CHAPTER 1

Research Strategies and the Control of Nuisance Variables

1.1 Introduction

Sir Ronald Fisher, the statistician, eugenicist, evolutionary biologist, geneticist, and father of modern experimental design, observed that experiments are "only experience carefully planned in advance, and designed to form a secure basis of new knowledge" (Fisher, 1935a, p. 8). The design of experiments to investigate scientific or research hypotheses involves a number of interrelated activities:

1. Formulation of statistical hypotheses that are germane to the scientific hypothesis. A **statistical hypothesis** is a statement about (a) one or more parameters of a population or (b) the functional form of a population. Statistical hypotheses are rarely identical to scientific hypotheses; they are testable formulations of scientific hypotheses.
2. Determination of the experimental conditions (independent variable) to be used, the measurement (dependent variable) to be recorded, and the extraneous conditions (nuisance variables) that must be controlled.
3. Specification of the number of subjects (experimental units) required and the population from which they will be sampled.[1]
4. Specification of the procedure for assigning the subjects to the experimental conditions.
5. Determination of the statistical analysis that will be performed.

[1] An **experimental unit** is that entity that is assigned to an experimental condition independently of other entities. An experimental unit may contain several observational units. For example, in an educational experiment, the experimental unit is often the classroom, but the individual students are the observational units. Administering an educational intervention to a classroom can result in nonindependence of the observational units. For a discussion of this problem, see Levin (1992).

In short, an **experimental design** identifies the independent, dependent, and nuisance variables and indicates the way in which the randomization and statistical aspects of an experiment are to be carried out.

Subject Matter and General Organization of This Book

Experimental design, the subject of this book, refers to a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan. Selecting an appropriate plan and performing the correct statistical analysis are important facets of scientific research. However, the most important facet—identifying relevant research questions—is outside the scope of this book. The reader should remember that a carefully conceived and executed design is of no value if the scientific hypothesis that led to the experiment is without merit. Careful planning should always precede the data collection phase of an experiment. Data collection is usually the most costly and time-consuming aspect of an experiment. Advanced planning helps to ensure that the data can be used to maximum advantage. No amount of statistical wizardry can salvage a badly designed experiment.

Chapters 1 to 3 provide an overview of important design concepts and analysis tools that are used throughout the remainder of the book. Chapter 3 describes a procedure developed by Ronald A. Fisher called the **analysis of variance.** The procedure is used to decompose the total variation displayed by a set of observations into two or more identifiable sources of variation. Analysis of variance enables researchers to interpret the variability in designed experiments. Fisher showed that by comparing the variability among subjects treated differently to the variability among subjects treated alike, researchers can make informed choices between competing hypotheses in science and technology. A detailed examination of each analysis of variance design begins in Chapter 4. This examination includes a description of the design, conditions under which the design is appropriate, assumptions associated with the design, a computational example, and advantages and disadvantages of the design.

Two kinds of computational algorithms are provided for the designs. The first, referred to as the *classical sum-of-squares approach,* uses scalar algebra and is suitable for calculators. The second, called the *cell means model approach,* uses matrix algebra and is more suitable for computers.[2] In Chapters 7 and 13, I provide a brief description of a third computational algorithm: the *regression model approach.*

1.2 Formulation of Plans for the Collection and Analysis of Data

Acceptable Research Hypotheses

Some questions currently cannot be subjected to scientific investigation. For example, the questions “Can three or more angels dance on the head of a pin?” and “Does life exist in

[2] Readers who are interested only in the traditional approach to the analysis of variance can, without loss of continuity, omit the material on the cell means model.

more than one galaxy in the universe?" cannot be answered because no procedures now exist for observing either angels or life on other galaxies. Scientists confine their research hypotheses to questions that can be answered by procedures that are available or that can be developed. Thus, the question concerning the existence of life on other galaxies currently cannot be investigated, but with continuing advances in space science, it is likely that eventually the question will be answered.

An experiment involves the manipulation of one or more variables by a researcher to determine the effect of this manipulation on another variable. Questions that provide the impetus for experimental research should be reducible to the form, *if A*, *then B*. For example, *if* albino rats are exposed to microwave radiation, *then* their food consumption will decrease. This research hypothesis can be investigated because procedures are available both for manipulating the radiation level and for measuring the food consumption of rats.

Much research departs from this pattern because nature rather than the researcher manipulates the independent variable. It would be unethical, for example, to study the effects of prenatal malnutrition on IQ by deliberately providing pregnant women with inadequate diets. Instead, the question is investigated by locating children whose mothers were malnourished during pregnancy and then comparing their IQs with those of children whose mothers were not malnourished. Research strategies in which the independent variable is not manipulated by the researcher include surveys, case studies, and naturalistic observation. These research strategies pose special problems for researchers who want to make causal inferences, as I discuss in Section 1.3.

Distinction Between Independent and Dependent Variables

In the radiation example cited earlier, the presence or absence of radiation is the independent variable—the variable that is manipulated by the researcher. More generally, an **independent variable** is any suspected causal event that is under investigation. The terms *independent variable* and *treatment* are interchangeable. A **dependent variable** is the measurement that is used to assess the effects, if any, of manipulating the independent variable. In the radiation example, the dependent variable is the amount of food consumed by the rats.

Selection of the Independent Variable

The independent variable in the radiation example is the presence or absence of radiation. The treatment has two levels. If the researcher is interested in the nature of the relationship between the radiation dose and food consumption, three or more levels of radiation must be used. The levels could be 0 microwatts, 20,000 microwatts, 40,000 microwatts, and 60,000 microwatts of radiation. This treatment is an example of a **quantitative independent variable** in which different treatment levels are different amounts of the independent variable.

When the independent variable is quantitative, the levels of the variable are generally chosen so that they are equally spaced. Usually there is little interest in the exact values of the treatment levels used in the experiment. In the radiation example, the research hypothesis also could be investigated using three other levels of radiation—say, 25,000, 50,000, and 75,000 microwatts in addition to the 0 microwatt control level. The treatment levels should cover a sufficiently wide range so that the effects of the independent variable can be detected if such effects exist. In addition, the number and spacing of the levels should be sufficient to define the shape of the function that relates the independent and dependent variables. Selection of the appropriate levels of the independent variable can be based on the results of previous experiments or on theoretical considerations. It may be beneficial to carry out a small pilot experiment to identify the most appropriate treatment levels. A pilot experiment also is useful for determining the number of experimental units required to test the statistical hypothesis.

Under the conditions described in Chapters 3 and 4, the levels of a quantitative independent variable can be selected randomly from a population of treatment levels. If this procedure is followed, a researcher can extrapolate from the results of the experiment to treatment levels that are not included in the experiment. If the treatment levels are not randomly sampled, the results of an experiment apply only to the specific levels included in the experiment.

Often a different type of independent variable is used. For example, if the treatment levels are unmodulated radiation, amplitude-modulated radiation, and pulse-modulated radiation, the treatment is called a **qualitative independent variable.** The different treatment levels represent different *kinds* rather than different *amounts* of the independent variable. The particular levels of a qualitative independent variable used in an experiment are generally of specific interest to a researcher. And the levels chosen are usually dictated by the research hypothesis.

Selection of the 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 the radiation example, other dependent variables that could be measured include the following:

1. Activity level of the rats in an activity cage
2. Body temperature of the rats
3. Emotionality of the rats as evidenced by their amount of defecation and urination
4. Problem-solving ability
5. Weight change of the rats
6. Speed of running in a straight-alley maze
7. Visual discrimination capacity

Several independent variables can be used 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.[3] The selection of the most fruitful variables to

[3] For a discussion of these designs, see R. J. Harris (2001); Lattin, Carroll, and Green (2003); Meyers, Gamst, and Guarono (2006); Stevens (2002); and Todman and Dugard (2007).

measure should be determined by a consideration of the sensitivity, reliability, distribution, and practicality of the possible dependent variables. From previous experience, a researcher may know that one dependent variable is more sensitive than another to the effects of a treatment or that one 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 can minimize the amount of time and effort required to investigate a research hypothesis.

Other factors to consider in selecting a dependent variable are whether the population distributions for the various treatment levels are approximately normal and whether the populations have equal variances. I have more to say about these factors in Chapter 3 when I discuss the assumptions underlying the analysis of variance (**ANOVA**). 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 with equal variances, a researcher should select the variable that is most easily measured.

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 affect the dependent variable. As the name implies, the effects of nuisance variables are of no interest per se. There are many potential sources of nuisance variables. For example, the calibration of equipment may change during the course of an experiment; the presentation of instructions may vary slightly from subject to subject; errors may occur in measuring or recording a subject's response; environmental factors such as room illumination, noise level, and room temperature may not be constant for all subjects; and subjects may experience lapses in attention, concentration, and interest.

In the radiation example, potential nuisance variables include the sex of the rats, differences in the weights 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. If not controlled, nuisance variables can affect the outcome of an experiment. For example, if rats in the radiated groups suffer from some undetected disease, differences among the groups will reflect the effects of the disease in addition to radiation effects—if such effects exist.

The effect of a nuisance variable can take several forms. For example, a nuisance variable can systematically distort results in a particular direction, in which case the effect is called **bias.** Alternatively, a nuisance variable can increase the variability of the phenomenon being measured and thereby increase the error variance. **Error variance** is variability among observations that cannot be attributed to the effects of the independent variable. You also can think of error variance as differences in the performance of subjects who are treated alike. Sometimes a nuisance variable systematically distorts results in a particular direction and increases the error variance—the worst-case scenario.

Nuisance variables are undesired sources of variation and hence are threats to drawing valid inferences from research. Other threats to valid inference making are described in Sections 1.5 and 1.6.

1.3 Research Strategies

Research is performed for the following purposes: (1) to explore, (2) to describe or classify, (3) to establish relationships, or (4) to establish causality. Over the years, researchers have developed a variety of research strategies to accomplish these purposes. These strategies include the experiment, quasi-experiment, survey, case study, and naturalistic observation.

Experiments

An experiment enables a researcher to test a hypothesized relationship between an independent variable and a dependent variable by manipulating the independent variable. Experiments are usually performed in an environment that permits a high degree of control of nuisance variables. Such environments rarely duplicate real-life situations, but an experiment is still a useful way of obtaining knowledge. An **experiment** is characterized by the (1) manipulation by the researcher of one or more independent variables, (2) use of controls such as randomly assigning subjects or experimental units to the experimental conditions, and (3) careful observation or measurement of one or more dependent variables. The first and second characteristics—manipulation of an independent variable and the use of controls such as randomization—distinguish experiments from nonexperimental research strategies. The manipulation of one or more independent variables also is necessary for inferring causality. We infer that A causes Y if the following are true: A precedes Y (temporal precedence of A); whenever A is present, Y occurs (sufficiency of A); and A must be present for Y to occur (necessity of A).[4]

As noted earlier, this book is concerned with two aspects of experiments: the plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan. Because the statistical analysis procedures for experiments also are applicable to other research strategies, I briefly describe some of these strategies next.

Quasi-Experiments

Quasi-experiments are similar to experiments except that the subjects are not randomly assigned to the independent variable. Quasi-experiments are used instead of experiments when random assignment is not possible or when, for practical or ethical reasons, it is necessary to use preexisting or naturally occurring groups such as subjects with a particular illness or subjects who have been sexually abused.[5]

An example of a well-designed quasi-experiment is the Newburgh-Kingston Caries-Fluorine Study (Hilleboe, 1956). This study was designed to determine the effect of adding fluoride to a community water supply. The cities studied, Newburgh and Kingston, New York, are located on the Hudson River about 35 miles apart. Beginning on May 2, 1945, sodium fluoride was added to the drinking water of Newburgh to bring the fluoride content from about 0.1 part per million to about 1.2 parts per million. The fluoride concentration

[4] Causality is a complex concept. For other definitions and views of causality, see Pearl (2000); Shadish (2010); Shadish, Cook, and Campbell (2002); Sobel (2008); and West and Thoemmes (2010).

[5] For an excellent treatment of quasi-experimental designs, see Shadish et al. (2002).

of Kingston's water remained at about 0.1 part per million. In the year prior to adding fluoride to Newburgh's water supply, baseline data on the prevalence of tooth decay were obtained for school children aged 6 to 12. Baseline pediatric examinations also were performed on smaller samples. The baseline data in the two communities were similar for both tooth decay and general health. The effect of adding fluoride to Newburgh's water supply was evaluated 10 years later by examining more than 2000 children aged 6 to 16 in the two communities. For the 6- to 9-year-olds, the reduction in tooth decay in Newburgh relative to the rate in Kingston was 57%. For older children in Newburgh who had not used fluoridated water all their lives, the reduction was 41%. The tooth decay rate in Newburgh also was similar to that in Aurora, Illinois. Aurora has a naturally occurring fluoride level of about 1.2 parts per million—the same as that in Newburgh—and is known for its low level of tooth decay. The data from this quasi-experiment provide strong support for the efficacy of fluoridated water in preventing tooth decay.

The interpretation of the results of the Newburgh-Kingston study is relatively straightforward. The interpretation of the results of most quasi-experiments is often less straightforward because it is difficult to rule out all variables other than the independent variable as explanations for an observed difference. Researchers in the Newburgh-Kingston study attempted to rule out as many nuisance variables as possible. They chose two communities of comparable size on the Hudson River. And the communities had similar naturally occurring levels of fluoride in their water supplies. Because the communities are only 35 miles apart, they have similar climates and weather conditions. The potential variable of differences in the general health of children in the two communities was ruled out by a pediatric examination. Also, the tooth decay rate obtained with artificially fluoridated water in Newburgh was found to be similar to the rate in Aurora, which has about the same naturally occurring fluoride level.

There is always the possibility that some variable other than the higher fluoride level was responsible for the observed difference in tooth decay between Newburgh and Kingston. However, every effort, short of random assignment, was made to eliminate other variables as explanations for the observed difference. Random assignment is the best safeguard against undetected nuisance variables. As a general principle, the difficulty of unambiguously interpreting the outcome of research varies inversely with the degree of control that a researcher is able to exercise over randomization.

Surveys

Surveys rely on the technique of self-report to obtain information about such variables as people's attitudes, opinions, behaviors, and demographic characteristics. The data are collected by means of an interview or a questionnaire. Although surveys cannot establish causality, they can explore, describe, classify, and establish relationships among variables.

A survey enables a researcher to collect a considerable amount of information about a large number of people. If the survey respondents are representative of a population of interest, the results of the survey can be generalized to the population. Unfortunately, survey respondents are often not representative. For example, many people refuse to give phone interviews or respond to Internet questionnaires, and the return rate for questionnaires received in the mail is typically between 10% and 45%. In all likelihood, people who do not cooperate differ in significant ways from those who do. There are other problems with surveys. Some people tend to give socially acceptable answers or answers that they think the interviewer wants to hear. Also, people may have incomplete or inaccurate memories for past events. Despite these problems, surveys can be efficient, useful sources of information: witness the success of the U.S. census, Gallup Poll, Harris Poll, and Roper Poll.

Case Studies

In a **case study,** a researcher observes selected aspects of a subject's behavior over a period of time. The subject is usually a person, but it may be a setting such as a business, school, or neighborhood. Often the subject possesses an unusual or noteworthy condition. For example, Luria (1968) studied the Russian mnemonist Shereshevskii, who used mnemonic tricks and devices to remember phenomenal amounts of material. Significant discoveries also may result from studying less remarkable subjects. Jean Piaget's theory of intellectual development, for example, evolved from his intensive observation of his own three children. He presented tasks in a nonstandard manner to one child at a time in informal settings and observed the child's verbal and motor responses. Piaget did not attempt to systematically manipulate preselected independent variables, nor did he focus on just one or two dependent variables. Instead, his approach was quite flexible, which allowed him to alter his procedures and follow up on new hypotheses. His flexible case study approach uncovered knowledge about children's cognitive development that might not have been discovered by a more rigid experimental approach.

Case studies can lead to interesting insights that merit further investigation. However, case studies are particularly susceptible to the effects of nuisance variables. Furthermore, questions arise about the degree to which the findings generalize to other populations.

Naturalistic Observation

Naturalistic observation involves observing individuals or events in their natural setting without using manipulative interventions or measuring techniques that might intrude on the setting. Naturalistic observation is a passive form of research in the sense that the individual being observed determines the events that are available to be recorded. The researcher is an unobtrusive recorder of the ongoing events. Because a researcher can focus on only a finite number of events, decisions must be made concerning the events that will be observed. As in a case study, the researcher has the freedom to shift his or her focus to those events that seem most interesting. The data from naturalistic observations may be difficult to analyze, as when the researcher records a running description of a behavior, or easy to analyze, as when a frequency count of a behavior is made.

Naturalistic observation is one of the oldest methods for studying individuals and events. In some sciences, most notably astronomy, the strategy has led to extremely accurate predictions. Classic examples of naturalistic observation are Charles Darwin's voyages on the HMS *Beagle* as he compiled the data that led to the theory of evolution and Jane Goodall's (1971, 1986) study of chimpanzees in their natural habitat in Tanzania, which gave us a new appreciation for this highly social animal.

As a research strategy, naturalistic observation has two advantages over more controlled strategies such as the experiment. First, findings from naturalistic observations generalize readily to other real-life situations. Second, the strategy avoids the reactive arrangements problem that is described in Section 1.5. This problem is avoided because subjects are unaware that their behavior is being studied; hence, they do not react in an unnatural way as they might if they were aware that they were being studied. Unfortunately, there are some serious limitations associated with naturalistic observation. Although the strategy is useful for describing what happened, it does not yield much information about why something happened. To find out why something happened, it is necessary to tamper with the natural course of events. Also, the strategy is an inefficient way to answer "What if?" questions because the event of interest may occur infrequently or not at all in a natural setting.

In this section, I described five widely used research strategies. The strategies are presented in order of decreasing control of the independent and dependent variables. Research always involves a series of trade-offs—a theme I return to time and again. As our control of the independent and dependent variables decreases, our ability to unambiguously interpret the outcome of the research decreases, but our ability to generalize the results to the real world increases.

1.4 Other Research Strategies

The classification scheme for research strategies that I have described is widely used, but it is not exhaustive. There are numerous other ways of classifying research strategies. Each discipline tends to develop its own nomenclature and categories. This section describes some other ways of categorizing research strategies.

Ex Post Facto Studies

The term **ex post facto study** (after-the-fact study) refers to any nonexperimental research strategy in which subjects are singled out because they have already been exposed to a particular condition or because they exhibit a particular characteristic. In such studies, the researcher does not manipulate the independent variable or assign the experimental conditions to the subjects. The retrospective cohort study and the case-control study, described in the following section, are examples of ex post facto studies.

Retrospective and Prospective Studies

Retrospective and **prospective studies** are nonexperimental research strategies in which the independent and dependent variables occur before or after, respectively, the beginning of the study. Retrospective studies use historical records to look backward in time; prospective studies look forward in time. A retrospective study is particularly useful for studying the relationship between variables that occur infrequently or variables whose occurrence is difficult to predict. For example, much of our knowledge about the health effects of ionizing radiation came from studying persons exposed to the World War II bombings of Hiroshima and Nagasaki. A retrospective study also is useful when there is a long time interval between a presumed cause and effect. For example, a decade or more can pass between exposure to a carcinogen and the clinical detection of cancer.

There are two types of retrospective studies: retrospective cohort studies and case-control studies. In a **retrospective cohort study,** also called a **historical cohort study,** records are used to identify two groups of subjects: those who have and those who have not been exposed to the independent variable. Once the exposed and nonexposed groups have been identified, they are compared in terms of the frequency of occurrence of the dependent variable. Consider, for example, McMichael, Spirtas, and Kupper's (1974) study of workers in the rubber industry. Employment records were used to identify 6678 workers who were alive on January 1, 1964. The mortality experience of these workers over the following 9-year period was compared with the mortality experience of persons in the same age and gender categories in the U.S. population. The researchers found that the rubber workers had much higher death rates from cancers of the stomach, prostate, and hematopoietic tissues.

In a **case-control study,** also called a **case-referent study,** records are used to identify two groups of subjects: those who exhibit evidence of the dependent variable, called *cases,* and those who do not, called *controls.* The cases and controls are then compared in terms of their previous exposure to the independent variable. Consider, for example, the study by Clarke, Morgan, and Newman (1982), who investigated the relationship between cigarette smoking and cancer of the cervix. One hundred eighty-one women with cervical cancer (cases) and 905 women without cervical cancer (controls) were interviewed to determine their smoking histories. The researchers found that a much larger proportion of the cases than the controls had smoked cigarettes.

Neither the retrospective cohort study nor the case-control study can establish a causal relationship. However, the research strategies can suggest interesting relationships that warrant experimental investigation. In the retrospective cohort study, more than one dependent variable can be investigated, but only one independent variable can be studied at a time. In the case-control study, multiple independent variables can be investigated, but only one dependent variable can be studied at a time. Despite these and other limitations, both research strategies have been particularly useful in the health sciences.

As noted earlier, a **prospective study,** also called a **follow-up study, longitudinal study,** or **cohort study,** is a nonexperimental research strategy in which the independent and dependent variables are observed after the onset of the investigation. Subjects are classified as exposed or nonexposed based on whether they have been exposed to a naturally occurring independent variable. The exposed and nonexposed groups are then followed for a period of time, and the incidence of the dependent variable is recorded. A classic example is the Framingham Study (T. Gordon & Kannel, 1970), which attempted to identify factors related to the dependent variable of cardiovascular disease. In the study, more than 5000 persons living in Framingham, Massachusetts, who did not have clinical evidence of atherosclerotic heart disease were examined at 2-year intervals for more than 30 years. The study identified several factors, including hypertension, elevated serum cholesterol, and cigarette smoking, that were related to cardiovascular disease.

Prospective studies have advantages over retrospective studies. First, the purported cause (independent variable) clearly precedes the effect (dependent variable); second, the amount and quality of information are not limited by the availability of historical records or the recollections of subjects; and third, measures of the incidence of the dependent variable can be computed. But prospective studies have some serious limitations, too. If the dependent variable is a rare event, a prohibitively large sample may be required to find a sufficient number of subjects who develop the rare event. Also, the investigation of a chronic process using a prospective study may require years to complete. Unfortunately, lengthy studies often suffer from logistic problems such as keeping in touch with the subjects and turnover of the research staff. The distinguishing features of retrospective and prospective studies are summarized in Table 1.4-1.

Table 1.4-1 ■ Distinguishing Features of Retrospective and Prospective Studies

	Time of Occurrence of Independent and Dependent Variables	
	Prior to Initiation of Study	After Initiation of Study
Subject Classified on Basis of Independent Variable	Retrospective cohort study (historical cohort study)	Prospective study (follow-up study, longitudinal study, cohort study)
Subject Classified on Basis of Dependent Variable	Case-control study (case-referent study)	

Longitudinal and Cross-Sectional Studies

The term **longitudinal study** refers to any research strategy in which the same individuals are observed at two or more times. Usually the time interval between observations is fairly long. For example, in the Framingham Study mentioned earlier, subjects were examined at 2-year intervals for more than 30 years in an attempt to identify factors related to cardiovascular disease.

A longitudinal study involves studying the same individuals over time. Identifying changes in individuals over time is not difficult, but identifying the cause of the changes can be a problem because it is difficult to control all nuisance variables over an extended period of time. As a result, a researcher is often faced with competing explanations for the observed changes. The longer the study, the more numerous the competing explanations. There are other problems with longitudinal studies. Over the course of a long study, subjects move, die, or decide to drop out of the study. Often the attrition rates for the groups being followed are different, which introduces another source of bias. Also, longitudinal studies tend to be more expensive and require a longer commitment of a researcher's time than cross-sectional studies, which are described next.

A **cross-sectional study** is any research strategy in which two or more cohorts are observed at the same time. As used here, a **cohort** denotes a person or group of people who have experienced a significant life event such as a birth, marriage, or illness during a given time interval—say, a calendar year or a decade. The Newburgh-Kingston Caries-Fluorine Study mentioned earlier involved several cohort comparisons: children living in Newburgh versus those living in Kingston and 6- to 9-year-olds versus older children.

Cross-sectional studies tend to be less expensive than longitudinal studies, and they provide more immediate results. Also, attrition of subjects is less likely to be a problem in cross-sectional studies. However, as mentioned earlier in discussing the Newburgh-Kingston Caries-Fluorine Study, there is always the possibility that even in a well-designed cross-sectional study, variables other than those under investigation are responsible for the observed difference in the dependent variable. As noted earlier, random assignment is the best safeguard against undetected nuisance variables.

Longitudinal-Overlapping and Time-Lag Studies

The two research strategies described in this section combine features of longitudinal and cross-sectional studies. A **longitudinal-overlapping study,** also called a **sequential study,** can be used to compress the time required to perform a longitudinal study. Suppose that a researcher wants to observe children at 2-year intervals from ages 5 through 13. A longitudinal study would require 8 years. This time can be compressed to 4 years by observing a group of 5-year-olds and a second group of 9-year-olds. The 5-year-old children are observed at ages 5, 7, and 9; the 9-year-old children are observed at ages 9, 11, and 13. Note the overlapping age: Both groups include 9-year-olds. The layout for this study is diagrammed in Figure 1.4-1, where O_1, O_2, and O_3 denote the first, second, and third observations of the children in each group, respectively. In addition to cutting the length of the study in half, this research strategy enables a researcher to compare 5- and 9-year-olds after completing the first set of observations. This comparison would be delayed for 4 years in a longitudinal study. The earlier discussion of the advantages and disadvantages of cross-sectional studies is applicable to a longitudinal-overlapping study.

	Subject's Age	1st Obs.	Subject's Age	2nd Obs.	Subject's Age	3rd Obs.
Group$_1$	5	O_1	7	O_2	9	O_3
Group$_2$	9	O_1	11	O_2	13	O_3

Figure 1.4-1 ■ Simplified layout for a longitudinal-overlapping study, where O_1, O_2, and O_3 denote, respectively, the first, second, and third observations (Obs.) on the children in Group$_1$ and Group$_2$.

In a **time-lag study,** observations are made at two or more times but different subjects (cohorts) are measured at each time. Consider, for example, the annual administration of the Scholastic Aptitude Test to high school juniors and seniors. For a number of years, the test score means for seniors have been declining. This example of a time-lag study shares some of the characteristics of longitudinal and cross-sectional studies. The test scores are obtained at two or more times, as in a longitudinal study, but as in a cross-sectional study, different senior classes are observed at each testing period. The layout for this study is diagrammed in Figure 1.4-2, where the groups represent five senior classes that are each observed once and O_i denotes one of the $i = 1, \ldots, 5$ observations.

Time-Series and Single-Case Studies

A **time-series study** involves making multiple observations of one or more subjects or cohorts before and after the introduction of an independent variable. The independent variable may or may not be controlled by the researcher. Consider a study to determine the effect of banning the importation of assault rifles in 2005 on the incidence of homicides

	Year	1st Obs.	2nd Obs.	3rd Obs.	4th Obs.	5th Obs.
$\text{Group}_{1(\text{Seniors})}$	2005	O_1				
$\text{Group}_{2(\text{Seniors})}$	2006		O_2			
$\text{Group}_{3(\text{Seniors})}$	2007			O_3		
$\text{Group}_{4(\text{Seniors})}$	2008				O_4	
$\text{Group}_{5(\text{Seniors})}$	2009					O_5

Figure 1.4-2 ▪ Simplified layout for a time-lag study, where $O_1, \ldots, O_5$ denote observations (Obs.) on members of five senior classes denoted by Group_1 through Group_5.

and suicides. One way to evaluate the effect of the ban is to compare the number of homicides and suicides in 2004 with the number in 2005. Suppose that the data in Figure 1.4-3(a) are obtained. Because of the reduction from 2004 to 2005, one might infer that the ban reduced the number of homicides and suicides. However, other nuisance variables such as an unusually cool summer could have been responsible for the reduction. A time-series study would provide stronger evidence for or against the effectiveness of banning the importation of assault rifles. Following this approach, a researcher would record the number of homicides and suicides for several years before and after the ban and note trends in the data. Consider the hypothetical data in Figures 1.4-3(b–d). Figure 1.4-3(b) suggests that the decrease in the number of homicides and suicides from 2004 to 2005 reflected nothing more than random year-to-year variation. Figure 1.4-3(c) suggests that the ban had only a temporary effect. Figure 1.4-3(d) suggests that the ban had no effect because similar reductions were observed during the years prior to and after the ban. These hypothetical examples illustrate the importance of obtaining multiple observations so that change can be viewed within a context.

A **single-case study,** not to be confused with the case studies described in Section 1.3, has many of the characteristics of a time-series study. However, in a single-case study, multiple observations of a single subject are made before and after the introduction of an independent variable, and the researcher controls the independent variable.

Single-case studies were widely used in the behavioral sciences in the late 1880s and early 1900s. Examples include the pioneering work of Ebbinghaus (1850–1909) on forgetting, Wundt's (1832–1920) research on sensory and perceptual processes, and Titchener's (1867–1927) measurement of sensory thresholds. Researchers began to use large samples and random assignment in the 1920s and 1930s, primarily because of the influence of R. A. Fisher (1890–1962). B. F. Skinner's (1904–1990) research on schedules of reinforcement in the 1940s and 1950s rekindled an interest in single-case studies. This research strategy has proven to be particularly useful in assessing the effects of an intervention in clinical psychology research.[6]

The simplest single-case study uses an A-B design. The letter A denotes a baseline phase during which no intervention is in effect; the letter B denotes the intervention phase.

[6] Barlow, Nock, and Hersen (2009); Kazden (1982); and Morgan and Morgan (2009) provide in-depth discussions of single-case studies.

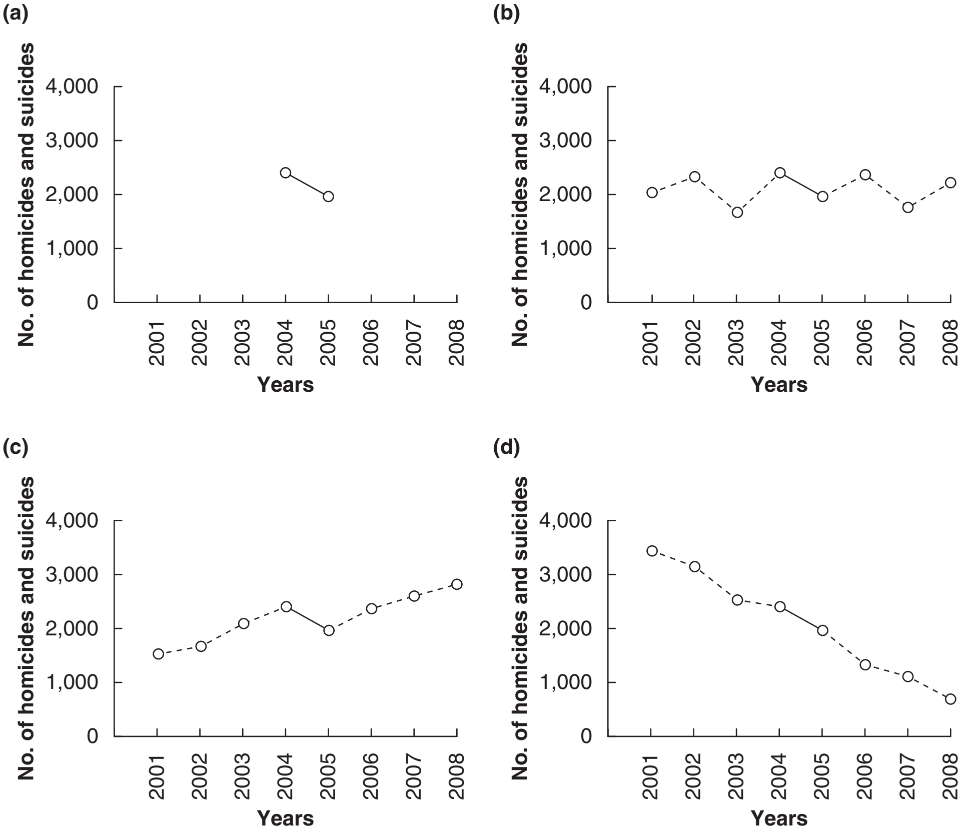


Figure 1.4-3 ▪ Part (a) shows the decrease in the number of homicides and suicides following a ban on the importation of assault rifles in 2004. A time-series study can place the data in perspective. The hypothetical data in part (b) suggest that the decrease in the number of homicides and suicides from 2004 to 2005 reflected random year-to-year variation, part (c) suggests that the ban had a temporary effect, and part (d) suggests that the ban had no effect.

The baseline phase serves three purposes: It provides data about a subject's performance prior to instituting an intervention, it provides a basis for predicting a subject's future performance in the absence of an intervention, and it indicates the normal variability in the subject's performance.

Consider an experiment to reduce the occurrence of thumb sucking of a 6-year-old named Bill. Bill usually sucked his thumb at bedtime while his mother read to him. During the baseline phase that lasted 3 days, Bill's mother read to him while an older sibling recorded the percent of story-reading time during which Bill sucked his thumb. During the treatment phase, when Bill began sucking his thumb, his mother would stop reading and remain quiet until Bill removed his thumb from his mouth. By the end of the seventh treatment day, Bill had stopped sucking his thumb when his mother read to him. The layout for this study is diagrammed in Figure 1.4-4, where O_i denotes one of the $i = 1, \ldots, n$ observations of the dependent variable.

	Baseline (A Phase)	Treatment (B Phase)
Subject	$O_1, O_2, \ldots, O_i$	$O_{i+1}, O_{i+2}, \ldots, O_n$

Figure 1.4-4 ■ Simplified layout for a single-case study, where $O_1, O_2, \ldots, O_i$ denote observations on a subject during the baseline period (A phase) and O_{i+1}, $O_{i+2}, \ldots, O_n$ denote observations during the treatment period (B phase). Any difference between the A and B phases in the mean of the observations or change in the trend of the observations is attributed to the intervention.

In this example, the treatment appears to be related to the cessation of thumb sucking. But there is always the possibility that coincidental changes in a nuisance variable were completely or partly responsible for the cessation of thumb sucking. Statistical regression, which is described in Section 1.5, is a potential nuisance variable in this kind of research because the behavior that is to be altered is one that occurs frequently. Because of statistical regression, there is a tendency for the frequency of behaviors that have a high rate of occurrence to decrease in the absence of any intervention, as well as a tendency for the frequency of behaviors that have a low rate of occurrence to increase. In the thumb-sucking example, a stronger case for the efficacy of the treatment could be made if thumb sucking reappears when the treatment is withdrawn—that is, when Bill's mother continues reading even though Bill sucks his thumb. This modified design with the sequence of events

$$\text{baseline} \rightarrow \text{treatment} \rightarrow \text{baseline}$$

is diagrammed in Figure 1.4-5. Note that there are two opportunities to observe the effects of the treatment: the transition from the baseline to the treatment (A-B) and the transition from the treatment to the baseline (B-A). The presence of two transitions in the A-B-A design decreases the probability that changes in the dependent variable are the result of coincidental changes in a nuisance variable. A problem with this design is that the experiment ends on a baseline phase—a phase during which thumb sucking is expected to reappear. The solution to this problem is to reintroduce the B phase following the second A phase so that the experiment ends with the intervention phase. The design is called an A-B-A-B design and is shown in Figure 1.4-6. This design has the added advantage of providing three transitions: from A to B, from B to A, and from A to B. Hence there are three opportunities to evaluate the efficacy of the treatment. In a single-subject study, the use of one or more reversals in which a treatment is withdrawn to see whether the dependent variable returns to the baseline level can raise ethical questions. For example, if a treatment is successful in stopping an autistic child from repeatedly hitting his or her head against a wall, the withdrawal of the treatment and the subsequent return to head banging could result in physical injury to the child. In this example, the withdrawal of the treatment would be unacceptable.

	Baseline (A Phase)	Treatment (B Phase)	Baseline (A Phase)
Subject	$O_1, O_2, \ldots, O_i$	$O_{i+1}, O_{i+2}, \ldots, O_{i'}$	$O_{i'+1}, O_{i'+2}, \ldots, O_n$

Figure 1.4-5 ▪ Simplified layout for a single-case study, where $O_1, O_2, \ldots, O_n$ denote observations on a subject during a sequence of A-B-A phases.

	Baseline (A Phase)	Treatment (B Phase)	Baseline (A Phase)	Treatment (B Phase)
Subject	$O_1, O_2, \ldots, O_i$	$O_{i+1}, O_{i+2}, \ldots, O_{i'}$	$O_{i'+1}, O_{i'+2}, \ldots, O_{i''}$	$O_{i''+1}, O_{i''+2}, \ldots, O_n$

Figure 1.4-6 ▪ Simplified layout for a single-case study, where $O_1, O_2, \ldots, O_n$ denote observations on a subject during a sequence of A-B-A-B phases.

I have described a variety of research strategies in Section 1.3 and this section. In the next two sections, I briefly examine some threats to drawing valid inferences from research. In Section 1.7, I describe some general approaches to controlling nuisance variables and minimizing threats to valid inference making.

1.5 Threats to Valid Inference Making

Two goals of research are to draw valid conclusions about the effects of an independent variable and to make valid generalizations to populations and settings of interest. Shadish, Cook, and Campbell (2002), drawing on the earlier work of Campbell and Stanley (1966), have identified four categories of threats to these goals:[7]

1. **Statistical conclusion validity** is concerned with threats to valid inference making that result from random error and the ill-advised selection of statistical procedures.
2. **Internal validity** is concerned with correctly concluding that an independent variable is, in fact, responsible for variation in the dependent variable.
3. **Construct validity** of causes and effects is concerned with the possibility that operations that are meant to represent the manipulation of a particular independent variable can be construed in terms of other variables.
4. **External validity** is concerned with the generalizability of research findings to and across populations of subjects and settings.

This book is concerned with three of the threats to valid inference making: threats to statistical conclusion validity, internal validity, and external validity. In the discussion that

[7] The list of categories and threats to valid inference making are taken from Campbell and Stanley (1966) and Shadish et al. (2002). Responsibility for the interpretation of items in their lists is mine.

follows, I focus on the three threats. The reader is encouraged to consult the original sources: Campbell and Stanley (1966) and Shadish et al. (2002). The latter book should be read by all researchers who, for one reason or another, are unable to randomly assign subjects to treatment conditions.

Threats to Statistical Conclusion Validity

1. **Low statistical power.** A researcher may fail to reject a false null hypothesis because the sample size is inadequate, irrelevant sources of variation are not controlled or isolated, or inefficient test statistics are used.

2. **Violated assumptions of statistical tests.** Test statistics require the tenability of certain assumptions. If these assumptions are not met, incorrect inferences may result. This threat is discussed in Section 3.5.

3. **Fishing for significant results and the error rate problem.** With certain test statistics, the probability of drawing one or more erroneous conclusions increases as a function of the number of tests performed. This threat to valid inference making is discussed in detail in Chapter 5.

4. **Reliability of measures.** The use of a dependent variable that has low reliability may inflate the estimate of the error variance and result in not rejecting a false null hypothesis.

5. **Reliability of treatment implementation.** Failure to standardize the administration of treatment levels may inflate the estimate of the error variance and result in not rejecting a false null hypothesis.

6. **Random irrelevancies in the experimental setting.** Variation in the environment (physical, social, etc.) in which a treatment level is administered may inflate the estimate of the error variance and result in not rejecting a false null hypothesis.

7. **Random heterogeneity of respondents.** Idiosyncratic characteristics of the subjects may inflate the estimate of the error variance and result in not rejecting a false null hypothesis.

Threats to Internal Validity

1. **History.** Events other than the administration of a treatment level that occur between the time a subject is assigned to the treatment level and the time the dependent variable is measured may affect the dependent variable.

2. **Maturation.** Processes not related to the administration of a treatment level that occur within subjects simply as a function of the passage of time (growing older, stronger, larger, more experienced, etc.) may affect the dependent variable.

3. **Testing.** Repeated testing of subjects may result in familiarity with the testing situation or acquisition of information that can affect the dependent variable.

4. **Instrumentation.** Changes in the calibration of a measuring instrument, shifts in the criteria used by observers and scorers, or unequal intervals in different ranges of a measuring instrument can affect the measurement of the dependent variable.

5. **Statistical regression.** When the measurement of the dependent variable is not perfectly reliable, there is a tendency for extreme scores to regress or move toward the mean. **Statistical regression** operates to (a) increase the scores of subjects originally found to score low on a test, (b) decrease the scores of subjects originally found to score high on a test, and (c) not affect the scores of subjects at the mean of the test. The amount of statistical regression is inversely related to the reliability of the test.

6. **Selection.** Differences among the dependent-variable means may reflect prior differences among the subjects assigned to the various levels of the independent variable.

7. **Mortality.** The loss of subjects in the various treatment conditions may alter the distribution of subject characteristics across the treatment groups.

8. **Interactions with selection.** Some of the foregoing threats to internal validity may interact with selection to produce effects that are confounded with or indistinguishable from treatment effects. Among these are selection history effects and selection maturation effects. For example, selection maturation effects occur when subjects with different maturation schedules are assigned to different treatment levels.

9. **Ambiguity about the direction of causal influence.** In some types of research—for example, correlational studies—it may be difficult to determine whether *X* is responsible for the change in *Y* or vice versa. This ambiguity is not present when *X* is known to occur before *Y*.

10. **Diffusion or imitation of treatments.** Sometimes the independent variable involves information that is selectively presented to subjects in the various treatment levels. If the subjects in different levels can communicate with one another, differences among the treatment levels may be compromised.

11. **Compensatory rivalry by respondents receiving less desirable treatments.** When subjects in some treatment levels receive goods or services generally believed to be desirable and this becomes known to subjects in treatment levels that do not receive those goods and services, social competition may motivate the subjects in the latter group, the control subjects, to attempt to reverse or reduce the anticipated effects of the desirable treatment levels. Saretsky (1972) named this the "John Henry" effect in honor of the steel driver who, upon learning that his output was being compared with that of a steam drill, worked so hard that he outperformed the drill and died of overexertion.

12. **Resentful demoralization of respondents receiving less desirable treatments.** If subjects learn that the treatment level to which they have been assigned received less desirable goods or services, they may experience feelings of resentment and demoralization. Their response may be to perform at an abnormally low level, thereby increasing the magnitude of the difference between their performance and that of subjects assigned to the desirable treatment level.

Threats to External Validity

1. **Interaction of testing and treatment.** Results obtained under conditions of repeated testing may not generalize to situations that do not involve repeated testing. A pretest, for example, may sensitize subjects to a topic and, by focusing attention on the topic, enhance the effectiveness of a treatment. The opposite effect also can occur. A pretest may diminish subjects' sensitivity to a topic and thereby reduce the effectiveness of a treatment.

2. **Interaction of selection and treatment.** The constellation of factors that affect the availability of subjects to participate in an experiment may restrict the generalizability of results to populations that share the same constellation of factors. For example, if volunteers were used in an experiment, the results may generalize to only volunteer populations.

3. **Interaction of setting and treatment.** The unique characteristics of the setting in which results are obtained may restrict the generalizability of the results to settings that share the same characteristics. Results obtained in a classroom, for example, may not generalize to an assembly line.

4. **Interaction of history and treatment.** Occasionally results are obtained on the same day as a particularly noteworthy event. These results may be different from results that would have been obtained in the absence of the noteworthy event.

5. **Reactive arrangements.** Subjects who are aware that they are being observed may behave differently than subjects who are not aware that they are being observed.

6. **Multiple-treatment interference.** When subjects are exposed to more than one treatment, the results may generalize to only populations that have been exposed to the same combination of treatments.

1.6 Other Threats to Valid Inference Making

Experimenter-Expectancy Effect

Controlling nuisance variables in research with human subjects is particularly challenging. Experiments with human subjects are social situations in which one person behaves under the scrutiny of another. The two people in this social situation have expectations about each other, communicate with each other, and form impressions about each other. The power of the subjects in the situation is always unequal: The researcher requests a behavior and the subject behaves. The researcher's overt request may be accompanied by other more subtle requests and messages. For example, body language, tone of voice, and facial expressions can communicate the researcher's expectations and desires concerning the outcome of an experiment. Such communications can affect a subject's performance. Rosenthal (1963) has reported that researchers tend to obtain from their subjects—whether human or animal—the data they want or expect to obtain. A researcher's expectations and desires also can influence the way he or she records, analyzes, and interprets data. According to Rosenthal (1969, 1978), observational or recording errors are usually small and unintentional. However, when such errors occur, more often than not they are in the direction of supporting the researcher's hypothesis. Sheridan (1976) has reported that researchers are much more likely to recompute and double-check results that conflict with their hypotheses than results that support their hypotheses. The effect of a researcher's expectations and desires on the outcome of an experiment is called the **experimenter-expectancy effect.**

Demand Characteristics

The experimenter-expectancy effect is one source of bias in an experiment; another source is what Orne (1962) has called demand characteristics. **Demand characteristics** refer to any aspect of the experimental environment or procedure that leads a subject to make inferences about the purpose of an experiment and to respond in accordance with—or in some cases, contrary to—the perceived purpose. Subjects are inveterate problem solvers. When they are told to perform a task, the majority will try to figure out what is expected of them and perform accordingly. Demand characteristics can result from rumors about an experiment, what subjects are told when they sign up for an experiment, the laboratory environment, or the communication that occurs during the course of an experiment. Demand characteristics influence a subject's perceptions of what is appropriate or expected.

Subject-Predisposition Effects

As I have discussed, an experimenter's expectations and motives can influence a subject's performance, and subjects often respond in ways that they think are appropriate or expected by the researcher. There is another source of bias in an experiment. Subjects, because of past experience, personality, and so on, come to experiments with a predisposition to respond in a particular way. I describe four kinds of subject-predisposition effects.

Cooperative-subject effect. The first predisposition is that of the cooperative subject whose main concern is to please the researcher and be a "good subject." Cooperative subjects are particularly susceptible to the experimenter-expectancy effect. They try, consciously or unconsciously, to provide data that support the researcher's hypothesis. This subject predisposition is called the **cooperative-subject effect.**

Screw you effect. A second group of subjects tends to be uncooperative and may even try to sabotage the experiment. Masling (1966) has called this predisposition the "**screw you effect.**" It can result from resentment over being required to participate in an experiment, from a bad experience in a previous experiment such as being deceived or made to feel inadequate, or from a dislike for the course or the professor associated with the experiment. Uncooperative subjects may try, consciously or unconsciously, to provide data that do not support the researcher's hypothesis.

Evaluation apprehension. A third group of subjects are apprehensive about being evaluated. Subjects with **evaluation apprehension** (Rosenberg, 1965) aren't interested in the experimenter's hypothesis, much less in sabotaging the experiment. Instead, their primary concern is to gain a positive evaluation from the researcher. The data they provide are colored by a desire to appear intelligent, well adjusted, and so on and to avoid revealing characteristics that they consider undesirable.

Faithful subjects. A fourth group of subjects have been labeled **faithful subjects** (Fillenbaum, 1966). Faithful subjects try to put aside their own hypotheses about the purpose of an experiment and to follow the researcher's instructions to the letter. Often they are motivated by a desire to advance scientific knowledge. The data produced by overly cooperative or uncooperative subjects or by subjects with evaluation apprehension can cause a researcher to draw a wrong conclusion. The data of faithful subjects, however, are not contaminated by such predispositions; faithful subjects simply try to do exactly what they are told to do.

Placebo Effect

The last source of bias that I describe is the placebo effect. A **placebo** is an inert substance or neutral stimulus that is administered to subjects as if it was the actual treatment condition. When subjects begin an experiment, they are not entirely naive. They have ideas, understandings, and perhaps a few misunderstandings about what will happen. If subjects expect that an experimental condition will have a particular effect, they are likely to behave in a manner consistent with their expectation. For example, subjects who believe that a medication will relieve a particular symptom may report feeling better even though they have received a chemically inert substance instead of the medication. Any change in the dependent variable attributable to receiving a placebo is called the **placebo effect.**

In the previous sections, I described a variety of threats to valid inference making: threats to statistical conclusion validity, internal validity, and external validity; the experimenter-expectancy effect; demand characteristics; subject-predisposition effects; and the placebo effect. This list of threats is far from complete. For a fuller discussion of threats to valid inference making, the reader should consult Shadish et al. (2002) and Rosenthal (1979). In the following section, I describe some procedures for controlling nuisance variables and minimizing threats to valid inference making.

1.7 Controlling Nuisance Variables and Minimizing Threats to Valid Inference Making

General Approaches to Control

Four general approaches are used to control nuisance variables. One approach is to hold the nuisance variable constant for all subjects. Examples are using only male rats of the same weight and presenting all instructions to subjects by means of an iPad, computer, or DVD player. Although a researcher may attempt to hold all nuisance variables constant, inevitably some variable will escape 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 are distributed over the entire experiment and thus do not selectively affect just one or a limited number of treatment levels. Random assignment has two other purposes. It permits the computation of an unbiased estimate of **error effects**—those effects not attributable to the manipulation of the independent variable—and it helps to ensure that the error effects are statistically independent. Through random assignment, a researcher creates two or more groups of subjects that at the time of assignment are probabilistically similar on the average. When random assignment is used, a researcher increases the magnitude of random variation among observations to minimize bias, which is the distortion of results in a particular direction. Random variation can be taken into account in evaluating the outcome of an experiment; it is more difficult to account for bias.

A third approach to controlling nuisance variables is to include the variable as one of the factors in the experimental design. This approach is illustrated in Section 2.2.

The 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 13) to statistically remove the effects of a nuisance variable. This use of statistical control is referred to as the *analysis of covariance.*

Some Specific Approaches to Control

In addition to the four general approaches just described, a variety of other procedures are used to control nuisance variables and minimize threats to valid inference making.

Single-blind procedure. In a **single-blind experiment,** subjects are not informed about the nature of their treatment or, when feasible, the purpose of the experiment. A single-blind procedure helps to minimize the effects of demand characteristics. Sometimes the purpose of an experiment cannot be withheld from subjects because of informed consent requirements that are imposed on the researcher. (Informed consent requirements are discussed in Section 1.8.)

Double-blind procedure. In a **double-blind experiment,** neither the subjects nor the researcher are informed about the nature of the treatment that the subjects receive. For example, in a drug study, the dose levels and placebo can be coded so that those administering the drug and those receiving the drug cannot identify the condition that is administered. A double-blind procedure helps to minimize experimenter-expectancy effects and demand characteristics.

Partial-blind procedure. Many treatments are of such a nature that they are easily identified by a researcher. In this case, a **partial-blind procedure** can be used in which the researcher does not know until just before administering the treatment level which level will be administered. In this way, experimenter-expectancy effects are minimized until the administration of the treatment level.

Deception. Deception occurs when subjects are not told the relevant details of an experiment or when they are told that the experiment has one purpose when in fact the purpose is really something else. Deception is used to direct a subject's attention away from the purpose of an experiment so as to minimize the effects of demand characteristics. Deception should never be used without a prior careful analysis of the ethical ramifications. (Ethical issues are discussed in Section 1.8.)

Disguised-experiment technique (unobtrusive experimentation). In the disguised-experiment technique, the subjects are not aware that they are participating in an experiment. The naturalistic-observation research strategy described in Section 1.3 is an example of this approach to minimizing the bias that might result from reactive arrangements and demand characteristics.

Multiple researchers. In some research areas, the characteristics of a researcher such as appearance, personality, inexperience, and so on can affect the results that are obtained. These researcher characteristics can seriously limit the generalizability of results. If several researchers are used, the researchers can be included as one of the variables in the experiment, and the significance of the variable can be evaluated.

Debriefing. It is a common practice to hold a postexperimental meeting with subjects at which time details of the experiment are shared. During this debriefing, subjects can be quizzed concerning their beliefs and expectations about the experiment. Information obtained at this time can be used to determine whether demand characteristics could have affected the results of the experiment.

Experimenter-expectancy control groups. The magnitude of the experimenter-expectancy effect can be determined by using several groups of researchers. One group of researchers is led to expect one experimental outcome, a second group is led to expect the opposite outcome, and a third group is led to believe that the treatment will have no effect on the dependent variable. Unfortunately, this procedure can be costly because it involves using numerous researchers and subjects.

Unrelated-experiment technique. The unrelated-experiment technique is designed to disguise the purpose of an experiment and minimize subject demand characteristics by separating the presentation of the independent variable from the measurement of the dependent variable. This technique requires subjects to participate in two experiments. In the first experiment, the subjects receive the independent variable. Later, the subjects are contacted and asked to participate in a second experiment at which time the dependent variable is measured. The researcher conveys the impression that the second experiment has no relationship to the first experiment.

Quasi-control group. This procedure uses a second control group, called a quasi-control group, to assess the effects of demand characteristics. The **quasi-control group** is exposed to all of the instructions and conditions that are given to the experimental group except that the treatment condition of interest is not administered. This group, unlike a regular control group, does not receive a placebo. Following the presentations of the instructions, the quasi-control subjects are asked to produce the data that they would have produced if they had actually received the treatment condition.

In a double-blind experiment, the quasi-control procedure can be carried one step further: Subjects can be asked to pretend that they have received the treatment condition and to behave accordingly—that is, to be simulators. At the conclusion of the experiment, the researcher is asked to identify the real subjects, control subjects, and simulators. Comparisons among the groups can be useful in detecting experimenter-expectancy effects and demand characteristics.

Yoked control procedure. Researchers would like the experiences of subjects in an experiment to be identical except for the independent variable. Unfortunately, it is difficult to keep the experiences of all subjects identical when a subject's behavior determines aspects of the experimental situation—for example, the number of shocks received in a learning experiment. A **yoked control procedure** allows a researcher to match two subjects on some important aspects of the experience they have in an experiment. In this procedure, two subjects—an active subject and a passive subject—are simultaneously exposed to the same experimental condition, but the behavior of only the active subject affects the outcome. Both members of the pair are subjected to the consequences of the active subject's behavior. For example, yoked active and passive subjects receive a shock each time the active subject makes an incorrect response, thus controlling the variable of number of shocks received.

1.8 Ethical Treatment of Subjects

In recent years, the research community has witnessed a renewed resolve to protect the rights and interests of humans and animals. Codes of ethics for research with human subjects have been adopted by a number of professional societies. Of particular interest are those of the American Educational Research Association (2011), American Evaluation Association (2008), American Psychological Association (2002), American Sociological Association (1999), and American Statistical Association (1999). These codes specify what is required and what is forbidden. In addition, they point out the ideal practices of the profession as well as ethical pitfalls. The 1970s saw the passage of laws to govern the conduct of research with human subjects. One law, which was originally enforced by the U.S. Department of Health, Education, and Welfare (HEW), now the Department of Health and Human Services (HHS), requires that all research funded by HHS involving human subjects be reviewed by an institutional review board (Weinberger, 1974, 1975). As a result, most institutions that conduct research have human subjects committees that screen all research proposals. These committees can disapprove research proposals or require additional safeguards for the welfare of subjects.

In addition to codes of ethics of professional societies, legal statutes, and peer review, perhaps the most important regulatory force within society is the individual researcher's ethical code. Researchers should be familiar with the codes of ethics and statutes relevant to their research areas and incorporate them into their personal codes of ethics.

Space does not permit an extensive examination of ethical issues here. For this the reader should consult the references above and the thorough and balanced treatment by Diener and Crandall (1978). However, I cannot leave the subject without listing some general guidelines.

1. A researcher should be knowledgeable about issues of ethics and values, take these into account in making research decisions, and accept responsibility for decisions and actions that have been taken. The researcher also is responsible for the ethical behavior of collaborators, assistants, and employees who have parallel obligations.

2. Subjects should be informed of aspects of research that might be expected to influence their willingness to participate. Failure to make full disclosure places an added responsibility on the researcher to protect the welfare and dignity of the subject. Subjects should understand that they have the right to decline to participate in an experiment and to withdraw at any time; pressure should not be used to gain cooperation.

3. Research subjects should be protected from physical and mental discomfort, harm, and danger. If risk of such consequences exists, a researcher must inform the subject of this. If harm does befall a subject, the researcher has an obligation to remove or correct the consequences.

4. Special care should be taken to protect the rights and interests of less powerful subjects such as children, minorities, patients, the poor, and prisoners.

5. Research deception should never be used without a prior careful ethical analysis. When the methodological requirements of a study demand concealment or deception, the researcher should take steps to ensure the subject's understanding of the reason for this action and afterward restore the quality of the relationship that existed. Where scientific or other compelling reasons require that this information be withheld, the researcher acquires a special responsibility to ensure that there are no damaging consequences for the subject.

6. Private information about subjects may be collected only with their consent. All such research information is confidential. Publication of research results should be in a form that protects the subject's identity unless the subject agrees otherwise.

7. After data are collected, the researcher must provide the subjects with information regarding the nature of the study and relevant findings.

8. Results of research should be reported accurately and honestly, without omissions that might affect their interpretation.

A number of guides for research with animals have been published. Those engaged in such research should be familiar with the American Psychological Association's (1996) *Guidelines for Ethical Conduct in the Care and Use of Animals.*

1.9 Review Exercises[8]

1. Terms to remember:

   a. statistical hypothesis (1.1)[9]
   b. experimental unit (1.1)
   c. experimental design (1.1)
   d. analysis of variance (1.2)
   e. independent variable (1.2)
   f. dependent variable (1.2)
   g. quantitative independent variable (1.2)
   h. qualitative independent variable (1.2)
   i. ANOVA (1.2)
   j. nuisance variable (1.2)
   k. bias (1.2)
   l. error variance (1.2)
   m. experiment (1.3)
   n. quasi-experiment (1.3)
   o. survey (1.3)
   p. case study (1.3)
   q. naturalistic observation (1.3)
   r. ex post facto study (1.4)
   s. retrospective and prospective studies (1.4)
   t. retrospective cohort study (1.4)
   u. case-control study (1.4)
   v. longitudinal study (1.4)
   w. cross-sectional study (1.4)
   x. cohort (1.4)
   y. longitudinal-overlapping study (1.4)
   z. time-lag study (1.4)
   aa. time-series study (1.4)
   ab. single-case study (1.4)
   ac. statistical conclusion validity (1.5)
   ad. internal validity (1.5)
   ae. construct validity (1.5)
   af. external validity (1.5)
   ag. statistical regression (1.5)
   ah. demand characteristics (1.6)
   ai. cooperative-subject effect (1.6)
   aj. screw you effect (1.6)
   ak. evaluation apprehension effect (1.6)
   al. faithful subject (1.6)
   am. placebo effect (1.6)
   an. error effects (1.7)
   ao. single-blind experiment (1.7)
   ap. double-blind experiment (1.7)
   aq. partial-blind procedure (1.7)
   ar. quasi-control group (1.7)
   as. yoked control procedure (1.7)

*2. [1.1] For each of the following, identify the experimental unit (EU) and the observational unit (OU).

   *a. Fraternities at a large state university were randomly sampled and the members asked to complete several scales of the California Psychological Inventory.

[8] Problems or portions thereof for which answers are given in Appendix F are denoted by *.

[9] The numbers in parentheses indicate the section in which the term is first described.

*b. Cars at a roadblock were stopped at random and the occupants searched for illegal drugs.

c. Twenty students in an introductory psychology class were selected by random sampling and asked to participate in an experiment.

d. The time to run a straight-alley maze was recorded for each of five randomly sampled rats from 10 cages.

e. Telephone numbers obtained by random sampling from a directory were called and the respondents asked their political preference.

*3. [1.2] Which of the following are acceptable research hypotheses?

*a. Right-handed people tend to be taller than left-handed people.

*b. Behavior therapy is more effective than hypnosis in helping smokers kick the habit.

c. Most clairvoyant people are able to communicate with beings from outer space.

d. Rats are likely to fixate an incorrect response if it is followed by an intense noxious stimulus.

*4. [1.2] For each of the following studies, identify the (i) independent variable, (ii) dependent variable, and (iii) possible nuisance variables.

*a. Televised scenes portraying physical, cartoon, and verbal violence were shown to 20 preschool children. The facial expressions of the children were videotaped and then classified by judges.

*b. Power spectral analyses of changes in cortical electroencephalogram (EEG) were made during a 1- to 5-hour period of morphine administration in 10 female Sprague-Dawley rats.

c. The effects of four amounts of flurazepam on hypnotic suggestibility in men and women were investigated.

d. The keypecking rates of 20 female Silver King pigeons on fixed ratio reinforcement schedules of FR10, FR50, and FR100 were recorded.

*5. [1.2] For the independent variables in Exercise 4, indicate (i) which are quantitative and (ii) which are qualitative.

6. [1.3] (a) List the ways in which experiments and quasi-experiments differ.

(b) Why wasn't the Newburgh-Kingston Caries-Fluorine Study an experiment?

7. [1.3] Describe how you would design the study described in Exercise 4a (a) as an experiment and (b) as a naturalistic observation study.

*8. [1.3–1.4] (i) Classify each of the following according to the most descriptive or definitive category in Sections 1.3 and 1.4: Use only one classification. (ii) What features of the studies prompted your classification?

*a. The effect of participation in the Boy Scouts, the independent variable, on the propensity for assuming leadership roles as an adult was investigated for a random sample of 400 men who were lifelong residents of Columbus, Ohio, and between the ages of 30 and 60. The subjects were classified as having held or not held a leadership role during the previous 5-year period. Records were then used to determine those men who had participated in the Boy Scouts.

*b. In a study of 86 lonely people, it was found that they display some of the characteristics of shy people: Lonely people disclose less personal information about themselves to opposite-sex friends than do nonlonely people, and they use inappropriate levels (too intimate or too impersonal) of self-disclosure in initial interactions.

*c. Two hundred thirty-two sixth graders took a test that measured arithmetic achievement. Two hundred of the students were matched on the basis of their achievement scores. One member of each pair was randomly assigned to participate in a conventional arithmetic instruction program; the other member of the pair participated in an experimental arithmetic program. At the end of the semester, it was found that the arithmetic achievement scores of the students who participated in the experimental arithmetic program were higher than those of the other sample of students, $t(99) = 2.358$, $p = .020$, $g = .54$.

*d. Job performance ratings of graduates of a police academy were obtained for six classes from 2005 to 2010.

e. Cabdrivers in a large city were classified as expert drivers ($n = 14$), average drivers ($n = 33$), or poor drivers ($n = 11$), the independent variable, based on company records of their earnings for the previous 6-month period. All of the drivers were men between the ages of 26 and 45 and had driven a cab for at least 5 years. According to employment tests, expert drivers were superior to average and poor drivers in their ability to perceive large meaningful patterns and to do so with such speed that it appeared almost intuitive. Furthermore, expert drivers organized their knowledge of the city hierarchically, from large geographic areas down to smaller neighborhoods.

f. According to a national survey, the mean number of movies attended per month by 14-year-old boys in the United States is 4.8, with a standard deviation of 1.3. Juvenile court records in Houston, Texas, indicate that the corresponding statistics for a random sample of 31 boys who appeared in court are $\overline{Y} = 4.9$ and $\hat{\sigma} = 1.6$. The researcher concluded that the dispersion of the movie attendance distribution for boys who appeared in the Houston juvenile court, one of the dependent variables, was greater than that for the nation at large, $\chi^2(30) = 45.44$, $p < .05$.

g. A survey by the Centers for Disease Control and Prevention in Atlanta, Georgia, found that 27.6% of 15-year-old girls in 1999 had had premarital sex at least once. The comparable percentages for 2003 and 2008 were 53% and 77.2%, respectively.

h. The relationship between birth order and participation in dangerous sports such as hang gliding, auto racing, and boxing was investigated. Records at Florida State University were screened to obtain 50 men who were first-born, second-born, and so on and to identify their recreational activities while at the university.

i. Pediatricians in Oklahoma provided the names of 421 new mothers. The mothers' infant feeding practices were subsequently determined. Eight years later, elementary school records for 372 of these children indicated that the breast-fed babies had a higher level of performance in school than did those who had been bottle-fed.

j. Employment records were used to identify 86 men who had worked for a company in Cleveland, Ohio, that manufactured chemicals used as fire retardants. A second group of men, $n = 89$, was identified who worked for two other companies in Cleveland and had no exposure to the chemicals. Evidence of primary thyroid dysfunction was found in four of the exposed men; none of the unexposed men showed evidence of thyroid dysfunction.

*9. [1.5] Identify potential threats to internal validity for these studies.

*a. Exercise 8a *b. Exercise 8b

c. Exercise 8d d. Exercise 8g

e. Exercise 8h f. Exercise 8i

*10. [1.5] Identify potential threats to external validity in these studies.

*a. Exercise 8c *b. Exercise 8e

c. Exercise 8f d. Exercise 8j

*11. [1.6] For the experiments in Exercise 8, indicate those for which the following are potential threats to valid inference making.

*a. Experimenter-expectancy effect b. Demand characteristics

c. Subject-predisposition effects

12. [1.7] Two approaches to controlling nuisance variables and minimizing threats to valid inference making are holding the nuisance variable constant and using random assignment or random sampling. Indicate which experiments in Exercise 8 used these approaches and which approach was used.

13. [1.8] Section 1.8 lists eight general guidelines for the ethical treatment of subjects. Recognizing that all of the guidelines are important, select the five that you think are the most important and rank order them (assign 1 to the most important guideline). What do your selection and rankings reveal about your own ethical code?

CHAPTER 2

Experimental Designs: An Overview

2.1 Introduction

A variety of research strategies were described in Chapter 1. In this chapter, I describe some basic experimental designs that are used with these research strategies. Recall from Section 1.1 that an experimental design is a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan. This chapter focuses on the assignment of subjects to experimental conditions and on the general features of some basic designs. The statistical analysis associated with the designs is presented in Chapters 4 to 16.

2.2 Overview of Some Basic Experimental Designs

t Test for Independent-Samples Design

One of the simplest experimental designs is the randomization and analysis plan that is used with a ***t* test for independent samples.** A two-sample t statistic is often used to test the null hypothesis that the difference between two population means is equal to some value, usually zero. Consider an experiment to help cigarette smokers break the habit. The independent variable is two kinds of therapy; the dependent variable is the number of cigarettes smoked per day 6 months after therapy. For notational convenience, the two kinds of therapy are called treatment A. The levels of treatment A that correspond to the specific therapies are denoted by the lowercase letter a and a subscript: a_1 denotes cognitive behavioral therapy, and a_2 denotes hypnosis. A particular but unspecified level of treatment A is denoted by a_j, where j ranges over the values 1 and 2. The number of cigarettes smoked per day 6 months after therapy by subject i in treatment level j is denoted by Y_{ij}.

The null and alternative hypotheses for the cigarette smoking experiment are, respectively,

$$H_0\colon \mu_1 - \mu_2 = 0$$

$$H_1\colon \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 denote the mean cigarette consumption of the respective populations. The Greek letter μ (mu) is pronounced “mew.” Assume that 30 smokers who want to stop smoking are available to participate in the experiment. I want to assign $n = 15$ smokers to each of the $p = 2$ treatment levels so that each possible assignment has the same probability. This can be done by numbering the smokers from 1 to 30 and drawing numbers from the random numbers table in Appendix Table E.1. The first n numbers drawn between 1 and 30 are assigned to treatment level a_1; the remaining subjects are assigned to a_2. The layout for the experiment is shown in Figure 2.2-1. The subjects who received treatment level a_1 are called Group$_1$; those who received treatment level a_2 are called Group$_2$. The two sample means, $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$, reflect the effects of the treatment levels that were administered to the subjects.[1] The computational procedures and assumptions associated with a t test for independent samples are discussed in Section 4.2.

		Treat. Level	Dep. Var.
Group$_1$	Subject$_1$	a_1	Y_{11}
	Subject$_2$	a_1	Y_{21}
	⋮	⋮	⋮
	Subject$_{15}$	a_1	$Y_{15,\,1}$
			$\overline{Y}_{\cdot 1}$
Group$_2$	Subject$_1$	a_2	Y_{12}
	Subject$_2$	a_2	Y_{22}
	⋮	⋮	⋮
	Subject$_{15}$	a_2	$Y_{15,\,2}$
			$\overline{Y}_{\cdot 2}$

Figure 2.2-1 ■ Layout for a t test for independent-samples design. The treatment level is denoted by Treat. Level; the dependent variable is denoted by Dep. Var. Thirty subjects are randomly assigned to two levels of treatment A with the restriction that 15 subjects are assigned to each level. The mean cigarette consumptions for subjects in treatment levels a_1 and a_2 are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$, respectively.

[1] A sample mean for the jth treatment level is obtained by summing over the $i = 1, \ldots, n$ subjects in the jth treatment level and dividing by n—that is, $\sum_{i=1}^{n} Y_{ij}/n = \overline{Y}_{\cdot j}$. Notice that the i subscript in Y_{ij} has been replaced by a dot in $\overline{Y}_{\cdot j}$. The dot indicates that summation was performed over the i subscript.

Completely Randomized Design

The t test for independent-samples design involves randomly assigning subjects to two levels of a treatment. A completely randomized analysis of variance design, described next, extends this design strategy to two or more treatment levels. Consider an experiment to evaluate the effectiveness of three therapies in helping cigarette smokers break the habit. The null and alternative hypotheses for the experiment are, respectively,

$$H_0\colon \mu_1 = \mu_2 = \mu_3$$

$$H_1\colon \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j',\ j \neq j'$$

The null hypothesis can be tested by using a **completely randomized analysis of variance design.** This design is denoted by the letters CR-*p,* where CR stands for *completely randomized* and p is the number of levels of the treatment. In this example, p is equal to 3. For convenience, the three kinds of therapy are called treatment A. The three levels of treatment A are denoted by the lowercase letter a and a subscript: a_1 is behavioral therapy, a_2 is hypnosis, and a_3 is a medication delivered by means of a patch applied to a smoker's back.

Assume that 45 smokers who want to stop smoking are available to participate in the experiment. The subjects are randomly assigned to the treatment levels with the restriction that 15 subjects are assigned to each therapy. The phrase *randomly assigned* is important. Randomization distributes the idiosyncratic characteristics of the subjects over the three treatment levels so that they do not selectively bias the outcome of the experiment. And randomization helps to prevent the experimenter's personal biases from being introduced into the experiment. As I discuss in Chapter 3, randomization also helps to obtain an unbiased estimate of the random error variation in the experiment, and it helps to ensure that the error effects are statistically independent. The layout for the experiment is shown in Figure 2.2-2. A comparison of the layout in this figure with that in Figure 2.2-1 for a t test for independent-samples design reveals that they are the same except that a completely randomized design can have more than two treatment levels. When $p = 2$, the layouts for the designs are identical.

Thus far, I have identified the null hypothesis that I want to test, $\mu_1 = \mu_2 = \mu_3$, and described the manner in which the subjects are assigned to the three treatment levels. In the following paragraphs, I discuss the composite nature of an observation, describe the experimental design model equation for a CR-p design, and examine the meaning of the terms *treatment effect* and *error effect.*

Experimental design model equation. An observation, denoted by Y_{ij} for subject i in treatment level $j,$ can be thought of as a composite that reflects the effects of (1) the independent variable, (2) individual characteristics of the subject or experimental unit, (3) chance fluctuations in the subject's performance, (4) measurement and recording errors that occur during data collection, and (5) any other nuisance variables such as environmental conditions that have not been controlled. Consider the cigarette consumption of subject 2 in treatment level a_2 in Figure 2.2-2. Suppose that 6 months after therapy, this subject is smoking three cigarettes a day ($Y_{22} = 3$). What factors have affected the value of

		Treat. Level	Dep. Var.
Group$_1$	Subject$_1$	a_1	Y_{11}
	Subject$_2$	a_1	Y_{21}
	$\vdots$	$\vdots$	$\vdots$
	Subject$_{15}$	a_1	$Y_{15,1}$
			$\bar{Y}_{\cdot 1}$
Group$_2$	Subject$_1$	a_2	Y_{12}
	Subject$_2$	a_2	Y_{22}
	$\vdots$	$\vdots$	$\vdots$
	Subject$_{15}$	a_2	$Y_{15,2}$
			$\bar{Y}_{\cdot 2}$
Group$_3$	Subject$_1$	a_3	Y_{13}
	Subject$_2$	a_3	Y_{23}
	$\vdots$	$\vdots$	$\vdots$
	Subject$_{15}$	a_3	$Y_{15,3}$
			$\bar{Y}_{\cdot 3}$

Figure 2.2-2 ▪ Layout for a completely randomized design (CR-3 design). The treatment level is denoted by Treat. Level; the dependent variable is denoted by Dep. Var. Forty-five subjects are randomly assigned to three levels of treatment A, with the restriction that 15 subjects are assigned to each level. The mean cigarette consumptions for subjects in treatment levels a_1, a_2, and a_3 are denoted by $\bar{Y}_{\cdot 1}$, $\bar{Y}_{\cdot 2}$, and $\bar{Y}_{\cdot 3}$, respectively.

Y_{22}? One factor is the effectiveness of the therapy received—hypnosis in this case. Other factors are the subject's cigarette consumption prior to therapy, the subject's level of motivation to stop smoking, and the weather during the previous 6 months, to mention only a few. In summary, Y_{22} is a composite that reflects (1) the effects of treatment level a_2, (2) effects unique to the subject, (3) effects attributable to chance fluctuations in the subject's behavior, (4) errors in measuring and recording the subject's cigarette consumption, and (5) any other effects that have not been controlled.

My conjectures about Y_{22} or any other observation can be expressed more formally by an **experimental design model equation.** The model equation for the smoking experiment is

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)} \quad (i = 1, \ldots, n; j = 1, \ldots, p)$$

where

- Y_{ij} is the observation for subject *i* in treatment level *j.*
- μ is the population grand mean of μ_1, μ_2, and μ_3. You can think of μ as the average value around which the treatment means vary; μ is a constant for the 45 scores in the experiment.
- α_j (alpha) is the treatment effect for population *j* and is equal to $\mu_j - \mu$, the deviation of the grand mean from the *j*th population mean. The treatment effect reflects the effects of the *j*th therapy and is a constant for the 15 scores in treatment level a_j.
- $\varepsilon_{i(j)}$ (epsilon) is the error effect associated with Y_{ij} and is equal to $Y_{ij} - \mu - \alpha_j$. The error effect represents effects unique to subject *i,* effects attributable to chance fluctuations in subject *i*'s behavior, and any other effects that have not been controlled such as environmental conditions—in other words, all effects not attributable to treatment level a_j. The notation *i*(*j*) indicates that the *i*th subject appears only in treatment level *j;* subject *i* is said to be nested within the *j*th treatment level.[2]

According to the model equation for this completely randomized design, each observation is the sum of three parameters μ, α_j, and $\varepsilon_{i(j)}$. The values of the parameters are unknown, but in Section 3.2, I show how they can be estimated from sample data.

The meanings of the terms *grand mean,* μ, and *treatment effect,* α_j, in the model equation seem fairly clear; the meaning of *error effect,* $\varepsilon_{i(j)}$, requires a bit more explanation. Why do observations, Y_{ij}s, in the same treatment level vary from one subject to the next? This variation must be due to differences among the subjects and to other uncontrolled variables because the parameters μ and α_j in the model equation are constants for all subjects in the same treatment level. To put it another way, observations in the same treatment level are different because the error effects, $\varepsilon_{i(j)}$s, for the observations are different. Recall that error effects reflect idiosyncratic characteristics of the subjects—those characteristics that differ from one subject to another—and any other variables that have not been controlled. Researchers attempt to minimize the size of error effects by holding constant sources of variation that might contribute to the error effects and by the judicial choice of an experimental design. Designs described in the following sections permit a researcher to isolate and remove some sources of variation that would ordinarily be included in the error effect.

An experimental design model is an example of a **linear model.** A linear model consists of two parts: a linear model equation, for example, $Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)}$, and assumptions about the model parameters. The assumptions for this model are described in Section 3.3. The model is called a linear model because the observation, Y_{ij}, is equal to a linear combination of the model parameters: $\mu + \alpha_j + \varepsilon_{i(j)}$.

t Test for Dependent-Samples Design

The two designs just described require the use of independent samples. Two samples are independent if, for example, a researcher samples randomly from two populations or uses

[2] Nesting of subjects and treatments is discussed in Section 11.1.

a random procedure to assign subjects to two groups. Dependent samples, on the other hand, can be obtained by any of the following procedures:

1. Observing each subject under each treatment level in the experiment—that is, obtaining **repeated measures** on the subjects.
2. Forming sets of subjects who are similar with respect to a variable that is correlated with the dependent variable. This procedure is called **subject matching.**
3. Obtaining sets of identical twins or littermates and assigning one member of the pair randomly to one treatment level and the other member to the other treatment level.
4. Obtaining pairs of subjects who are matched by mutual selection—for example, husband and wife pairs or business partners.

In behavioral, medical, and educational research, the subjects are often people whose aptitudes and experiences differ markedly. Individual differences are inevitable, but it is often possible to isolate or partition out a portion of these effects so that they do not appear in estimates of the error effects. One design for accomplishing this is a ***t* test for dependent samples.** As the name suggests, the design uses dependent samples. A *t* test for dependent samples also uses a more complex randomization and analysis plan than a *t* test for independent samples, but the added complexity is usually accompanied by greater power[3]—a point that I develop when I discuss a randomized block analysis of variance design in the next section.

Let's reconsider the cigarette smoking experiment. It is reasonable to assume that the difficulty in breaking the smoking habit is related to the number of cigarettes that a person smokes per day. The design of the experiment can be improved by isolating this variable. Suppose that instead of randomly assigning 30 subjects to the treatment levels, I form pairs of subjects such that the subjects in each pair have similar cigarette consumptions prior to the experiment. The subjects in each pair constitute a **block** of matched subjects. A simple way to match the subjects is to rank them in terms of the number of cigarettes they smoke per day. The subjects ranked 1 and 2 are assigned to block 1, those ranked 3 and 4 are assigned to block 2, and so on. In this example, 15 blocks of matched subjects can be formed. After all of the blocks have been formed, the two subjects in each block are randomly assigned to the two kinds of therapy. The layout for this experiment is shown in Figure 2.2-3. If my hunch is correct—that the difficulty in breaking the smoking habit is related to the number of cigarettes that a person smokes per day—this design should result in a greater likelihood of rejecting the null hypothesis:

$$H_0\colon \mu_{\cdot 1} - \mu_{\cdot 2} = 0$$

than does a *t* test for independent samples. Later I show that the increased power to reject the null hypothesis results from isolating the nuisance variable of number of cigarettes smoked per day and thereby reducing the size of the error effects.

[3] Power refers to the probability of rejecting a false null hypothesis (see Section 2.5).

	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.
Block$_1$	a_1	Y_{11}	a_2	Y_{12}
Block$_2$	a_1	Y_{21}	a_2	Y_{22}
Block$_3$	a_1	Y_{31}	a_2	Y_{32}
⋮	⋮	⋮	⋮	⋮
Block$_{15}$	a_1	$Y_{15,\,1}$	a_2	$Y_{15,\,2}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$

Figure 2.2-3 ▪ Layout for a *t* test for dependent samples, where each block contains two subjects whose cigarette consumptions prior to the experiment were similar. The two subjects in a block are randomly assigned to the treatment levels. The mean cigarette consumptions for subjects in treatment levels a_1 and a_2 are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$, respectively.

Randomized Block Design

Earlier you learned that the layout and randomization procedures for a *t* test for independent samples and a completely randomized analysis of variance design are the same except that a completely randomized design can have more than two treatment levels. The same comparison can be drawn between a *t* test for dependent samples and a randomized block analysis of variance design. A **randomized block analysis of variance design** is denoted by the letters RB-*p,* where RB stands for *randomized block* and *p* is the number of levels of the treatment.

Suppose that in the cigarette smoking experiment, I want to evaluate the effectiveness of three kinds of therapy in helping smokers break the habit. The three kinds are behavioral therapy, denoted by a_1; hypnosis, denoted by a_2; and a medication, denoted by a_3. I suspect that the difficulty in breaking the smoking habit is related to the number of cigarettes that a person smokes per day. I can use the blocking procedure described in connection with a *t* test for dependent samples to isolate and control this nuisance variable. If a sample of 45 smokers is available, I can form 15 blocks that contain three subjects who have had similar consumptions of cigarettes prior to the experiment. The dependent variable for a subject in block *i* and treatment level *j* is denoted by Y_{ij}. The layout for the experiment is shown in Figure 2.2-4. A comparison of the layout in this figure with that in Figure 2.2-3 for a *t* test for dependent-samples design reveals that they are the same except that the randomized block design has $p = 3$ treatment levels. When $p = 2$, the layouts for the designs are identical.

A randomized block design enables a researcher to test two null hypotheses:

H_0: $\mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$ (Treatment population means are equal.)

H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{15\cdot}$ (Block population means are equal.)

	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	
Block$_1$	a_1	Y_{11}	a_2	Y_{12}	a_3	Y_{13}	$\overline{Y}_{1\cdot}$
Block$_2$	a_1	Y_{21}	a_2	Y_{22}	a_3	Y_{23}	$\overline{Y}_{2\cdot}$
Block$_3$	a_1	Y_{31}	a_2	Y_{32}	a_3	Y_{33}	$\overline{Y}_{3\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Block$_{15}$	a_1	$Y_{15,\,1}$	a_2	$Y_{15,\,2}$	a_3	$Y_{15,\,3}$	$\overline{Y}_{15\cdot}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$		$\overline{Y}_{\cdot 3}$	

Figure 2.2-4 ■ Layout for a randomized block design (RB-3 design), where each block contains three matched subjects whose cigarette consumptions prior to the experiment were similar. The subjects in a block are randomly assigned to the treatment levels. The mean cigarette consumptions for subjects in treatment levels a_1, a_2, and a_3 are denoted by $\overline{Y}_{\cdot 1}$, $\overline{Y}_{\cdot 2}$, and $\overline{Y}_{\cdot 3}$, respectively; the mean cigarette consumptions for subjects in Block$_1$, Block$_2, \ldots,$ Block$_{15}$ are denoted by $\overline{Y}_{1\cdot}$, $\overline{Y}_{2\cdot}, \ldots,$ $\overline{Y}_{15\cdot}$, respectively.

The first hypothesis states that the population means for the three therapies are equal. The second hypothesis, which is usually of little interest, states that the population means for the 15 levels of the nuisance variable, cigarette consumption prior to the experiment, are equal. I expect a test of this null hypothesis to be significant. If the nuisance variable of the number of cigarettes smoked prior to the experiment does not account for an appreciable proportion of the total variation in the experiment, little has been gained by isolating the effects of the variable. Before exploring this point, I describe the experimental design model equation for an RB-p design.

Experimental design model equation. The model equation for the smoking experiment is

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij} \quad (i = 1, \ldots, n; j = 1, \ldots, p)$$

where

Y_{ij} is the observation for the subject in block i and treatment level j.

μ is the population grand mean of $\mu_{11}, \mu_{21}, \ldots, \mu_{15,\,3}$. You can think of μ as the average value around which the treatment and block means vary; μ is a constant for the 45 observations in the experiment.

α_j is the treatment effect for population j and is equal to $\mu_{\cdot j} - \mu$, the deviation of the grand mean from the jth population treatment mean. The treatment effect reflects the effects of the jth therapy and is a constant for the 15 observations in treatment level a_j.

π_i (pi) is the block effect for population i and is equal to $\mu_{i\cdot} - \mu$, the deviation of the grand mean from the ith population block mean. The block effect reflects the effects of smoking a certain number of cigarettes per day prior to therapy.

ε_{ij} is the error effect associated with Y_{ij} and is equal to $Y_{ij} - \mu - \alpha_j - \pi_i$. The error effect represents effects unique to subject i in treatment level j, effects attributable to chance fluctuations in subject i's behavior, and any other effects that have not been controlled such as environmental conditions—in other words, all effects not attributable to treatment level j and block i.[4]

According to the equation for this randomized block design, each observation is the sum of four parameters: μ, α_j, π_i, and ε_{ij}. The error effect is that portion of an observation that remains after the grand mean, treatment effect, and block effect have been subtracted from it; that is, $\varepsilon_{ij} = Y_{ij} - \mu - \alpha_j - \pi_i$. The sum of the squared error effects for this randomized block design,

$$\Sigma\Sigma\varepsilon_{ij}^2 = \Sigma\Sigma(Y_{ij} - \mu - \alpha_j - \pi_i)^2$$

will be smaller than the sum for the completely randomized design,

$$\Sigma\Sigma\varepsilon_{ij}^2 = \Sigma\Sigma(Y_{ij} - \mu - \alpha_j)^2$$

if π_i is greater than zero for one or more blocks. As I show in Section 3.3, the F statistic that is used to test the null hypothesis in analysis of variance can be thought of as the ratio of error and treatment effects,

$$F = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})}$$

where $f(\ )$ denotes a function of the effects in parentheses. It is apparent from an examination of this ratio that the smaller the sum of the squared error effects, the larger the F statistic and, hence, the greater the probability of rejecting a false null hypothesis. Thus, by isolating a nuisance variable that accounts for an appreciable portion of the total variation in a randomized block design, a researcher is rewarded with a more powerful test of a false null hypothesis.

As you have seen, blocking with respect to the nuisance variable, the number of cigarettes smoked per day, enables me to isolate this variable and remove it from the error effects. But what if the nuisance variable does not account for any of the variation in the experiment? In other words, what if all of the block effects are equal to zero ($\mu_{i\cdot} - \mu = 0$ for all i)? Then the sum of the squared error effects for the randomized block and the completely randomized designs will be equal, and the effort used to form blocks of matched subjects in the randomized block design will be for naught. The larger the correlation between the nuisance variable and the dependent variable, the more likely it is that the block effects account for an appreciable proportion of the total variation in the experiment.

[4] For now, I ignore the possibility that cigarette consumption interacts with type of therapy. Section 8.3 describes an experimental design model equation that includes a treatment-block interaction term $(\alpha\pi)_{ji}$.

Latin Square Design

The Latin square design described in this section derives its name from an ancient puzzle that was concerned with the number of different ways that Latin letters can be arranged in a square matrix so that each letter appears once in each row and once in each column. An example of a 3 × 3 Latin square is shown in Figure 2.2-5. I use the letter a and subscripts in place of Latin letters. The **Latin square design** is denoted by the letters LS-p, where LS stands for *Latin square* and p is the number of levels of the treatment. A Latin square design enables a researcher to isolate the effects of not one but two nuisance variables. The levels of one nuisance variable are assigned to the rows of the square; the levels of the other nuisance variable are assigned to the columns. The levels of the treatment are assigned to the cells of the square.

	c_1	c_2	c_3
b_1	a_1	a_2	a_3
b_2	a_2	a_3	a_1
b_3	a_3	a_1	a_2

Figure 2.2-5 ▪ Three-by-three Latin square, where a_j denotes one of the $j = 1, \ldots, p$ levels of treatment A, b_k denotes one of the $k = 1, \ldots, p$ levels of nuisance variable B, and c_l denotes one of the $l = 1, \ldots, p$ levels of nuisance variable C. Each level of treatment A appears once in each row and once in each column as required for a Latin square.

Let's return to the cigarette smoking experiment. With a Latin square design, I can isolate the effects of cigarette consumption and the effects of a second nuisance variable—say, the length of time in years that a person has smoked. The advantage of being able to isolate two nuisance variables comes at a price. The randomization procedures for a Latin square design, which are described in Chapter 14, are more complex than those for a randomized block design. Also, the number of rows and columns of a Latin square must each equal the number of treatment levels, which is three in this example. I can assign three levels of cigarette consumption to the rows of the Latin square: b_1 is less than one pack per day, b_2 is one to three packs per day, and b_3 is more than three packs per day. The other nuisance variable, the duration of the smoking habit in years, can be assigned to the columns of the square: c_1 is less than 1 year, c_2 is 1 to 5 years, and c_3 is more than 5 years. The dependent variable for the ith subject in the jth treatment level, kth row, and lth column is denoted by Y_{ijkl}. The layout of the design is shown in Figure 2.2-6.

The Latin square design lets me test three null hypotheses:

H_0: $\mu_{1..} = \mu_{2..} = \mu_{3..}$ (Treatment population means are equal.)

H_0: $\mu_{.1.} = \mu_{.2.} = \mu_{.3.}$ (Row population means are equal.)

H_0: $\mu_{..1} = \mu_{..2} = \mu_{..3}$ (Column population means are equal.)

		Treat. Comb.	Dep. Var.
Group_1	Subject_1	$a_1b_1c_1$	Y_{1111}
	⋮	⋮	⋮
	Subject_5	$a_1b_1c_1$	Y_{5111}
			$\overline{Y}_{\cdot 111}$
Group_2	Subject_1	$a_1b_2c_3$	Y_{1123}
	⋮	⋮	⋮
	Subject_5	$a_1b_2c_3$	Y_{5123}
			$\overline{Y}_{\cdot 123}$
Group_3	Subject_1	$a_1b_3c_2$	Y_{1132}
	⋮	⋮	⋮
	Subject_5	$a_1b_3c_2$	Y_{5132}
			$\overline{Y}_{\cdot 132}$
Group_4	Subject_1	$a_2b_1c_2$	Y_{1212}
	⋮	⋮	⋮
	Subject_5	$a_2b_1c_2$	Y_{5212}
			$\overline{Y}_{\cdot 212}$
	⋮	⋮	⋮
Group_9	Subject_1	$a_3b_3c_1$	Y_{1331}
	⋮	⋮	⋮
	Subject_5	$a_3b_3c_1$	Y_{5331}
			$\overline{Y}_{\cdot 331}$

Figure 2.2-6 ■ Layout for a Latin square design (LS-3 design) that is based on the Latin square in Figure 2.2-5. The treatment combination is denoted by Treat. Comb. Treatment *A* represents three kinds of therapy, nuisance variable *B* represents the number of cigarettes smoked per day, and nuisance variable *C* represents the length of time in years that a person has smoked. Subjects in Group_1, for example, received behavioral therapy (a_1), smoked less than one pack of cigarettes per day (b_1), and had smoked for less than 1 year (c_1). The mean cigarette consumptions for the subjects in the nine groups are denoted by $\overline{Y}_{\cdot 111}$, $\overline{Y}_{\cdot 123}, \ldots, \overline{Y}_{\cdot 331}$.

The first hypothesis states that the population means for the three therapies are equal. The second and third hypotheses make similar assertions about the population means for the two nuisance variables: number of cigarettes smoked per day and duration of the smoking habit in years. Tests of these nuisance variables are expected to be significant. As discussed earlier, if the nuisance variables do not account for an appreciable proportion of the total variation in the experiment, little has been gained by isolating the effects of the variables.

Experimental design model equation. The model equation for this version of our smoking experiment is

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + \varepsilon_{\text{Pooled}} \qquad (i = 1, \ldots, n; j = 1, \ldots, p; k = 1, \ldots, p; l = 1, \ldots, p)$$

where

Y_{ijkl} is the observation for subject i in treatment level j, row k, and column l.

μ is the population grand mean of $\mu_{111}, \mu_{123}, \ldots, \mu_{331}$. You can think of μ as the average value around which the treatment, row, and column means vary; μ is a constant for all scores in the experiment.

α_j is the treatment effect for population j and is equal to $\mu_{j\cdot\cdot} - \mu$, the deviation of the grand mean from the jth population treatment mean. The treatment effect reflects the effects of the jth therapy and is a constant for the scores in treatment level a_j.

β_k (beta) is the row effect for population k and is equal to $\mu_{\cdot k \cdot} - \mu$, the deviation of the grand mean from the kth population row mean. The row effect reflects the effects of smoking a certain number of cigarettes per day prior to therapy.

γ_l (gamma) is the column effect for population l and is equal to $\mu_{\cdot\cdot l} - \mu$, the deviation of the grand mean from the lth population column mean. The column effect reflects the effects of smoking for a certain number of years prior to therapy.

$\varepsilon_{\text{Pooled}}$ is the pooled error effect associated with Y_{ijkl} and is equal to $Y_{ijkl} - \mu - \alpha_j - \beta_k - \gamma_l$. The nature of this pooled error effect is discussed in Chapter 14.

According to the equation for this Latin square design, each observation is the sum of five parameters: μ, α_j, β_k, γ_l, and $\varepsilon_{\text{Pooled}}$. The sum of the squared error effects for this Latin square design,

$$\sum\sum\sum \varepsilon^2_{\text{Pooled}} = \sum\sum\sum (Y_{ijkl} - \mu - \alpha_j - \beta_k - \gamma_l)^2$$

will be smaller than the sum for the randomized block design,

$$\sum\sum \varepsilon^2_{ij} = \sum\sum (Y_{ij} - \mu - \alpha_j - \pi_i)^2$$

if the combined effects of $\sum \beta^2_k$ and $\sum \gamma^2_l$ are greater than $\sum \pi^2_i$.

Building block designs. Thus far, I have described three of the simplest analysis of variance designs: completely randomized design, randomized block design, and Latin square design. I call these three ANOVA designs **building block designs** because all complex ANOVA designs can be constructed by modifying or combining these simple designs. Furthermore, the randomization procedures, data analysis procedures, and assumptions for complex ANOVA designs are extensions of those for the three building block designs. The following section provides a preview of a factorial design that is constructed from two completely randomized designs.

Completely Randomized Factorial Design

Factorial designs differ from those described previously because two or more treatments can be evaluated simultaneously in a single experiment.[5] The simplest factorial design from the standpoint of randomization procedures, data analysis, and assumptions is based on a completely randomized analysis of variance design and hence is called a **completely randomized factorial design.** A two-treatment, completely randomized factorial design is denoted by the letters CRF-*pq,* where the letters CR denote the building block design, F indicates that the design is a factorial design, and *p* and *q* stand for the number of levels of treatments *A* and *B,* respectively.

Consider an experiment to evaluate the effects of two treatments on the speed of reading. Let treatment *A* consist of two levels of room illumination: a_1 is 15-foot candles and a_2 is 30-foot candles. Treatment *B* consists of three levels of type size: b_1 is 9-point type, b_2 is 12-point type, and b_3 is 15-point type. This design is designated by the letters CRF-23, where 2 is the number of levels of treatment *A* and 3 is the number of levels of treatment *B.* The layout for the design is obtained by combining the treatment levels of a CR-2 design with those of a CR-3 design so that each treatment level of the CR-2 design appears once with each level of the CR-3 design. The resulting CRF-23 design has $2 \times 3 = 6$ **treatment combinations** as follows: a_1b_1, a_1b_2, a_1b_3, a_2b_1, a_2b_2, a_2b_3. When treatment levels are combined in this way, the treatments are said to be **completely crossed.** Completely crossed treatments are a characteristic of all completely randomized factorial designs. Assume that 30 sixth-graders are available to participate in the experiment. The children are randomly assigned to the six treatment combinations, with the restriction that five children receive each combination. The layout of the design is shown in Figure 2.2-7.

Experimental design model equation. The model equation for the experiment is

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{i(jk)} \quad (i = 1, \ldots, n; j = 1, \ldots, p; k = 1, \ldots, q)$$

where

Y_{ijk} is the observation for subject *i* in treatment combination a_jb_k.

μ is the population grand mean of μ_{11}, $\mu_{12}, \ldots, \mu_{23}$. You can think of μ as the average value around which the treatment *A* and *B* means vary; μ is a constant for the 30 observations in the experiment.

[5] The distinction between a treatment and a nuisance variable is in the mind of the researcher. A nuisance variable is included in a design to improve the efficiency and power of the design; a treatment is included because it is related to the scientific hypothesis that a researcher wants to test. This distinction has important implications for the statistical analysis, as you will learn.

		Treat. Comb.	Dep. Var.
Group_1	Subject_1	a_1b_1	Y_{111}
	$\vdots$	$\vdots$	$\vdots$
	Subject_5	a_1b_1	Y_{511}
			$\overline{Y}_{\cdot 11}$
Group_2	Subject_1	a_1b_2	Y_{112}
	$\vdots$	$\vdots$	$\vdots$
	Subject_5	a_1b_2	Y_{512}
			$\overline{Y}_{\cdot 12}$
Group_3	Subject_1	a_1b_3	Y_{113}
	$\vdots$	$\vdots$	$\vdots$
	Subject_5	a_1b_3	Y_{513}
			$\overline{Y}_{\cdot 13}$
$\vdots$		$\vdots$	$\vdots$
Group_6	Subject_1	a_2b_3	Y_{123}
	$\vdots$	$\vdots$	$\vdots$
	Subject_5	a_2b_3	Y_{523}
			$\overline{Y}_{\cdot 23}$

Figure 2.2-7 ■ Layout for a completely randomized factorial design (CRF-23 design) where 30 subjects are randomly assigned to six combinations of treatments *A* and *B,* with the restriction that five subjects are assigned to each combination. Treatment *A* represents two levels of room illumination; treatment *B* represents three levels of type size. Subjects in Group_1, for example, read in a room with 15-foot candles of illumination (a_1), and the material was typed using 9-point type (b_1). The mean reading speeds for the subjects in the six groups are denoted by $\overline{Y}_{\cdot 11}$, $\overline{Y}_{\cdot 12}$,..., $\overline{Y}_{\cdot 23}$.

α_j is the treatment effect for population a_j and is equal to $\mu_{j\cdot} - \mu$, the deviation of the grand mean from the *j*th population treatment mean. This treatment effect reflects the effects of the *j*th level of room illumination and is a constant for the 15 scores in treatment level a_j.

β_k is the treatment effect for population b_k and is equal to $\mu_{\cdot k} - \mu$, the deviation of the grand mean from the *k*th population treatment mean. This treatment effect reflects the effects of the *k*th level of type size and is a constant for the 10 scores in treatment level b_k.

$(\alpha\beta)_{jk}$ is the interaction effect for populations a_j and b_k and is equal to $\mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu$. Interaction effects are discussed in Chapter 9.

$\varepsilon_{i(jk)}$ is the error effect associated with Y_{ijk} and is equal to $Y_{ijk} - \mu - \alpha_j - \beta_k - (\alpha\beta)_{jk}$. The error effect represents effects unique to subject *i,* effects attributable to chance fluctuations in subject *i*'s behavior, and any other effects that have not been controlled—in other words, all effects not attributable to treatment level a_j, treatment level b_k, and the interaction of a_j and b_k.

Null hypotheses. This design lets me test three null hypotheses:

H_0: $\mu_{1\cdot} = \mu_{2\cdot}$ (Treatment *A* population means are equal.)

H_0: $\mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$ (Treatment *B* population means are equal.)

H_0: $\mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k'} = 0$ for all j and k (All *AB* interaction effects equal zero.)

The last hypothesis is unique to factorial designs. It states that the joint effects (interaction) of treatments *A* and *B* are equal to zero for all combinations of the two treatments. Two treatments are said to interact if differences in performance under the levels of one treatment are different at two or more levels of the other treatment. Figure 2.2-8 illustrates two possible outcomes of the reading experiment: Part (a) illustrates the presence of an interaction, and part (b) illustrates the absence of an interaction. When two treatments interact as in Figure 2.2-8(a), a graph of the population means always reveals at least two nonparallel lines for one or more segments of the lines. I say more about interactions in Chapter 9.

Comparison of CR-*p* and CRF-*pq* designs. Earlier I observed that a completely randomized design is the building block design for a completely randomized factorial design. The similarities between the two designs become apparent when you compare the

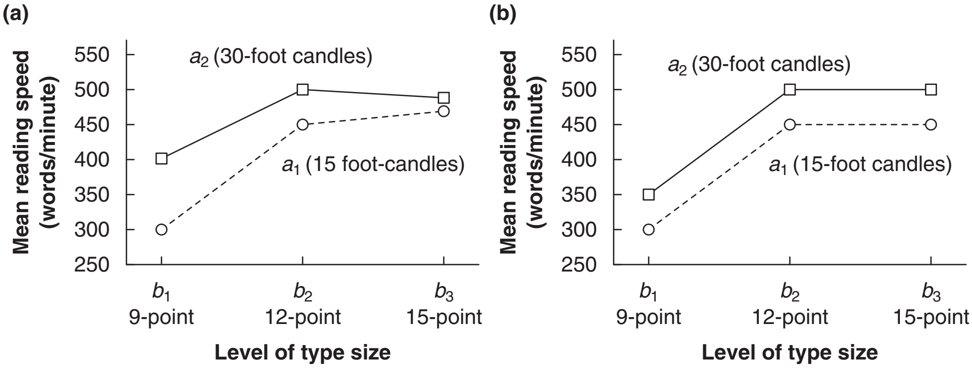


Figure 2.2-8 ■ Two possible outcomes of the reading experiment. Part (a) illustrates an interaction between treatments *A* and *B;* part (b) illustrates the absence of an interaction. Nonparallelism of the lines for one or more segments of the lines signifies interaction.

randomization procedures for the designs. For both designs, the subjects are randomly assigned to the treatment levels (combinations), with the restriction that the same number of subjects is assigned to each level (combination).[6] When I describe the assumptions for the two designs in Chapters 3 and 9, we find additional similarities.

2.3 Classification of Analysis of Variance Designs

In the last section, I described four of the simpler analysis of variance (ANOVA) designs. As you will see, this discussion only scratched the surface. There is a bewildering array of analysis of variance designs available to researchers. Furthermore, there is no universally accepted nomenclature for analysis of variance designs; some designs have as many as five different names.[7] And most of the design nomenclatures do not indicate which ANOVA designs share similar randomization plans and layouts. My design nomenclature in Table 2.3-1 is based on the concept of building block designs. Recall that all complex designs can be constructed by modifying or combining three simple designs: completely

Table 2.3-1 ■ Classification of ANOVA Designs

Analysis of Variance Designs	Abbreviated Designation
I. Systematic designs	
II. Randomized designs with one treatment	
A. Experimental units randomly assigned to treatment levels	
1. Completely randomized design	CR-p
B. Experimental units assigned to relatively homogeneous blocks or groups prior to random assignment to treatment levels	
1. Balanced incomplete block design	BIB-p
2. Crossover design	CO-p
3. Generalized randomized block design	GRB-p
4. Graeco-Latin square design	GLS-p
5. Hyper-Graeco-Latin square design	HGLS-p
6. Latin square design	LS-p
7. Lattice balanced incomplete block design	LBIB-p
8. Lattice partially balanced incomplete block design	LPBIB-p

(Continued)

[6] As discussed in Chapter 5, the assignment of the same number of subjects to each treatment level of a completely randomized design is desirable but not necessary.

[7] For example, a completely randomized design has been called a one-way classification design, single-factor experiment, randomized group design, simple randomized design, and single variable experiment.

Table 2.3-1 ▪ Classification of ANOVA Designs (Continued)

9. Lattice unbalanced incomplete block design	LUBIB-p
10. Partially balanced incomplete block design	PBIB-p
11. Randomized block design	RB-p
12. Youden square design	YBIB-p
III. Randomized designs with two or more treatments	
A. Factorial designs: designs in which all treatments are crossed	
1. Designs without confounding	
a. Completely randomized factorial design	CRF-pq
b. Generalized randomized block factorial design	GRBF-pq
c. Randomized block factorial design	RBF-pq
2. Design with group-treatment confounding	
a. Split-plot factorial design	SPF-$p{\cdot}q$
3. Designs with group-interaction confounding	
a. Latin square confounded factorial design	LSCF-p^k
b. Randomized block completely confounded factorial design	RBCF-p^k
c Randomized block partially confounded factorial design	RBPF-p^k
4. Designs with treatment-interaction confounding	
a. Completely randomized fractional factorial design	CRFF-p^{k-i}
b. Graeco-Latin square fractional factorial design	GLSFF-p^k
c. Latin square fractional factorial design	LSFF-p^k
d. Randomized block fractional factorial design	RBFF-p^{k-i}
B. Hierarchical designs: designs in which one or more treatments are nested	
1. Designs with complete nesting	
a. Completely randomized hierarchical design	CRH-$pq(A)$
b. Randomized block hierarchical design	RBH-$pq(A)$
2. Designs with partial nesting	
a. Completely randomized partial hierarchical design	CRPH-$pq(A)r$
b. Randomized block partial hierarchical design	RBPH-$pq(A)r$
c. Split-plot partial hierarchical design	SPPH-$p{\cdot}qr(B)$

(Continued)

Table 2.3-1 ■ (Continued)

IV. Designs with one or more covariates	
A. Designs that include a covariate have the letters AC added to the abbreviated designation.	
1. Completely randomized analysis of covariance design	CRAC-*p*
2. Completely randomized factorial analysis of covariance design	CRFAC-*pq*
3. Latin square analysis of covariance design	LSAC-*p*
4. Randomized block analysis of covariance design	RBAC-*p*
5. Split-plot factorial analysis of covariance design	SPFAC-*p*·*q*

Note: Abbreviated designations are discussed in the text.

randomized design (CR-*p*), randomized block design (RB-*p*), and Latin square design (LS-*p*). These three designs provide the organizational structure for the design nomenclature and classification system that is outlined in Table 2.3-1. The letter *p* in the table denotes the number of levels of a treatment. If a design includes a second or third treatment, the number of their levels is denoted by *q* and *r,* respectively.

The category *systematic designs* in Table 2.3-1 is of historical interest only. According to Leonard and Clark (1939), agricultural field research using systematic designs on a practical scale dates back to 1834. Prior to the work of R. A. Fisher in the 1920s and 1930s, as well as that of J. Neyman and E. S. Pearson on the theory of statistical inference, investigators used systematic schemes instead of random procedures to assign plots of land or other suitable experimental units to treatment levels—hence the designation systematic designs for these early field experiments. Over the past 100 years, systematic designs have fallen into disuse because designs that use random assignment are more likely to provide valid estimates of treatment effects and error effects, and they can be analyzed with the powerful tools of statistical inference such as the analysis of variance.

The impetus for the development of better research procedures came from the need to improve agricultural techniques. Today the experimental design nomenclature is replete with terms from agriculture. Modern principles of experimental design, particularly the random assignment of experimental units to treatment levels, received general acceptance as a result of the work of Fisher (1935b) and Fisher and MacKenzie (1922, 1923). Experimental designs that use the randomization principle are called **randomized designs.** The randomized designs in Table 2.3-1 are subdivided into several distinct categories based on (1) the number of treatments, (2) whether the subjects are subdivided into homogeneous blocks or groups prior to assigning them to treatment levels, (3) the nature of any confounding, (4) the use of crossed versus nested treatments, and (5) the use of a covariate.

A quick perusal of Table 2.3-1 reveals why researchers sometimes have difficulty selecting an appropriate ANOVA design; there are a lot of designs from which to choose. Because of the wide variety of designs available, it is important to identify them clearly in research reports. One often sees statements such as “a two-treatment factorial design was used.” It should be evident that a more precise description is required. This description could refer to any of the 11 factorial designs in Table 2.3-1.

In Section 2.2, I briefly described 4 of the 34 types of ANOVA designs in Table 2.3-1. These descriptions highlighted the ways in which the designs differ: (1) randomization, (2) experimental design model equation, (3) number of treatments, (4) inclusion of a nuisance variable as a factor in the experiment, and (5) power. In Chapters 4 and 8 to 16, I discuss other ways in which designs differ: (1) use of crossed or nested treatments or a combination of the two, (2) presence or absence of confounding, and (3) use of a covariate. I also identify the following common threads that run through the various designs:

1. All complex designs can be constructed from the three building block designs: completely randomized design, randomized block design, and Latin square design.
2. There are only four kinds of variation in ANOVA: total variation, between-groups variation, within-groups variation, and interaction variation.
3. All error terms involve either within-groups variation or interaction variation.
4. The numerator of an F statistic should always estimate one more source of variation than the denominator, and that source of variation should be the one that is being tested.

2.4 Selecting an Appropriate Design

Questions to Consider in Selecting an Appropriate Design

Considering the variety of analysis of variance designs available, it is not surprising that some researchers approach the selection of an appropriate design with trepidation. Selection of the best design for a particular research problem requires a familiarity with (1) the research area and (2) the designs that are available. In selecting a design, the following questions need to be considered.

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

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 of specific designs in Chapters 4 and 8 to 15 should go a long way toward demystifying the selection of an appropriate analysis of variance design.

*2.5 Review of Statistical Inference

Scientific and Statistical Hypotheses

People are by nature inquisitive. We ask questions, develop hunches, and sometimes put our hunches to the test. Over the years, a formalized procedure for testing hunches has evolved—the scientific method. The procedure involves (1) observing nature, (2) asking questions, (3) formulating hypotheses, (4) conducting experiments, and (5) developing theories and laws. Let's examine the third step: formulating hypotheses.

A **scientific hypothesis** is a testable supposition that is tentatively adopted to account for certain facts and to guide in the investigation of others. It is a statement about nature that requires verification. Four examples of scientific hypotheses are (1) the child-rearing practices of parents affect the personalities of their offspring, (2) college students who are active in student government have higher IQs than students who are not involved in student government, (3) cigarette smoking is associated with high blood pressure, and (4) children who feel insecure engage in overt aggression more frequently than do children who feel secure. These hypotheses have three characteristics in common with all scientific hypotheses: (1) They are intelligent, informed guesses about phenomena of interest; (2) they can be reduced to the form of an *if-then* statement—for example, "*If* John smokes, *then* he will show signs of high blood pressure"; and (3) their truth or falsity can be determined by observation or experimentation.

Statistical inference is a form of reasoning in which a rational decision about a scientific hypothesis can be made on the basis of incomplete information. Rational decisions often can be made without resorting to statistical inference, as when a scientific hypothesis concerns some limited phenomenon that is directly observable—for example, "This rat is running." The truth or falsity of this hypothesis can be determined by observing the rat. Many scientific hypotheses, however, refer to phenomena that cannot be directly observed or to populations that are so large that it is impossible or impractical to view all of their elements—for example, "All rats run under condition *X*." It is impossible to observe the entire population of rats under condition *X*. Likewise it is impossible to observe all parents rearing their children, all students who are active in student government, all smokers, or all insecure children. If a scientific hypothesis cannot be evaluated directly by observing all of the elements of a population, then it may be possible to evaluate the hypothesis indirectly by statistical inference. This evaluation involves observing a sample from the population of interest and making a rational decision about the probable truth or falsity of the scientific hypothesis. Classical statistical inference encompasses two complementary topics: hypothesis testing and confidence interval estimation. I consider hypothesis testing first.

The first step in evaluating a scientific hypothesis is to express the hypothesis in the form of a statistical hypothesis. You learned in Chapter 1 that a **statistical hypothesis** is a statement about one or more parameters of a population or the functional form of a population. For example, "$\mu \leq 115$" is a statistical hypothesis; it states that the population mean is less than or equal to 115. Another statistical hypothesis can be formulated that states that

*This section provides an elementary review of statistical inference. It assumes a prior exposure to both descriptive and inferential statistics. Readers who have a good grasp of statistical inference can omit this section. Those who want a more in-depth review can consult *Statistics: An Introduction* (Kirk, 2008).

the mean is greater than 115—that is, $\mu > 115$. These hypotheses, $\mu \leq 115$ and $\mu > 115$, are mutually exclusive and exhaustive; if one is true, the other must be false. They are examples, respectively, of the **null hypothesis,** denoted by H_0, and the **alternative hypothesis,** denoted by H_1. The null hypothesis is the one whose tenability is actually tested. If on the basis of this test the null hypothesis is rejected, then only the alternative hypothesis remains tenable. According to convention, the alternative hypothesis is formulated so that it corresponds to the researcher's scientific hunch. The process of choosing between the null and alternative hypotheses is called **hypothesis testing.**

The Role of Logic in Evaluating a Scientific Hypothesis

Logic plays a key role in evaluating a scientific hypothesis. This evaluation involves a chain of deductive and inductive logic that begins and ends with the scientific hypothesis. The chain is diagrammed in Figure 2.5-1. First, by means of deductive logic, the scientific hypothesis and its negation are expressed as two mutually exclusive and exhaustive statistical hypotheses that make predictions about one or more population parameters or the functional form of a population. These predictions, called the null and alternative hypotheses, are made about the population mean, median, variance, and so on. The next step in the chain is to obtain a random sample from the population and estimate the population parameters of interest. A statistical test is then used to decide whether the sample data are consistent with the null hypothesis. If the data are not consistent with the null hypothesis, the null hypothesis is rejected; otherwise, it is not rejected.

A **statistical test** involves (1) a test statistic, (2) a set of hypothesis-testing conventions, and (3) a decision rule that leads to an inductive inference about the probable truth or falsity of the scientific hypothesis, which is the final link in the chain shown in Figure 2.5-1. If errors occur in the deductive or inductive links in the chain of logic, the statistical

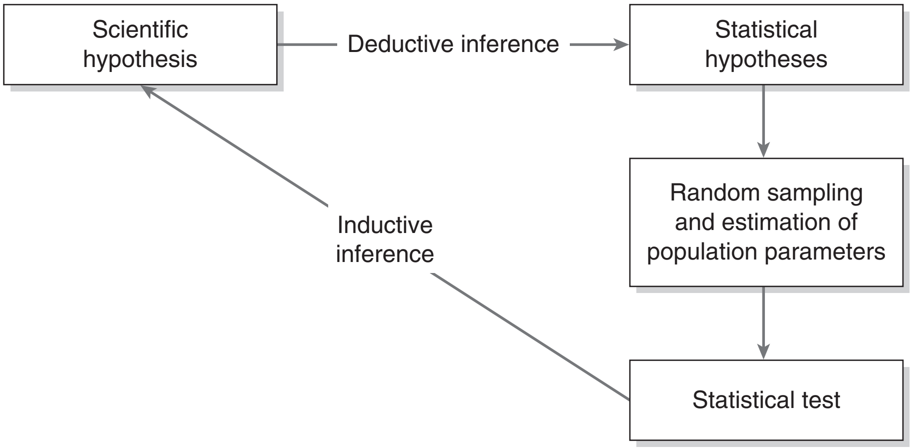


Figure 2.5-1 ■ The evaluation of a scientific hypothesis begins with a deductive inference and ends with an inductive inference concerning the probable truth or falsity of the scientific hypothesis.

hypothesis that is tested may have little or no bearing on the original scientific hypothesis, or the inference about the scientific hypothesis may be incorrect. Consider the scientific hypothesis that cigarette smoking is associated with high blood pressure. If this hypothesis is true, then a measure of central tendency such as mean blood pressure should be higher for the population of smokers than for nonsmokers. The statistical hypotheses are

$$H_0\colon \mu_1 - \mu_2 \leq 0$$

$$H_1\colon \mu_1 - \mu_2 > 0$$

where μ_1 and μ_2 denote the unknown population means for smokers and nonsmokers, respectively. The null hypothesis, $\mu_1 - \mu_2 \leq 0$, states in effect that the mean blood pressure of smokers is less than or equal to that of nonsmokers. The alternative hypothesis states that the mean blood pressure of smokers is greater than that of nonsmokers. These hypotheses follow logically from the original scientific hypothesis. Suppose the researcher formulates the statistical hypotheses in terms of population variances, for example,

$$H_0\colon \sigma_1^2 - \sigma_2^2 \leq 0$$

$$H_1\colon \sigma_1^2 - \sigma_2^2 > 0$$

where σ_1^2 and σ_2^2 denote the population variances of smokers and nonsmokers, respectively. A statistical test of this null hypothesis, which states that the variance of blood pressure for the population of smokers is less than or equal to the variance for nonsmokers, would have little bearing on the original scientific hypothesis. However, it would be relevant if the researcher was interested in determining whether the two populations differ in dispersion.

The reader should not infer that for any scientific hypothesis there is only one suitable null hypothesis. A null hypothesis that states that the correlation between the number of cigarettes smoked and blood pressure is equal to zero bears more directly on the scientific hypothesis than the one involving population means. If cigarette smoking is associated with high blood pressure, then there should be a positive correlation between cigarette consumption and blood pressure. The statistical hypotheses are

$$H_0\colon \rho \leq 0$$

$$H_1\colon \rho > 0$$

where ρ denotes the population correlation coefficient for cigarette consumption and blood pressure. So we see that both creativity and deductive skill are required to formulate relevant statistical hypotheses.

Sampling Distributions, the Central Limit Theorem, and Test Statistics

At this point in the review, I need to examine three concepts that play key roles in statistical inference: sampling distributions, the central limit theorem, and test statistics.

Sampling distribution. Inferential statistics are concerned with reasoning from a sample to the population—from the particular to the general. Such reasoning is based on a knowledge of the sample-to-sample variability of a statistic—that is, on its sampling behavior. Before data have been collected, I can speak of a sample statistic such as $\overline{Y}$ in terms of probability. Its value is yet to be determined and will depend on which observations happen to be randomly selected from the population. Thus, at this stage of an investigation, a sample statistic is a **random variable,**[8] because it is computed from observations obtained by random sampling. Like any random variable, a sample statistic has a probability distribution that gives the probability associated with each value of the statistic over all possible samples of the same size that could be drawn from the population. The distribution of a statistic is called a **sampling distribution** to distinguish it from the probability distribution for, say, an observation. In the discussion that follows, I focus on the sampling distribution of the mean.

Central limit theorem. The characteristics of the sampling distribution of the mean are succinctly stated in the **central limit theorem,** one of the most important theorems in statistics. In one form, the theorem states that if random samples are obtained from a population with mean μ and finite standard deviation σ, then as the sample size *n* increases, the distribution of $\overline{Y}$ approaches a normal distribution with mean μ and standard deviation (standard error) equal to $\sigma/\sqrt{n}$. Probably the most important point is that regardless of the shape of the sampled population, the means of sufficiently large samples will be nearly normally distributed. Just how large is sufficiently large? The answer depends on the shape of the sampled population; the more a population departs from the normal shape, the larger *n* must be. For most populations encountered in the behavioral sciences and education, a sample size of 50 to 100 is sufficient to produce a nearly normal sampling distribution of $\overline{Y}$. The tendency for the sampling distribution of a statistic to approach the normal distribution as *n* increases helps to explain why the normal distribution is so important in statistics.

Test statistics. It is important to distinguish between sample statistics and **test statistics.** The former are used to describe characteristics of samples or to estimate population parameters; the latter are used to test hypotheses about population parameters. An example of a test statistic is the *t* statistic:

$$t = \frac{\overline{Y} - \mu_0}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

where $\overline{Y}$ is the mean of a random sample from the population of interest, μ_0 is the hypothesized value of the population mean, $\hat{\sigma}$ is an estimator of the unknown population standard deviation, *n* is the size of the sample used to used to compute $\overline{Y}$ and $\hat{\sigma}$, and $\hat{\sigma}_{\overline{Y}} = \hat{\sigma}/\sqrt{n}$. $\hat{\sigma}_{\overline{Y}}$ is an estimator of the population standard error of the mean, $\sigma_{\overline{Y}}$.

Like all statistics, *t* has a sampling distribution. If the null hypothesis is true and *Y* is approximately normally distributed or *n* is large, *t* is distributed as the *t* distribution. The

[8] A random variable is a numerical quantity whose value is determined by the outcome of a random experiment.

distribution is unimodal and symmetrical about a mean of zero. The variance of the *t* distribution depends on the sample size or, more specifically, degrees of freedom. The term **degrees of freedom,** denoted by *df* or ν (Greek letter nu and pronounced "new"), refers to the number of scores whose values are free to vary, as I show in Section 3.2. For now I simply note that the degrees of freedom for the *t* statistic described above are equal to $\nu = n - 1$, and the variance of the *t* distribution is equal to $\nu/(\nu - 2)$. The *t* distribution is actually a family of distributions whose shapes depend on the number of degrees of freedom. A comparison of three members of the *t* family and the standard normal distribution is shown in Figure 2.5-2.

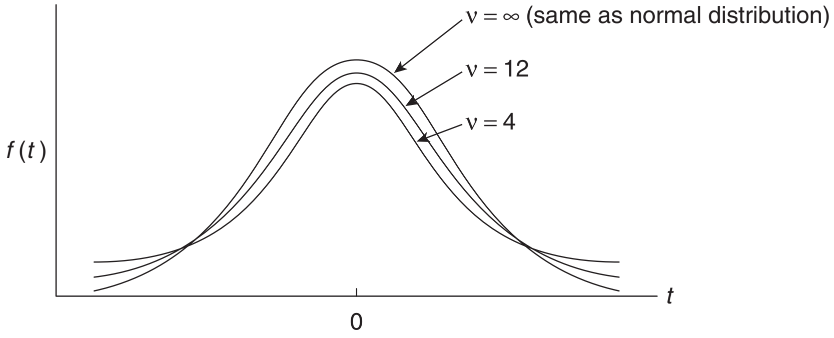


Figure 2.5-2 ▪ Graph of the *t* distribution for 4, 12, and ∞ degrees of freedom. The *t* distribution for $\nu = \infty$ is identical to the normal distribution.

Hypothesis Testing Using a One-Sample *t* Test Statistic

Suppose that I am interested in testing the scientific hypothesis that on the average, college students who are active in student government at Big Ten universities have higher IQs than college students who are not involved in student government. The corresponding statistical hypothesis is $\mu > \mu_0$, where μ denotes the unknown population mean of students who are active in student government and μ_0 denotes the population mean of college students who are not involved. Also assume that the mean IQ of noninvolved college students, μ_0, is known to equal 115. The first step in testing the scientific hypothesis is to state the null and alternative hypotheses:

$$H_0\text{: } \mu \leq 115$$

$$H_1\text{: } \mu > 115$$

where μ_0 has been replaced by 115, the known population mean IQ of college students who are not involved in student government. As written, the null hypothesis is inexact because it states a range of possible values for the population mean—all values less than or equal to 115. However, one exact value is specified—$\mu = 115$—and that is the value actually tested. If the null hypothesis $\mu = 115$ can be rejected, then any null hypothesis in which μ is less than 115 also can be rejected. This follows because if $\mu = 115$ is considered improbable because the sample mean exceeds 115, any population mean less than 115 is considered even less probable.

The second step in testing a scientific hypothesis is to specify the test statistic. A relatively small number of test statistics are used to evaluate hypotheses about population parameters. As you see in Section 3.1, the principal ones are denoted by z, t, χ^2 (chi square), and F. A test statistic is called a ***z* statistic** if its sampling distribution is the standard normal distribution, a test statistic is called a ***t* statistic** if its sampling distribution is a t distribution, and so on. The choice of a test statistic is determined by (1) the hypothesis to be tested, (2) information about the population that is known, and (3) assumptions about the population that appear to be tenable. Which test statistic should be used to test the hypothesis $\mu \leq 115$? Because the hypothesis concerns the mean of a single population, the population standard deviation is unknown, and the population is assumed to be approximately normal, the appropriate test statistic is

$$t = \frac{\overline{Y} - \mu_0}{\hat{\sigma} / \sqrt{n}}$$

The next step in the hypothesis testing process is to choose a sample size. I want the sample to be large enough but not too large. Later I show that there is a rational way of estimating the size of a sample that will be large enough to reject a false null hypothesis. For now, I resort to the time-honored practice of arbitrarily specifying a sample size that I think is large enough—say, 61. It turns out that this sample size is not large enough. Once the test statistic and sample size have been specified, the sampling distribution can be specified: It is the t sampling distribution for $n - 1 = 60$ degrees of freedom.

When a decision about a population is based on information from a sample, there is always the risk of making an error. I might decide that $\mu > 115$ when in fact $\mu \leq 115$. The fourth step in the hypothesis-testing process is to specify an acceptable risk of making this kind of error—that is, rejecting the null hypothesis when it is true. According to hypothesis-testing conventions, a probability of .05 is usually the largest risk a researcher should be willing to take of rejecting a true null hypothesis—deciding, for example, that $\mu > 115$ when in fact $\mu \leq 115$. Such a probability, called a **level of significance,** is denoted by the Greek letter α. For $\alpha = .05$ and H_1: $\mu > 115$, the region for rejecting the null hypothesis, called the **critical region,** is shown in Figure 2.5-3. The location and size of the critical region are determined, respectively, by the alternative hypothesis and α. A decision to adopt the .05, .01, or any other level of significance is based on hypothesis-testing conventions that have evolved since the 1920s. I return to the problem of selecting a significance level later.

The final step in testing H_0: $\mu \leq 115$ is to obtain a random sample of size 61 from the population of students who are active in student government at Big Ten universities, compute the test statistic, and make a decision. The **decision rule** is as follows:

> Reject the null hypothesis if the test statistic falls in the critical region; otherwise, do not reject the null hypothesis. If the null hypothesis is rejected, conclude that the mean IQ of students who are active in student government at Big Ten universities is higher than that of college students who are not involved; if the null hypothesis is not rejected, do not draw this conclusion.

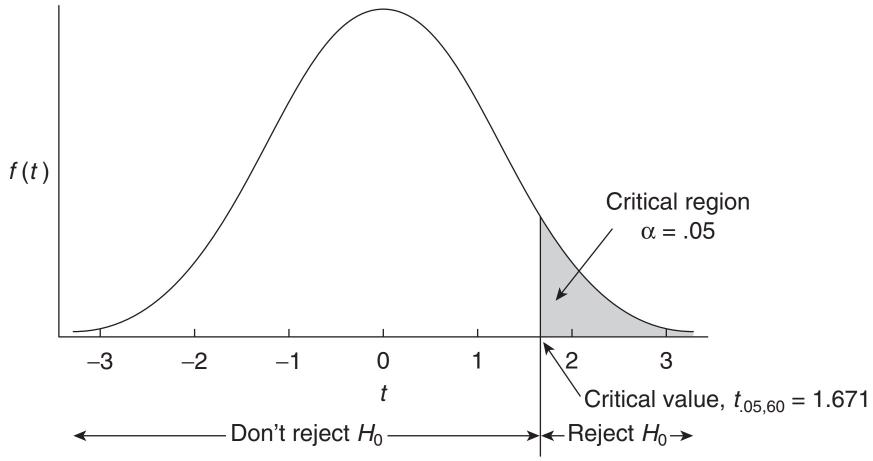


Figure 2.5-3 ■ Sampling distribution of the *t* statistic, given that the null hypothesis is true. The critical region, which corresponds in this example to the upper .05 portion of the sampling distribution, defines values of *t* that are improbable if the null hypothesis $\mu \leq 115$ is true. Hence, if the *t* statistic falls in the critical region, the null hypothesis is rejected. The value of *t* for $61 - 1 = 60$ degrees of freedom that cuts off the upper .05 portion of the sampling distribution is called the *critical value* and is denoted by $t_{.05,60}$. This value can be found in the *t* table in Appendix Table E.3 and is 1.671.

The procedures just described for testing H_0: $\mu \leq 115$ can be summarized in five steps and a decision rule:

Step 1. State the statistical hypotheses: H_0: $\mu \leq 115$
H_1: $\mu > 115$

Step 2. Specify the test statistic: $t = (\overline{Y} - \mu_0)/(\hat{\sigma}/\sqrt{n})$ because I want to test $\mu \leq 115$, σ is unknown, the sample is random, and I assume that the population of *Y* is approximately normal

Step 3. Specify the sample size: $n = 61$
and the sampling distribution: *t* distribution with $\nu = n - 1 = 60$ because σ is unknown and must be estimated from a sample, and I assume that the population distribution of *Y* is approximately normal

Step 4. Specify the level of significance: $\alpha = .05$

Step 5. Obtain a random sample of size *n,* compute *t,* and make a decision.

Decision rule:

Reject the null hypothesis if *t* falls in the upper 5% of the sampling distribution of *t;* otherwise, do not reject the null hypothesis. If the null hypothesis is rejected, conclude that the mean IQ of students who are active in student government at Big Ten universities is higher than that of college students who are not involved; if the null hypothesis is not rejected, do not draw this conclusion.

Computational Example for *t* Test

I now illustrate the use of

$$t = \frac{\overline{Y} - \mu_0}{\hat{\sigma} / \sqrt{n}}$$

where

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}}$$

to test the hypothesis $\mu \leq \mu_0$. Recall that $\overline{Y}$ is the mean of a random sample from the population of interest, μ_0 is the hypothesized value of the population mean, $\hat{\sigma}$ estimates the unknown population standard deviation, *n* is the size of the sample used to compute $\overline{Y}$ and $\hat{\sigma}$, and μ is the unknown mean of the population.

Assume that a random sample of 61 students who are active in student government has been obtained from the population of college students at the Big Ten universities and that the sample estimates of the population mean and standard deviation are, respectively, 117 and 12.5. The number 117 is called a **point estimate** of μ; it is the best guess I can make about the unknown value of μ. How improbable is a sample mean of 117 if the population mean is really 115? Would a mean of 117 occur five or fewer times in 100 experiments by chance? Stated another way, is it reasonable to believe that the population mean is really less than or equal to 115 if I have obtained a sample mean of 117? To answer this question, I compute a *t* statistic. If the *t* falls in the upper 5% of the sampling distribution of *t,* I have reason to believe that μ is not equal to 115. Such a result would occur five or fewer times in 100 by chance.

The *t* statistic for the example is

$$t = \frac{\overline{Y} - \mu_0}{\hat{\sigma} / \sqrt{n}} = \frac{117 - 115}{12.5 / \sqrt{61}} = \frac{2}{1.60} = 1.25$$

According to the *t* table in Appendix E.3, the value of *t* that cuts off the upper .05 region of the sampling distribution for 61 – 1 = 60 degrees of freedom is 1.671. This value of *t,* called the **critical value,** is denoted by $t_{.05,\ 60}$, where the subscript .05 refers to the proportion of the sampling distribution that lies above the critical value and 60 is the degrees of freedom associated with the denominator of the *t* statistic. Because $t = 1.25$ is less than $t_{.05,\ 60} = 1.671$, the observed *t* falls short of the upper .05 critical region. The critical region is shown in Figure 2.5-3. According to my decision rule, I fail to reject the null hypothesis and therefore conclude that the sample data do not support the hypothesis that the mean IQ of students who are active in student government at Big Ten universities is higher than that of college students who are not active in student government. Two points need to be emphasized. First, I have not proven that the null hypothesis is true—only that the evidence does not warrant its rejection. Second, my conclusion has been restricted to the population from which I sampled, namely, college students at Big Ten universities.

Three explanations can be advanced to account for my failure to reject the null hypothesis: (1) The null hypothesis is true and shouldn't be rejected; (2) the null hypothesis is false, but the t test lacked sufficient power to reject the null hypothesis; or (3) the null hypothesis is false, but the particular sample belied this fact—I obtained an unrepresentative sample from the population of students who are active in student government. In the following sections, I examine the second explanation for not rejecting the null hypothesis and discuss a number of concepts that round out my review of hypothesis testing.

One-Tailed and Two-Tailed Tests

A statistical test for which the critical region is in either the upper or the lower tail of the sampling distribution is called a **one-tailed test.** If the critical region is in both the upper and lower tails of the sampling distribution, the statistical test is called a **two-tailed test.**

A one-tailed test is used whenever the researcher makes a **directional prediction** about the phenomenon of interest—for example, that the mean IQ of students who are active in student government is higher than that of noninvolved students. The statistical hypotheses associated with this scientific hypothesis are

$$H_0\text{: } \mu \leq \mu_0$$

$$H_1\text{: } \mu > \mu_0$$

These hypotheses are called **directional hypotheses** or **one-sided hypotheses.** The region for rejecting the null hypothesis is shown in Figure 2.5-3. If the scientific hypothesis stated that students who are active in student government have lower IQs than noninvolved students, the following statistical hypotheses would be appropriate:

$$H_0\text{: } \mu \geq \mu_0$$

$$H_1\text{: } \mu < \mu_0$$

The region for rejecting this null hypothesis is shown in Figure 2.5-4(a). To reject the null hypothesis, an observed t statistic would have to be less than or equal to the critical value $-t_{.05,60} = -1.671$.

Often researchers do not have sufficient information to make a directional prediction about a population parameter; they simply believe the parameter is not equal to the value

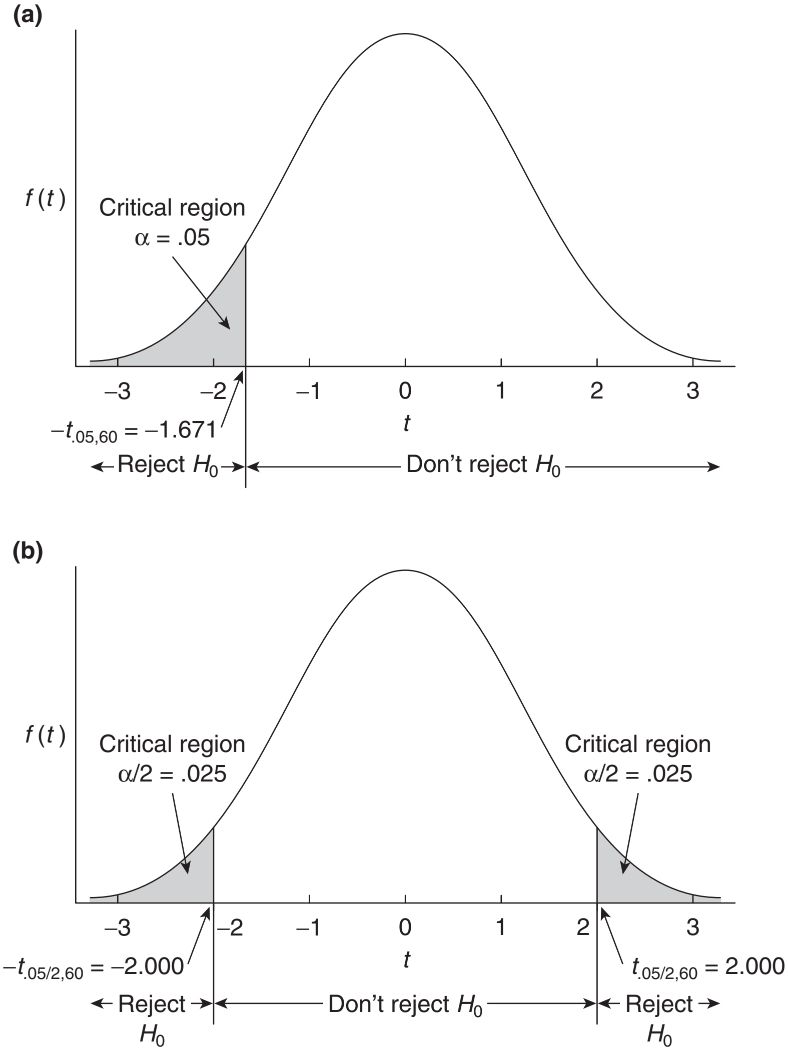


Figure 2.5-4 ■ (a) Critical region of the t statistic for a one-tailed test; H_0: $\mu \geq \mu_0$; H_1: $\mu < \mu_0$; $\alpha = .05$. (b) Critical regions for a two-tailed test; H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$; $\alpha = .025 + .025 = .05$.

specified by the null hypothesis. This situation calls for a two-tailed test. The statistical hypotheses for a two-tailed test have the following form:

$$H_0\colon \mu = \mu_0$$

$$H_1\colon \mu \neq \mu_0$$

These hypotheses are called **nondirectional hypotheses** or **two-sided hypotheses.** For two-sided hypotheses, the regions for rejecting the null hypothesis are in both the upper and the lower tails of the sampling distribution. The two-tailed critical regions are shown in Figure 2.5-4(b).

In summary, a one-sided or directional hypothesis is called for when the researcher's original hunch is expressed in such terms as "more than," "less than," "increased," or "decreased." Such a hunch indicates that the researcher has quite a bit of knowledge about the research area. This knowledge could come from previous research, a pilot study, or perhaps a theory. If the researcher is interested only in determining whether the independent variable affects the dependent variable without specifying the direction of the effect, a two-sided or nondirectional hypothesis should be used. Generally, significance tests in the behavioral sciences are two-tailed because most researchers lack the information necessary to formulate directional predictions.

One- and Two-Tailed Tests and Power

How does the choice of a one- or two-tailed test affect the probability of rejecting a false null hypothesis? A researcher is more likely to reject a false null hypothesis with a one-tailed test than with a two-tailed test if the critical region has been placed in the correct tail. A one-tailed test places all of the α area, say .05, in one tail of the sampling distribution. A two-tailed test divides the $\alpha = .05$ area between the two tails with .025 in one tail and .025 in the other tail. To illustrate, assume that $\alpha = .05$ and the following hypotheses for a two-tailed test have been advanced:

$$H_0\text{: } \mu = \mu_0$$

$$H_1\text{: } \mu \neq \mu_0$$

If the t statistic falls in either the upper or lower .025 region of the sampling distribution, the result is said to be significant at the .05 level of significance because .025 + .025 = .05. The values of t for 60 degrees of freedom that cut off the upper and lower .05/2 = .025 regions are $t_{.05/2,\,60} = 2.000$ and $-t_{.05/2,\,60} = -2.000$, respectively. An observed t statistic is significant at the .05 level if its value is greater than or equal to 2.000 or less than or equal to -2.000 or, more simply, if its absolute value, $|t|$, is greater than or equal to 2.000.

Now consider the hypotheses for a one-tailed test where the researcher believes that the population mean is less than μ_0.

$$H_0\text{: } \mu \geq \mu_0$$

$$H_1\text{: } \mu < \mu_0$$

Again, let $\alpha = .05$. If the t statistic falls in the lower tail of the sampling distribution—that is, if t is less than or equal to -1.671—the result is said to be significant at the .05 level of significance. The critical regions and critical values for the one- and two-tailed tests are shown in Figures 2.5-4(a) and (b), respectively. From an inspection of these figures, it should be apparent that the difference $\overline{Y} - \mu_0$ necessary to reach the critical region for a two-tailed test is larger than that required for a one-tailed test. Consequently, a researcher is less likely to reject a false null hypothesis with a two-tailed test than with a one-tailed test.

The term **power** refers to the probability of rejecting a false null hypothesis. A one-tailed test is more powerful than a two-tailed test if the researcher's hunch about the true difference $\mu - \mu_0$ is correct—that is, if the alternative hypothesis places the critical region in the correct tail of the sampling distribution. If the directional hunch is incorrect, the rejection region will be in the wrong tail, and the researcher will most certainly fail to reject the null hypothesis even though it is false. A researcher is rewarded for making a correct directional prediction and penalized for making an incorrect directional prediction. In the absence of sufficient information for using a one-tailed test, the researcher should play it safe and use a two-tailed test.

Type I and Type II Errors

When the null hypothesis is tested, a researcher's decision will be either correct or incorrect. An incorrect decision can be made in two ways. The researcher can reject the null hypothesis when it is true; this is called a **Type I error.** Alternatively, the researcher can fail to reject the null hypothesis when it is false; this is called a **Type II error.** Likewise, a correct decision can be made in two ways. If the null hypothesis is true and the researcher does not reject it, a **correct acceptance** has been made; if the null hypothesis is false and the researcher rejects it, a **correct rejection** has been made. The two kinds of correct decisions and the two kinds of errors are illustrated in Table 2.5-1.

The probability of making a Type I error is determined by the researcher when the level of significance, α, is specified. If α is specified as .05, the probability of making a Type I error is .05. The level of significance also determines the probability of a correct acceptance of a true null hypothesis because this probability is equal to $1 - \alpha$.

The probability of making a Type II error is denoted by β. The probability of making a correct rejection, the power, is denoted by $1 - \beta$. These two probabilities are determined by a number of variables: (1) the level of significance adopted, α; (2) the size of the sample, *n;* (3) the size of the population standard deviation, σ; and (4) the magnitude of the difference between μ and μ_0. The two probabilities also are affected by the researcher's decision to use

Table 2.5-1 ■ Decision Outcomes Categorized

		True Situation	
		H_0 true	H_0 false
Researcher's Decision	Fail to reject H_0	Correct acceptance Probability = $1 - \alpha$	Type II error Probability = β
	Reject H_0	Type I error Probability = α	Correct rejection Probability = $1 - \beta$

a one- or two-tailed test. To compute the probability of making a Type II error and power, it is necessary either to know μ, the true population mean, or to specify a value of μ that is sufficiently different from μ_0 to be of practical value. The latter approach is usually necessary because in any practical hypothesis-testing situation, μ is unknown. Also, the population standard deviation is rarely known. Sample data can be used to estimate this parameter.

Hypothesis testing involves a number of conventions. As you have seen, one convention is to set the probability of a Type I error equal to or less than .05. Another convention is to design an experiment so that the probability of a Type II error is equal to or less than .20. If $\beta = .20$, the power of the test is $1 - \beta = .80$. A power of .80 is considered by many researchers to be the minimum acceptable power. If the probability of rejecting a false null hypothesis is less than .80, many researchers would choose not to perform the experiment or would redesign the experiment so that its power is greater than or equal to .80. As you will see, there are ways to increase power. Before examining these, I compute the power of the test of the hypothesis that the mean IQ of college students who are active in student government is higher than that of students who are not involved. The statistical hypotheses are H_0: $\mu \leq 115$ and H_1: $\mu > 115$.

To compute the power of the *t* test in the student government example, it is necessary to know μ, the value of the population mean, or specify a value of μ that is sufficiently different from μ_0 to be worth detecting. I'll denote the value of the population mean that I am interested in detecting by μ'. Suppose that this value is $\mu' = 117.5$. I am saying in effect that any IQ difference less than $|\ \mu' - \mu_0\ | = |\ 117.5 - 115\ | = 2.5$ IQ points is too small to be of **practical significance.** Recall that for this example, $\hat{\sigma} = 12.5$, $n = 61$, and $t_{.05,\,60} = 1.671$. To compute an estimate of power, I need one more bit of information—the value of $\bar{Y}$ that cuts off the upper .05 region of the null hypothesis sampling distribution. I'll denote this mean by $\bar{Y}_{.05}$. I can compute $\bar{Y}_{.05}$ by rearranging the terms in the formula $t_{.05,\,60} = (\bar{Y}_{.05} - \mu_0)/(\hat{\sigma}/\sqrt{n})$ as follows:

$$\begin{aligned}\bar{Y}_{.05} &= \mu_0 + t_{.05,\,60}(\hat{\sigma}/\sqrt{n})\\ &= 115\ + (1.671)(12.5/\sqrt{61})\\ &= 117.67\end{aligned}$$

Thus, a mean of 117.67 cuts off the upper .05 region of the null hypothesis sampling distribution. In Figure 2.5-5, $\bar{Y}_{.05} = 117.67$ falls on the boundary between the reject and nonreject regions.

Estimates of the sizes of the regions corresponding to a Type II error and a correct rejection (labeled $\hat{\beta}$ and $1 - \hat{\beta}$ in Figure 2.5-5) can be obtained by computing a *t* statistic for the difference $\bar{Y}_{.05} - \mu' = 117.67 - 117.5$. The *t* statistic is

$$t = \frac{\bar{X}_{.05} - \mu'}{\hat{\sigma}/\sqrt{n}} = \frac{117.67 - 117.5}{12.5/\sqrt{61}} = \frac{0.17}{1.60} = 0.11$$

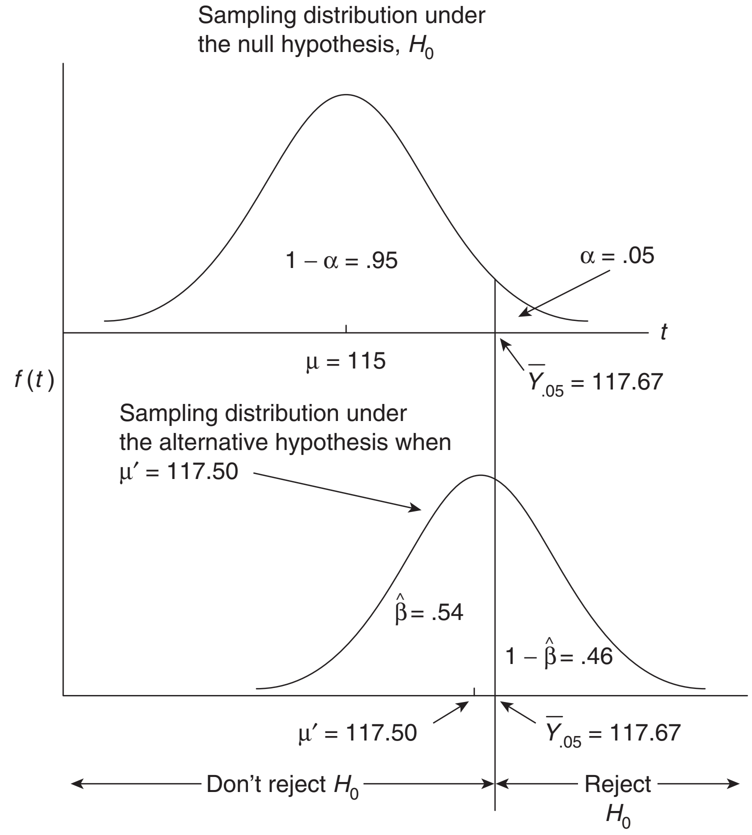


Figure 2.5-5 ▪ Regions that correspond to making a Type I error, α, and a Type II error, $\hat{\beta}$, if $\mu' = 117.50$. The mean that cuts off the upper .05 region of the sampling distribution under H_0 is denoted by $\bar{Y}_{.05}$ and is equal to 117.67. The value of $\bar{Y}_{.05}$ is obtained by rearranging the terms in the t formula as follows: $\bar{Y}_{.05} = \mu_0 + t_{.05,60}(\hat{\sigma} / \sqrt{n}) \;= 115 + 1.671\,(12.5 / \sqrt{61}) = 117.67$. If for a given n, H_0, and true H_1, the size of the α region is made smaller, the size of the $\hat{\beta}$ region increases.

The size of the area above $t = 0.11$, the $1 - \hat{\beta}$ region, can be obtained with the aid of a statistical calculator or computer. A simple way to find this area is to use Microsoft's Excel TDIST function. To access this function, select "Insert" in Excel's Menu bar and then the menu command "Function." You then select the TDIST function from the list of functions. After you access the TDIST function,

TDIST(x,deg_freedom,tails)

you replace "*x*" with the absolute value of the *t* statistic, "deg_freedom" with the degrees of freedom for the *t* statistic, and "tails" with 1 because the $1 - \hat{\beta}$ area always lies in only one of the distribution tails. The size of the $1 - \hat{\beta}$ area for the one-tailed $t = 0.11$ with 60 degrees of freedom is obtained from

TDIST(0.11,60,1)

and is equal to .46. Thus, if the mean IQ of students who are active in student government is $\mu' = 117.5$, the probability of making a correct rejection (power) is $1 - \hat{\beta} = .49$. The probability of making a Type II error ($\hat{\beta}$) is $1 - (1 - \hat{\beta}) = 1 - .46 = .54$. Figure 2.5-5 shows the regions corresponding to these two probabilities. A power of .46 is considerably less than .80, the minimum acceptable power according to convention. Table 2.5-2 summarizes the possible decision outcomes when $\mu = 115$ and when $\mu' = 117.5$. In this example, the size of the region corresponding to making a correct decision is larger when the null hypothesis is true ($1 - \alpha = .95$) than when the null hypothesis is false ($1 - \hat{\beta} = .46$). It also is apparent that the size of the region corresponding to making a Type I error ($\alpha = .05$) is much smaller than the probability of making a Type II error ($\hat{\beta} = .54$). In most research situations, the researcher follows the convention of setting $\alpha = .05$ or $\alpha = .01$. This convention of choosing a small numerical value for α is based on the notion that making a Type I error is bad and should be avoided. Unfortunately, as the probability of a Type I error is made smaller and smaller, the probability of a Type II error increases and vice versa. This can be seen from an examination of Figure 2.5-5. If the vertical line cutting off the upper α region is moved to the right or to the left, the region designated $\hat{\beta}$ in the lower distribution is made, respectively, larger or smaller.

Table 2.5-2 ■ Decision Outcomes Categorized

		True Situation	
		$\mu = 115$	$\mu' = 117.5$
Researcher's Decision	$\mu \le 115$	Correct acceptance $1 - \alpha = 1 - .05$ $= .95$	Type II error $\hat{\beta} = .54$
	$\mu > 115$	Type I error $\alpha = .05$	Correct rejection $1 - \hat{\beta} = 1 - .54$ $= .46$

As you have seen, four factors determine the power of a test: (1) the level of significance, α; (2) the size of the sample, *n;* (3) estimate of the population standard deviation, $\hat{\sigma}$; and (4) the magnitude of the difference between μ' and μ_0. The power of a test can be increased by making an appropriate change in any of these factors. For example, power can be increased by adopting a larger value for α, say .10 or .20; increasing the sample size; refining the experimental design or measuring procedure so as to decrease the population standard deviation estimate; and increasing the difference between μ' and μ_0 that is considered worth detecting. Often the simplest way to increase the power of a statistical test is to increase the sample size.

Hypothesis testing involves a number of conventions. I hope that this discussion has dispelled the magical aura that surrounds the .05 level of significance; its use in hypothesis testing is simply a convention. A level of significance is the probability of committing an error in rejecting a true null hypothesis. It says nothing about the importance or practical significance of a result.[9]

Effect Magnitude

Researchers want to answer two basic questions from their research: (1) Is an observed treatment effect real, or should it be attributed to chance? and (2) If the effect is real, how large is it? The first question concerning whether chance is a viable explanation for an observed treatment effect is usually addressed with a null hypothesis significance test. A significance test tells the researcher the probability of obtaining the effect or a more extreme effect if the null hypothesis is true. The test does not address the question of how large the effect is. This question is usually addressed with descriptive statistics and measures of effect magnitude. The most widely used measures of effect magnitude fall into one of two categories: measures of *effect size* and measures of *strength of association*. I describe Cohen's and Hedge's measures of effect size here and defer a discussion of measures of effect magnitude to Section 4.4.

In 1969, Cohen introduced the first effect-size measure that was explicitly labeled as such. His measure, denoted by *d*, expresses the size of the absolute difference $\mu - \mu_0$ in units of the population standard deviation,

$$d = \frac{|\mu - \mu_0|}{\sigma}$$

Cohen recognized that the size of the difference $\mu - \mu_0$ is influenced by the scale of measurement of the means. Cohen divided the difference between the means by σ to rescale the difference in units of the amount of variability in the data. What made Cohen's contribution unique is that he provided guidelines for interpreting the magnitude of *d*.

$d = 0.2$ is a small effect.

$d = 0.5$ is a medium effect.

$d = 0.8$ is a large effect.

[9] For an in-depth examination of practical significance, see Kirk (1996).

According to Cohen (1992), a medium effect of 0.5 is visible to the naked eye of a careful observer. Several surveys have found that 0.5 approximates the average size of observed treatment effects in various fields. A small effect of 0.2 is noticeably smaller than medium but not so small as to be trivial. A large effect of 0.8 is the same distance above medium as small is below it. These operational definitions turned his measure of effect size into a much more useful statistic. A sample estimator of Cohen's *d* is obtained by replacing μ with $\overline{Y}$ and σ with $\hat{\sigma}$.

$$d = \frac{|\overline{Y} - \mu_0|}{\hat{\sigma}}$$

For experiments with two sample means, Larry Hedges proposed a modification of Cohen's *d* as follows:

$$g = \frac{|\overline{Y}_1 - \overline{Y}_2|}{\hat{\sigma}_{\text{Pooled}}}$$

where

$$\hat{\sigma}^2_{\text{Pooled}} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)}$$

is a pooled estimator of the unknown population standard deviation. Hedges's *g* is interpreted the same as Cohen's *d*.

From Cohen's rule of thumb, a small effect for the student IQ data is one for which $|\mu - \mu_0| = |117.5 - 115| = 2.5$, because

$$d = \frac{|117.5 - 115|}{12.5} = \frac{2.5}{12.5} = 0.2$$

Similarly, medium and large effects correspond to $|121.25 - 115| = 6.25$ and $|125 - 115| = 10$, respectively, because

$$d = \frac{|121.25 - 115|}{12.5} = \frac{6.25}{12.5} = 0.5$$

$$d = \frac{|125 - 115|}{12.5} = \frac{10}{12.5} = 0.8$$

An effect size is a valuable supplement to the information provided by a null hypothesis significance test. It helps to provide a context for interpreting research results. A test that is significant at the .0001 level of significance loses its luster if the effect turns out to be trivial.

Reporting *p* Values

Most research reports and computer printouts contain a statistic called a **probability value,** or simply ***p* value.** A *p* value is the probability of obtaining a value of the test statistic that is equal to or more extreme than the one observed, given that the null hypothesis is true. Usually p values are obtained with the aid of a statistical calculator or computer statistical program. p values for a variety of sampling distributions also can be obtained with Microsoft's Excel program. To illustrate, I use the Excel TDIST function

TDIST(x,deg_freedom,tails)

described earlier to obtain the p value of the t statistic for the students who are active in student government. After the TDIST function has been accessed, I replace x with the value of t, which is 1.25, degrees of freedom with 60, and tails with 1 as follows:

TDIST(1.25,60,1)

Excel returns the p value of .108.

In reporting the results of hypothesis tests in the text portion of publications, the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010) recommends reporting in order the test statistic that was used, the degrees of freedom (in parentheses) associated with the test statistic, the value of the test statistic, the exact p value to two or three decimal places, and effect size. For example, in describing the results of the college student experiment, I could report that "the difference between the means of students who are active in student government and those who are not active, 117 – 115 = 2, was not statistically significant, $t(60) = 1.25, p = .108, d = 0.16$." The *Publication Manual* says to "report p value less than .001 as $p < .001$" (American Psychological Association, 2010, p. 114).

Earlier, I formulated a hypothesis-testing decision rule in terms of the test statistic and the critical region: Reject the null hypothesis if the test statistic falls in the critical region; otherwise, do not reject the null hypothesis. A decision rule also can be formulated in terms of a p value and a level of significance. The rule is as follows:

> Decision rule:
>
> Reject the null hypothesis if the p value is less than or equal to the preselected level of significance, that is, reject the null hypothesis if $p \leq \alpha$; otherwise, do not reject the null hypothesis.

The inclusion of a p value in a research report provides useful information because it enables a reader to discern those significance levels for which the null hypothesis could have been rejected. The p values provided in computer printouts are usually appropriate for two-sided hypotheses. If your null hypothesis is directional, the two-tailed p value given in the computer printout should be divided by 2. Of course, the p value for a one-sided hypothesis is only meaningful if the data are consistent with the alternative hypothesis. Before leaving the subject of p values, I want to emphasize that a p value is related to statistical significance; it says nothing about practical significance.

Confidence Interval Estimation

Null hypothesis significance testing is the dominant approach to statistical inference in psychology, education, and the medical sciences. There is a growing awareness among researchers that this approach has some shortcomings.[10] A null hypothesis significance test addresses the question, Is chance a likely explanation for the results that have been obtained? The test does not address the question, Are the results important or useful? There are other criticisms. For example, null hypothesis significance testing and scientific inference address different questions. In scientific inference, what you want to know is the conditional probability that the null hypothesis (H_0) is true, given that you have obtained a set of data (D); that is, Prob($H_0|D$). What null hypothesis significance testing tells you is the conditional probability of obtaining these data or more extreme data if the null hypothesis is true, Prob($D|H_0$). Obtaining data for which Prob($D|H_0$) is low does not imply that Prob($H_0|D$) also is low.

A third criticism of null hypothesis significance testing is that it is a trivial exercise. John Tukey (1991) said, "All we know about the world teaches us that the effects of A and B are always different—in some decimal place. Thus asking 'Are the effects different?' is foolish" (p. 100). Hence, because all null hypotheses are false, Type I errors cannot occur and statistically significant results are ensured if large enough samples are used. Bruce Thompson (1998) captured the essence of this view when he wrote, "Statistical testing becomes a tautological search for enough subjects to achieve statistical significance. If we fail to reject, it is only because we've been too lazy to drag in enough subjects" (p. 799). In the real world, all null hypotheses are false. Hence, a decision to reject simply means that the research methodology had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect.

The list of criticisms goes on. I mention one more. By adopting a fixed significance level such as $\alpha = .05$, a researcher turns a continuum of uncertainty about a true state of affairs into a dichotomous reject/do-not-reject decision. Researchers ordinarily react to a p value of .06 with disappointment and even dismay, but not p values of .05 or smaller. Rosnow and Rosenthal's (1989) comment is pertinent: "Surely, God loves the .06 nearly as much as the .05." Many psychologists believe that an emphasis on null hypothesis significance tests and p values distracts researchers from the main business of science—understanding and interpreting the outcomes of research. An alternative approach to statistical inference using confidence intervals is described next.

A **confidence interval** is a segment or interval on the real number line that has a high probability of including a population parameter. Confidence intervals can be either one- or two-sided. A one-sided confidence interval is constructed when the researcher has made a directional prediction about the population mean; otherwise, a two-sided interval is constructed.

[10] Cumming (2012), Kline (2004), and Nickerson (2000) provide in-depth discussions of these shortcomings.

The computation of a one-sided confidence interval is illustrated for the mean IQ of college students who are active in student government at Big Ten universities. Recall that the null hypothesis is

$$H_0\colon \mu \le 115$$

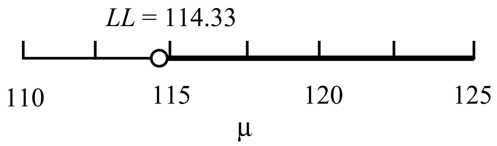
Also, $\bar{Y} = 117$, $\hat{\sigma} = 12.5$, $n = 61$, and $\alpha = .05$. A one-sided 95% confidence interval for the population mean is[11]

$$\bar{Y} - \frac{t_{.05,60}\hat{\sigma}}{\sqrt{n}} < \mu$$

$$117 - \frac{1.671(12.5)}{\sqrt{61}} < \mu$$

$$114.33 < \mu$$

The number 114.33 is the lower limit of the open confidence interval[12] and is denoted by *LL.* I can be fairly confident that the population mean is greater than 114.33. The degree of my confidence is represented by the **confidence coefficient** 100(1 – .05)% = 95%, where $\alpha = .05$ is the level of significance. It helps to visualize the confidence interval as a segment on the number line as follows:



The confidence interval indicates values of the parameter μ that are consistent with the observed sample statistic. It also contains a range of values for which the null hypothesis is nonrejectable at the .05 level of significance. To put it another way, the confidence interval can be used to test all one-sided hypotheses of interest, not just $H_0\colon \mu \le 115$. For example, I know that $H_0\colon \mu \le 113$ and $H_0\colon \mu \le 112$ would be rejected, but not $H_0\colon \mu \le 115$ or $H_0\colon \mu \le 116$. These decisions follow because 113 and 112 are not included in the confidence interval, whereas 115 and 116 are included.

[11] See Kirk (2008, pp. 294–295) for the derivation of the confidence interval.

[12] An interval in which the endpoint is not included is called an **open interval.**

A two-sided confidence interval can be constructed that is analogous to a two-tailed test of significance. The interval is given by

$$\bar{Y} - \frac{t_{\alpha/2,\nu}\hat{\sigma}}{\sqrt{n}} < \mu < \bar{Y} + \frac{t_{\alpha/2,\nu}\hat{\sigma}}{\sqrt{n}}$$

Suppose that I had proposed a two-sided null hypothesis, H_0: $\mu = 115$, for the students who are active in student government. A two-sided 100(1 – .05)% = 95% confidence interval for the population mean is

$$117 - \frac{2.000(12.5)}{\sqrt{61}} < \mu < 117 + \frac{2.000(12.5)}{\sqrt{61}}$$

$$113.80 < \mu < 120.20$$

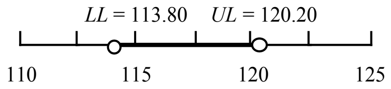
where the lower and upper confidence limits are, respectively, $LL = 113.80$ and $UL = 120.20$. I can be 95% confident that the open interval 113.80 to 120.20 contains the population mean. The *Publication Manual of the American Psychological Association* (American Psychological Association, 2010) says to express the confidence interval in the text portion of a publication as 95% CI[113.80, 120.20].

I could increase my confidence that the interval includes the population mean by replacing $t_{.05/2,60}$ with $t_{.01/2,60}$. The resulting interval

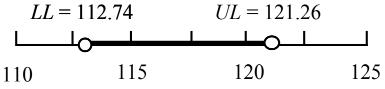
$$117 - \frac{2.660(12.5)}{\sqrt{61}} < \mu < 117 + \frac{2.660(12.5)}{\sqrt{61}}$$

$$112.74 < \mu < 121.26$$

is a 100(1 – .01)% = 99% confidence interval. Notice that as my confidence that I have captured μ increases, so does the size of the interval. This is illustrated in the following figures.



95% confidence interval for μ



99% confidence interval for μ

Confidence interval procedures and hypothesis-testing procedures involve the same assumptions. And both procedures can be used to test null hypotheses. However, confidence interval procedures provide more information about one's data than do hypothesis-testing procedures. A sample mean and confidence interval provide an estimate of the population parameter and a range of values—the error variation—that qualifies the estimate. A $100(1 - \alpha)\%$ confidence interval for μ contains all the values of μ_0 for which the null hypothesis would *not* be rejected at an α level of significance. All values of μ_0 outside the confidence interval would be rejected.

2.6 Review Exercises

1. Terms to remember:

a. *t* test for independent samples (2.2)
b. completely randomized design (2.2)
c. experimental design model equation (2.2)
d. linear model (2.2)
e. repeated measures (2.2)
f. subject matching (2.2)
g. *t* test for dependent samples (2.2)
h. block (2.2)
i. randomized block design (2.2)
j. Latin square design (2.2)
k. building block design (2.2)
l. completely randomized factorial design (2.2)
m. treatment combination (2.2)
n. completely crossed treatments (2.2)
o. randomized design (2.3)
p. scientific hypothesis (2.5)
q. statistical inference (2.5)
r. statistical hypothesis (2.5)
s. null hypothesis (2.5)
t. alternative hypothesis (2.5)
u. hypothesis testing (2.5)
v. statistical test (2.5)
w. random variable (2.5)
x. sampling distribution (2.5)
y. central limit theorem (2.5)
z. test statistic (2.5)
aa. degrees of freedom (2.5)
ab. *z* statistic (2.5)
ac. *t* statistic (2.5)
ad. level of significance (2.5)
ae. critical region (2.5)
af. decision rule (2.5)
ag. point estimate (2.5)
ah. critical value (2.5)
ai. one- and two-tailed tests (2.5)
aj. directional prediction (2.5)
ak. directional hypothesis (2.5)
al. nondirectional hypothesis (2.5)
am. power (2.5)
an. Type I and II errors (2.5)
ao. correct acceptance (2.5)
ap. correct rejection (2.5)
aq. practical significance (2.5)
ar. probability (*p*) value (2.5)
as. confidence interval (2.5)
at. confidence coefficient (2.5)
au. open interval (2.5)

*2. [2.2] For each of the following experiments or investigations, indicate (i) the type of experimental design, (ii) the null hypothesis (exclude nuisance variables), and (iii) the experimental design model equation.

*a. The effects of three kinds of instruction on first-grade students' tendency to help another child were investigated. Forty-two boys were randomly assigned to the three kinds of instructions denoted by a_1, a_2, and a_3 with the restriction that 14 boys were assigned to each kind of instruction. Boys in the a_1 instruction group (indirect responsibility group) were told that there was another boy alone in an adjoining room who had been told not to climb on a chair. Boys in the a_2 group were told the same story and in addition were told that they were being left in charge and to take care of anything that happened (direct responsibility group). All of the boys were given a simple task to perform. Shortly after the researcher left the room, there was a loud crash in the adjoining room followed by a minute of crying and sobbing. Boys in the a_3 group were given the same instructions as those in group a_2, but the sounds from the adjoining room included calls for help (second direct responsibility group). The behaviors of the boys were observed from behind a one-way mirror and rated in terms of the amount of help offered: 1 = offered no help, . . . , 5 = went to the adjoining room. (Experiment suggested by Staub, E. A child in distress: The effect of focusing of responsibility on children on their attempts to help. *Developmental Psychology*.)

*b. Forty-five executives were assigned to one of nine categories on the basis of their years of experience (b_1 is less than 3 years, b_2 is 3 to 6 years, b_3 is more than 6 years) and educational attainment (c_1 is less than 3 years of college, c_2 is a college graduate, c_3 has some graduate work). Five executives were in each category. The independent variable was type of training used to increase speed in composing complex business letters (a_1 is preparing an outline of a letter prior to dictating it, a_2 is making a list of the major points to be covered prior to dictating a letter, and a_3 is silently dictating a letter prior to dictating it). Treatment level a_1 was paired with b_1 and c_1, b_2 and c_3, and b_3 and c_2; a_2 was paired with b_1 and c_2, and so on. The dependent variable was the average time taken to dictate five letters following 2 weeks of practice with the assigned practice procedure.

*c. The effects of isolation at 90, 120, 150, and 180 days of age on subsequent combative behavior of Mongolian gerbils (*Meriones unguiculatus*) were investigated. Eighty gerbils were randomly assigned to one of the four isolation conditions with 20 in each condition. The number of combative encounters was recorded when the gerbils were 2 years old. It was hypothesized that the number of combative encounters would increase with earlier isolation.

d. Scores on the Conforming-Compulsive scale of the Millon Clinical Multiaxial Inventory are known to be positively correlated with the dependent variable in Exercise 2(a). Suppose that this scale was used to form blocks of three boys who had similar Conforming-Compulsive scores and that the three boys in each block were randomly assigned to the three treatment conditions.

e. Dreams of a random sample of 50 English-Canadian and 50 French-Canadian students were analyzed in terms of the proportion of achievement-oriented elements such as achievement imagery, success, and failure. The students were matched on the basis of reported frequency of dreaming. It was hypothesized that the French-Canadian students' dreams would contain a higher proportion of achievement-oriented elements.

f. Pictures of human faces posing six distinct emotions (treatment *A*) were obtained. The faces and their mirror reversals were split down the midline, and left-side and right-side composites were constructed. This resulted in 12 pictures. Six hundred introductory psychology students were randomly assigned to one of the 12 groups with 50 in each group. Each student rated one of the 12 pictures on a 7-point scale in terms of the intensity of the emotion expressed. It was hypothesized that the left-side composites would express the most intense emotion.

g. Ninety Sprague-Dawley rats performed a simple operant barpress response and were given partial reinforcement, partial reinforcement followed by continuous reinforcement, or partial reinforcement preceded by continuous reinforcement. The dependent variable was rate of responding following the training condition. The rats were randomly assigned to the experimental conditions; there were 30 rats in each condition.

3. [2.2] In your own words, describe what is meant by the terms (a) grand mean, (b) treatment effect, and (c) error effect.

*4. [2.2] *(a) Under what conditions is the sum of the squared error effects for the randomized block and Latin square designs less than the sum for the completely randomized design?

(b) Discuss the relative merits of completely randomized, randomized block, and Latin square designs.

*5. [2.2] List the treatment combinations for the following completely randomized factorial designs.

*a. CRF-24 *b. CRF-222

c. CRF-33 d. CRF-42

e. CRF-322

*6. [2.2] Construct block diagrams similar to those in Figures 2.2-1 through 2.2-7 for the following designs.

*a. *t* test for independent samples, $n = 10$ *b. CR-3 design, $n = 10$

*c. CRF-32 design, $n = 3$ d. CR-5 design, $n = 6$

e. *t* test for dependent samples, $n = 7$ f. RB-4 design, $n = 6$

g. CRF-222 design, $n = 3$ h. LS-3 design, $n = 3$

*7. [2.2] Construct block diagrams similar to those in Figures 2.2-1 through 2.2-7 for the designs in Exercise 2a–c.

8. [2.2] Construct block diagrams similar to those in Figures 2.2-1 through 2.2-7 for the designs in Exercise 2d–g.

9. [2.2] The following data on running time (in seconds) in a straight-alley maze were obtained in a CRF-33 design.

	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
$\overline{Y}_{\cdot jk} =$	9	8	5	7	7	5	6	5	5

Magnitude of reinforcement	Hours of deprivation
a_1 = small	$b_1 = 10$
a_2 = medium	$b_2 = 15$
a_3 = large	$b_3 = 20$

a. Graph the interaction.

b. Give a verbal description of the interaction.

*10. [2.3] What is the major difference between systematic designs and randomized designs?

*11. [2.3] Criticize the statement: The subjects were randomly assigned to the eight treatment combinations in a two-treatment factorial design.

12. [2.5] (a) Distinguish between a scientific hypothesis and a statistical hypothesis.

(b) According to convention,

i. Which of the statistical hypotheses corresponds to the researcher's scientific hunch?

ii. Which of the statistical hypotheses is actually tested?

*13. [2.5] Distinguish among the following concepts.

*a. Sample distribution, population distribution, and sampling distribution

*b. Sample statistic and test statistic

*14. [2.5] Suppose that a researcher has a hunch that pregnant women who use drugs have babies who weigh less at birth than do drug-free women. Let μ_1 and μ_2 denote the two population means, respectively.

*a. List the steps that you would use to test the null hypothesis. Let $\alpha = .05$ and $n_1 = n_2 = 50$. The two-sample test statistic is $t = (\overline{Y}_1 - \overline{Y}_2)/\sqrt{\hat{\sigma}^2_{\text{Pooled}}(1/n_1 + 1/n_2)}$, where $\hat{\sigma}^2_{\text{Pooled}} = [(n_1 - 1)\hat{\sigma}^2_1 + (n_2 - 1)\hat{\sigma}^2_2]/[(n_1 - 1) + (n_2 - 1)]$ and $\nu = n_1 + n_2 - 2$.

*b. State the decision rule.

*c. Draw the sampling distribution associated with the null hypothesis and indicate the region or regions that lead to rejection and nonrejection of the null hypothesis.

*d. Suppose that the researcher has obtained the following statistics: $\overline{Y}_1 = 5.5$, $\overline{Y}_2 = 7.0$, and $\hat{\sigma}_{\text{Pooled}} = 3.0$. Compute a t statistic and use Microsoft's Excel TDIST function to determine the p value of the statistic. Write a sentence that summarizes the results of the research.

*e. Compute and interpret the effect size where $g = \left|\overline{Y}_1 - \overline{Y}_2\right| / \hat{\sigma}_{\text{Pooled}}$.

*f. Use the formula

$$\mu_1 - \mu_2 < (\overline{Y}_1 - \overline{Y}_2) + t_{.05,98}\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

to compute a one-sided 95% confidence interval for $\mu_1 - \mu_2$; let $t_{.05,\,98} = 1.66$. Interpret the confidence interval. What does the confidence interval tell you about the tenability of the null hypothesis $\mu_1 - \mu_2 \geq 0$?

*g. If the researcher believed that the minimum difference between the population means that was worth detecting was $\delta_0' = \mu_1' - \mu_2' = -1.5$, estimate the power of the research methodology. Let $\overline{Y}_{.05} = \delta_0 - t_{.05,98}\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and $t = \dfrac{\overline{Y}_{.05} - \delta_0'}{\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $\delta_0 = \mu_1 - \mu_2 = 0$.

*h. Make a table like Table 2.5-2 that summarizes the sizes of the regions of the sampling distributions associated with the four possible decision outcomes.

15. [2.5] For the past several years, the mean arithmetic achievement score for ninth-graders has been 45 with a standard deviation, $\hat{\sigma}$, equal to 15. After participating in an experimental teaching program, a random sample of 27 students had a mean score of 52.5 Let μ denote the population mean of the children who participated in the experimental program.

a. List the steps that you would use to test a two-sided null hypothesis. Let $\alpha = .05$.

b. State the decision rule.

c. Draw the sampling distribution associated with the null hypothesis and indicate the region or regions that lead to rejection and nonrejection of the null hypothesis. Compute and interpret the effect size.

d. Compute a t statistic and use Microsoft's Excel TDIST function to determine the p value of the statistic. Write a sentence summarizing the results of the research.

e. Compute and interpret the effect size where $d = \left|\overline{Y} - \mu_0\right| / \hat{\sigma}$.

f. Use the formula

$$\overline{Y} - \frac{t_{.05/2,26}\hat{\sigma}}{\sqrt{n}} < \mu < \overline{Y} + \frac{t_{.05/2,26}\hat{\sigma}}{\sqrt{n}}$$

to compute a two-sided 95% confidence interval for μ. Interpret the confidence interval. What does the confidence interval tell you about the tenability of the null hypothesis $\mu_1 - \mu_2 = 0$?

g. If the researcher believed that the minimum population mean that was worth detecting was $\mu' = 52.5$, estimate the power of the research methodology.

h. Make a table like Table 2.5-2 that summarizes the sizes of the regions of the sampling distributions associated with the four possible decision outcomes.

*16. [2.5] Indicate the type of error or correct decision for each of the following.

*a. A true null hypothesis was rejected.

*b. The researcher failed to reject a false null hypothesis.

c. The null hypothesis is false and the researcher rejected it.

d. The researcher did not reject a true null hypothesis.

e. A false null hypothesis was rejected.

f. The researcher rejected the null hypothesis when he or she should have failed to reject it.

17. List the ways that a researcher can increase the power of an experiment. What are their relative merits?

18. [2.5] The effect of playing video racing games or neutral games on cognitions associated with risk taking was investigated. The games were played on a Sony PlayStation. Sales rankings in computer magazines were used to select the most popular games in each category. Forty-seven men at Ludwig-Maximilians University, Munich, Germany, were randomly assigned to the two types of games. The dependent variable was a paper-and-pencil measure of risk-related cognitions. The means, standard deviations, and sample sizes for the men who played the video racing games and the neutral games were, respectively, $\overline{Y}_1 = 7.54, \hat{\sigma}_1 = 1.3$, and $n_1 = 24$ and $\overline{Y}_2 = 6.41$, $\hat{\sigma}_2 = 1.2$, and $n_1 = 23$. (Suggested by Fletcher, P., & Kubitzki, J. Virtual driving and risk taking: Do racing games increase risk-taking cognitions, affect, and behaviors? *Journal of Experimental Psychology, Applied.*)

a. List the steps that you would use to test the hypothesis that playing video racing games increases risk-related cognitions. Let $\alpha = .05$.

b. State the decision rule.

c. Draw the sampling distribution associated with the null hypothesis and indicate the region or regions that lead to rejection or nonrejection of the null hypothesis.

d. Compute a t statistic and use Microsoft's Excel TDIST function to determine the p value of the statistic. Write a sentence that summarizes the results of the research.

e. Compute and interpret the effect size where $g = \left|\overline{Y}_1 - \overline{Y}_2\right| / \hat{\sigma}_{\text{Pooled}}$.

f. Use the formula

$$(\overline{Y}_1 - \overline{Y}_2) - t_{.05,45}\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \mu_1 - \mu_2$$

to compute a one-sided 95% confidence interval for $\mu_1 - \mu_2$. Interpret the confidence interval. What does the confidence interval tell you about the tenability of the null hypothesis $\mu_1 - \mu_2 \leq 0$?

g. If the researcher believed that the minimum difference between the population means that was worth detecting was $\delta_0' = \mu_1' - \mu_2' = 1.0$, estimate the power of the research methodology. Let $\overline{Y}_{.05} = \delta_0 + t_{.05,45}\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and $t = \dfrac{\overline{Y}_{.05} - \delta_0'}{\sqrt{\hat{\sigma}^2_{\text{Pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $\delta_0 = \mu_1 - \mu_2 = 0$.

h. Make a table like Table 2.5-2 that summarizes the sizes of the regions of the sampling distributions associated with the four possible decision outcomes.

*19. [2.5] Use Microsoft's Excel TDIS function, TDIS(x,deg_freedom,tails), to determine the p value for the following t statistics.

*a. $t = 2.463$, $n = 32$, H_0: $\mu \leq \mu_0$ *b. $t = 2.761$, $n = 39$, H_0: $\mu = \mu_0$

c. $t = 3.553$, $n = 46$, H_0: $\mu \leq \mu_0$ d. $t = 1.659$, $n = 42$, H_0: $\mu = \mu_0$

*20. [2.5] Use Microsoft's Excel TINV function, TINV(probability,deg_freedom), to determine the value of t that cuts off the critical region for the following significance levels. The TINV function provides two-tailed t values. For one-tailed significance levels, use 2α in place of α.

*a. $\alpha = .05$, $n = 48$, H_0: $\mu \leq \mu_0$ *b. $\alpha = .01$, $n = 33$, H_0: $\mu = \mu_0$

c. $\alpha = .01$, $n = 52$, H_0: $\mu \leq \mu_0$ d. $\alpha = .001$, $n = 35$, H_0: $\mu = \mu_0$

*21. [2.5] What advantages do confidence interval procedures have over null hypothesis-testing procedures?