CHAPTER 1

Data and Statistics

CONTENTS

1.1 Introduction 1
1.2 Observations and Variables 6
1.3 Types of Measurements for Variables 10
1.4 Distributions 12
1.5 Numerical Descriptive Statistics 19
1.6 Exploratory Data Analysis 32
1.7 Bivariate Data 39
1.8 Populations, Samples, and Statistical Inference — A Preview 43
1.9 Data Collection 44
1.10 Chapter Summary 46
1.11 Chapter Exercises 51

1.1 INTRODUCTION

To most people the word **statistics** conjures up images of vast tables of confusing numbers, volumes and volumes of figures pertaining to births, deaths, taxes, populations, and so forth, or figures indicating baseball batting averages or football yardage gained flashing across television screens. This is so because in common usage the word *statistics* is synonymous with the word *data*. In a sense this is a reasonably accurate impression because the discipline of statistics deals largely with principles and procedures for collecting, describing, and drawing conclusions from data. Therefore

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00001-9

© 2010, Elsevier Inc. All rights reserved.

it is appropriate for a text in statistical methods to start by discussing what data are, how data are characterized, and what tools are used to describe a set of data. The purpose of this chapter is to

1. provide the definition of a set of data,
2. define the components of such a data set,
3. present tools that are used to describe a data set, and briefly
4. discuss methods of data collection.

Definition 1.1 *A set of* ***data*** *is a collection of observed values representing one or more characteristics of some objects or units.*

■ Example 1.1: A typical data set

Every year, the National Opinion Research Center (NORC) publishes the results of a personal interview survey of U.S. households. This survey is called the General Social Survey (GSS) and is the basis for many studies conducted in the social sciences. In the 1996 GSS, a total of 2904 households were sampled and asked over 70 questions concerning lifestyles, incomes, religious and political beliefs, and opinions on various topics. Table 1.1 lists the data for a sample of 50 respondents on four of the questions asked. This table illustrates a typical midsized data set. Each of the rows corresponds to a particular respondent (labeled 1 through 50 in the first column). Each of the columns, starting with column two, are responses to the following four questions:

1. AGE: The respondent's age in years
2. SEX: The respondent's sex coded 1 for male and 2 for female
3. HAPPY: The respondent's general happiness, coded:

   1 for "Not too happy"
   2 for "Pretty happy"
   3 for "Very happy"
4. TVHOURS: The average number of hours the respondent watched TV during a day

This data set obviously contains a lot of information about this sample of 50 respondents. Unfortunately this information is hard to interpret when the data are presented as shown in Table 1.1. There are just too many numbers to make any sense of the data — and we are only looking at 50 respondents! By summarizing some aspects of this data set, we can obtain much more usable information and perhaps even answer some specific questions. For example, what can we say about the overall frequency of the various levels of happiness? Do some respondents watch a lot of TV? Is there a relationship between the age of the respondent and his or her general happiness? Is there a relationship between the age of the respondent and the number of hours of TV watched?

Table 1.1 Sample of 50 Responses to the 1996 GSS

Respondent	AGE	SEX	HAPPY	TVHOURS	Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0	26	53	1	1	2
2	25	2	1	0	27	26	2	2	0
3	43	1	2	4	28	89	2	2	0
4	38	1	2	2	29	65	1	1	0
5	53	2	3	2	30	45	2	2	3
6	43	2	2	5	31	64	2	3	5
7	56	2	2	2	32	30	2	2	2
8	53	1	2	2	33	75	2	2	0
9	31	2	1	0	34	53	2	2	3
10	69	1	3	3	35	38	1	2	0
11	53	1	2	0	36	26	1	2	2
12	47	1	2	2	37	25	2	3	1
13	40	1	3	3	38	56	2	3	3
14	25	1	2	0	39	26	2	2	1
15	60	1	2	2	40	54	2	2	5
16	42	1	2	3	41	31	2	2	0
17	24	2	2	0	42	44	1	2	0
18	70	1	1	0	43	36	2	2	3
19	23	2	3	0	44	74	2	2	0
20	64	1	1	10	45	74	2	2	3
21	54	1	2	6	46	37	2	3	0
22	64	2	3	0	47	48	1	2	3
23	63	1	3	0	48	42	2	2	6
24	33	2	2	4	49	77	2	2	2
25	36	2	3	0	50	75	1	3	0

We will return to this data set in Section 1.10 after we have explored some methods of summarizing and making sense of data sets like this one. As we develop more sophisticated methods of analysis in later chapters, we will again refer to this data set.[1] ■

Definition 1.2 *A **population** is a data set representing the entire entity of interest.*

For example, the decennial census of the United States yields a data set containing information about all persons in the country at that time (theoretically all households correctly fill out the census forms). The number of persons per household as listed in the census data constitutes a population of family sizes in the United States.

[1]The GSS is discussed on the following Web page: http://www.icpsr.umich.edu/GSS/.