Statistical Methods for the Social Sciences

Third Edition

Alan Agresti • Barbara Finlay

Chapter 1

Introduction

1.1 Introduction to Statistical Methodology

In the last several years, all social science disciplines have seen an increase in the use of statistical methods. There are many reasons for this. Research in the social sciences has taken on a more quantitative orientation. Like research in other sciences, it is becoming more strongly oriented toward analyzing empirical data. The computer revolution has made a greater variety of information easily available to researchers as well as to students and the general public. Computers have also made statistical methods themselves easier to use.

The increase in the use of statistics is evident in the changes in the content of articles published in major journals and reports prepared for use in government and private industry, in the styles of textbooks, and in the increasingly common requirement of academic departments that their majors take courses in statistics. A quick glance through recent issues of *American Political Science Review*, *American Sociological Review*, or other leading social science journals will reveal the fundamental role of statistics in social science research.

Job advertisements for social scientists commonly list a knowledge of statistics as an important work tool. A student preparing for a career as a social scientist needs to become familiar with basic statistical methodology. Or, as the joke goes that we recently heard, "What did the sociologist who passed statistics say to the sociologist who failed it? 'I'll have a Big Mac, fries, and a Coke.' "

In today's world, an understanding of statistics is essential in many professions, across the spectrum from medicine to business. Physicians and other health-related professionals evaluate results of studies investigating new drugs and therapies for treating disease. Managers analyze quality of products, determine factors that help predict sales of various products, and measure employee performance.

But statistics is an important tool today even for those who do not use statistical methods as part of their job. Every day we are exposed to an explosion of information, from advertising, news reporting, political campaigning, surveys about opinions on controversial issues, and other communications containing statistical arguments. Statistics helps us make sense of this information and better understand our world. Even if you never use statistical methods in your career, we think that you will find many of the ideas in this text helpful in understanding the information that you will encounter.

We realize that you are probably not reading this book in hopes of becoming a statistician, and you may not even plan on working in social science research. In fact, you may suffer from math phobia and feel fear at what lies ahead. Please be assured that you can read this book and learn the major concepts and methods of statistics with very little knowledge of mathematics. To understand this book, logical thinking and perseverance are much more important than mathematics. And don't be frustrated if learning comes slowly and you need to read a chapter a few times before it starts to make sense. Just as you would not expect to take a single course in a foreign language and be able to speak that language fluently, the same is true with the language of statistics. On the other hand, once you have completed even a portion of this text, you will have a much better understanding of how to make sense of quantitative information.

Data

Information-gathering is at the heart of all sciences. The social sciences use a wide variety of information-gathering techniques that provide the ***observations*** used in statistical analyses. These techniques include questionnaire surveys, telephone surveys, content analysis of newspapers and magazines, planned experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze information already observed and recorded for other purposes, such as police records, census materials, and hospital files.

The observations gathered through such processes are collectively called ***data***. Data consist of measurements on the characteristics of interest. We might measure, for instance, characteristics such as political party affiliation, annual income, marital status, race, and opinion about the legalization of abortion.

Statistics

Statistics consists of a body of methods for collecting and analyzing data.

These methods for collecting and analyzing data help us to evaluate the world in an objective manner.

Purposes of Using Statistical Methods

Let's now be more specific about the objectives of using statistical methods. Statistics provides methods for

1. ***Design***: Planning and carrying out research studies.
2. ***Description***: Summarizing and exploring data.
3. ***Inference***: Making predictions or generalizing about phenomena represented by the data.

Design refers to ways of determining how best to obtain the required data. The design aspects of a study might consider, for instance, how to conduct a survey, including the construction of a questionnaire and selection of a sample of people to participate in it. ***Description*** and ***inference*** are the two elements of ***statistical analysis***—ways of analyzing the data obtained as a result of the design.

This book deals primarily with statistical analysis. This is not to suggest that statistical design is unimportant. If a study is poorly designed or if the data are improperly collected or recorded, then the conclusions may be worthless or misleading, no matter how good the statistical analysis. Methods for statistical design are covered in detail in textbooks on research methods (e.g., Babbie, 1995).

Description—describing and exploring data—includes ways of summarizing and exploring patterns in the data using measures that are more easily understood by an observer. The main purpose is to take what, to the untrained observer, are meaningless reams of data and present them in an understandable and useful form.

The raw data are a complete listing of measurements for each characteristic under study. For example, an analysis of family size in New York City might start with a list of the sizes of all families in the city. Such bulks of data, however, are not easy to assess—we simply get bogged down in numbers.

For presentation of statistical information to readers, instead of listing *all* observations we use numbers that summarize the *typical* family size in the collection of data. Or we present a graphical picture of the data. These summary descriptions, called ***descriptive statistics***, are much more meaningful for most purposes than the complete data listing. In addition, these descriptions and related explorations of the data may reveal patterns to be investigated more fully in future studies.

Inference consists of ways of making predictions based on the data. For instance, in a recent survey of 750 Americans conducted by the Gallup organization, 24% indicated a belief in reincarnation. Can we use this information to predict the percentage of the entire population of 260 million Americans that believe in reincarnation? A method presented later in this book allows us to predict that for this much larger group, the percentage believing in reincarnation falls between 21% and 27%. Predictions made using data are called ***statistical inferences***.

Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, "Are women politically more liberal than men?" "Is the imposition of the death penalty associated with a reduction in violent crime?" "Does student performance in secondary schools depend on the amount of money spent per

student, the size of the classes, or the teachers' salaries?" Statistical methods help us study such issues.

1.2 Description and Inference

We have seen that statistics consists of methods for *designing* studies and methods for *analyzing* data collected for the studies. Statistical methods for analyzing data include descriptive methods for summarizing the data and inferential methods for making predictions. A statistical analysis is classified as ***descriptive*** or ***inferential***, according to whether its main purpose is to describe the data or make predictions. To explain this distinction in more detail, we next define the *population* and *sample*.

Populations and Samples

The objects on which one makes measurements are called the ***subjects*** for the study. Usually the subjects are people, but might instead be families, schools, cities, or companies, for instance.

Population and Sample

The ***population*** is the total set of subjects of interest in a study. A ***sample*** is the subset of the population on which the study collects data.

The ultimate goal of any study is to learn about populations. But it is usually necessary, and more practical, to study only samples from those populations. For example, the Gallup and Harris polling organizations usually select samples of 750–1500 Americans to collect information about political and social beliefs of the population of *all* Americans.

Descriptive Statistics

Descriptive statistical methods summarize the information in a collection of data.

We use descriptive statistics to summarize basic characteristics of a sample. We might, for example, describe the typical family size in New York City by computing the *average* family size. The main purpose of descriptive statistics is to explore the data and to reduce them to simpler and more understandable terms without distorting or losing much of the available information. Summary graphs, tables, and numbers such as averages and percentages are easier to comprehend and interpret than are long listings of data.

Inferential Statistics

Inferential statistical methods provide predictions about characteristics of a population, based on information in a sample from that population.

Example 1.1 illustrates the use of inferential statistics.

Example 1.1 Opinion About Handgun Control

The first author of this text is a resident of Florida, a state with a relatively high crime rate. He would like to know the percentage of Florida residents who favor controls over the sales of handguns. The population of interest is the collection of more than 10 million adult residents in Florida. Since it is impossible for him to discuss the issue with everyone, he can study results from a poll of 834 residents of Florida conducted in 1995 by the Institute for Public Opinion Research at Florida International University. In that poll, 54% of the sampled subjects said that they favored controls over the sales of handguns.

This poll collected data for 834 residents. He is interested, however, not just in those 834 people but in the *entire population* of all adult Florida residents. Inferential statistics can provide a prediction about this larger population using the sample data. An inferential method presented in Chapter 5 predicts that the population percentage favoring control over sales of handguns falls between 50% and 58%. Even though the sample is very small compared to the population size, he can conclude, for instance, that probably a slim majority of Florida residents favor handgun control. □

Using inferential statistical methods with properly chosen samples, we can determine characteristics of entire populations quite well by selecting samples that are small relative to the size of the population.

Parameters and Statistics

Parameters and Statistics

A ***parameter*** is a numerical summary of the population. A ***statistic*** is a numerical summary of the sample data.

Example 1.1 dealt with estimating the percentage of Florida residents who support gun control. The parameter of interest was the true, but unknown, population percentage favoring gun control. The inference about this parameter was based on a statistic—the percentage of the 834 Florida residents in the sample who favor gun control, namely, 54%. Since this number *describes* a characteristic of the sample, it is an example of a descriptive statistic. The value of the parameter to which the inference refers, namely, the population percentage in favor of gun control, is unknown. In summary, we use known sample statistics in making inferences about unknown population parameters.

(Students should note that, in statistical usage, the term *parameter* does not have its usual meaning of "limit" or "boundary.")

The primary focus of most research studies is the parameters of the population, not the statistics calculated for the particular sample selected. The sample and statistics describing it are important only insofar as they provide information about the unknown parameters. We would want a prediction about *all* Floridians, not only the 834 subjects in the sample.

An important aspect of statistical inference involves reporting the likely *accuracy* of the sample statistic that predicts the value of a population parameter. An inferential statistical method predicts how close the *sample* value of 54% is likely to be to the true (unknown) percentage of the *population* favoring gun control. A method from Chapter 5 determines that a sample of size 834 yields accuracy within about 4%; that is, the true population percentage favoring gun control falls within 4% of the sample value of 54%, or between 50% and 58%.

When data exist for an entire population, there is no need to use inferential statistical methods, since one can then calculate exactly the parameters of interest. For example, place of residence and home ownership are observed for virtually all Americans during census years. When the population of interest is small, we would normally study the records of the entire population instead of only a sample. In studying the voting records of members of the U.S. Senate on bills concerning defense appropriations, for example, we could obtain data on votes for all senators on all such bills.

In most social science research, it is impractical to collect data for the entire population, due to monetary and time limitations. It is usually unnecessary to do so, in any case, since good precision for inferences about population parameters results from relatively small samples, such as the 750–1500 subjects that most polls take. This book explains why this is so.

1.3

Defining Populations

Inferential statistical methods require specifying clearly the population to which the inferences apply. Sometimes the population is a clearly defined set of subjects. In Example 1.1, it was the collection of adult Florida residents. Often, however, the generalizations refer to a *conceptual* population—a population that does not actually exist but that one can hypothetically conceptualize.

For example, suppose a team of researchers tests a new drug designed to relieve severe depression. They plan to analyze results for a sample of patients suffering from depression to make inferences about the conceptual population of all individuals who might suffer depressive symptoms now or sometime in the future. Or a consumer organization may evaluate gas mileage for a new model of an automobile by observing the average number of miles per gallon for five sample autos driven on a standardized 100-mile course. Inferences then refer to the performance on this course for the conceptual population of all autos of this model that will be or could hypothetically be manufactured.

A caution is due here. Investigators often try to generalize to a broader population than the one to which the sample results can be statistically extended. A psychologist

may conduct an experiment using a sample of students from an introductory psychology course. With statistical inference, the sample results generalize to the population of all students in the class. For the results to be of wider interest, however, the psychologist might claim that the conclusions generalize to *all* college students, to *all* young adults, or even to a more heterogeneous group. These generalizations may well be wrong, since the sample may differ from those populations in fundamental ways, such as in racial composition or average socioeconomic status.

For instance, in her 1987 book *Women in Love*, Shere Hite presented results of a survey she conducted of adult women in the United States. One of her conclusions was that 70% of women who had been married at least five years have extramarital affairs. She based this conclusion on responses to questionnaires returned from a sample of 4500 women, which sounds impressively large. However, the questionnaire was mailed to about 100,000 women. We cannot know whether this sample of 4.5% of the women who responded is representative of the 100,000 who received the questionnaire, much less the entire population of adult American women. Thus, it is dangerous to try to make an inference to the larger population.

You should carefully assess the scope of conclusions in research articles, political and government reports, advertisements, and the mass media. Evaluate critically the basis for the conclusions by noting the makeup of the sample upon which the inferences are built. Chapter 2 discusses some desirable and undesirable types of samples.

In the past quarter century, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 5.

1.3 The Role of Computers in Statistics

Even as you read this book, the computer industry continues its ceaseless growth. New and more powerful personal computers and workstations are reaching the market, and these computers are becoming more accessible to people who are not technically trained. An important aspect of this expansion is the development of highly specialized software.

Versatile and user-friendly software is now readily available for analyzing data using descriptive and inferential statistical methods. The development of this software has provided an enormous boon to the use of sophisticated statistical methods.

Statistical Software

Statistical Package for the Social Sciences (SPSS), SAS (SAS Institute, Inc.), and Minitab are among popular statistical software found on college campuses. It is much easier to apply statistical methods using these software than using old-fashioned hand calculation. The accuracy of computations is greatly improved, since hand calculations often result in mistakes or crude answers, especially when the data set is large. Moreover, some modern statistical methods presented in this text are too complex to be done by hand.

STATISTICAL ANALYSIS IN PSYCHOLOGY AND EDUCATION

Sixth Edition

of students in a particular university from a knowledge of the mean computed on the sample of 100 and to estimate the error involved in this statement, we use procedures from inferential statistics. The application of these procedures provides information about the accuracy of the sample mean as an estimate of the population mean; that is, it indicates the degree of assurance we may place in the inferences we draw from the sample to the population.

In this section no discussion is advanced on methods of drawing samples or the conditions which these methods must satisfy to allow the drawing of valid inferences from the sample to the population. Further, no precise meaning has been assigned to the term "error." These topics will be elaborated at a later stage.

1.6 PARAMETERS AND ESTIMATES

A clear distinction is usually drawn between parameters and estimates. A *parameter* is a property descriptive of the population. The term *estimate* refers to a property of a sample drawn at random from a population. The sample value is presumed to be an estimate of a corresponding population parameter. Suppose, for example, that a sample of 1,000 adult male Canadians of a given age range is drawn from the total population, the height of the members of the sample measured, and a mean value, 68.972 inches, obtained. This value is an estimate of the population parameter which would have been obtained had it been possible to measure all the members in the population. Usually parameters or population values are unknown. We estimate them from our sample values. The distinction between parameter and estimate is reflected in statistical notation. A widely used convention in notation is to employ Greek letters to represent parameters and Roman letters to represent estimates. Thus the symbol σ, the Greek letter sigma, may be used to represent the standard deviation in the population, the standard deviation being a commonly used measure of variation. The symbol s may be used as an estimate of the parameter σ. This convention in notation is applicable only within broad limits. By and large we shall adhere to this convention in this book, although in certain instances it will be necessary to depart from it. By common practice and tradition a Greek letter may be used on occasion to denote a sample statistic.

1.7 VARIABLES AND THEIR CLASSIFICATION

The term *variable* refers to a property whereby the members of a group or set differ one from another. The members of a group may be individuals and may be found to differ in sex, age, eye color, intelligence, auditory acuity, reaction time to a stimulus, attitudes toward a political issue, and many other ways. Such properties are variables. The term *constant* refers to a property whereby the members of a group do not differ one from another. In a sense a constant is a particular type of variable; it is a variable which does not vary from one member of a group to another or within a particular set of defined conditions.

Labels or numerals may be used to describe the way in which one member of a group is the same as or different from another. With variables like sex, racial origin,

religious affiliation, and occupation, labels are employed to identify the members which fall within particular classes. An individual may be classified as male or female; of English, French, or Dutch origin; Protestant or Catholic; a shoemaker or a farmer; and so on. The label identifies the class to which the individual belongs. Sex for most practical purposes is a two-valued variable, individuals being either male or female. Occupation, on the other hand, is a multivalued variable. Any particular individual may be assigned to any one of a large number of classes. With variables like height, weight, intelligence, and so on, measuring operations may be employed which enable the assignment of descriptive numerical values. An individual may be 72 inches tall, weigh 190 pounds, and have an IQ of 90.

The particular values of a variable are referred to as *variates*, or *variate values*. To illustrate, in considering the height of adult males, height is the variable, whereas the height of any particular individual is a variate, or variate value.

In dealing with variables which bear a functional relationship one to another, the distinction may be drawn between *dependent* and *independent* variables. Consider the expression $Y = f(X)$. This expression says that a given variable Y is some unspecified function of another variable X. The symbol f is used generally to express the fact that a functional relationship exists, although the precise nature of the relationship is not stated. In any particular case the nature of the relationship may be known; that is, we may know precisely what f means. Under these circumstances, for any given value of X a corresponding value of Y can be calculated: that is, given X and a knowledge of the functional relationship, Y can be predicted. It is customary to speak of Y, the predicted variable, as the dependent variable because the prediction of it depends on the value of X and the known functional relationship, whereas X is spoken of as the independent variable. Given an expression of a kind $Y = X^3$ for any given value of X, an exact value of Y can readily be determined. Thus if X is known, Y is also known exactly. Many of the functional relationships found in statistics permit probabilistic and not exact prediction to occur. Such relationships may provide the most probable value of Y for any given value of X, but do not permit the making of perfect predictions.

A distinction may be drawn between *continuous* and *discrete* (or *discontinuous*) variables. A continuous variable may take any value within a defined range of values. The possible values of the variable belong to a continuous series. Between any two values of the variable an indefinitely large number of in-between values may occur. Height, weight, and chronological time are examples of continuous variables. A discontinuous or discrete variable can take specific values only. Size of family is a discontinuous variable. A family may comprise 1, 2, 3, or more children, but values between these numbers are not possible. The values obtained in rolling a die are 1, 2, 3, 4, 5, and 6. Values between these numbers are not possible. Although the underlying variable may be continuous, all sets of real data in practice are discontinuous or discrete. Convenience and errors of measurement impose restrictions on the refinement of the measurement employed.

Another classification of variables is possible which is of some importance and is of particular interest to statisticians. This classification is based on differences in the type of information which different operations of classification or measurement yield. To illustrate, consider the following situations. An observer using direct in-

spection may rank-order a group of individuals from the tallest to the shortest according to height. On the other hand, he may use a foot rule and record the height of each individual in the group in feet and inches. These two operations are clearly different, and the nature of the information obtained by applying the two operations is different. The former operation permits statements of the kind: individual *A* is taller or shorter than individual *B*. The latter operation permits statements of how much taller or shorter one individual is than another. Differences along these lines serve as a basis for a classification of variables, the class to which a variable belongs being determined by the nature of the information made available by the measuring operation used to define the variable. Four broad classes of variables may be identified. These are referred to as (1) nominal, (2) ordinal, (3) interval, and (4) ratio variables.

A *nominal variable* is a property of the members of a group defined by an operation which permits the making of statements only of equality or difference. Thus we may state that one member is the *same as* or *different from* another member with respect to the property in question. Statements about the ordering of members, or the equality of differences between members, or the number of times a particular member is greater than or less than another are not possible. To illustrate, individuals may be classified by the color of their eyes. Color is a nominal variable. The statement that an individual with blue eyes is in some sense "greater than" or "less than" an individual with brown eyes is meaningless. Likewise the statement that the difference between blue eyes and brown eyes is equal to the difference between brown eyes and green eyes is meaningless. The only kind of meaningful statement possible with the information available is that the eye color of one individual is the same as or different from the eye color of another. A nominal variable may perhaps be viewed as a primitive type of variable, and the operations whereby the members of a group are classified according to such a variable constitute a primitive form of measurement. In dealing with nominal variables, numerals may be assigned to represent classes, but such numerals are labels, and the only purpose they serve is to identify the members within a given class.

An *ordinal variable* is a property defined by an operation which permits the rank ordering of the members of a group; that is, not only are statements of equality and difference possible, but also statements of the kind *greater than* or *less than*. Statements about the equality of differences between members or the number of times one member is greater than or less than another are not possible. If a judge is required to order a group of individuals according to aggressiveness, or cooperativeness, or some other quality, the resulting variable is ordinal in type. Many of the variables used in psychology are ordinal.

An *interval variable* is a property defined by an operation which permits the making of statements of equality of intervals, in addition to statements of sameness or difference or greater than or less than. An interval variable does not have a "true" zero point, although a zero point may for convenience be arbitrarily defined. Fahrenheit and Celsius temperature measurements constitute interval variables. Consider three objects, *A*, *B*, and *C*, with temperatures 12°, 24°, and 36°, respectively. It is appropriate to say that the difference between the temperature of *A* and *B* is equal to the difference in the temperature of *B* and *C*. It is appropriate also to say that the difference between

the temperature of A and C is twice the difference between the temperature of A and B or B and C. It is not appropriate to say that B has twice the temperature of A, or that C has three times the temperature of A. In common usage, if the temperature yesterday was 64° and today it was 32°, we would not say that it was twice as hot yesterday, or that the temperature was twice as great, as it was today. Calendar time is also an interval variable with an arbitrarily defined zero point.

A *ratio variable* is a property defined by an operation which permits the making of statements of equality of ratios in addition to all other kinds of statements discussed above. This means that one variate value, or measurement, may be spoken of as double or triple another, and so on. An absolute 0 is always implied. The numbers used represent distances from a natural origin. Length, weight, and the numerosity of aggregates are examples of ratio variables. One object may be twice as long as another, or three times as heavy, or four times as numerous. Many of the variables used in the physical sciences are of the ratio type. In psychological work, variables which conform to the requirements of ratio variables are uncommon.

The essential difference between a ratio and an interval variable is that for the former the measurements are made from a true zero point, whereas for the latter the measurements are made from an arbitrarily defined zero point or origin. Because of this, for a ratio variable, ratios may be formed directly from the variate values themselves, and meaningfully interpreted. For an interval variable, ratios may be formed from differences between the variate values. The differences constitute a ratio variable, because the process of subtraction eliminates, or cancels out, the arbitrary origin. Differences are the same regardless of the location of the zero point or origin.

Statistical methods exist for the analysis of data composed of nominal variables, ordinal variables, and interval and ratio variables. From the viewpoint of practical statistical work in psychology and education the distinction between interval and ratio variables is perhaps unimportant, and it is convenient to think of three, and not four, classes of variables, with three corresponding classes of statistical method. Procedures for the analysis of interval and ratio variables constitute by far the largest, and most important, class of statistical methods.

The importance of the concept of a variable cannot be overemphasized. All statistics is concerned with variables and the relations between them. Variables are the stuff of which statistics is made. Variables may be added together, partitioned into additive bits, related one to another, and interrelated in complex networks.

An investigator may study variables, and their interrelations, as they exist in nature; for example, age, IQ, examination marks, blood pressure, and anxiety level. Such investigations are sometimes called correlational studies. Some of the more important variables in nature are nominal, such as being male or female and being alive or dead. On the other hand, in many investigations one or more variables may be created by the investigator who decides the values the variable will take and the frequency of occurrence of these values. Such investigations are called experiments. For example, a simple experiment may involve two groups of subjects; a particular dosage of a drug is administered to one group, but not to the other. Then some measure of motor performance is obtained from the members of the two groups. Two variables are involved here: the independent variable—receiving or not receiving the drug—and

the dependent variable—motor performance. The purpose of the experiment is to explore the relation between these two variables, that is, to study how motor performance *depends* on the presence or the absence of the drug. Note that the independent variable is a simple nominal variable with two categories denoting group membership. A subject is or is not a member of the group receiving the drug. The basic point here is that in both correlational and experimental studies, variables and the relations between them are the object of inquiry. Thus the investigator studies how variables are related, whether these variables are naturally occurring or are created intentionally by him.

1.8 DATA ANALYSIS

Since the first edition of this book in 1959, remarkable changes have occurred in computational methods. Also, enormous changes are anticipated in the future, as increasing computational power is incorporated in smaller computers at decreasing cost. These technical developments have eliminated much of the drudgery associated with statistical work. Many earlier statisticians devoted substantial parts of their working lives to arithmetical labor. Much of this was elementary. Many methods that were devised to reduce computational time are now obsolete. These have been removed from the various editions of this book, except in a few instances where they may assist the student in understanding some aspect of the statistic under discussion.

The speed and ease with which numerical computation can now be done has led to the increased frequency of use of complex statistical methods, many of which were known earlier but were not often used because of the arithmetical labor required. In this context, multivariate statistical methods deserve mention. These are methods that require the analysis of data comprising several, perhaps many, variables. In psychology, education, and elsewhere, the attempt to explain or predict a particular phenomenon may entail studying processes that are highly complex and that involve the functioning of collections of interacting variables. Inspection of current research literature suggests that the frequency of use of multivariate methods is increasing rapidly.

The ready availability of small and powerful computers has enhanced, not diminished, the importance of understanding statistical concepts. Also, this development has added to the importance of understanding the task a particular method is designed to accomplish, the problems it purportedly solves, and the assumptions it requires. This understanding can in many instances be readily acquired by the student without any extended exploration of either the mathematical apparatus underlying a method or the details of the computational procedures employed.

Earlier statisticians with primitive computing devices spent much time and labor exploring by various methods, graphic and otherwise, the information contained in collections of data. This investment of effort not uncommonly led to a useful intuitive understanding of what the data had to say. Now that powerful computing devices are available, this intuitive level of understanding frequently is not attained. Data are fed to computers, and solutions emerge. Not uncommonly the investigator only partially understands the data. A strong case can be made for an initially rather simpleminded, exploratory approach to data analysis. In the study of many sets of data, simple

preliminary insights into the story the data have to tell are useful in guiding subsequent forms of analysis. The computer may, of course, be used to assist such exploratory data analyses.

Tukey (1977) stresses exploratory data analysis. Van Dantzig (1978) provides a simple and readable overview of the topic. Exploratory data analysis is descriptive, involving simple data manipulations, summary methods, and graphic descriptions. Its purpose is to make the nature and structure of the data understandable to the investigator, and, thereby, to lead to the subsequent use of appropriate statistical models. Tukey speaks of exploratory data analysis as analogous to detective work; it may compel the investigator to notice aspects of the data that were not expected. All scientists know that careful exploration of sets of data will on occasion lead to findings they did not anticipate. The importance of an experiment may go beyond the original intent.

Tukey distinguishes exploratory from confirmatory data analysis. By confirmatory data analysis he means the confirmation, or disconfirmation, of the original hypothesis that led to the experiment in the first place. Tukey argues that exploratory and confirmatory data analyses complement each other.

The present writer has long held the view that in the analysis of any set of data, some understanding of the data, obtained perhaps by calculating a few simple statistics, is a necessary preliminary to the planning of more elaborate analyses. Such analyses frequently involve complex models and assumptions. The availability of powerful computers, which automatically apply so-called canned programs, enhances the importance of simple methods that assist the investigator in getting to know the data.

BASIC TERMS AND CONCEPTS

Population
Population: finite; infinite
Sample
Descriptive statistics
Sampling statistics
Parameter
Estimate
Variable
Variate value
Variable: dependent; independent
Variable: continuous; discrete
Variable: nominal; ordinal; interval; ratio
Correlational study
Experiment

Statistical Methods for the Social Sciences

Third Edition

Alan Agresti ▪ Barbara Finlay

Chapter 2

Sampling and Measurement

The ultimate goals of social science research are to understand, explain, and make inferences about social phenomena. To do this, we need data. *Descriptive* statistical methods provide ways of summarizing the data. *Inferential* statistical methods use sample data to make predictions about populations. To make inferences, we must decide which subjects of the population to sample. Selecting a sample that is likely to be representative of the population is a primary topic of this chapter.

We must convert our ideas about social phenomena into actual data through measurement. The development of ways to measure abstract concepts such as prejudice, love, intelligence, and status is one of the most difficult problems of social research. Moreover, the problems related to finding valid and reliable measures of concepts have consequences for statistical analysis of the data. In particular, invalid or unreliable data-gathering instruments render the statistical manipulations of the data meaningless.

The first section of this chapter discusses some statistical aspects of measurement, such as the different types of data. The second and third sections discuss the principal methods for selecting the sample that provides the measurements.

2.1 Variables and Their Measurement

Statistical methods provide a way to deal with *variability*. Variation occurs among people, schools, towns, and the various subjects of interest to us in our everyday lives. For

instance, variation occurs from person to person in characteristics such as income, IQ, political party preference, religious beliefs, marital status, and musical talent. We shall see that the nature and the extent of the variability has important implications both on descriptive and inferential statistical methods.

Variables

A characteristic measured for each subject in a sample is called a *variable*. The name refers to the fact that values of the characteristic vary among subjects in a sample or population.

Variable

A ***variable*** is a characteristic that can vary in value among subjects in a sample or population.

Each subject has a particular value for a variable, but different subjects may possess different values. Examples of variables are gender (with values female and male), age at last birthday (with values 0, 1, 2, 3, and so on), religious affiliation (Protestant, Roman Catholic, Jewish, Other, None), number of children in a family (0, 1, 2, ...), and political party preference (Democrat, Republican, Independent). The possible values the variable can assume form the ***scale*** for measuring the variable. For gender, for instance, that scale consists of the two labels, female and male.

The valid statistical methods for analyzing a variable depend on the scale for its measurement. We treat a numerical-valued variable such as annual income (in thousands of dollars) differently than a variable with a scale consisting of labels, such as political preference (with scale Democrat, Republican, Independent). We next introduce two ways to classify variables that determine the valid statistical methods. The first refers to whether the measurement scale consists of labels or numbers. The second refers to the number of levels in that scale.

Qualitative and Quantitative Data

Data are called ***qualitative*** when the scale for measurement is a set of unordered categories. For example, marital status, with categories (single, married, divorced, widowed), is qualitative. For Canadians, the province of one's residence is qualitative, with the categories Alberta, British Columbia, and so on. Other qualitative variables are religious affiliation (with categories such as Catholic, Jewish, Muslim, Protestant, Other, None), gender (female, male), political party preference (Democrat, Republican, Independent), and marriage form of a society (monogamy, polygyny, polyandry). For each variable, the categories are unordered; the scale does not have a "high" or "low" end.

For qualitative variables, distinct categories differ in quality, not in quantity or magnitude. Although the different categories are often called the *levels* of the scale, no level

is greater than or smaller than any other level. Names or labels such as "Alberta" and "British Columbia" identify the categories, but those names do not represent different magnitudes of the variable.

When the possible values of a variable do differ in magnitude, the variable is called ***quantitative***. Each possible value of a quantitative variable is *greater than* or *less than* any other possible value. Such comparisons result from variables having a numerical scale. Examples of quantitative variables are a subject's annual income, number of years of education completed, number of siblings, and number of times arrested.

The set of categories for a qualitative variable is called a ***nominal scale***. For instance, a variable pertaining to one's mode of transportation to work might use the nominal scale consisting of the categories (car, bus, subway, bicycle, walk). A set of numerical values for a quantitative variable is called an ***interval*** scale. Interval scales have a specific numerical distance or "interval" between each pair of levels. Annual income is usually measured on an interval scale; the interval between $40,000 and $30,000, for instance, equals $10,000. We can compare outcomes in terms of how much larger or how much smaller one is than the other, a comparison that is not relevant for a nominal scale.

A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural *ordering* of values, but undefined interval distances between the values. Examples are social class (classified into upper, middle, lower), political philosophy (measured as very liberal, slightly liberal, moderate, slightly conservative, very conservative), and government spending on the environment (classified as too little, about right, too much). These scales are not nominal, because the categories are naturally ordered. The levels are said to form an ***ordinal*** scale.

Ordinal scales consist of a collection of ordered categories. Although the categories have a clear ordering, the distances between them are unknown. For example, a person categorized as very liberal is *more* liberal than a person categorized as slightly liberal, but there is no numerical value for *how much more* liberal that person is.

Both nominal and ordinal scales consist of a set of categories. Each observation falls into one and only one category. Variables having categorical scales are called ***categorical variables***. While the categories have a natural ordering for an ordinal scale, they are unordered for a nominal scale. For the categories (Catholic, Jewish, Muslim, Protestant, Other, None) for religious affiliation, it does not make sense to think of one category as being higher or lower than another.

The various scales refer to the actual measurement of social phenomena and not to the phenomena themselves. *Place of residence* may indicate the geographic place name of one's residence (nominal), the distance of that residence from a point on the globe (interval), the size of one's community (interval or ordinal), or other kinds of sociological variables.

Quantitative Nature of Ordinal Data

As we've discussed, data from nominal scales are qualitative—distinct levels differ in quality, not in quantity. Data from interval scales are quantitative: distinct levels have differing magnitudes of the characteristic of interest. The position of ordinal scales

on the quantitative–qualitative classification is fuzzy. Because their scale consists of a set of categories, they are often treated as qualitative, being analyzed using methods for nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: each level has a *greater* or *smaller* magnitude of the characteristic than another level.

Some statistical methods apply specifically to ordinal variables. Often, though, statisticians take advantage of the quantitative nature of ordinal scales by assigning numerical scores to categories. That is, they often treat ordinal data as interval in order to use the more sophisticated methods available for quantitative data. For instance, course grades (such as A, B, C, D, E) are ordinal, but we treat them as interval when we assign numbers to the grades (such as 4, 3, 2, 1, 0) to compute a grade point average. Treating ordinal data as interval requires good judgment in assigning scores, and it is often accompanied by a "sensitivity analysis" of checking whether substantive results differ for differing choices of the scores. The quantitative treatment of ordinal data has benefits in the variety of methods available for data analysis, particularly for data sets with many variables.

Statistical Methods and Type of Measurement

The main reason for distinguishing between qualitative and quantitative data is that different statistical methods apply to each type of data. Some methods are designed for qualitative variables and others are designed for quantitative variables.

It is not possible to analyze qualitative data using methods for quantitative variables. If a variable has only a nominal scale, for instance, one cannot use methods for interval data, since the levels of the scale do not have numerical values. One cannot apply quantitative statistical methods based on interval scales to qualitative variables such as religious affiliation or county of residence. For instance, the *average* is a statistical summary for quantitative data, since it uses numerical values; one can compute the average for a variable having an interval scale, such as income, but not for a variable having a nominal scale, such as religious affiliation.

On the other hand, it is always possible to treat a variable in a less quantitative manner. For example, suppose age is measured using the ordered categories under 18, 18–40, 41–65, over 65. This variable is quantitative, but one could treat it as qualitative either by ignoring the ordering of these four categories or by using unordered levels such as working age, nonworking age. Normally, though, we apply statistical methods specifically appropriate for the actual scale of measurement, since they use the characteristics of the data to the fullest. You should measure variables at as high a level as possible, because a greater variety of methods apply with higher-level variables.

Discrete and Continuous Variables

We now present one other way of classifying variables that helps determine which statistical method is most appropriate for a data set. This classification refers to the number of values in the measurement scale.

Discrete and Continuous Variables

A variable is ***discrete*** if it can take on a finite number of values and ***continuous*** if it can take an infinite continuum of possible real number values.

Examples of discrete variables are number of children (measured for each family), number of murders in the past year (measured for each census tract), and number of visits to a physician in past year (measured for each subject). Any variable phrased as "the number of . . ." is discrete, since one can list all the possible values [0, 1, 2, 3, 4, . . .] for the variable. (Strictly speaking, there could be an infinite number of values for such a variable, namely, all the nonnegative integers. As long as the possible values do not form a continuum, the variable is still said to be discrete.)

Examples of continuous variables are height, weight, age, and the amount of time it takes to read a passage of a book. It is impossible to write down all the distinct potential values of a continuous variable, since they form a continuum. The amount of time needed to read a book, for example, could take on the value 8.6294473 . . . hours.

With discrete variables, one cannot subdivide the basic unit of measurement. For example, 2 and 3 are possible values for the number of children in a family, but 2.571 is not. On the other hand, a collection of values for a continuous variable can always be refined; that is, between any two possible values, there is always another possible value. For example, an individual does not age in discrete jumps. Between 20 and 21 years of age, there is 20.5 years (among other values); between 20.5 and 21, there is 20.7. At some well-defined point during the year in which a person ages from 20 to 21, that person is 20.3275 years old, and similarly for every other real number between 20 and 21. A continuous, infinite collection of age values occurs between 20 and 21 alone.

Qualitative variables are discrete, having a finite set of unordered categories. In fact, all categorical variables, nominal or ordinal, are discrete. Quantitative variables can be discrete or continuous; age is continuous, and number of times arrested is discrete.

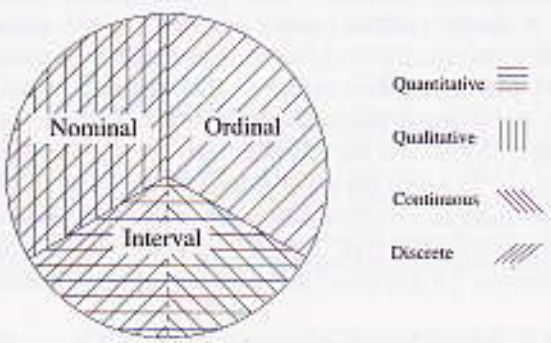
The distinction between discrete and continuous variables is often blurry in practice, because of the way variables are actually measured. Continuous variables must be rounded when measured, so we measure them as though they are discrete. We usually say that an individual is 20 years old whenever that person's age is somewhere between 20 and 21. Other variables of this type are prejudice, intelligence, motivation, and other internalized attitudes or orientations. Such variables are assumed to vary continuously, but measurements of them describe, at best, rough sections of the underlying continuous distributions. A scale of prejudice may have discrete units from 0 to 10, but each discrete value is assumed to include all values within a certain continuous range of the degree of prejudice.

On the other hand, some variables, though discrete, may take on a very large number of different values. In measuring annual family income in thousands of dollars, the potential values are 0, 1, 2, 3, . . ., up to some very large highest value. Statistical

methods for continuous variables are often simpler than methods for discrete variables. Thus, statisticians treat discrete variables that can assume many different values as if they were continuous. For example, they treat variables such as income and college entrance examination score as continuous variables. The discrete–continuous distinction is, in practice, a distinction between variables that can take lots of values, such as income, and variables that take relatively few values, such as number of times married.

You need to understand the discrete–continuous classification, qualitative–quantitative classification, and nominal–ordinal–interval scale classification, because each statistical method refers to a particular type of data. For instance, some methods (such as summarizing data using an average) require quantitative data, and some of these methods also require the variable to be continuous.

Figure 2.1 summarizes the types of data and their connections. Variables having a nominal scale are qualitative. Variables having an interval scale are quantitative. Variables having an ordinal scale are sometimes treated as quantitative and sometimes as qualitative. Variables having a nominal or ordinal scale take values in a set of categories, and are categorical. Categorical variables are discrete. Variables having an interval scale can be either discrete or continuous.



Note: Ordinal data are treated sometimes as qualitative and sometimes as quantitative

Figure 2.1 Summary of Quantitative–Qualitative, Nominal–Ordinal–Interval, Continuous–Discrete Classifications

2.2 Randomization

Inferential statistical methods use sample statistics to make predictions about population parameters. The quality of the inferences depends crucially on how well the sample represents the population. This section introduces an important sampling method that incorporates ***randomization***, the mechanism for ensuring that the sample representation is adequate for inferential methods.

Simple Random Sampling

Subjects of a population to be sampled could be individuals, families, schools, houses, cities, hospitals, records of reported crimes, and so on. *Simple random sampling* is a method of sampling for which every possible sample has equal chance of occurring. This provides fairness and also permits inference about the population sampled. In fact, most inferential statistical methods assume randomization of the sort provided by simple random sampling.

Let n denote the number of subjects in the sample, called the ***sample size***.

Simple Random Sample

A ***simple random sample*** of n subjects from a population is one in which each possible sample of that size has the same probability of being selected.

For instance, suppose that a survey interviewer administers a questionnaire to one randomly selected adult subject of each of several separate households. A particular household contains four adults—mother, father, aunt, and uncle—identified as M, F, A, and U. A simple random sample of $n = 1$ of the adults is one in which each of the four adults is equally likely to be interviewed. The selection might be made, for example, by placing the four names on four identical ballots and selecting one blindly from a hat. For a simple random sample of $n = 2$ adults, each possible sample of size two is equally likely. The six potential samples are (M, F), (M, A), (M, U), (F, A), (F, U), and (A, U). To select the sample, we blindly select two ballots from the hat.

A simple random sample is often just called a ***random sample***. The "simple" adjective distinguishes this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

How to Select a Simple Random Sample

Before we can select a random sample, we need a list of all subjects in the population. This list is called the ***sampling frame***. The most common method for selecting a random sample from the sampling frame uses a *random number table* to ensure that each subject has an equal chance of selection.

Random Number Table

A ***random number table*** is a table containing a sequence of numbers that is computer generated according to a scheme whereby each digit is equally likely to be any of the integers $0, 1, 2, \ldots, 9$.

Table 2.1 shows a section of a random number table. The numbers fluctuate according to no set pattern. Any particular number has the same chance of being a 0, 1, 2, . . .,

or 9. The numbers are chosen independently, so any one digit has no influence on any other. If the first digit in a row of the table is a 9, for instance, the next digit is still just as likely to be a 9 as a 0 or 1 or any other number.

TABLE 2.1 Part of a Table of Random Numbers

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	10480	15011	01536	02011	81647	91646	69179	14194
2	22368	46573	25595	85393	30995	89198	27982	53402
3	24130	48360	22527	97265	76393	64809	15179	24830
4	42167	93093	06243	61680	07856	16376	39440	53537
5	37570	39975	81837	16656	06121	91782	60468	81305
6	77921	06907	11008	42751	27756	53498	18602	70659
7	99562	72905	56420	69994	98872	31016	71194	18738
8	96301	91977	05463	07972	18876	20922	94595	56869
9	89579	14342	63661	10281	17453	18103	57740	84378
10	85475	36857	53342	53988	53060	59533	38867	62300
11	28918	69578	88231	33276	70997	79936	56865	05859
12	63553	40961	48235	03427	49626	69445	18663	72695
13	09429	93969	52636	92737	88974	33488	36320	17617
14	10365	61129	87529	85689	48237	52267	67689	93394
15	07119	97336	71048	08178	77233	13916	47564	81056
16	51085	12765	51821	51259	77452	16308	60756	92144
17	02368	21382	52404	60268	89368	19885	55322	44819
18	01011	54092	33362	94904	31273	04146	18594	29852
19	52162	53916	46369	58586	23216	14513	83149	98736
20	07056	97628	33787	09998	42698	06691	76988	13602

Source: Abridged from William H. Beyer, ed., *Handbook of Tables for Probability and Statistics*, 2nd ed., © The Chemical Rubber Co., 1968, Used by permission of the Chemical Rubber Co.

We illustrate the use of this table for selecting a simple random sample of $n = 100$ students from a university student body of size 30,000. The sampling frame is a list of these students, such as a student directory. We select the students by using five-digit sequences to identify them, as follows:

1. Assign the numbers 00001 to 30000 to the students in the sampling frame, using 00001 for the first student in the list, 00002 for the second student in the list, and so on.
2. Starting at any point in the random number table, choose successive five-digit numbers until you obtain 100 distinct numbers between 00001 and 30000.
3. Include in the sample the students with assigned numbers equal to the random numbers selected.

For example, using the first column of five-digit numbers in Table 2.1, the first three random numbers are 10480, 22368, and 24130; thus, the first three students selected are those numbered 10480, 22368, and 24130 in the listing.

In selecting the 100 five-digit numbers, we skip numbers greater than 30000, such as the next seven five-digit numbers in Table 2.1, since no student in the sampling frame has an assigned number that large. After using the first column of five-digit numbers, we move to the next column of numbers and continue. If the population size were between 1000 and 9999, we would use only four digits at a time. The column (or row) from which we begin selecting the numbers does not matter, since the numbers have no set pattern.

The reason for using random sampling is that it reduces the chance of selecting a sample that is seriously biased in some way, thus leading to inaccurate inferences about the population. Everyone has the same chance of inclusion in the sample.

Probability and Nonprobability Sampling

Simple random sampling is a type of ***probability sampling*** method. Such methods can specify the probability that any particular sample will be selected. For simple random sampling, each distinct possible sample of n subjects has the same probability of selection. With probability samples one can apply inferential statistical methods, since the derivation of those methods requires knowing the probabilities of the possible samples. ***Nonprobability sampling*** methods are ones for which it is not possible to specify the probabilities of the possible samples. Inferences using such samples are of unknown reliability.

One of the most common nonprobability sampling methods is ***volunteer sampling***. As the name implies, in this method subjects volunteer themselves for the sample. Inherent in this method is the danger that the sample will poorly represent the population and will yield misleading conclusions.

For instance, a mail-in questionnaire published in *TV Guide* posed the question, "Should the President have the Line Item Veto to eliminate waste?" Of those who responded, 97% said yes. For the same question posed to a random sample, 71% said yes (D. M. Wilbur, *The Public Perspective*, 1993).

A good example of volunteer sampling is visible almost any day on U.S. television. It's become a trend on many TV news and entertainment programs to ask viewers to offer their opinions on an issue of the moment by calling a 900 number. The problem is that the viewers who respond are unlikely to be a representative cross section, but will be those people who watch that program and who happen to feel strongly enough to call. Individuals possessing one specific opinion on that issue might be much more likely to respond than individuals holding a different opinion.

For instance, one night the ABC program *Nightline* asked viewers whether the United Nations should continue to be located in the United States. Of more than 186,000 callers, 67% wanted the United Nations out of the United States. At the same time, a scientific poll using a random sample of about 500 respondents estimated the true percentage wanting the United Nations out of the United States to be about 28%. Even though the random sample is much smaller, it is far more trustworthy since it greatly reduces the chance of bias. From statistical inferential methods (from Chapter 5) with this random sample, the true percentage of the population of all Americans who want the United Nations out of the United States is between about 24% and 32%.

A large sample does not help with volunteer sampling—the bias remains. In 1936, the newsweekly *Literary Digest* sent over 10 million questionnaires in the mail to predict the outcome of the presidential election. The questionnaires went to a relatively wealthy segment of society (those having autos or telephones), and fewer than 25% were returned. The journal used these to predict an overwhelming victory by Alfred Landon over Franklin Roosevelt. The opposite result was predicted by George Gallup with a much smaller sample in the first scientific public poll taken for this purpose. (In fact, Roosevelt won with 62% of the vote.)

Another example of nonprobability sampling is what we might call the *streetcorner interview*: An interviewer stands at a specific location and conducts interviews by stopping whoever passes by. Severe biases may arise as a function of the time and location of the interview and the judgment of the interviewer in deciding whom to interview. For example, working people might be underrepresented if the interviews are conducted on weekdays between 9:00 A.M. and 5:00 P.M., and the racial or socioeconomic composition of the sample may be biased if the interviewer conducts the interviews in an upscale shopping mall.

Inferential statistical methods utilize assumptions about the probability that any particular sample is selected. They are not valid for data obtained with nonprobability samples, for which such probabilities are unknown.

Samples Based on Experimental Designs

In many sciences, data result from a planned *experiment*. The scientist has experimental control over the subjects' values on factors that can influence the variable of interest in the study, such as temperature, pressure, humidity, and so forth. Data obtained in such studies are called ***experimental data***. By contrast, data obtained from surveys are called ***observational data***. The researcher measures subjects' responses on the variables of interest, but has no experimental control over the subjects.

A major purpose of many experiments is to compare responses of subjects on some outcome measure, under different conditions. For instance, those conditions might be different drugs for treating some illness or different combinations of chemicals for manufacturing some product. The conditions are called ***treatments***. To obtain experimental data, the researcher needs a plan for assigning subjects to the different conditions being compared. These plans are called ***experimental designs***.

For instance, in the late 1980s, the Physicians' Health Study Research Group at Harvard Medical School designed a five-year randomized study to analyze whether regular intake of aspirin reduces mortality from cardiovascular disease. Of about 22,000 physicians, half were randomly chosen to take an aspirin every other day, and the remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared for the two groups.

By using randomization to determine who received the placebo and who received the aspirin, the researchers knew that the groups would roughly balance on factors that could affect heart attack rates, such as age and overall quality of health. If the physicians could decide on their own which treatment to take, the groups might have been out of balance on some important factor. Perhaps, for instance, younger physicians

would have been more likely to select the aspirin regimen; then, a lower heart attack rate among the aspirin group could simply represent younger subjects being less likely to suffer heart attacks.

Imbalance between groups is always a danger with observational studies, making it difficult to compare groups. For instance, if white students have a higher average score than black students on some standardized test, a variety of other unmeasured variables might account for that difference, such as parents' education or income or other environmental factors.

In social research, unfortunately, it is rarely possible to conduct controlled experiments. One cannot randomly assign subjects to treatments such as race or gender. This is also sometimes true in other sciences, particularly medical sciences. Consider, for instance, a study of whether passive smoking (being exposed to secondhand cigarette smoke on a regular basis) leads to higher rates of lung cancer. An experimental study might take a sample of children, randomly select half of them for placement in an environment where they are passive smokers, and place the other half in an environment where they are not exposed to smoke. Then, perhaps 60 years later the observation is whether each has developed lung cancer. Clearly, for many reasons, including time and ethics, it is not possible to conduct such an experimental study.

Regardless of whether a study is observational or experimental, randomization is an important feature in any study that involves making inferences. This randomization could take the form of randomly selecting a sample for an observational study, or randomly allocating subjects to different conditions for an experimental study.

2.3 Sampling and Nonsampling Variability

Even if a study wisely uses randomization in selecting a sample, the results of the study still depend on which sample of subjects is actually selected. Two researchers who select separate random samples from some population may have very little overlap, if any, between the two sample memberships. Therefore, the values of sample statistics will differ for the two samples, and the respective inferences based on these samples may differ.

Sampling Error

For instance, the Gallup and Harris organizations might each take a random sample of 1000 Americans, in order to estimate the percentage of Americans who give the president's performance in office a favorable rating. Based on the samples they select, Gallup might report an approval rating of 53%, whereas Harris might report one of 58%. This difference could reflect slightly different question wording, but even if the questions are worded exactly the same, the percentages would probably differ somewhat because the samples are different. For conclusions based on statistical methods to be worthwhile, one must determine the potential *sampling error*—the extent to which

the value of a statistic may differ from the parameter it predicts because of the way results vary from sample to sample.

Sampling Error

The ***sampling error*** of a statistic is the error that occurs when a statistic based on a sample estimates or predicts the value of a population parameter.

Suppose that the true population percentage giving the president a favorable rating is 56%. Then the Gallup organization, which predicted 53%, had a sampling error of $53\% - 56\% = -3\%$; the Harris organization, which predicted 58%, had a sampling error of $58\% - 56\% = 2\%$. In practice, of course, the sampling error is unknown, since the true values of population parameters are unknown. However, the methods of this text allow us to predict the size of the sampling error. For samples of size 1000, for instance, the sampling error for estimating percentages is usually no greater than 3% or 4%.

Random sampling guards against a systematic bias in the sampling error, such as a tendency to underestimate consistently or overestimate consistently the true parameter values. It also allows us to gauge the likely size of the sampling error. Variability also occurs in the values of sample statistics with nonrandom sampling, but the extent of that variability is not predictable as it is with random samples.

Other Sources of Variability

Other factors besides sampling error can introduce variability into results from samples and possibly cause bias. For instance, the sampling frame may suffer from ***undercoverage***. It may lack representation from some groups in the population of interest to us. A telephone survey will not reach homeless people or prison inmates or people not having a telephone. If its sampling frame consists of the names in a telephone directory, it will not reach those having an unlisted number. For many variables, responses by the homeless or by those with unlisted numbers might well tend to be considerably different from those actually sampled, leading to biased results.

Some subjects who are supposed to be in the sample may refuse to participate, or it may not be possible to reach them. This results in the problem of ***nonresponse***, which is a serious one for many surveys. If only half the intended sample was actually observed, we should worry about whether the half not observed differ from those observed in a way that causes bias in the overall results. Even if we select the sample randomly, the results are questionnable if there is substantial nonresponse, say, over 20%. Even in censuses, which are supposed to record data for all people in the country, some people are not observed or simply fail to cooperate.

In an interview, characteristics of the interviewer or other factors may affect the response in a way that introduces ***response bias***. Respondents might lie if they think their

response to a question is socially unacceptable. They may be more likely to give the response that they think the interviewer would prefer. An example is provided by a study by Lynn Sanders, a political scientist at the University of Chicago, on the effect of the race of the interviewer. Following a phone interview, respondents were asked whether they thought the interviewer was black or white (all were actually black). Perceiving a white interviewer resulted in more conservative opinions. For example, 14% agreed that "American society is fair to everyone" when they thought the interviewer was black, but 31% agreed to the same statement when posed by an interviewer the respondent thought was white (*Washington Post*, National Weekly Edition, June 26, 1995).

The way a variable is measured can have a large impact on the types of results observed. For instance, the wording of a question in a survey can greatly affect the responses. A Roper Poll was designed to determine the percentage of Americans who express some doubt that the Holocaust occurred. In response to the question, "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" 22% said it was possible the Holocaust never happened. The Roper organization later admitted that the question was worded in a confusing manner. When they asked, "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" only 1% said it was possible it never happened (*Newsweek*, July 25, 1994).

Finally, even the order in which questions are asked can influence the results dramatically. Crosson (1994) described a study that, during the Cold War, asked "Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?" and "Do you think Russia should let American newspaper reporters come in and send back whatever they want?" The percentage of yes responses to the first question was 36% when it was asked first and 73% when it was asked second.

In a fine book summarizing the potential difficulties with conducting and interpreting survey research, Crosson (1994) makes several recommendations. In particular, she notes that any newspaper report or TV story should say who sponsored and conducted the research, indicate how the questions were worded, and tell how the sample was selected and how large it was. She notes, "As a general rule, the less information that is available about the way a poll was conducted, the less it can be trusted."

Missing Data

A problem encountered in almost all large studies is ***missing data***. Some subjects do not provide responses for some of the variables measured. Standard software ignores cases for which observations are missing for at least one of the variables used in an analysis. This can result in much wasted information, however, and statisticians have recently developed methods that replace missing observations by predicted values based on patterns in the data. Most of this work is beyond the scope of this text, but we refer readers to Little and Rubin (1989) for a good introduction to this important topic.

STATISTICAL ANALYSIS IN PSYCHOLOGY AND EDUCATION

Sixth Edition

CHAPTER 4

MEASURES OF CENTRAL LOCATION

4.1 INTRODUCTION

Chapter 2 discussed the organization of collections of numbers in the form of frequenc distributions and how such distributions could be portrayed in graphic form. We no proceed to a consideration of how a collection of numbers, whether arranged in th form of a frequency distribution or not, may be described. How may indices or measur that are descriptive of the properties of collections of numbers be defined? Frequentl we wish to compare one collection of numbers with another collection. How is or collection the same as or different from another? What indices or measures can b used to make such comparisons possible? One property of a collection of numbers *central location*, and several indices, or measures, may be used to describe it.

The term *central location* refers to a central reference value which is usual close to the point of greatest concentration of the measurements and may in son sense be thought to typify the whole set. Measures of central location in common u are the mode, median, and arithmetic mean. Other less frequently used measures a the geometric mean and the harmonic mean. By far the most widely used measure central location is the arithmetic mean. This statistic is an appropriate measure central location for interval and ratio variables. The median and mode are sometim viewed as appropriate measures for ordinal and nominal variables, respectively,

though they can also be used with interval and ratio variables. The geometric mean and the harmonic mean have specialized applications.

The word *average* is commonly used as a general term and the equivalent of the term central location. When used in this way the arithmetic mean, the median, and the like are particular instances of it and are simply different kinds of averages. The word average is also used in common practice to refer to the arithmetic mean, as in the phrase "first year average."

4.2 THE ARITHMETIC MEAN

By definition the arithmetic mean is the sum of a set of measurements divided by the number of measurements in the set. Consider the following measurements: 7, 13, 22, 9, 11, 4. The sum of these measurements is 66. The arithmetic mean is, therefore, 66 divided by 6, or 11.

In general, if N measurements are represented by the symbols $X_1, X_2, X_3, \ldots, X_N$, the arithmetic mean in algebraic language is

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{4.1}$$

The symbol $\bar{X}$, spoken of as X bar, is used to denote the arithmetic mean of the values of X. The Greek letter sigma, $\sum_{i=1}^{N}$, describes the operation of summing the N measurements. The summation extends from $i = 1$ to $i = N$. Quite commonly the arithmetic mean is written simply as

$$\bar{X} = \frac{\sum X}{N} \tag{4.2}$$

The limits of the summation are omitted. The summation is understood to extend over all available values of X.

A bar above a symbol always denotes an arithmetic mean. Thus $\bar{Y}$ denotes the mean of a variable Y. The use of a bar to denote the mean is a widely used and preferred notational practice.

The reader should note also that $\Sigma X = N\bar{X}$. Thus the sum of a variable X is N times the mean of X. An awareness of this simple fact is useful in a variety of situations.

4.3 CALCULATING THE MEAN FROM FREQUENCY DISTRIBUTIONS

Consider a situation where different values of X occur more than once. The arithmetic mean is then obtained by multiplying each value of X by the frequency of its occurrence, adding together these products, and then dividing by the total number of measurements. Consider the following measurements: 11, 11, 12, 12, 12, 13, 13, 13, 13, 13, 14,

14, 15, 15, 15, 16, 16, 17, 17, 18. The value 11 occurs with a frequency of 2, 12 with a frequency of 3, 13 with a frequency of 5, and so on. These data may be written as follows:

X_i	f_i	f_iX_i
18	1	18
17	2	34
16	2	32
15	3	45
14	2	28
13	5	65
12	3	36
11	2	22
Total	20	280

This is a frequency distribution with a class interval of 1. The symbol f_i is used to denote the frequency of occurrence of the particular value X_i. Multiplying each value X_i by the frequency of its occurrence and adding together the products f_iX_i, we obtain the sum 280. The arithmetic mean is then 280 divided by 20, or 14.0.

In general, where $X_1, X_2, X_3, \ldots, X_k$ occur with frequencies $f_1, f_2, f_3, \ldots, f_k$, where k is the number of *different* values of X, the arithmetic mean

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k}{N} = \frac{\sum_{i=1}^{k} f_iX_i}{N} \tag{4.3}$$

Observe that here the summation is over k terms, the number of *different* values of the variable X. Observe also that $\sum_{i=1}^{N} X_i = \sum_{i=1}^{k} f_iX_i$. The above discussion suggests a simple method for calculating the mean from data grouped in the form of a frequency distribution regardless of the size of the class interval. The midpoint of the interval may be used to represent all values falling within the interval. We assume that the variable X takes values corresponding to the midpoints of the intervals, and these are weighted by the frequencies. We multiply the midpoints of the intervals by the frequencies, sum these products, and divide this sum by N to obtain the mean. More explicitly, the steps involved are as follows. *First*, calculate the midpoints of all intervals. *Second*, multiply each midpoint by the corresponding frequency. *Third*, sum the products of midpoints by frequencies. *Fourth*, divide this sum by N to obtain the mean. To illustrate, consider Table 4.1.

The midpoints of the intervals X_i appear in column 2. The frequencies f_i appear in column 3. The products of the midpoints by the frequencies f_iX_i are shown in column 4. The sum of these products $\sum_{i=1}^{k} f_iX_i$ is 1,612, N is 76, and the mean $\bar{X}$ is obtained by dividing 1,612 by 76 and is 21.21.

TABLE 4.1
Calculating th
scores

	1	
Class interval	M	
45–49		
40–44		
35–39		
30–34		
25–29		
20–24		
15–19		
10–14		
5–9		
0–4		
Total		

$\sum_{i=1}^{k} f_iX_i = 1{,}612$

4.4 DEVI

In statistical
X_i and the m

Here x_i, lowe
from the mea
Y, are sometim
about the me
scores instea
or $\bar{Y}$. Deviati

4.5 SOM
OF THE A

The arithmet
these is a ve
set from thei
22, 9, 11, ar

TABLE 4.1
Calculating the mean for distribution of test scores

1 Class interval	2 Midpoint X_i	3 Frequency f_i	4 Frequency × midpoint fX_i
45–49	47	1	47
40–44	42	2	84
35–39	37	3	111
30–34	32	6	192
25–29	27	8	216
20–24	22	17	374
15–19	17	26	442
10–14	12	11	132
5–9	7	2	14
0–4	2	0	0
Total		76	1,612

$\sum_{i=1}^{k} fX_i = 1{,}612 \qquad \bar{X} = \frac{1{,}612}{76} = 21.21$

4.4 DEVIATION FROM THE MEAN

In statistical work frequent use is made of the difference between a particular score X_i and the mean. Such a difference is spoken of as a deviation from the mean. Thus

$$x_i = X_i - \bar{X}$$

Here x_i, lowercase x_i, is used to denote such a deviation. Similarly y denotes a deviation from the mean of Y. Measurements or scores in their original form, denoted by X or Y, are sometimes known as *raw scores*. Measurements or scores expressed as deviations about the mean, denoted by x or y, are called *deviation scores*. The use of deviation scores instead of raw scores involves a change of origin. Raw scores have a mean $\bar{X}$ or $\bar{Y}$. Deviation scores have a mean of 0.

4.5 SOME PROPERTIES OF THE ARITHMETIC MEAN

The arithmetic mean has a number of interesting and useful properties. The first of these is a very simple property. *The sum of deviations of all the measurements in a set from their arithmetic mean is 0*. The arithmetic mean of the measurements 7, 13, 22, 9, 11, and 4 is 11. The deviations of these measurements from this mean are -4,

2, 11, −2, 0, and −7. The sum of these deviations is 0. Proof of this result is as follows:

$$\sum_{i=1}^{N} (X_i - \bar{X}) = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \bar{X} = N\bar{X} - N\bar{X} = 0 \tag{4.4}$$

Since $\bar{X} = \left(\sum_{i=1}^{N} X\right) \Big/ N$, it follows that $\sum_{i=1}^{N} X = N\bar{X}$. Also, adding $\bar{X}$, the mean, N times is the same as multiplying $\bar{X}$ by N; thus if $\bar{X}$ is 11 and N is 6, we observe that $11 + 11 + 11 + 11 + 11 + 11 = 6 \times 11 = 66$.

In many situations in statistics, use is made of the square of deviations from the mean; that is, quantities of the kind $(X_i - \bar{X})^2$ are considered. A second useful property of the mean involves the sum of squares of deviations from the mean, that is, the quantity, $\sum_{i=1}^{N} (X_i - \bar{X})^2$. *The sum of squares of deviations from the arithmetic mean is less than the sum of squares of deviations from any other value.* The deviations of the measurements 7, 13, 22, 9, 11, 4 from the mean 11 are −4, 2, 11, −2, 0, −7. The squares of these deviations are 16, 4, 121, 4, 0, 49. The sum of squares is 194. Had any other origin been selected, the sum of squares of deviations would be greater than the sum of squares about the mean. Select a different origin, say, 13. The deviations are −6, 0, 9, 4, 2, −9. Squaring these, we have 36, 0, 81, 16, 4, 81. The sum of these squares is 218, which is greater than the sum of squares about the mean. Selection of any other origin will demonstrate the same result.

This property of the mean indicates that it is the centroid, or center of gravity, of the set of measurements. Indeed, the mean is the central value about which the sum of squares of deviations is a minimum. This result may be readily demonstrated. Consider deviations from an origin $\bar{X} + c$, where $c \neq 0$. A deviation of an observation from this origin is

$$X_i - (\bar{X} + c) = (X_i - \bar{X}) - c \tag{4.5}$$

Squaring and summing over N observations, we obtain

$$\sum_{i=1}^{N} [X_i - (\bar{X} + c)]^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2 + \sum_{i=1}^{N} c^2 - 2c \sum_{i=1}^{N} (X_i - \bar{X}) \tag{4.6}$$

Because the sum of deviations about the mean is 0, the third term on the right is 0. Also c^2 summed N times is Nc^2, and we write

$$\sum_{i=1}^{N} [X_i - (\bar{X} + c)]^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2 + Nc^2 \tag{4.7}$$

This expression states that the sum of squares of deviations about an origin $\bar{X} + c$ may be viewed as comprising two parts, the sum of squares of deviations about the mean $\bar{X}$ and Nc^2. The quantity Nc^2 is always positive. Hence the sum of squares of deviations about an origin $\bar{X} + c$ will always be greater than the sum of squares about

$\bar{X}$. Thus the sum of squares of deviations about the arithmetic mean is less than the sum of squares of deviations about any other value.

Any mean calculated on a sample of size N is an estimate of a population mean, which is the value that would have been obtained were it possible to measure all members of the population. The distinction between sample and population values was discussed in Chapter 1. The mean has the property that for most distributions it is a more accurate, or more efficient, estimate of the population mean than other measures of central location, such as the median and mode, are of the population values they purport to estimate. It is subject to less error. This is one reason why it is the most frequently used measure of central location. Proof of this result is beyond the scope of this book.

Reference has been made to a number of properties of the arithmetic mean. What importance attaches to these properties, or why should they be discussed? The fact that the sum of deviations about the mean is 0 greatly simplifies many forms of algebraic manipulation. Any term involving the sum of deviations about the mean will vanish. The fact that the sum of squares of deviations about the mean is a minimum in effect implies an alternative definition of the mean; namely, *the mean is that measure of central location about which the sum of the squares is a minimum*. In effect, the mean is a measure of central location in the *least-squares* sense. The method of least squares is of considerable importance in statistics and is used, for example, in the fitting of lines and curves. The mean may be regarded as a point located by the method of least squares. The fact that the sample mean provides a better estimate of a population parameter than other measures of central location is of primary importance. Throughout statistics we are concerned with the problem of making statements about population values from our knowledge of sample values. Obviously, the more accurate these statements are, the better.

4.6 THE MEDIAN

Another commonly used measure of central location is the *median*. The median is a value such that half the observations fall above it and half below it. Consider the following values of X arranged in rank order, where R corresponds to the rank and N is an odd number:

X	2	7	16	19	20	25	27
R	1	2	3	4	5	6	7

In this example the median is 19. It corresponds to the middle rank. Three observations fall above it and three below it. If another observation, say 31, is added, then N is an even number, and the median by common convention is arbitrarily taken as the arithmetic mean of the two middle values, 19 and 20, that is, $(19 + 20)/2$ or 19.5.

With some data problems arise in calculating the median. Consider the following values of X:

X	7	7	7	8	8	8	9	9	10	10
R	1	2	3	4	5	6	7	8	9	10

For these 10 observations we are required to locate a point such that half the observations fall above that point and half below. Two different procedures may be used. The choice of procedure depends on whether the variable is viewed as *continuous* or *discrete*. For the above data, if the variable is continuous, the three 8s may be assumed to occupy the interval 7.5 to 8.5. The median is then obtained by linear interpolation. In this instance we interpolate two-thirds of the way into the interval to obtain a point above and below which half of the observations fall. The median is then 7.5 + .67 = 8.17. If the variable is not continuous but is discrete and assumes only integral values, then the score corresponding to the middle rank, which is $(N + 1)/2$, is taken as the median. For the illustrative data above if the data are discrete, the median is 8.

When certain values of the variable occur more than once, and the variable is *continuous* and not discrete, the median is calculated by the method used in calculating the median from data grouped in the form of a frequency distribution, as described in Section 4.7.

4.7 CALCULATING THE MEDIAN FROM FREQUENCY DISTRIBUTIONS

In calculating the median from data grouped in the form of a frequency distribution, the problem is to determine a value of the variable such that one-half the observations fall above this value and the other half below. The method will be illustrated with reference to the data in Table 4.2.

TABLE 4.2
Frequency distribution of psychological test scores

1 Class interval	2 Frequency	3 Cumulative frequency
45–49	1	76
40–44	2	75
35–39	3	73
30–34	6	70
25–29	8	64
20–24	17	56
15–19	26	39
10–14	11	13
5–9	2	2
0–4	0	0
Total	76	

First, record the cumulative frequencies as shown in column 3. *Second*, determine $N/2$, one-half the number of cases, in this example 38. *Third*, find the class interval in which the 38th case, the middle case, falls. The 38th case falls within the interval 15 to 19, and the exact limits of this interval are 14.5 and 19.5. Clearly, the 38th case falls very close to the top of this interval because we know from an examination of our cumulative frequencies that 39 cases fall below the top of this interval, that is, below 19.5. *Fourth*, interpolate between the exact limits of the interval to find a value above and below which 38 cases fall. To interpolate, observe that 26 cases fall within the limits 14.5 and 19.5, and we assume that these 26 cases are uniformly distributed in rectangular fashion between these exact limits. Now to arrive at the 38th, or middle, case we require 25 of the 26 cases within this interval, because $2 + 11 + 25 = 38$. This means that we must find a point between 14.5 and 19.5 such that 25 cases fall below and 1 case above this point. The proportion of the interval we require is $\frac{25}{26}$, which is $\frac{25}{26} \times 5$ units of score, or 4.81. We add this to the lower limit of the interval to obtain the median, which is 14.50 + 4.81, or 19.31.

Let us summarize the steps involved:

1. Compute the cumulative frequencies.
2. Determine $N/2$, one-half the number of cases.
3. Find the class interval in which the middle case falls, and determine the exact limits of this interval.
4. Interpolate to find a value on the scale above and below which one-half the total number of cases falls. This is the median.

For the student who has difficulty in following the above, a simple formula may be employed.

$$\text{Median} = L + \frac{N/2 - F}{f_m} h \tag{4.8}$$

where L = exact lower limit of interval containing the median
F = sum of all frequencies below L
f_m = frequency of interval containing median
N = number of cases
h = class interval

In the present example $L = 14.5$, $F = 13$, $f_m = 26$, $N = 76$, and $h = 5$. We have

$$\text{Median} = 14.5 + \frac{\frac{76}{2} - 13}{26} \times 5 = 19.31$$

This method assumes that the observations within the interval containing the median are uniformly distributed over the range of that interval, and simple linear interpolation is appropriate.

Data for a discrete variable may be encountered which have been arranged in the form of a frequency distribution. With such data the median is the midpoint of the interval containing the median.

4.8 PROPERTIES OF THE MEDIAN

The reader will recall that the arithmetic mean has the property that the sum of squares of deviations from it is less than the sum of squares of deviations about any other value. In effect the mean $\bar{X}$ is a value such that $\Sigma(X - \bar{X})^2$ is a minimum. The median has an analogous property. The sum of absolute deviations (deviations without sign) about the median is less than the sum of absolute deviations about any other value. If we denote an absolute deviation from the median as $|X - \text{mdn}|$, then the median is a value such that $\Sigma|X - \text{mdn}|$ is a minimum.

Stavig (1978) points out that if a set of discrete values is treated as continuous, the median so calculated may not satisfy the requirement that $\Sigma|X - \text{mdn}|$ is a minimum. Consider the observations 7, 7, 7, 8, 8, 8, 9, 9, 10, 10. If the variable is viewed as discrete, the median is 8 and the sum of absolute deviations about it is 9. If the variable is treated as continuous, the median is 8.17 and the sum of absolute deviations is 10.83. Why does this discrepancy in the sum of absolute deviations occur? If the variable is viewed as continuous, any consideration of the sum of absolute deviations of the *original values* from the median is simply incorrect. An underlying continuous scale has been assumed. The median is a point on that scale corresponding to the middle rank. Every other rank has, however, a corresponding point on this scale. Consequently the appropriate sum of absolute deviations is the sum of absolute deviations of these points on the underlying scale from the median. Values corresponding to each rank on this scale may be readily calculated using a formula given by Stavig (1978), which is a modification of formula (4.8): $\hat{X} = L + [(R - .5 - F)/f_w]h$. Here $\hat{X}$ is the score on the underlying scale, R is the rank from 1 to N without regard for ties, and the remaining terms are as in formula (4.8). Note that this formula is the same as formula (4.8) except that $(R - .5)$ has been substituted for $N/2$. In the present example values of $\hat{X}$ corresponding to different values of R are 6.67, 7.00, 7.33, 7.67, 8.00, 8.33, 8.67, 9.00, 9.33, and 9.67. The median is 8.17. The sum of absolute deviations about this value is 8.33.

4.9 THE MODE

Another measure of central location is the *mode*. In situations where different values of X occur more than once, the mode is the most frequently occurring value. Consider the observations 11, 11, 12, 12, 12, 13, 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 16, 16, 17, 17, 18. Here the value 13 occurs five times, more frequently than any other value; hence the mode is 13.

In situations where all values of X occur with equal frequency, where that frequency may be equal to or greater than 1, no modal value can be calculated. Thus for the set of observations 2, 7, 16, 19, 20, 25, and 27 no mode can be obtained. Similarly, the observations 2, 2, 2, 7, 7, 7, 16, 16,16, 19, 19, 19, 20, 20, 20, 25,

25, 25, 27, 27, 27 do not permit the calculation of a modal value. All values occur with a frequency of 3.

In the case where two adjacent values of X occur with the same frequency, which is larger than the frequency of occurrence of other values of X, the mode may be taken rather arbitrarily as the mean of the two adjacent values of X. Consider the observations 11, 11, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 14, 15, 15, 16, 16, 17, 18. Here the values 13 and 14 both occur with a frequency of 4, which is greater than the frequency of occurrence of the remaining values. The mode may be taken as (13 + 14)/2, or 13.5.

Where two nonadjacent values of X occur such that the frequencies of both are greater than the frequencies in adjacent intervals, then each value of X may be taken as a mode and the set of observations may be spoken of as bimodal. Consider the observations 11, 11, 12, 12, 12, 13, 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 16, 16, 16, 17, 17, 18. Here the value 13 occurs five times, and this is greater than the frequency of occurrence of the adjacent values. Also 15 occurs four times, and this is also greater than the frequency of occurrence of the adjacent values. This set of observations may be said to be bimodal.

With data grouped in the form of a frequency distribution, the mode is taken as the midpoint of the class interval with the largest frequency.

The mode is a statistic of limited practical value. It does not lend itself to algebraic manipulation. For distributions with two or more modes, such modes are obviously not measures of central location. The mode can be considered a measure of central location only for distributions that taper off systematically toward the extremities.

4.10 COMPARISON OF THE MEAN, MEDIAN, AND MODE

The arithmetic mean may be regarded as an appropriate measure of central location for interval and ratio variables. All the particular values of the variable are incorporated in its calculation. The median is an ordinal statistic. Its calculation is based on the ordinal properties of the data. If the observations are arranged in order, the median is the middle value. Its calculation does not incorporate all the particular values of the variable, but merely the fact of their occurrence above or below the middle value. Thus the sets of numbers 5, 7, 20, 24, 25 and 10, 15, 20, 52, 63 have the same median, namely, 20, although their means are quite different. The mode, the value or class with the greatest frequency, is a nominal statistic. Its calculation does not depend on particular values of the variable or their order, but merely on their frequency of occurrence.

A comparison of the mean, median, and mode may be made when all three have been calculated for the same frequency distribution. If the frequency distribution is represented graphically, the mean is a point on the horizontal axis which corresponds to the centroid, or center of gravity, of the distribution. If a cutout of the distribution is made from heavy cardboard and balanced on a knife edge, the point of balance will be the mean. The median is a point on the horizontal axis where the ordinate divides the total area under the curve into two equal parts. Half the area falls to the left and

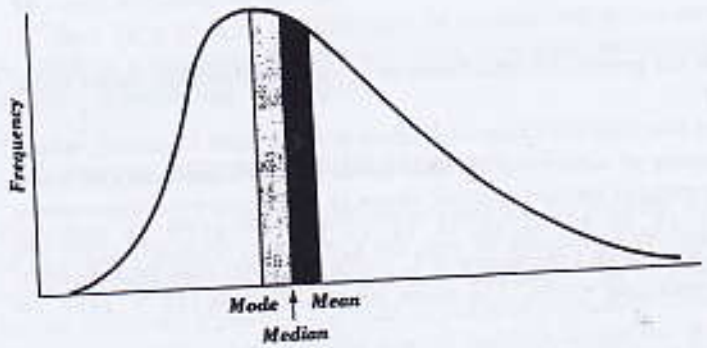


FIGURE 4.1
Relation between the mean, median, and mode in a positively skewed frequency distribution.

half to the right of the ordinate at the median. The mode is a point on the horizontal axis which corresponds to the highest point of the curve.

If the frequency distribution is symmetrical, the mean, median, and mode coincide. If the frequency distribution is skewed, these three measures do not coincide. Figure 4.1 shows the mean, median, and mode for a positively skewed frequency distribution. We note that the mean is greater than the median, which in turn is greater than the mode. If the distribution is negatively skewed the reverse relation holds.

A question may be raised regarding the appropriate choice of a measure of central location. In practical situations this question is rarely in doubt. The arithmetic mean is usually to be preferred to either the median or the mode. It is rigorously defined, easily calculated, and readily amenable to algebraic treatment. It provides also a better estimate of the corresponding population parameter than either the median or the mode.

The median is, however, to be preferred in some situations. Observations may occur which appear to be atypical of the remaining observations in the set. Such observations may greatly affect the value of the mean. Consider the observations 2, 3, 3, 4, 7, 9, 10, 11, 86. Observation 86 is quite atypical of the remaining observations, and its presence greatly affects the value of the mean. The mean is 15, a value greater than eight of the nine observations. The median is 7. Under circumstances such as this it may prove advisable in treating the data to use statistical procedures that are based on the ordinal properties of the data in preference to procedures that incorporate all the particular values of the variable and may be grossly affected by atypical values. The median, an ordinal statistic, may under such circumstances be preferred to the mean. In the above example the set of observations is grossly asymmetrical. If the distribution of the variables shows gross asymmetry, the median may be the preferred statistic, because, regardless of the asymmetry of the distribution, it can always be interpreted as the middle value.

For a strictly nominal variable the mode, the most frequently occurring class or

value, is the only "most typical" statistic that can be used. It is rarely used with interval, ratio, and ordinal variables where means and medians can be calculated.

4.11 OTHER MEASURES OF CENTRAL LOCATION

The arithmetic mean is obtained by adding together all N measurements in a set and dividing by N. Thus the mean is a particular function of a set of measurements. Other functions may be defined. For example, all N measurements may be multiplied together and the Nth root of the product obtained. This is the *geometric mean*, denoted by GM. Thus

$$\text{GM} = \sqrt[N]{X_1 \cdot X_2 \cdot \cdots \cdot X_N}$$

With two numbers, say 3 and 6, the geometric mean is $\sqrt{3 \times 6} = \sqrt{18} = 4.24$. The geometric mean of the numbers 2, 3, 8, 10 is $\text{GM} = \sqrt[4]{2 \times 3 \times 8 \times 10} = \sqrt[4]{480} = 4.68$. Note that $\text{GM}^4 = 4.68^4 = 480$.

The geometric mean is used as a measure of central location when the measurements are ratios, and the variable meets the criteria required for ratio measurement. It has limited use in psychology and education, but has frequent use in areas such as economics, demography, and sociology. It has a number of applications in this book although it may not be identified as the geometric mean. For a simple example, expressions of the kind $\sqrt{\Sigma x^2 \Sigma y^2}$ are geometric means.

The geometric mean has a simple geometric interpretation. Numbers such as 3 and 6 can be represented as distances between points. The product of two numbers can be represented as an area. Thus a rectangular figure 3 inches wide and 6 inches long has an area of 18 square inches. The square root of 18, or 4.24, is the dimension of a square whose area is equal to that of the rectangle 3 by 6. Likewise the product of three numbers is a volume, and the cube root of their product is the dimension of a cube with the same volume. This argument can be generalized to any value of N. When viewed in geometric terms the geometric mean is the dimension of a symmetric N-dimensional "cube" with the same volume as an N-dimensional figure whose volume is $\text{GM}^N = X_1 \cdot X_2 \cdot \cdots \cdot X_N$.

BASIC TERMS AND CONCEPTS

Average
Arithmetic mean
Raw score
Deviation score
Sum of deviations about mean

CHAPTER 2

FREQUENCY DISTRIBUTIONS AND THEIR GRAPHIC REPRESENTATION

2.1 INTRODUCTION

The data obtained from the conduct of experiments or surveys are frequently collections of numbers. Simple inspection of a collection of numbers will ordinarily communicate very little to the understanding of the investigator. Some form of classification and description of these numbers is required to assist interpretation and to enable the information which the numbers contain to emerge. Under certain circumstances advantages attach to the classification of data in the form of frequency distributions. Such classification may help the investigator to understand important features of the data. This chapter discusses the arrangement of data in the form of frequency distributions, the graphic representation of frequency distributions, and the ways in which one frequency distribution may differ from another. Chapters 3, 4, and 5 to follow, discuss the statistics used to describe the properties of frequency distributions, or the properties of the collections of numbers which these distributions comprise.

2.2 FREQUENCY DISTRIBUTIONS

A coin is tossed 10 times, and the following results are obtained: H H T H T H H H T H. Here the number of times heads occurs, the frequency of heads, is 7, and the

number of times tails occurs, the frequency of tails, is 3. These data could be arranged in the form

	f
H	7
T	3
Total	10

The symbol f denotes the frequency. This arrangement of the data is a frequency distribution. It is an arrangement of the data that shows how often heads and tails occur.

A die is rolled 24 times and the following results recorded:

6 3 1 4 1 6
5 2 4 3 5 5
4 1 5 2 2 6
5 3 3 4 5 5

The numbers appearing when a die is rolled constitute a variable X, which is limited to the values 1, 2, 3, 4, 5, and 6. In the above data, 6 occurs 3 times, 5 occurs 7 times, and so on. These data may be arranged as follows:

X	f
6	3
5	7
4	4
3	4
2	3
1	3
Total	24

This arrangement is a frequency distribution. It shows the frequency of occurrence of the values 1, 2, 3, 4, 5, and 6.

Consider the data of Table 2.1. These are the IQs of 100 children obtained from the administration of a psychological test. These scores range from 67 to 134. By counting the number of times each score occurs, an arrangement of the data as shown in Table 2.2 is obtained. This arrangement shows how many times each score occurs, and is a frequency distribution. Note, however, that the number of groupings of scores is large. Usually it is advisable to reduce the number of classes by arranging the data into arbitrarily defined groupings of the variable; thus all scores within the range 65

TABLE 2.1
Intelligence quotients made by 100 pupils on a mental test

109	111	82	105	134
113	90	79	100	117
80	90	121	75	93
99	90	92	96	82
101	104	80	81	83
104	93	109	72	110
111	91	109	111	81
122	83	92	101	77
99	103	93	91	67
108	93	84	84	100
102	84	96	89	81
107	95	91	107	102
109	93	82	103	116
86	78	73	104	104
103	108	76	94	108
72	87	121	80	127
105	103	106	119	90
93	89	110	103	100
99	79	117	114	117
93	82	98	89	119

to 69, that is, all scores with the values 65, 66, 67, 68, and 69, may be grouped together. All scores within the ranges 70 to 74, 75 to 79, and so on, may be similarly grouped. Such groupings of data are usually done by entering a tally mark for each score opposite the range of the variable within which it falls and counting those tally marks to obtain the number of cases within the range. This procedure is shown in Table 2.3.

The arbitrarily defined groupings of the variable are called *class intervals*. In Table 2.3 the class interval is 5. This arrangement of data is also a frequency distribution, and the number of cases falling within each class interval is a frequency. The only difference between Tables 2.2 and 2.3 is in the class interval, which is 1 in the former table and 5 in the latter.

In general, a frequency distribution is any arrangement of the data that shows the frequency of occurrence of different values of the variable or the frequency of occurrence of values falling within arbitrarily defined ranges of the variable known as class intervals.

2.3 CONVENTIONS REGARDING CLASS INTERVALS

In the arrangement of data with a class interval of 1, as shown in Table 2.2, the original observations are retained and may be reconstructed directly from the frequency

TABLE 2.2
Frequency distribution of intelligence quotients of Table 2.1 with as many classes as score values

Score	f	Score	f	Score	f	Score	f
134	1	117	3	100	3	83	2
133	...	116	1	99	3	82	4
132	...	115	...	98	1	81	3
131	...	114	1	97	...	80	3
130	...	113	1	96	2	79	2
129	...	112	...	95	1	78	1
128	...	111	3	94	1	77	1
127	1	110	2	93	7	76	1
126	...	109	4	92	2	75	1
125	...	108	3	91	3	74	...
124	...	107	2	90	4	73	1
123	...	106	1	89	3	72	2
122	1	105	2	88	...	71	...
121	2	104	4	87	1	70	...
120	...	103	5	86	1	69	...
119	2	102	2	85	...	68	...
118	...	101	2	84	3	67	1

TABLE 2.3
Frequency distribution of the intelligence quotients of Table 2.1

Class interval	Tally	Frequency
130–134	/	1
125–129	/	1
120–124	///	3
115–119	𝍸 /	6
110–114	𝍸 //	7
105–109	𝍸 𝍸 //	12
100–104	𝍸 𝍸 𝍸 /	16
95–99	𝍸 //	7
90–94	𝍸 //// 𝍸 //	17
85–89	𝍸	5
80–84	𝍸 𝍸 𝍸	15
75–79	𝍸 /	6
70–74	///	3
65–69	/	1
Total		100

distribution without loss of information. If the class interval is greater than 1, say, 3, 5, or 10, some loss of information regarding individual observations is incurred; that is, the original observations cannot be reproduced exactly from the frequency distribution. If the class interval is large in relation to the total range of the set of observations, this loss of information may be appreciable. If the class interval is small, the classification of data in the form of a frequency distribution may lead to very little gain in convenience over the utilization of the original observations.

The rules listed below are widely used in the selection of class intervals and lead in most cases to a convenient handling of the data.

1. Select a class interval of such a size that between 10 and 20 such intervals will cover the total range of the observations. For example, if the smallest observation in a set were 7 and the largest 156, a class interval of 10 would be appropriate and would result in an arrangement of the data into 16 intervals. If the smallest observation were 2 and the largest 38, a class interval of 3 would result in an arrangement of 14 intervals. If the observations ranged from 9 to 20, a class interval of 1 would be convenient.
2. Select class intervals with a range of 1, 3, 5, 10, or 20 points. These will meet the requirements of most sets of data.
3. Start the class interval at a value which is a multiple of the size of that interval. For example, with a class interval of 5, the intervals should start with the values 5, 10, 15, 20, etc.
4. Arrange the class intervals according to the order of magnitude of the observations they include, the class interval containing the largest observations being placed at the top.

2.4 EXACT LIMITS OF THE CLASS INTERVAL

Where the variable under consideration is continuous, and not discrete, we select a unit of measurement and record our observations as discrete values. When we record an observation in discrete form and the variable is a continuous one, we imply that the value recorded represents a value falling within certain limits. These limits are usually taken as one-half unit above and below the value reported. Thus when we report a measurement to the nearest inch, say, 16 inches, we mean that if a more accurate form of measurement had been used, the value obtained would fall within the limits 15.5 and 16.5 inches.

Strictly speaking the limits are 15.5 to 16.499, where the latter figure is a recurring decimal, but for convenience we write the limits as 15.5 to 16.5. Similarly, a measurement made to the nearest tenth part of an inch, say, 31.7 inches, is understood to fall within the limits 31.65 and 31.75 inches. In a reaction-time experiment a particular observation measured to the nearest thousandth of a second might be, say, .196 second. This assumes that had a more accurate timing device been used the measurement would have been found to fall somewhere within the limits .1955 and .1965 second.

TABLE 2.4
Class intervals, exact limits, and midpoints for frequency distribution of intelligence quotients

1 Class interval	2 Exact limits	3 Midpoint of interval	4 Frequency
130–134	129.5–134.5	132.0	1
125–129	124.5–129.5	127.0	1
120–124	119.5–124.5	122.0	3
115–119	114.5–119.5	117.0	6
110–114	109.5–114.5	112.0	7
105–109	104.5–109.5	107.0	12
100–104	99.5–104.5	102.0	16
95–99	94.5–99.5	97.0	7
90–94	89.5–94.5	92.0	17
85–89	84.5–89.5	87.0	5
80–84	79.5–84.5	82.0	15
75–79	74.5–79.5	77.0	6
70–74	69.5–74.5	72.0	3
65–69	64.5–69.5	67.0	1
Total			100

Class intervals are usually recorded to the nearest unit and thereby reflect the accuracy of measurement. For various reasons it is frequently necessary to think in terms of so-called *exact limits* of the class interval. These are sometimes spoken of as *class boundaries*, or *end values*, and sometimes as *real limits*. Consider the class interval 95 to 99 in Table 2.3. We grouped within this interval all measurements taking the values 95, 96, 97, 98, and 99. The limits of the lower values are 94.5 and 95.5, while those of the upper value are 98.5 and 99.5. The total range, or exact limits, which the interval is presumed to cover is then clearly 94.5 and 99.5, which means all values greater than or equal to 94.5 and less than 99.5.

The above discussion is applicable to continuous variables only. With discrete variables no distinction need be made between the class interval and the exact limits of the interval, the two being identical.

Table 2.4 shows the frequency distribution of the IQs of Table 2.1. Column 1 shows the class interval as usually written, while column 2 records the exact limits. In practice, of course, the exact limits are rarely recorded as in Table 2.4.

2.5 DISTRIBUTION OF OBSERVATIONS WITHIN THE CLASS INTERVAL

The grouping of data in class intervals results in a loss of information regarding the individual observations themselves. Scores may differ one from another within a limited

LANGUAGE ASSESSMENT

Principles and Classroom Practices

Longman

longman.com

H. DOUGLAS BROWN

CHAPTER 2

PRINCIPLES OF LANGUAGE ASSESSMENT

This chapter explores how principles of language assessment can and should be applied to formal tests, but with the ultimate recognition that these principles also apply to assessments of all kinds. In this chapter, these principles will be used to evaluate an existing, previously published, or created test. Chapter 3 will center on how to use those principles to design a good test.

How do you know if a test is effective? For the most part, that question can be answered by responding to such questions as: Can it be given within appropriate administrative constraints? Is it dependable? Does it accurately measure what you want it to measure? These and other questions help to identify five cardinal criteria for "testing a test": practicality, reliability, validity, authenticity, and washback. We will look at each one, but with no priority order implied in the order of presentation.

PRACTICALITY

An effective test is **practical**. This means that it

- is not excessively expensive,
- stays within appropriate time constraints,
- is relatively easy to administer, and
- has a scoring/evaluation procedure that is specific and time-efficient.

A test that is prohibitively expensive is impractical. A test of language proficiency that takes a student five hours to complete is impractical—it consumes more time (and money) than necessary to accomplish its objective. A test that requires individual one-on-one proctoring is impractical for a group of several hundred test-takers and only a handful of examiners. A test that takes a few minutes for a student to take and several hours for an examiner to evaluate is impractical for most classroom situations. A test that can be scored only by computer is impractical if the test takes place a thousand miles away from the nearest computer. The value and quality of a test sometimes hinge on such nitty-gritty, practical considerations.

Here's a little horror story about practicality gone awry. An administrator of a six-week summertime short course needed to place the 50 or so students who had enrolled in the program. A quick search yielded a copy of an old English Placement Test from the University of Michigan. It had 20 listening items based on an audiotape and 80 items on grammar, vocabulary, and reading comprehension, all multiple-choice format. A scoring grid accompanied the test. On the day of the test, the required number of test booklets had been secured, a proctor had been assigned to monitor the process, and the administrator and proctor had planned to have the scoring completed by later that afternoon so students could begin classes the next day. Sounds simple, right? Wrong.

The students arrived, test booklets were distributed, and directions were given. The proctor started the tape. Soon students began to look puzzled. By the time the tenth item played, everyone looked bewildered. Finally, the proctor checked a test booklet and was horrified to discover that the wrong tape was playing; it was a tape for another form of the same test! Now what? She decided to randomly select a short passage from a textbook that was in the room and give the students a dictation. The students responded reasonably well. The next 80 non-tape-based items proceeded without incident, and the students handed in their score sheets and dictation papers.

When the red-faced administrator and the proctor got together later to score the tests, they faced the problem of how to score the dictation—a more subjective process than some other forms of assessment (see Chapter 6). After a lengthy exchange, the two established a point system, but after the first few papers had been scored, it was clear that the point system needed revision. That meant going back to the first papers to make sure the new system was followed.

The two faculty members had barely begun to score the 80 multiple-choice items when students began returning to the office to receive their placements. Students were told to come back the next morning for their results. Later that evening, having combined dictation scores and the 80-item multiple-choice scores, the two frustrated examiners finally arrived at placements for all students.

It's easy to see what went wrong here. While the listening comprehension section of the test was apparently highly practical, the administrator had failed to check the materials ahead of time (which, as you will see below, is a factor that touches on unreliability as well). Then, they established a scoring procedure that did not fit into the time constraints. In classroom-based testing, time is almost always a crucial practicality factor for busy teachers with too few hours in the day!

RELIABILITY

A **reliable** test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. The issue of reliability of a test may best be addressed by considering a number of factors that may contribute to the unreliability of a test. Consider the

following possibilities (adapted from Mousavi, 2002, p. 804): fluctuations in the student, in scoring, in test administration, and in the test itself.

Student-Related Reliability

The most common learner-related issue in reliability is caused by temporary illness, fatigue, a "bad day," anxiety, and other physical or psychological factors, which may make an "observed" score deviate from one's "true" score. Also included in this category are such factors as a test-taker's "test-wiseness" or strategies for efficient test taking (Mousavi, 2002, p. 804).

Rater Reliability

Human error, subjectivity, and bias may enter into the scoring process. **Inter-rater reliability** occurs when two or more scorers yield inconsistent scores of the same test, possibly for lack of attention to scoring criteria, inexperience, inattention, or even preconceived biases. In the story above about the placement test, the initial scoring plan for the dictations was found to be unreliable—that is, the two scorers were not applying the same standards.

Rater-reliability issues are not limited to contexts where two or more scorers are involved. **Intra-rater reliability** is a common occurrence for classroom teachers because of unclear scoring criteria, fatigue, bias toward particular "good" and "bad" students, or simple carelessness. When I am faced with up to 40 tests to grade in only a week, I know that the standards I apply—however subliminally—to the first few tests will be different from those I apply to the last few. I may be "easier" or "harder" on those first few papers or I may get tired, and the result may be an inconsistent evaluation across all tests. One solution to such intra-rater unreliability is to read through about half of the tests before rendering any final scores or grades, then to recycle back through the whole set of tests to ensure an even-handed judgment. In tests of writing skills, rater reliability is particularly hard to achieve since writing proficiency involves numerous traits that are difficult to define. The careful specification of an analytical scoring instrument, however, can increase rater reliability (J. D. Brown, 1991).

Test Administration Reliability

Unreliability may also result from the conditions in which the test is administered. I once witnessed the administration of a test of aural comprehension in which a tape recorder played items for comprehension, but because of street noise outside the building, students sitting next to windows could not hear the tape accurately. This was a clear case of unreliability caused by the conditions of the test administration. Other sources of unreliability are found in photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the conditions of desks and chairs.

Test Reliability

Sometimes the nature of the test itself can cause measurement errors. If a test is too long, test-takers may become fatigued by the time they reach the later items and hastily respond incorrectly. Timed tests may discriminate against students who do not perform well on a test with a time limit. We all know people (and you may be included in this category!) who "know" the course material perfectly but who are adversely affected by the presence of a clock ticking away. Poorly written test items (that are ambiguous or that have more than one correct answer) may be a further source of test unreliability.

VALIDITY

By far the most complex criterion of an effective test—and arguably the most important principle—is **validity,** "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Gronlund, 1998, p. 226). A valid test of reading ability actually measures reading ability—not 20/20 vision, nor previous knowledge in a subject, nor some other variable of questionable relevance. To measure writing ability, one might ask students to write as many words as they can in 15 minutes, then simply count the words for the final score. Such a test would be easy to administer (practical), and the scoring quite dependable (reliable). But it would not constitute a valid test of writing ability without some consideration of comprehensibility, rhetorical discourse elements, and the organization of ideas, among other factors.

How is the validity of a test established? There is no final, absolute measure of validity, but several different kinds of evidence may be invoked in support. In some cases, it may be appropriate to examine the extent to which a test calls for performance that matches that of the course or unit of study being tested. In other cases, we may be concerned with how well a test determines whether or not students have reached an established set of goals or level of competence. Statistical correlation with other related but independent measures is another widely accepted form of evidence. Other concerns about a test's validity may focus on the consequences—beyond measuring the criteria themselves—of a test, or even on the test-taker's perception of validity. We will look at these five types of evidence below.

Content-Related Evidence

If a test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behavior that is being measured, it can claim content-related evidence of validity, often popularly referred to as **content validity** (e.g., Mousavi, 2002; Hughes, 2003). You can usually identify content-related evidence observationally if you can clearly define the achievement that you are measuring. A test of tennis competency that asks someone to run a 100-yard

dash obviously lacks content validity. If you are trying to assess a person's ability to speak a second language in a conversational setting, asking the learner to answer paper-and-pencil multiple-choice questions requiring grammatical judgments does not achieve content validity. A test that requires the learner actually to speak within some sort of authentic context does. And if a course has perhaps ten objectives but only two are covered in a test, then content validity suffers.

Consider the following quiz on English articles for a high-beginner level of a conversation class (listening and speaking) for English learners.

English articles quiz

> *Directions: The purpose of this quiz is for you and me to find out how well you know and can apply the rules of article usage. Read the following passage and write* a/an, the, *or* 0 *(no article) in each blank.*
>
> Last night, I had (1)_____ very strange dream. Actually, it was (2)_____ nightmare! You know how much I love (3)_____ zoos. Well, I dreamt that I went to (4)_____ San Francisco zoo with (5)_____ few friends. When we got there, it was very dark, but (6)_____ moon was out, so we weren't afraid. I wanted to see (7)_____ monkeys first, so we walked past (8)_____ merry-go-round and (9)_____ lions' cages to (10)_____ monkey section.
>
> (The story continues, with a total of 25 blanks to fill.)

The students had had a unit on zoo animals and had engaged in some open discussions and group work in which they had practiced articles, all in listening and speaking modes of performance. In that this quiz uses a familiar setting and focuses on previously practiced language forms, it is somewhat content valid. The fact that it was administered in written form, however, and required students to read the passage and write their responses makes it quite low in content validity for a listening/speaking class.

There are a few cases of highly specialized and sophisticated testing instruments that may have questionable content-related evidence of validity. It is possible to contend, for example, that standard language proficiency tests, with their context-reduced, academically oriented language and limited stretches of discourse, lack content validity since they do not require the full spectrum of communicative performance on the part of the learner (see Bachman, 1990, for a full discussion). There is good reasoning behind such criticism; nevertheless, what such proficiency tests lack in content-related evidence they may gain in other forms of evidence, not to mention practicality and reliability.

Another way of understanding content validity is to consider the difference between **direct** and **indirect** testing. Direct testing involves the test-taker in actually performing the target task. In an indirect test, learners are not performing the

task itself but rather a task that is related in some way. For example, if you intend to test learners' oral production of syllable stress and your test task is to have learners mark (with written accent marks) stressed syllables in a list of written words, you could, with a stretch of logic, argue that you are indirectly testing their oral production. A direct test of syllable production would have to require that students actually produce target words orally.

The most feasible rule of thumb for achieving content validity in classroom assessment is to test performance directly. Consider, for example, a listening/speaking class that is doing a unit on greetings and exchanges that includes discourse for asking for personal information (name, address, hobbies, etc.) with some form-focus on the verb *to be*, personal pronouns, and question formation. The test on that unit should include all of the above discourse and grammatical elements and involve students in the actual performance of listening and speaking.

What all the above examples suggest is that content is not the *only* type of evidence to support the validity of a test, but classroom teachers have neither the time nor the budget to subject quizzes, midterms, and final exams to the extensive scrutiny of a full construct validation (see below). Therefore, it is critical that teachers hold content-related evidence in high esteem in the process of defending the validity of classroom tests.

Criterion-Related Evidence

A second form of evidence of the validity of a test may be found in what is called criterion-related evidence, also referred to as **criterion-related validity**, or the extent to which the "criterion" of the test has actually been reached. You will recall that in Chapter 1 it was noted that most classroom-based assessment with teacher-designed tests fits the concept of criterion-referenced assessment. In such tests, specified classroom objectives are measured, and implied predetermined levels of performance are expected to be reached (80 percent is considered a minimal passing grade).

In the case of teacher-made classroom assessments, criterion-related evidence is best demonstrated through a comparison of results of an assessment with results of some other measure of the same criterion. For example, in a course unit whose objective is for students to be able to orally produce voiced and voiceless stops in all possible phonetic environments, the results of one teacher's unit test might be compared with an independent assessment—possibly a commercially produced test in a textbook—of the same phonemic proficiency. A classroom test designed to assess mastery of a point of grammar in communicative use will have criterion validity if test scores are corroborated either by observed subsequent behavior or by other communicative measures of the grammar point in question.

Criterion-related evidence usually falls into one of two categories: concurrent and predictive validity. A test has **concurrent validity** if its results are supported by other concurrent performance beyond the assessment itself. For example, the validity of a high score on the final exam of a foreign language course will be substantiated

by actual proficiency in the language. The **predictive validity** of an assessment becomes important in the case of placement tests, admissions assessment batteries, language aptitude tests, and the like. The assessment criterion in such cases is not to measure concurrent ability but to assess (and predict) a test-taker's likelihood of future success.

Construct-Related Evidence

A third kind of evidence that can support validity, but one that does not play as large a role for classroom teachers, is construct-related validity, commonly referred to as **construct validity**. A construct is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Constructs may or may not be directly or empirically measured—their verification often requires inferential data. "Proficiency" and "communicative competence" are linguistic constructs; "self-esteem" and "motivation" are psychological constructs. Virtually every issue in language learning and teaching involves theoretical constructs. In the field of assessment, construct validity asks, "Does this test actually tap into the theoretical construct as it has been defined?" Tests are, in a manner of speaking, operational definitions of constructs in that they operationalize the entity that is being measured (see Davidson, Hudson, & Lynch, 1985).

For most of the tests that you administer as a classroom teacher, a formal construct validation procedure may seem a daunting prospect. You will be tempted, perhaps, to run a quick content check and be satisfied with the test's validity. But don't let the concept of construct validity scare you. An informal construct validation of the use of virtually every classroom test is both essential and feasible.

Imagine, for example, that you have been given a procedure for conducting an oral interview. The scoring analysis for the interview includes several factors in the final score: pronunciation, fluency, grammatical accuracy, vocabulary use, and sociolinguistic appropriateness. The justification for these five factors lies in a theoretical construct that claims those factors to be major components of oral proficiency. So if you were asked to conduct an oral proficiency interview that evaluated only pronunciation and grammar, you could be justifiably suspicious about the construct validity of that test. Likewise, let's suppose you have created a simple written vocabulary quiz, covering the content of a recent unit, that asks students to correctly define a set of words. Your chosen items may be a perfectly adequate sample of what was covered in the unit, but if the lexical objective of the unit was the communicative use of vocabulary, then the writing of definitions certainly fails to match a construct of communicative language use.

Construct validity is a major issue in validating large-scale standardized tests of proficiency. Because such tests must, for economic reasons, adhere to the principle of practicality, and because they must sample a limited number of domains of language, they may not be able to contain all the content of a particular field or skill. The TOEFL®, for example, has until recently not attempted to sample oral production, yet oral production is obviously an important part of academic success in a university

course of study. The TOEFL's omission of oral production content, however, is ostensibly justified by research that has shown positive correlations between oral production and the behaviors (listening, reading, grammaticality detection, and writing) actually sampled on the TOEFL (see Duran et al., 1985). Because of the crucial need to offer a financially affordable proficiency test and the high cost of administering and scoring oral production tests, the omission of oral content from the TOEFL has been justified as an economic necessity. (Note: As this book goes to press, oral production tasks are being included in the TOEFL, largely stemming from the demands of the professional community for authenticity and content validity.)

Consequential Validity

As well as the above three widely accepted forms of evidence that may be introduced to support the validity of an assessment, two other categories may be of some interest and utility in your own quest for validating classroom tests. Messick (1989), Gronlund (1998), McNamara (2000), and Brindley (2001), among others, underscore the potential importance of the consequences of using an assessment. Consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its impact on the preparation of test-takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use.

As high-stakes assessment has gained ground in the last two decades, one aspect of consequential validity has drawn special attention: the effect of test preparation courses and manuals on performance. McNamara (2000, p. 54) cautions against test results that may reflect socioeconomic conditions such as opportunities for coaching that are "differentially available to the students being assessed (for example, because only some families can afford coaching, or because children with more highly educated parents get help from their parents)." The social consequences of large-scale, high-stakes assessment are discussed in Chapter 6.

Another important consequence of a test falls into the category of *washback*, to be more fully discussed below. Gronlund (1998, pp. 209–210) encourages teachers to consider the effect of assessments on students' motivation, subsequent performance in a course, independent learning, study habits, and attitude toward school work.

Face Validity

An important facet of consequential validity is the extent to which "students view the assessment as fair, relevant, and useful for improving learning" (Gronlund, 1998, p. 210), or what is popularly known as **face validity**. "Face validity refers to the degree to which a test *looks* right, and *appears* to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers" (Mousavi, 2002, p. 244).

OXFORD
APPLIED
LINGUISTICS

Language Testing in Practice

Lyle F. Bachman
Adrian S. Palmer

Oxford University Press

2 Test usefulness: Qualities of language tests

Introduction

The most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness. This may seem so obvious that it need not be stated. But what makes a test useful? How do we know if a test will be useful before we use it? Or if it has been useful after we have used it? Stating the question of usefulness this way implies that simply using a test does not make it useful. By stating the obvious and by questioning it, we wish to point out that although usefulness is of unquestioned importance, it has not been defined precisely enough to provide a basis for either designing and developing a test or for determining its usefulness after it has been developed.

We believe that test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use. We thus regard a model of test usefulness as the essential basis for quality control throughout the entire test development process. We would further argue that all test development and use should be informed by a model of test usefulness. In this chapter we propose a model of test usefulness that includes six test qualities—reliability, construct validity, authenticity, interactiveness, impact, and practicality. We also propose three principles that we believe are the basis for operationalizing our model of usefulness in the development and use of language tests. This model, along with the three principles, provides a basis for answering the question, 'How useful is this particular test for its intended purpose(s)?' We first describe the model and principles, and then discuss each of the six qualities of test usefulness, providing examples to illustrate each. In Chapter 7 we provide specific questions that might be asked, during the test design and development process, for evaluating these qualities for specific testing situations.

Test usefulness

The traditional approach to describing test qualities has been to discuss these as more or less independent characteristics, emphasizing the need to

maximize them all. This has led some language testers to what we see as the extreme and untenable position that maximizing one quality leads to the virtual loss of others. Language testers have been told that the qualities of reliability and validity are essentially in conflict (for example, Underhill 1982; Heaton 1988), or that it is not possible to design test tasks that are authentic and at the same time reliable (for example, Morrow 1979, 1986). A much more reasonable position, expressed by Hughes (1989), is that although there is a tension among the different test qualities, this need not lead to the total abandonment of any. It is our view that rather than emphasizing the tension among the different qualities, test developers need to recognize their complementarity. We would thus argue that test developers need to find an appropriate balance among these qualities, and that this will vary from one testing situation to another. This is because what constitutes an appropriate balance can be determined only by considering the different qualities in combination as they affect the overall usefulness of a particular test.

Our notion of usefulness can be expressed as in Figure 2.1.

Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality

Figure 2.1: Usefulness

This is a representation of our view that test usefulness can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test.[1] We believe that a basis for operationalizing this view of usefulness in the development and use of language tests is provided by the three principles that follow. (We provide detailed discussions of how these principles can be operationalized in test design and development in Chapters 7 and 9.)

Principle 1 It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.

Principle 2 The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

Principle 3 Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation.

These principles reflect our belief that, in order to be useful, any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain (i.e. situation or context in which the test taker will be using the language outside of the test itself) in mind. (We will refer to this domain as a 'target language use', or TLU, domain, and the tasks in the TLU domain as 'TLU tasks'. This is discussed

in greater detail on pages 44–5 in Chapter 3.) Usefulness thus cannot be evaluated in the abstract, for all tests. We can describe the notion of test usefulness in terms of the six test qualities, and outline general considerations and procedures for assessing these. We cannot, however, offer general prescriptions about either what the appropriate balance among the different qualities should be or what are minimum acceptable levels. This can only be done for a given test and testing situation.

Evaluating the overall usefulness of a given test is essentially subjective, since this involves value judgments on the part of the test developer. In a large-scale test that will be used for making important decisions about large numbers of individuals, for example, the test developer may want to design the test and test tasks so as to achieve the highest possible levels of reliability and validity. In a classroom test, on the other hand, the teacher may want to utilize test tasks that will provide higher degrees of authenticity, interactiveness, and impact.

Test qualities

In considering the specific qualities that determine the overall usefulness of a given test, we believe it is essential to take a systemic view, considering tests as part of a larger societal or educational context. In this discussion, we will focus on the use of tests in educational programs, which will include many components, such as teaching materials and learning activities, as well as tests. The main difference between tests and other components of an instructional program, in our view, is in their purpose. While the primary purpose of other components is to promote learning, the primary purpose of tests is to measure. Tests can serve pedagogical purposes, to be sure, but this is not their *primary* function. Four of the qualities that we will discuss with respect to tests are shared by other components of a learning program. Thus, we can consider the authenticity of a particular language sample that may be used for instruction, the interactiveness of a particular learning task, the impact of a given learning activity, or the practicality of a particular teaching approach for a given situation. Two of the qualities—reliability and validity—are, however, critical for tests, and are sometimes referred to as essential *measurement* qualities. This is because these are the qualities that provide the major justification for using test scores—numbers—as a basis for making inferences or decisions.

Reliability

Reliability is often defined as consistency of measurement. A reliable test score will be consistent across different characteristics of the testing situation. Thus, reliability can be considered to be a function of the consistency

of scores from one set of tests and test tasks to another. If we think of test tasks as sets of task characteristics, as described in the next chapter, then reliability can be considered to be a function of consistencies across different sets of test task characteristics. (Test task characteristics are discussed in Chapter 3.) This can be represented as in Figure 2.2.

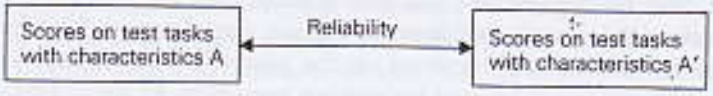


Figure 2.2: Reliability

In this figure, the double-headed arrow is used to indicate a correspondence between two sets of task characteristics (A and A') which differ only in incidental ways. For example, if the same test were to be administered to the same group of individuals on two different occasions, in two different settings, it should not make any difference to a particular test taker whether she takes the test on one occasion and setting or the other. Or suppose, for example, we had developed two forms of a test that were intended to be used interchangeably, it should not make any difference to a particular test taker which form of the test she takes; she should obtain the same score on either form. Thus, in a test designed to rank order individuals from highest to lowest, if the scores obtained on the different forms do not rank individuals in essentially the same order, then these scores are not very consistent, and would be considered to be unreliable indicators of the ability we want to measure. Similarly, in a test designed to distinguish individuals who are at or above a particular mastery level of ability from those who are below it, if the scores obtained from the two forms do not identify the same individuals as 'masters' and 'non-masters', then this test would be unreliable for making such classification decisions. Another example would be if we used several different raters to rate a large number of compositions. In this case, a given composition should receive the same score irrespective of which particular rater scored it. If some raters rate more severely than others, then the ratings of different raters are not consistent, and the scores obtained could not be considered to be reliable.

Reliability is clearly an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. At the same time, we need to recognize that it is not possible to eliminate inconsistencies entirely. What we can do, however, is try to minimize the effects of those potential sources of inconsistency that are under our control, through test design. Of the many factors that can affect test performance, the characteristics of the test tasks are at least partly under our control. Thus, in designing and developing language tests, we try to minimize variations in the test task

characteristics that do not correspond to variations in TLU tasks. In Chapters 7 and 9 we discuss ways in which we can take reliability considerations into account in order to reduce inconsistencies across test tasks, as we design a test and evaluate its potential usefulness.

In addition to using test design to minimize variations in test task characteristics, we need to estimate their effects on test scores, so as to determine how successful we have been. (Procedures for investigating and demonstrating reliability are discussed in the Suggested Readings at the end of this chapter.)

Construct validity

Construct validity pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores. When we interpret scores from language tests as indicators of test takers' language ability, a crucial question is, 'To what extent can we justify these interpretations?' The clear implication of this question is that as test developers and test users we must be able to provide adequate justification for any interpretation we make of a given test score.[2] That is, we need to demonstrate, or justify, the validity of the interpretations we make of test scores, and not simply assert or argue that they are valid.

In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else. In order to provide such evidence, we must define the construct we want to measure. For our purposes, we can consider a *construct* to be the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task. The term *construct validity* is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure. Construct validity also has to do with the domain of generalization to which our score interpretations generalize. The domain of generalization is the set of tasks in the TLU domain to which the test tasks correspond. At the very least we want our interpretations about language ability to generalize beyond the testing situation itself to a particular TLU domain. These two aspects of the construct validity of score interpretations are represented visually in Figure 2.3.

This figure indicates that test scores are to be interpreted appropriately as indicators of the ability we intend to measure *with respect to a specific domain of generalization*. Thus, when we consider the construct validity of a score interpretation, we need to consider both the construct definition and the characteristics of the test task. We need to consider the characteristics of the test task for two reasons. First we need to determine the extent to which the test task corresponds to tasks in the TLU domain, or the domain of generalization. (This correspondence is discussed below as

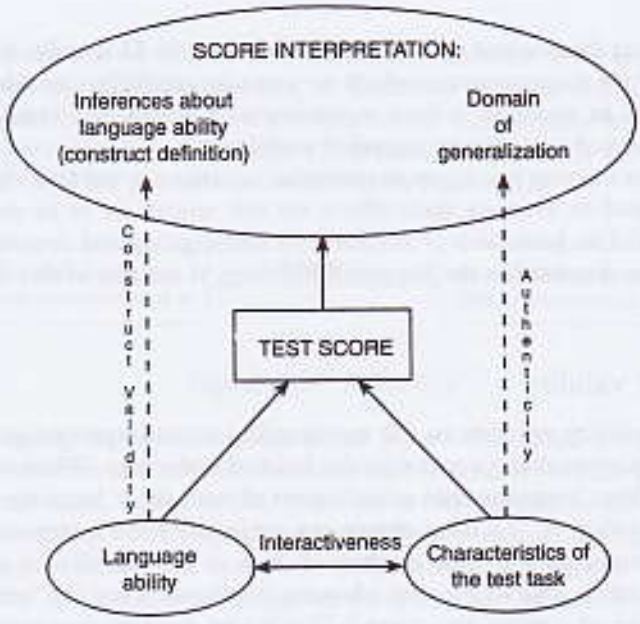


Figure 2.3: Construct validity of score interpretations

'authenticity'.) A second reason is to determine the degree to which the test task engages the test taker's areas of language ability (discussed below under 'interactiveness').

Construct validation is the on-going process of demonstrating that a particular interpretation of test scores is justified, and involves, essentially, building a logical case in support of a particular interpretation and providing evidence justifying that interpretation.[3] Several types of evidence (for example, content relevance and coverage, concurrent criterion relatedness, predictive utility) can be provided in support of a particular score interpretation, as part of the validation process, and these are discussed in the Suggested Readings at the end of this chapter. In Chapters 7 and 9 we discuss ways in which we can logically take validity considerations into account as we design a test and evaluate its potential usefulness.

It is important for test developers and users to realize that test validation is an on-going process and that the interpretations we make of test scores can never be considered absolutely valid. Justifying the interpretations we make on the basis of language test scores begins with test design and continues with the gathering of evidence to support our intended interpretations. However, even when we have provided evidence in support of a particular set of interpretations, we need to recognize that these must be viewed as tenuous. For this reason, we should not give the impression that a given interpretation is 'valid' or 'has been validated'.

Summary of reliability and construct validity

The primary purpose of a language test is to provide a *measure* that we can interpret as an indicator of an individual's language ability. The two measurement qualities, reliability and construct validity, are thus essential to the usefulness of any language test. Reliability is a necessary condition for construct validity, and hence for usefulness. However, reliability is not a sufficient condition for either construct validity or usefulness. Suppose, for example, that we needed a test for placing individuals into different levels in an academic writing course. A multiple-choice test of grammatical knowledge might yield very consistent or reliable scores, but this would not be sufficient to justify using this test as a placement test for a writing course. This is because grammatical knowledge is only one aspect of the ability to use language to perform academic writing tasks. In this case, defining the construct to include only one area of language knowledge is inappropriately narrow, since the construct involved in the TLU domain—ability to perform academic writing tasks—involves other areas of language knowledge, as well as metacognitive strategies, and may involve topical knowledge and affective responses as well.

Authenticity[4]

In Chapter 3 we argue that, in order to justify the use of language tests, we need to be able to demonstrate that performance on language tests corresponds to language use in specific domains other than the language test itself. One aspect of demonstrating this pertains to the correspondence between the characteristics of TLU tasks and those of the test task. It is this correspondence that is at the heart of authenticity, and we would describe a test task whose characteristics correspond to those of TLU tasks as relatively authentic. We define *authenticity* as the degree of correspondence of the characteristics of a given language test task to the features of a TLU task. This relationship is shown in Figure 2.4 (this corresponds to arrow 'B' in Figure 1.1 in Chapter 1).

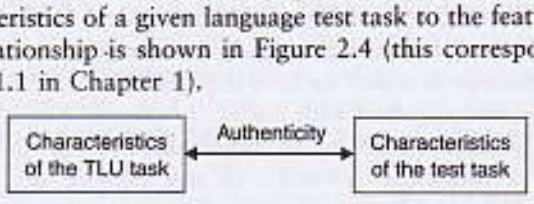


Figure 2.4: Authenticity

Authenticity as a critical quality of language tests has not generally been discussed in language testing textbooks, even though it has been debated among language testing researchers now for over a decade. We consider authenticity to be an important test quality because it relates the test task to the domain of generalization to which we want our score interpretations