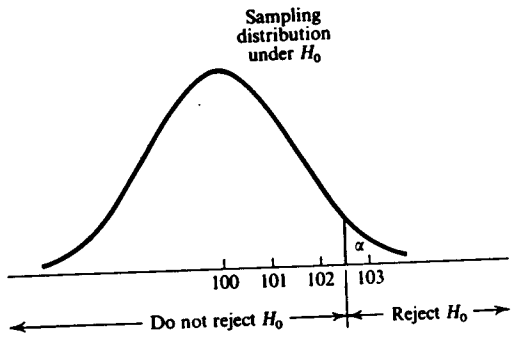

Figure 1.5-2 Regions of the sampling distribution of z that lead to rejection or nonrejection of H_0 according to decision rules specified previously.

sample mean had been 102.5 instead of 102, the decision would have been to reject H_0. This can be demonstrated as follows:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{102.5 - 100.0}{1.5} = 1.67.$$

According to the normal distribution table in Appendix D.3, the probability of obtaining a sample mean of 102.5 if the true mean is really 100 is less than .05. Thus, either a rare and improbable event has occurred or the true parameter is not 100. This example was fabricated to illustrate the steps involved in testing a statistical hypothesis. In real life most hypotheses are concerned not with a single population but with differences among two or more populations. The steps that have been described in connection with a single population are also applicable to tests involving two or more populations. Procedures and assumptions associated with testing statistical hypotheses with respect to two or more populations are described in Chapters 2 and 3. An excellent survey of hypothesis testing and statistical inference is given by Clark (1963).

TYPE I AND II ERRORS

In carrying out the decision process outlined above, the experimenter may make a correct decision or he may commit an error. If he decides to reject H_0 when the population mean is really equal to 100, he has committed a type I error. On the other hand, if he decides not to reject H_0 when the population mean is really equal to, say, 103, he has committed a type II error. In summary, the two possible errors an experimenter may make are

Type I error. Reject H_0 (tested hypothesis) when it is true. The probability α is the risk of making a type I error.

Type II error. Fail to reject H_0 when it is false. The risk of making a type II error is designated as β.

The regions corresponding to the probability of making a type I error (α) and a type II error (β) are shown in Figure 1.5-3.

It is apparent from Figure 1.5-3 that the probability of making a type I error is determined by an experimenter when he specifies α. This probability can be made as small as an experimenter wishes. It should be noted from the figure that as the area corresponding to α is made smaller, the area designated as β becomes larger. Thus the two types of errors are interrelated. The probability of committing a type II error is determined by α, magnitude of difference between the true parameter and parameter under H_0, size of population error variance, and size of sample (n). If, in the hypothesis-testing example described previously, the statistic is equal to 102, a decision is made not to reject H_0. This decision may be correct or incorrect, depending on the value of the parameter. If

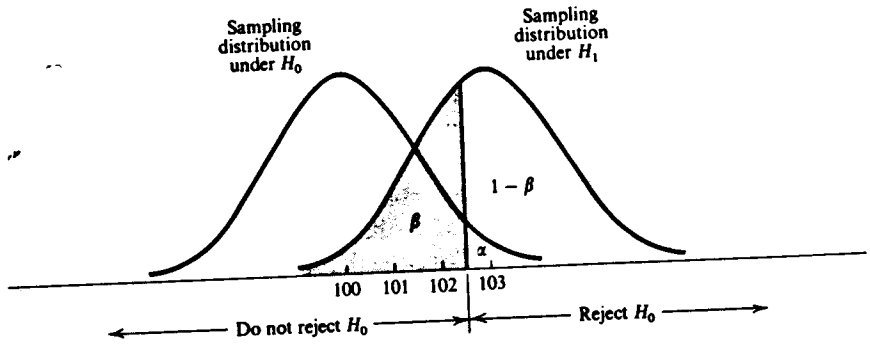

Figure 1.5-3 Regions corresponding to probabilities of making type I error (α) and type II error (β). The region corresponding to a type I error is determined by the experimenter when he specifies α and H_1. If, for a given H_0 and true alternative, α is made smaller, the probability of making a type II error is increased.

the parameter is equal to or less than 100, a correct decision has been made. On the other hand, if the parameter is equal to 103, a type II error has been made.

POWER OF A TEST

If the parameter is equal to 100, the probability of making a correct decision is $1 - \alpha$. If, on the other hand, the parameter is equal to 103, the probability of making a correct decision is $1 - \beta$. This latter probability is called the power of the test. It is simply the probability of deciding that H_0 is wrong, given a decision rule and the true value under H_1. The possible decision outcomes can be categorized as shown in Table 1.5-1.

TABLE 1.5-1 Decision Outcomes Categorized

		True Situation	
		$\mu = 100$	$\mu = 103$
Decision	$\mu = 100$	correct decision $= 1 - \alpha$	type II error $= \beta$
	$\mu = 103$	type I error $= \alpha$	correct decision $= 1 - \beta$

An experimenter attempts to select an experimental design and set of decision rules that will result in the highest power for a given type I

error. How can the power of an experimental methodology be increased for a given type I error rate? Two approaches were suggested in Section 1.3. One approach is to increase the size of the sample. A second approach is to use an experimental design that provides for a more precise estimate of treatment effects and a smaller error variance.

If information concerning the true parameter is available, the probability of committing a type II error can be determined. Generally, however, the value of the parameter is unknown. In practice, an experimenter can specify various possible values of the parameter of interest to him and then compute the probability of committing a type II error and $1 - \beta$, given that the specified value of μ is true. Let us assume that the obtained sample statistic is equal to 102.5 and that an experimenter is interested in determining the probability of correctly rejecting H_0 if the population mean is really equal to 103. The probabilities β and $1 - \beta$ can be determined from

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{102.5 - 103.0}{15/\sqrt{100}} = -.33.$$

According to the normal distribution table in Appendix D.3, the probabilities β and $1 - \beta$ are .37 and .63, respectively. The location of the regions corresponding to β and $1 - \beta$ are shown in Figure 1.5-3. In this example the probability of making a correct decision if $\mu = 103$ is only .63, whereas the corresponding probability if $\mu = 100$ is .95. The probabilities associated with the possible outcomes of our decision rule are summarized in Table 1.5-2.

TABLE 1.5-2 Probabilities Associated with the Decision Process

		True Situation	
		$\mu = 100$	$\mu = 103$
Decision	$\mu = 100$	$1 - \alpha = .95$	type II error $\beta = .37$
	$\mu = 103$	type I error $\alpha = .05$	$1 - \beta = .63$

SELECTION OF A LEVEL FOR α

In the preceding hypothetical example, the probability of a type I error (α) is much lower than the corresponding type II error (β). Experimenters in the behavioral sciences frequently set the type I error rate at .05 or .01. This convention is based primarily on the notion that a type I

error is very bad and is to be avoided. In the present example, the decision rule is biased in favor of deciding that the population mean is equal to 100 rather than, say, 103. In many research situations, the cost of a type I error may be large relative to the cost of a type II error. For example, to commit a type I error in concluding that a particular medication arrests the production of cancer cells and therefore can be used in place of other medical procedures is a serious matter. On the other hand, falsely deciding that the medication does not arrest the production of cancer cells (type II error) would result in withholding the medication from the public and would probably lead to further research. In such a context, a type II error is less undesirable than a type I error. However, in another context, concluding that an experimental effect is not significant may result in an experimenter discontinuing a promising line of research whereas a type I error would mean further exploration into a *blind alley*. Faced with these two alternatives, many experimenters might prefer to make a type I rather than a type II error. It is apparent from the foregoing discussion that the *loss function* associated with the two errors must be known before a rational choice concerning α can be made. However, experimenters in the behavioral sciences are generally unable to specify the losses associated with the two errors of inference. Therein lies the problem. The problem is resolved by falling back on accepted conventions. The principal benefit of statistical decision theory—that of using decision rules having optimum properties for a given purpose—is seldom enjoyed by experimenters in the behavioral sciences. A general introduction to the meaning of optimal solutions to problems is given by Ackoff, Gupta, and Minas (1962).

It is hoped that the preceding discussion helps to dispel the magic that seems so inextricably tied to the .05 and .01 levels of significance. The use of the .05 or .01 level of significance in hypothesis testing is a convention. When either level is achieved by a test, it signals that an *improbable* event has occurred or that the hypothesis under test has led to a poor prediction. A test of significance provides information concerning the probability of committing an error in rejecting the null hypothesis. It is one bit of information required in making a decision concerning a research hypothesis. A test of significance embodies no information concerning loss-values associated with the decision, the experimenter's prior personal convictions concerning the hypotheses, or the importance or usefulness of the obtained results. Various problems associated with the uncritical use of significance tests in research have been examined in detail by Bakan (1966). Bayesian statistical theory represents an attempt to incorporate prior information into the decision process, information that is not utilized within the classical theories of Neyman-Pearson and Fisher. A rapprochement involving the best features of classical theory, decision theory, and Bayesian theory is to be hoped for. Binder (1964) in a review mentioned one modification of classical theory that incorporates Bayesian theory. A general introduction to Bayesian theory can be found in Edwards, Lindman, and Savage (1963) and additional references in Binder (1964).

Hypothesis-testing procedures should be viewed as tools that aid an experimenter in interpreting the outcome of research. Such procedures should not be permitted to replace the judicial use of logic by an alert analytic experimenter. In particular, the technique of analysis of variance described in this book should be considered an aid in summarizing data. It should be used to help an experimenter understand what went on in an experiment; it is not an end in itself.

2.5 THE *F* RATIO IN ANALYSIS OF VARIANCE

An F ratio in analysis of variance provides a test of the hypothesis that all treatment population means are equal. That is,

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k = \mu.$$

This null hypothesis is equivalent to the hypothesis that

$$H_0: \beta_j = 0 \qquad \text{for all } j.$$

We have seen that when H_0 is true,

$$\mathrm{E}(\mathrm{MS_{BG}}) = \sigma_\varepsilon^2$$

and

$$\mathrm{E}(\mathrm{MS_{WG}}) = \sigma_\varepsilon^2.$$

When the null hypothesis is false and the alternative hypothesis that

$$H_1: \beta_j \neq 0 \qquad \text{for some } j$$

is true,

$$\mathrm{E}(\mathrm{MS_{BG}}) > \mathrm{E}(\mathrm{MS_{WG}}).$$

If the null hypothesis is true, we know from Section 2.1 that the random variables $\mathrm{MS_{BG}}/\sigma_\varepsilon^2$ and $\mathrm{MS_{WG}}/\sigma_\varepsilon^2$ are both distributed as chi-square variables divided by their respective degrees of freedom. Thus, if the null hypothesis is true, if $\mathrm{MS_{BG}}$ and $\mathrm{MS_{WG}}$ are statistically independent, and if the population variances are homogeneous, the ratio

$$\frac{\mathrm{MS_{BG}}}{\mathrm{MS_{WG}}} = \frac{\sigma_\varepsilon^2 \chi^2_{(\nu_1)}/\nu_1}{\sigma_\varepsilon^2 \chi^2_{(\nu_2)}/\nu_2} = F_{(\nu_1, \nu_2)}$$

is distributed as the F distribution, with $\nu_1 = k - 1$ and $\nu_2 = kn - k$ degrees of freedom. It can be stated without proof that the mean $\bar{X}_j$ and the variance $\hat{\sigma}_j^2$ estimates are statistically independent provided that the population is normally distributed. Hence $\mathrm{MS_{BG}}$ and $\mathrm{MS_{WG}}$ are independent as long as the k samples of observations are independently drawn from normally distributed populations. The probability of obtaining an F as large as that observed in an experiment if the null hypothesis is true can be determined from a table of F given in Appendix D.

An F ratio, as defined above, always provides a one-tailed test of H_0. Ratios less than 1.0 have no meaning with respect to H_0. Such ratios may occur as a result of the operation of chance, for both numerator and denominator are subject to sampling error. Ratios less than 1.0 may also occur because of failure to randomize some important factor properly in the experimental design or because some of the assumptions concerning the linear model for the design are inappropriate. In summary, if two sets of assumptions are tenable—those associated with the derivation of the

distribution of F and those associated with the mathematical model for a particular design—the F ratio can be used to test the hypothesis that all treatment population means are equal.

2.6 EFFECTS OF FAILURE TO MEET ASSUMPTIONS IN ANALYSIS OF VARIANCE

The emphasis in the previous sections of this chapter has been on the assumptions necessary for the mathematical justification of hypothesis-testing procedures using the F distribution. What are the consequences of failure to meet these assumptions? Cochran and Cox (1957, 91) stated that failure to meet the assumptions affects both the significance level of a test and the sensitivity of a test. For example, a test performed at the .05 level may actually be made at the .04 or .07 level. Also, a loss in sensitivity results when the assumptions are not fulfilled because it is often possible to construct a more powerful test than that using the F ratio if the correct model can be specified. Fortunately, the F distribution is very robust with respect to violation of many of the assumptions associated with its mathematical derivation. The effects of failure to meet certain assumptions associated with the F distribution and the mathematical model for a design are discussed in the following paragraphs. Cochran (1947) has pointed out that it is impossible to be certain that all required assumptions are exactly satisfied by a set of data. Thus analysis of variance must be regarded as approximate rather than exact. However, it is generally possible, by a careful examination of the data, to detect cases in which a standard analysis will lead to gross errors in interpreting the outcome of an experiment.

ASSUMPTION OF NORMALLY DISTRIBUTED POPULATION

One of the requirements in order for an F ratio to be distributed as the F distribution is that the numerator and denominator of the ratio are independent. If scores are randomly sampled from a normal population, this requirement is satisfied.

An assumption of both the fixed-effects and random-effects models is that the errors ε_{ij} are normally distributed for each treatment population. Because the only source of variation within a treatment population are the errors, the assumption of normally distributed ε_{ij}'s is equivalent to the assumption of normally distributed scores.

A population of scores can depart from the normal distribution in terms of either skewness or kurtosis, or in both skewness and kurtosis.

Studies by Pearson (1931) and Norton, as cited by Lindquist (1953), indicate that the F distribution is relatively unaffected by lack of symmetry of treatment populations. It is also relatively unaffected by kurtosis except in extreme cases of very leptokurtic or platykurtic populations. For the fixed-effects model, an experimenter need not be concerned if the k populations exhibit a moderate departure from the normal distribution provided that the k populations are homogeneous in form, for example, all treatment populations positively skewed and slightly leptokurtic. In general, unless the departure from normality is so extreme that it can be readily detected by visual inspection of the data, the departure will have little effect on the probability associated with the test of significance. It may be possible to transform nonnormally distributed scores so as to achieve normality, under conditions described in Section 2.7.

ASSUMPTION OF HOMOGENEITY OF POPULATION-ERROR VARIANCES

The F distribution is robust with respect to violation of the assumption of homogeneity of population-error variances provided that the number of observations in the samples is equal (Cochran, 1947; Norton as cited by Lindquist, 1953). However, for samples of unequal size, violation of the homogeneity assumption can have a marked effect on the test of significance. According to Box (1953, 1954a), the nature of the bias for this latter case may be positive or negative.

Several statistics are available for testing the homogeneity assumption that

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma_\varepsilon^2.$$

The alternative to the above null hypothesis is

$$H_1: \text{some } \sigma_j^2\text{'s are unequal.}$$

A test statistic proposed by Bartlett (1937) is

$$B = \frac{2.30259}{C}\left[v \log_{10} \mathrm{MS}_{\text{error}} - \sum_{j=1}^{k} (v_j \log_{10} \hat{\sigma}_j^2)\right],$$

where

$$C = 1 + \frac{\sum_{j=1}^{k} \frac{1}{v_j} - \frac{1}{v}}{3(k-1)},$$

v_j = degrees of freedom for $\hat{\sigma}_j^2$, v = degrees of freedom for $\mathrm{MS}_{\text{error}}$ equal to $\Sigma_{j=1}^{k} v_j$, $\hat{\sigma}_j^2$ = unbiased estimate of population variance for the jth population given by

$$\hat{\sigma}_j^2 = \left[\sum_1^n X^2 - \frac{\left(\sum_1^n X\right)^2}{n}\right]/(n-1),$$

$MS_{error} = \Sigma_{j=1}^k \hat{\sigma}_j^2/v$, and k = number of variances. For values of $v_j \geq 5$, B is approximately distributed as the χ^2 distribution, with $k-1$ degrees of freedom. If $v_j < 5$, tables prepared by Merrington and Thompson (1946) may be used.

Two other tests are computationally simpler than Bartlett's test and provide an adequate test of the assumption of homogeneity of variance. The simpler of the two tests, which was proposed by Hartley (1940, 1950), uses the statistic F_{max},

$$F_{max} = \frac{\text{largest of } k \text{ variances}}{\text{smallest of } k \text{ variances}} = \frac{\hat{\sigma}_{j\,\text{largest}}^2}{\hat{\sigma}_{j\,\text{smallest}}^2},$$

with degrees of freedom equal to k and $n-1$, where k is the number of variances and n is the number of observations within each treatment level. The distribution of F_{max} is given in Table D.10. The hypothesis of homogeneity of variance is rejected if F_{max} is greater than the tabled value for $F_{max,\alpha}$. If the n's for the treatment levels differ only slightly, the largest of the n's can be used for purposes of determining the degrees of freedom for this test. This procedure leads to a slight positive bias in the test, that is, in rejecting the hypothesis of homogeneity more frequently than it should be rejected.

The other relatively simple test of homogeneity of variance is that proposed by Cochran (1941). This test statistic is given by

$$C = \frac{\hat{\sigma}_{j\,\text{largest}}^2}{\sum_{j=1}^k \hat{\sigma}_j^2},$$

where $\hat{\sigma}_{j\,\text{largest}}^2$ is the largest of the k treatment variances and $\Sigma_{j=1}^k \hat{\sigma}_j^2$ is the sum of all of the variances. The degrees of freedom for this test are equal to k and $n-1$ as defined for the F_{max} test. The sampling distribution of C is given in Table D.11.

Since the F distribution is so robust with respect to violation of the assumption of homogeneity of error variance, it is not customary to test this assumption routinely. Both the Hartley and the Cochran tests have adequate sensitivity for testing the assumption in situations where heterogeneity is suspected. If variances are heterogeneous, a transformation of scores as described in Section 2.7 may produce homogeneity.

It should be noted that all three tests described here are sensitive to departures from normality as well as heterogeneity of variances (Box and Anderson, 1955). For a description of a test that is relatively insensitive to departures from normality, see Odeh and Olds (1959).

ASSUMPTION OF ADDITIVITY OF EFFECTS

A basic assumption of the experimental designs described in this book is that a score is the *sum* of the effects in the linear model. If the assumption of additivity of effects is not tenable, it may be possible to achieve additivity by a suitable transformation of the scores.

References that provide additional discussion of the assumptions in analysis of variance may be found in the papers by Eisenhart (1947) and Cochran (1947).

2.7 TRANSFORMATIONS

A transformation is any systematic alteration in a set of scores whereby certain characteristics of the set are changed and other characteristics remain unchanged. Three major reasons for using transformations in analysis of variance are

1. To achieve homogeneity of error variance.
2. To achieve normality of treatment-level distributions (or within-cell distributions).
3. To obtain additivity of treatment effects.

Because the F distribution is relatively unaffected by lack of normality and heterogeneity of variance, the first two reasons for performing a transformation are less compelling than the third. Obtaining additivity of effects is particularly important in designs such as a randomized block design in which a *residual* mean square (abbreviated MS_{res}) is used as an estimate of experimental error. For example, if treatment levels and blocks are not additive, the expected value of the residual mean square is

$$E(MS_{res}) = \sigma_\varepsilon^2 + \sigma_{\beta\pi}^2$$

instead of

$$E(MS_{res}) = \sigma_\varepsilon^2,$$

where $\sigma_{\beta\pi}^2$ refers to the interaction of treatment levels and blocks. Interaction in this context is said to be present when the dependent variable that is measured under the k treatment levels behaves differently for different blocks of subjects. The expected value of the treatment mean square for a fixed-effects model is

$$E(MS_B) = \sigma_\varepsilon^2 + n\sigma_\beta^2.$$

If the null hypothesis is true, then, according to Section 2.5, the numerator and denominator of the ratio

$$F = \frac{MS_B}{MS_{res}}$$

should provide independent estimates of the same population error variance, σ_ε^2. It is apparent from an examination of the expected values of the two mean squares that this can occur only if $\sigma_{\beta\pi}^2 = 0$. We shall return to this point in Chapter 5.

Fortunately, a transformation that accomplishes any one of the objectives listed above will usually accomplish the other two objectives. In general, a transformation can be used whenever there is a relationship between the means and variances of the treatment levels and whenever the form of the treatment level distributions is homogeneous. It is not always possible to find an appropriate transformation for a set of data. For example, if any of the following conditions are present, no transformation exists that will make the data more suitable for analysis of variance: (1) means of treatment levels are approximately equal but variances are heterogeneous, (2) means of treatment levels vary independently of variances, or (3) variances are homogeneous but treatment level distributions are heterogeneous in form. If no transformation is appropriate, and if the departures from normality and homogeneity are gross, an experimenter may be able to use one of the nonparametric statistics for k treatment levels described in Chapter 13. Although these statistics require less stringent assumptions than analysis of variance, they are less powerful and provide less information concerning the outcome of an experiment. It should also be noted that the nonparametric procedures described in Chapter 13 provide a test of the hypothesis that $k \geq 2$ population distributions of *unspecified* form are exactly alike. In order to test hypotheses concerning population means, the homogeneity assumptions of analysis of variance must be tenable. This point is discussed in Section 13.1. Another alternative that may be available to the experimenter is to select a different criterion measure. The choice of a dependent variable in the behavioral sciences is often arbitrary; a different choice may fulfill the requirements of additivity, normality, and homogeneity.

A number of procedures exist for determining which transformation is appropriate for a set of data. Several methods are described by Olds, Mattson, and Odeh (1956) and by Tukey (1949b). One procedure is to follow general rules concerning situations in which a given transformation is often successful. This approach will be emphasized in presenting each of the types of transformations. Alternative procedures for selecting a transformation will be described later.

SQUARE-ROOT TRANSFORMATION

For certain types of data, treatment level means and variances tend to be proportional, as in a Poisson distribution, where $\mu = \sigma^2$. This kind of distribution often results when the dependent variable is a frequency

count of events having a small probability of occurrence, for example, number of errors at each choice point in a relatively simple multiple T maze. The data can often be normalized for this type of situation by taking the square root of each of the scores. A transformed score X' is given by

$$X' = \sqrt{X}.$$

If any X is less than 10, a more appropriate transformation is given either by

$$X' = \sqrt{X + .5} \quad \text{or} \quad X' = \sqrt{X} + \sqrt{X + 1}.$$

The latter transformation has been recommended by Freeman and Tukey (1950). Tables of $\sqrt{X} + \sqrt{X + 1}$ are reproduced in Mosteller and Bush (1954). The effects of performing a square-root transformation are shown for the data in Table 2.7-1. An examination of the means and variances of the transformed scores shows that they are no longer proportional; additionally, the variances are more homogeneous. These transformed scores are more suitable than the original scores for an analysis of variance.

TABLE 2.7-1 Original and Transformed Scores

Original Scores			*Transformed Scores* $X' = \sqrt{X + .5}$		
b_1	b_2	b_3	b_1	b_2	b_3
3	6	12	1.87	2.55	3.54
0	4	6	.71	2.12	2.55
4	2	6	2.12	1.58	2.55
2	4	10	1.58	2.12	3.24
2	7	6	1.58	2.74	2.55
$\bar{X} = 2.2$	4.6	8.0	1.57	2.22	2.89
$S^2 = 2.2$	3.4	8.0	.28	.20	.22

LOGARITHMIC TRANSFORMATION

If treatment means and standard deviations tend to be proportional, a logarithmic transformation may be appropriate. A transformed score X' is given by

$$X' = \log_{10} X \quad \text{or} \quad X' = \log_{10}(X + 1).$$

The latter formula is used when some scores are zero or very small. Logarithmic transformations have been found to be useful when the dependent variable is some measure of reaction time and the data are positively skewed.

RECIPROCAL TRANSFORMATION

If the square of treatment means and standard deviations are proportional, a reciprocal transformation may be appropriate. A transformed score X' is given by

$$X' = \frac{1}{X} \quad \text{or} \quad X' = \frac{1}{X + 1}.$$

The latter formula should be used if any scores are equal to zero. A reciprocal transformation may be useful when the dependent variable is reaction time.

ANGULAR OR INVERSE SINE TRANSFORMATION

The angular transformation is given by

$$X' = 2 \arcsin \sqrt{X},$$

where X is expressed as a proportion. It is not necessary to solve for X' in the above formula; a table of values of X from .001 to .999 is given in Table D.13. The transformed values in Table D.13 are in radians. Bartlett (1947) suggests that $\frac{1}{2n}$ or $\frac{1}{4n}$ be substituted for $X =$ zero and $1 - \frac{1}{2n}$ or $1 - \frac{1}{4n}$ be substituted for $X = 1$, where n is the number of observations on which each proportion is based. An angular transformation may be useful when means and variances are proportional and the distribution has a binomial form. This condition may occur when the number of trials is fixed and X is the probability of a correct response that varies from one treatment level to another.

SELECTING A TRANSFORMATION

We have already described situations where particular transformations have been found to be successful. An alternative approach to selecting a transformation uses the fact that means and variances are unrelated for normally distributed treatment populations. The correct transformation to use for a set of data is the one that removes the relationship between the sample means and variances. This can be determined by graphing the means and variances on the x and y axes respectively, for each transformation and selecting the one that appears to remove the dependency relationship best. The correctness of the selected transformation can be verified by inspecting the transformed treatment distributions for normality and homogeneity of variances.

An additional procedure for selecting a transformation is to apply each of the transformations to the largest and smallest score in the treatment levels. The range within each treatment level is then determined and

the ratio of the largest to the smallest range is computed. The transformation that produces the smallest ratio is selected as the most appropriate one. This procedure is illustrated in Table 2.7-2 for the data in Table 2.7-1. On the basis of this procedure, a square-root formation would be selected for these data.

Once an appropriate transformation is selected and the data analyzed on the new scale, all inferences regarding treatment effects must be made with respect to the new scale. In most behavioral research situations, inferences based on log X's or $\sqrt{X}$'s, for example, are just as meaningful as inferences based on untransformed scores.

If additivity of treatment effects is the principal concern of an experimenter, the appropriateness of a particular transformation can be determined by a test of nonadditivity that is described in Section 5.3. This test provides a means of determining if treatment effects are additive for the untransformed scores and for any transformations that may be tried. A mathematically sophisticated exposition of general issues involved in the use of transformations is given by Box and Cox (1964).

TABLE 2.7-2 Transformations Applied to Largest and Smallest Scores in Table 2.7-1

	Treatment Levels			$\frac{\text{Range}_{\text{largest}}}{\text{Range}_{\text{smallest}}}$
	b_1	b_2	b_3	
Largest score (L)	4	7	12	
Smallest score (S)	0	2	6	
Range =	4	5	6	6/4 = 1.50
$\sqrt{L + .5}$	2.12	2.74	3.54	
$\sqrt{S + .5}$	.71	1.58	2.55	
Range =	1.41	1.16	.99	1.41/.99 = 1.42
log (L + 1)	.6990	.9031	1.1139	
log (S + 1)	.0000	.4771	.8451	
Range =	.6990	.4260	.2688	.6990/.2688 = 2.60
1/(L + 1)	.20	.12	.08	
1/(S + 1)	1.00	.33	.14	
Range =	.80	.21	.06	.80/.06 = 13.33