1

Statistics and Data

1.1. STATISTICS AS A SCIENCE

The past few decades have witnessed a tremendous increase in the amount of data being collected in our modern society. For example, data about individual spending habits and patterns, school and college achievement, aptitude or intelligence are collected frequently by various persons and organizations—e.g., banks, teachers, schools, instructors, colleges, clinicians, administrators. Accompanying this data buildup is also a great deal of interest in analyzing these data to address specific questions of interest using statistics.

Statistics can be defined as a science that helps design empirical studies to collect data, as well as to organize, classify, analyze, and interpret these data, in order to make decisions in situations that are often characterized by uncertainty. This uncertainty generally results from the at times very limited information contained in the data obtained through these empirical studies. Additionally, this information can be potentially highly variable across the examined individuals, cases, subjects, or units of analysis considered. Based upon these features, statistics can be regarded as a scientific discipline used in almost all aspects of modern society to greatly facilitate the process of learning from data.

In this book, we will be mainly interested in the application of basic statistics to a variety of empirical studies in the behavioral, biological, educational, medical, management, and social sciences. One can distinguish thereby between the following four stages of using statistics to conduct research in these sciences (e.g., Ott & Longnecker, 2010):

(i) defining the problem to be studied,
(ii) collecting the data that contains the necessary information about the problem,
(iii) summarizing the data, and
(iv) analyzing and modeling the data, interpreting and communicating results.

Each of these stages is essential for the process of applying statistics to address questions within the above-mentioned sciences. If any of these stages is bypassed or carried out in an inappropriate way, the outcome of the subsequent stage(s) may well be significantly affected if not totally compromised. Additionally, the ultimate results and interpretation of the study in question may be seriously misleading. As it turns out, this four-stage process closely resembles the scientific method that underlies scientific inquiry (e.g., Graziano & Raulin, 2009). The scientific method is essentially a set of principles and procedures used to advance any empirical science. It consists of the following phases:

(a) the formulation of a research goal,
(b) the design of an experiment or an observational study,
(c) the collection of data (observations), and
(d) the analysis and modeling of the data, and testing of research hypotheses.

We note that the phases (a) through (d) listed above closely resemble the statistical application stages (i) through (iv).

As an example, let us consider the following study, which we will return to frequently in this book. In this investigation, our main interest lies with examining the presence of possible gender differences in depression in older adults in the United States. To begin with, the detailed description of the study's goals accomplishes much of the above stage (i) of using statistics in the process of empirical research. We need to note also that since we cannot realistically examine each and every elderly person in the entire country, due to resource- and time-related limitations, we will have to rely on information obtained from just a subset of the population. This is a generally relevant issue that we will often make reference to as we advance through this book. Then in accordance with stage (ii), we collect data by using a well-established instrument (e.g., a scale) for measuring depression on a randomly selected group of say $n = 500$ elderly adults that represents well the population of aged persons in the United States. Section 1.2 of this chapter deals with general issues pertaining to this stage and with the selection of the group to be studied, which is called a *sample*. In stage (iii), we obtain summary indexes signifying for instance average depression score and degree of variability within each of the two genders from the sample studied (as well as conduct further activities, in general, which depend on the specific research questions being asked). Chapter 2 discusses matters pertaining to this stage in a more general context. Finally, in stage (iv), we may examine for instance the gender differences on these summary indexes (and/or possibly carry further activities, again depending on the research questions pursued). In addition, we may draw conclusions about those differences and communicate our results in

writing or verbally to an intended audience. Much of the remainder of this book will deal with the various statistical methods that could be appropriately used during this fourth stage.

This depression study example illustrates that when using statistics one is usually interested in *drawing conclusions about large groups of subjects*, typically referred to as populations, while *data are available from only portions of these groups*, referred to as samples. (We will provide further discussion on these two concepts of population and sample in a later section of this chapter.) In this sense, statistics deals with a *major inferential problem*—how to achieve trustworthy conclusions about a given large group (a population), based on limited data obtained from only a subgroup selected from it (a sample), which data in addition is most likely going to be varied. This variability is a particular feature for variables or subject characteristics in studies where statistics is applicable. Specifically, different subjects will usually give rise to (*quite*) *different scores* on considered measures or variables (e.g., depression in the above example). This is essentially what we meant when we stated above that the data will be varied. Hence, given this likely diversity of resulting scores, the task of statistics is to permit one to reach credible conclusions about the large group of subjects (population) that are of actual concern, based on a *multiplicity* of different scores on given variable(s) measured only on a small group of subjects (sample).

The above discussion leads us to the following stricter definition of major relevance to statistics and to the application of statistics in empirical research.

Definition: *Population* is the set of all possible measurements on a given variable of concern that are of interest in a specific research question.

We note that the precise definition of a population in a particular empirical setting may well depend on the research question and/or on the variable or subject characteristics of interest to be studied. In this book, a population will most often consist of scores on particular variable(s) that could be measured on subjects from a potentially very large group of persons, which is of substantive interest. At times, these may be scores for aggregates of subjects, such as schools or classes, hospitals or districts, companies or corporations. When no confusion arises, we will just refer to the large group of subjects themselves as a population, as is usually done in the behavioral and social science literature. (Part of the reason for the possible confusion is that in an empirical study one usually collects data on more than a single variable or subject characteristic. Rather than conceiving of multiple populations of relevance to the same study—one per variable of possible interest—one considers only one population, viz., the collection of all subjects to which inference is to be made.)

A closely related concept that we will also routinely use in this book is that of the part or subset of a population under consideration that is exhaustively studied. This is the notion of a *sample*. We define a sample as a selected

(drawn) subset of the population, in which every unit of analysis is examined or measured on the variable(s) of concern to a given empirical investigation. Similarly to a population, a sample may consist of subject aggregates, e.g., schools or classes, hospitals, neighborhoods, companies, cities, nursing homes.

To return for a moment to the aging example given above, we observe that depression is the variable of main concern in it. The scores on this measure of all elderly adults in the United States constitute the population of interest in the posed research question. In addition, the actually studied group of depression scores based on the $n = 500$ persons is the sample drawn from that population. We reiterate that often for convenience the set of all subjects that one wishes to make conclusions about is referred to as population, while the set of subjects drawn from it who actually participate in the study are referred to as sample. Alternatively, depending on the research question one can think of a population as consisting of aggregates of subjects (hospitals, companies, schools, etc.), and a sample as an appropriate subset of such aggregates. Similarly, as we will elaborate on in a later section of the book, most of the time in empirical research one is interested in studying not just one but multiple variables on a given sample of subjects and sometimes examining their interrelationships. To simplify subsequent discussions and terminology, when we make reference to a population in this book we will typically mean a set of subjects (or more specifically subject scores on variables of interest); then samples will be appropriately selected subsets of populations under investigation (or specifically of individual scores). We will return to this issue again in Chapter 2, where we will elaborate on it further.

1.2. COLLECTING DATA

As indicated above, obtaining data in a study represents the second stage in the outlined four-step process of using statistics in empirical research. Data are typically collected in a well-thought-out and planned process, leading to either conducting an experiment or nonexperimental study (also more generally referred to as an observational study), including the development of a survey, a questionnaire, or some other particular measurement device or instrument. This process itself also evolves through the following several steps (e.g., King & Minium, 2003):

(a) specifying the objective of the study, survey, or experiment,
(b) identifying the variable(s) of interest,
(c) choosing an appropriate design (for an experiment, or sample for an observational study), and
(d) obtaining the data.

To illustrate further this process, let us revisit again the earlier depression example in which the objective is the examination of gender differences in depression of older adults (defined as elderly persons aged 65 or older). This observation represents step (a) mentioned above. In step (b), we identify the score on a depression scale (measure) as the variable of interest. We then choose a measuring device or instrument(s), such as the Beck Depression Inventory (Beck et al., 1977), to collect subsequently these scores from the subjects (the units of analysis) in the drawn sample. In step (c), we decide how to select the subjects to be included into the actual study. Let us suppose for now that we were able to obtain a random sample of 250 men and 250 women from the population of elderly in the United States. In step (d), we administer the measure(s) (scale, test, instrument), and then score the results we obtain on it from each person in the sample.

Frequently, step (b) of the discussed process requires a detailed consideration of measures that can be of interest in the context of the study to be conducted. Experts in the substantive (subject-matter) field of application may be asked for their opinion on this issue during this step. Similarly, step (c) may involve time- and effort-consuming activities to come up with samples that are representative of the studied population(s). Due to the complexities involved in this activity, a special branch of statistics referred to as *sampling theory* has actually been developed to meet the demands of obtaining a representative sample (as well as carrying out subsequently appropriate related data analyses). We will touch upon some of these complexities and activities in a later chapter of the book. Finally, in step (d), one needs to ensure that specific and detailed instructions are precisely followed when administering the measure(s) of concern, as well as that their objective and accurate scoring is accomplished at the end of the data collection process. We note that if the study is indeed experimental, then further activities are also involved in steps (b) through (d). Although elaborating on these activities is beyond the scope of this introductory book, they are discussed in greater detail in more specialized experimental design sources and books (e.g., Kirk, 1989).

1.3. WHY STUDY STATISTICS?

There are a number of reasons that the study of statistics is very beneficial to advancing knowledge in the empirical sciences. For one, statistics can be used to summarize and interpret correctly published numerical data (e.g., data from surveys or various other forms of reports). Further, statistics can be used to help develop a critical and informed way of interpreting data and evaluating possible conclusions drawn from examined data sets. For instance, the media continually expose us to large bodies of information through news and

advertising agencies. Reports on economic conditions, political issues, surveys about opinions on diverse matters, and many other communications frequently have one feature in common—they all contain certain statistics to support related arguments. Statistics as a science is necessary to help us make sense of this vast amount of data and thereby better understand the world we live in.

Thus, statistics is indispensable in the empirical sciences where data on samples from studied populations are routinely made available. For instance, the application of statistics is essential for answering questions like ''Are observed group (treatment, program, intervention) differences in responding to a studied drug 'real' or are they only the result of random variation resulting from the fact that only subjects in a randomly drawn, relatively limited sample from the population under investigation were examined?''

The remainder of this book is concerned with discussing a number of basic statistical methods that can help scientists make progress in their empirical research. In addition to describing some of the technical details that underlie each method, we also will employ for analysis the popular statistical software package R. We will specifically use it in order to carry out data analysis and modeling on illustrative data sets as we proceed through each chapter in the book. This popular software is highly comprehensive, state-of-the-art, and obtainable free of charge. Indeed, the software package R can be located and downloaded via any available search engine by simply using the letter ''R'' or ''R-project'' as a keyword (and following the prompts; see also Venables, Smith, & The R Development Core Team, 2012). We believe that the package R is in many ways a software for the next decade and beyond.