CHAPTER 1

Data and Statistics

CONTENTS

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

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/.

Similarly, the weights of all steers brought to an auction by a particular rancher is a data set that is the population of the weights of that rancher's marketable steers.

Note that elements of a population are really measures rather than individuals. This means that there can be many different definitions of populations that involve the same collection of individuals. For example, the number of school-age children per household as listed in the census data would constitute a population for another study. As we shall see in discussions about statistical inference, it is important to define the population that we intend to study very carefully.

Definition 1.3 *A **sample** is a data set consisting of a portion of a population. Normally a sample is obtained in such a way as to be representative of the population.*

The Census Bureau conducts various activities during the years between each decennial census, such as the Current Population Survey. This survey samples a small number of scientifically chosen households to obtain information on changes in employment, living conditions, and other demographics. The data obtained constitute a sample from the population of all households in the country. If two steers were selected from a herd of steers brought to an auction by a rancher, these two steers would be considered a sample from the herd.

1.1.1 Data Sources

This book contains many examples and exercises consisting of data sets that are to be subjected to statistical analysis. Although the emphasis in this book is on the statistical analysis of data, we must emphasize that proper data collection is just as important as proper analysis. We touch briefly on issues of data collection in Section 1.9. There are many more detailed texts on this subject (for example, Scheaffer *et al*. 2006). Remember, even the most sophisticated analysis procedures cannot provide good results from bad data.

In general, data are obtained from two broad categories of sources:

- **Primary** data are collected as part of the study.
- **Secondary** data are obtained from published sources, such as journals, governmental publications, news media, or almanacs.

There are several ways of obtaining primary data. Data are often obtained from simple observation of a process, such as characteristics and prices of homes sold in a particular geographic location, quality of products coming off an assembly line, political opinions of registered voters in the state of Texas, or even a person standing on a street corner and recording how many cars pass each hour during the day. This kind of a study is called an **observational study**. Observational studies are often used to determine whether an association exists between two or more characteristics measured in the study. For example, a study to determine the relationship between high school student performance and the highest educational level of the student's parents would be based on an examination of student performance

and a history of the parents' educational experiences. No cause-and-effect relationship could be determined, but a strong association might be the result of such a study. Note that an observational study does not involve any intervention by the researcher.

Much primary data are obtained through the use of **sample surveys** such as Gallup polls or the Nielsen TV ratings. Such surveys normally represent a particular group of individuals and are intended to provide information on the characteristics and/or habits of such a group.

Often data used in studies involving statistics come from **designed experiments**. In a designed experiment researchers impose treatments and controls on the process and then observe the results and take measurements. For example, in a laboratory experiment rats may be subjected to various noise levels and the rapidity of their movements recorded. Designed experiments can be used to help establish causation between two or more characteristics. For example, a study could be designed to determine if high school student performance is affected by a nutritious breakfast. By choosing a proper design and conducting the experiment in a rigorous manner, an actual cause-and-effect relationship might be established. Data from designed experiments are considered a sample. For example, a study relating high school student performance to breakfast may use as few as 25 typical urban high school students. The results of the study would then be inferred to the population of all urban high school students. Chapter 10 provides an introduction to experimental designs.

1.1.2 Using the Computer

Today, comprehensive programs for conducting statistical and data analyses are available in general-use spreadsheet software, graphing calculators, and dedicated statistical software. A person rarely needs to write his or her own programs, since they already exist for almost all aspects of statistics. Because a large number of such packages are currently available, it is impossible to provide specific instructions for each package in a single book. Although a few exercises in the beginning of this book, especially those in Chapters 2–5, can be done manually or with the aid of calculators, most exercises even in these chapters, and all exercises in Chapters 8–11, will require the use of a computer. In some examples we have included generic instructions for effective computer usage.

For reasons of consistency and convenience we have used the SAS System almost exclusively for examples in this book. The SAS System is a very comprehensive software package, of which statistical analysis is only a minor portion. Because it is such a large system it may not be optimal for students to have on their personal computers. We assume that additional instructions will be available for the particular software you will be using. In a few instances, especially in the earlier chapters, output from several software packages are used for comparative purposes.

One common feature of almost every package is the way files containing the data are organized. A good rule of thumb is "one observation equals one row"; another is "one type of measurement (or variable) is one column." Consider the data in Table 1.1. Arranged in a spreadsheet or a text file, the data would appear much as in that table, except that the right half of the table is pasted below the left, to make 50 rows. Each row would correspond to a different respondent. Each column would correspond to a different item reported on that respondent.

Although the input files have a certain similarity, each software package has its own style of output. However, most will contain essentially the same results, although they may appear in a different order and may even have different labels. It is therefore important to study the documentation of any package being used. We should note that most computer outputs in this book have been abbreviated because the full default output often contains information not needed at that particular time, although in a few instances we have presented the full output for illustration purposes.

If a set of data represents an entire population, the techniques presented in this chapter can be used to describe various aspects of that population and a statistical analysis using these descriptors is useful solely for that purpose. However, as is more often the case, the data to be analyzed come from a sample. In this case, the descriptive statistics obtained may subsequently be used as tools for **statistical inference**. A general introduction to the concept of statistical inference is presented in Section 1.8, and most of the remainder of this text is devoted to that subject.

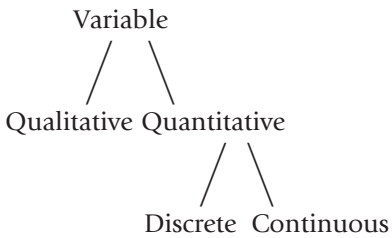
1.2 OBSERVATIONS AND VARIABLES

A data set is composed of information from a set of units. Information from a unit is known as an **observation**. An observation consists of one or more pieces of information about the unit; these are called **variables**. Some examples:

- In a study of the effectiveness of a new headache remedy, the units are individual persons, of which 10 are given the new remedy and 10 are given an aspirin. The resulting data set has 20 observations and two variables: the medication used and a score indicating the severity of the headache.
- In a survey for determining TV viewing habits, the units are families. Usually there is one observation for each of thousands of families that have been contacted to participate in the survey. The variables describe the programs watched as well as descriptions of the characteristics of the families.
- In a study to determine the effectiveness of a college admissions test (e.g., SAT) the units are the freshmen at a university. There is one observation per unit and the variables are the students' scores on the test and their first year's GPA.

Variables that yield nonnumerical information are called **qualitative** variables. Qualitative variables are often referred to as **categorical** variables. Those that yield

numerical measurements are called **quantitative** variables. Quantitative variables can be further classified as discrete or continuous. The diagram below summarizes these definitions:



Definition 1.4 *A **discrete variable** can assume only a countable number of values. Typically, discrete variables are frequencies of observations having specific characteristics, but all discrete variables are not necessarily frequencies.*

Definition 1.5 *A **continuous variable** is one that can take any one of an uncountable number of values in an interval. Continuous variables are usually measured on a scale and, although they may appear discrete due to imprecise measurement, they can conceptually take any value in an interval and cannot therefore be enumerated.*

In the field of statistical quality control, the term **variable data** is used when referring to data obtained on a continuous variable and **attribute data** when referring to data obtained on a discrete variable (usually the number of defectives or nonconformities observed).

In the preceding examples, the names of the headache remedies and names of TV programs watched are qualitative (categorical) variables. Headache severity scores is a discrete numeric variable, while the incomes of TV-watching families, and SAT and GPA scores are continuous quantitative variables.

We will use the data set in Example 1.2 to present greater detail on various concepts and definitions regarding observations and variables.

■ Example 1.2

In the fall of 2001, John Mode was offered a new job in a midsized city in east Texas. Obviously, the availability and cost of housing will influence his decision to accept, so he and his wife Marsha go to the Internet, find www.realtor.com, and after a few clicks find some 500 single-family residences for sale in that area. In order to make the task of investigating the housing market more manageable, they arbitrarily record the information provided on the first home on each page of six. This information results in a data set that is shown in Table 1.2.

The data set gives information on 69 homes, which comprise the *observations* for this data set. In this example, each property is a **unit**, often called a sample,

Table 1.2 Housing Data

Obs	zip	age	bed	bath	size	lot	exter	garage	fp	price
1	3	21	3	3.0	951	64904	Other	0	0	30000
2	3	21	3	2.0	1036	217800	Frame	0	0	39900
3	4	7	1	1.0	676	54450	Other	2	0	46500
4	3	6	3	2.0	1456	51836	Other	0	1	48600
5	1	51	3	1.0	1186	10857	Other	1	0	51500
6	2	19	3	2.0	1456	40075	Frame	0	0	56990
7	3	8	3	2.0	1368	.	Frame	0	0	59900
8	4	27	3	1.0	994	11016	Frame	1	0	62500
9	1	51	2	1.0	1176	6259	Frame	1	1	65500
10	3	1	3	2.0	1216	11348	Other	0	0	69000
11	4	32	3	2.0	1410	25450	Brick	0	0	76900
12	3	2	3	2.0	1344	.	Other	0	1	79000
13	3	25	2	2.0	1064	218671	Other	0	0	79900
14	1	31	3	1.5	1770	19602	Brick	0	1	79950
15	4	29	3	2.0	1524	12720	Brick	2	1	82900
16	3	16	3	2.0	1750	130680	Frame	0	0	84900
17	3	20	3	2.0	1152	104544	Other	2	0	85000
18	3	18	4	2.0	1770	10640	Other	0	0	87900
19	4	28	3	2.0	1624	12700	Brick	2	1	89900
20	2	27	3	2.0	1540	5679	Brick	2	1	89900
21	1	8	3	2.0	1532	6900	Brick	2	1	93500
22	4	19	3	2.0	1647	6900	Brick	2	0	94900
23	2	3	3	2.0	1344	43560	Other	1	0	95800
24	4	5	3	2.0	1550	6575	Brick	2	1	98500
25	4	5	4	2.0	1752	8193	Brick	2	0	99500
26	4	27	3	1.5	1450	11300	Brick	1	1	99900
27	4	33	2	2.0	1312	7150	Brick	0	1	102000
28	1	4	3	2.0	1636	6097	Brick	1	0	106000
29	4	0	3	2.0	1500	.	Brick	2	0	108900
30	2	36	3	2.5	1800	83635	Brick	2	1	109900
31	3	5	4	2.5	1972	7667	Brick	2	0	110000
32	3	0	3	2.0	1387	.	Brick	2	0	112290
33	4	27	4	2.0	2082	13500	Brick	3	1	114900
34	3	15	3	2.0	.	269549	Frame	0	0	119500
35	4	23	4	2.5	2463	10747	Brick	2	1	119900
36	4	25	3	2.0	2572	7090	Brick	2	1	119900
37	4	24	4	2.0	2113	7200	Brick	2	1	122900
38	4	1	3	2.5	2016	9000	Brick	2	1	123938
39	1	34	3	2.0	1852	13500	Brick	2	0	124900
40	4	26	4	2.0	2670	9158	Brick	2	1	126900

(*Continued*)

Table 1.2 *(Continued)*

Obs	zip	age	bed	bath	size	lot	exter	garage	fp	price
41	2	26	3	2.0	2336	5408	Brick	0	1	129900
42	4	31	3	2.0	1980	8325	Brick	2	1	132900
43	2	24	4	2.5	2483	10295	Brick	2	1	134900
44	2	29	5	2.5	2809	15927	Brick	2	1	135900
45	4	21	3	2.0	2036	16910	Brick	2	1	139500
46	3	10	3	2.0	2298	10950	Brick	2	1	139990
47	4	3	3	2.0	2038	7000	Brick	2	0	144900
48	2	9	3	2.5	2370	10796	Brick	2	1	147600
49	2	29	5	3.5	2921	11992	Brick	2	1	149990
50	2	8	3	2.0	2262	.	Brick	2	1	152550
51	4	7	3	3.0	2456	.	Brick	2	1	156900
52	4	1	4	2.0	2436	52000	Brick	2	1	164000
53	3	27	3	2.0	1920	226512	Frame	4	1	167500
54	4	5	3	2.5	2949	11950	Brick	2	1	169900
55	2	32	4	3.5	3310	10500	Brick	2	1	175000
56	4	29	3	3.0	2805	16500	Brick	2	1	179000
57	4	1	3	3.0	2553	8610	Brick	2	1	179900
58	4	1	3	2.0	2510	.	Other	2	1	189500
59	4	33	3	4.0	3627	17760	Brick	3	1	199000
60	2	25	4	2.5	3056	10400	Other	2	1	216000
61	3	16	3	2.5	3045	168576	Brick	3	1	229900
62	4	2	4	4.5	3253	54362	Brick	3	2	285000
63	2	2	4	3.5	4106	44737	Brick	3	1	328900
64	4	0	3	2.5	2993	.	Brick	2	1	313685
65	4	0	3	2.5	2992	14500	Other	3	1	327300
66	4	20	4	3.0	3055	250034	Brick	3	0	349900
67	4	18	5	4.0	3846	23086	Brick	4	3	370000
68	4	3	4	4.5	3314	43734	Brick	3	1	380000
69	4	5	4	3.5	3472	130723	Brick	2	2	395000

experimental, or observational unit.[2] The 11 columns of the table provide specific characteristics information for each home and compose the 11 *variables* of this data set. The variable definitions along with brief mnemonic descriptors commonly used in computers are as follows:

- `Obs`[3]: a sequential number assigned to each observation as it is entered into the computer. This is useful for identifying individual observations.

[2]These different types of units are not always synonymous. For example, an experimental unit may be an animal subjected to a certain diet while the observational units may be several determinations of the weight of the animal at different times. Unless otherwise specified, most of the methods presented in this book are based on the assumption that the three are synonymous and will usually be referred to as experimental units.

[3]The term Obs is used by the SAS System. Other computer software may use other notations.

- `zip`: the last digit of the postal service zip code. This variable identifies the area in which the home is located.
- `age`: the age of the home in years.
- `bed`: the number of bedrooms.
- `bath`: the number of bathrooms.
- `size`: the interior area of the home in square feet.
- `lot`: the size of the lot in square feet.
- `exter`: the exterior siding material.
- `garage`: the capacity of the garage; zero means no garage.
- `fp`: the number of fireplaces.
- `price`: the price of the home, in dollars.

The elements of each row define the observed values of the variables. Note that some values are represented by ".". In the SAS System, and other statistical computing packages, this notation specifies a missing value; that is, no information on that variable is available. Such missing values are an unavoidable feature in many data sets and occasionally cause difficulties in analyzing the data.

Brief mnemonic identifiers such as these are used by computer programs to make their outputs easier to interpret and are unique for a given set of data. However, for use in formulas we will follow mathematics convention, where variables are generically identified by single letters taken from the latter part of the alphabet. For example the letter Y can be used to represent the variable `price`. The same lowercase letter, augmented by a subscript identifying the observation number, is used to represent the value of the variable for a particular observation. Using this notation, y_i is the observed price of the ith house. Thus, $y_1 = 30000, y_2 = 39900, \ldots, y_{69} = 395000$. The set of observed values of `price` can be symbolically represented as $y_1, y_2, \ldots, y_{69}$, or $y_i, i = 1, 2, \ldots, 69$. The total number of observations is symbolically represented by the letter n; for the data in Table 1.2, $n = 69$. We can generically represent the values of a variable Y, as $y_i, i = 1, 2, \ldots, n$. We will most frequently use Y as the variable and y_i as observations of the variable of interest. ■

1.3 TYPES OF MEASUREMENTS FOR VARIABLES

We usually think of data as consisting of numbers, and certainly many data sets do contain numbers. In Example 1.2, for instance, the variable `price` is the asking price of the home, measured in dollars. This measurement indicates a definite metric or scale in the values of the variable `price`. Certainly a $200,000 house costs twice as much as a $100,000 house. As we will see later, not all variables that measure a quantity have this characteristic. However, not all data necessarily consist of numbers. For example, the variable `exter` is observed as either `brick`, `frame`, or `other`, a measurement that does not convey any relative value. Further, variables that are recorded as numbers do not necessarily imply a quantitative measurement. For example, the

variable `zip` simply locates the home in some specific area and has no quantitative meaning.

We can classify observations according to a standard measurement scale that goes from "strong" to "weak" depending on the amount or precision of information available in the scale. These measurement scales are discussed at some length in various publications, including Conover (1999). We present the characteristics of these scales in some detail since the nature of the data description and statistical inference is dependent on the type of variable being studied.

Definition 1.6 *The **ratio scale** of measurement uses the concept of a unit of distance or measurement and requires a unique definition of a zero value.*

Thus, in the ratio scale the difference between any two values can be expressed as some number of these units. Therefore, the ratio scale is considered the "strongest" scale since it provides the most precise information on the value of a variable. It is appropriate for measurements of heights, weights, birth rates, and so on. In the data set in Table 1.2, all variables except `zip` and `exter` are measured in the ratio scale.

Definition 1.7 *The **interval scale** of measurement also uses the concept of distance or measurement and requires a "zero" point, but the definition of zero may be arbitrary.*

The interval scale is the second "strongest" scale of measurement, because the "zero" is arbitrary. An example of the interval scale is the use of degrees Fahrenheit or Celsius to measure temperature. Both have a unit of measurement (degree) and a zero point, but the zero point does not in either case indicate the absence of temperature. Other popular examples of interval variables are scores on psychological and educational tests, in which a zero score is often not attainable but some other arbitrary value is used as a reference value.

We will see that many statistical methods are applicable to variables of either the ratio or interval scales in exactly the same way. We therefore usually refer to both of these types as **numeric variables**.

Definition 1.8 *The **ordinal scale** distinguishes between measurements on the basis of the relative amounts of some characteristic they possess. Usually the ordinal scale refers to measurements that make only "greater," "less," or "equal" comparisons between consecutive measurements.*

In other words, the ordinal scale represents a ranking or ordering of a set of observed values. Usually these ranks are assigned integer values starting with "1" for the lowest value, although other representations may be used. The ordinal scale does not provide as much information on the values of a variable and is therefore considered "weaker" than the ratio or interval scale.

For example, if a person were asked to taste five chocolate pies and rank them according to taste, the result would be a set of observations in the ordinal scale of measurement.

Table 1.3 Example of Ordinal Data

Pie	Rank
1	4
2	3
3	1
4	2
5	5

A set of data illustrating an ordinal variable is given in Table 1.3. In this data set, the "1" stands for the most preferred pie while the worst tasting pie receives the rank of "5." The values are used only as a means of arranging the observations in some order. Note that these values would not differ if pie number 3 were clearly superior or only slightly superior to pie number 4.

It is sometimes useful to convert a set of observed ratio or interval values to a set of ordinal values by converting the actual values to ranks. Ranking a set of actual values induces a loss of information, since we are going from a stronger to a weaker scale of measurement. Ranks do contain useful information and, as we will see (especially in Chapter 14), may provide a useful base for statistical analysis.

Definition 1.9 *The **nominal scale** identifies observed values by name or classification.*

A nominally scaled variable is also often called a categorical or qualitative variable. Although the names of the classifications may be represented by numbers, these are used merely as a means of identifying the classifications and are usually arbitrarily assigned and have no quantitative implications. Examples of nominal variables are sex, breeds of animals, colors, and brand names of products. Because the nominal scale provides no information on differences among the "values" of the variable, it is considered the weakest scale. In the data in Table 1.2, the variable describing the exterior siding material is a nominal variable.

We can convert ratio, interval, or ordinal scale measurements into nominal level variables by arbitrarily assigning "names" to them. For example, we can convert the ratio-scaled variable `size` into a nominally scaled variable, by defining homes with less than 1000 square feet as "cottages," those with more than 1000 but less than 3000 as "family-sized," and those with more than 3000 as "estates."

Note that the classification of scales is not always completely clear-cut. For example, the "scores" assigned by judges for track or gymnastic events are usually treated as possessing the ratio scale but are probably closer to being ordinal in nature.

1.4 DISTRIBUTIONS

Very little information about the characteristics of recently sold houses can be acquired by casually looking through Table 1.2. We might be able to conclude that most of the houses have brick exteriors, or that the selling price of houses ranges from $30,000 to $395,000, but a lot more information about this data set can be obtained through the use of some rather simple organizational tools.

To provide more information, we will construct **frequency distributions** by grouping the data into categories and counting the number of observations that fall into each one. Because we want to count each house only once, these categories (called classes) are constructed so they don't overlap. Because we count each observation only once, if we add up the number (called the frequency) of houses in all the classes, we get

the total number of houses in the data set. Nominally scaled variables naturally have these classes or categories. For example, the variable `exter` has three values, `Brick`, `Frame`, and `Other`. Handling ordinal, interval, and ratio scale measurements can be a little more complicated, but, as subsequent discussion will show, we can easily handle such data simply by correctly defining the classes.

Once the frequency distribution is constructed, it is usually listed in tabular form. For the variable `exter` from Table 1.2 we get the frequency distribution presented in Table 1.4. Note that one of our first impressions is substantiated by the fact that 48 of the 69 houses are brick while only 8 have frame exteriors. This simple summarization shows how the frequency of the exteriors is distributed over the values of `exter`.

Definition 1.10 *A **frequency distribution** is a listing of frequencies of all categories of the observed values of a variable.*

We can construct frequency distributions for any variable. For example, Table 1.5 shows the distribution of the variable `zip`, which despite having numeric values, is actually a categorical variable. This frequency distribution is produced by `Proc Freq` of the SAS System where the frequency distribution is shown in the column labeled `Frequency`. Apparently the area represented by zip code 4 has the most homes for sale.

Definition 1.11 *A **relative frequency distribution** consists of the **relative** frequencies, or proportions (percentages), of observations belonging to each category.*

The relative frequencies expressed as percents are provided in Table 1.5 under the heading `Percent` and are useful for comparing frequencies among categories. These relative frequencies have a useful interpretation: They give the chance or **probability** of getting an observation from each category in a blind or random draw. Thus if we

Table 1.4 Distribution of `exter`

exter	Frequency
Brick	48
Frame	8
Other	13

Table 1.5 Distribution of `zip`

		THE FREQ PROCEDURE		
zip	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	8.70	6	8.70
2	13	18.84	19	27.54
3	16	23.19	35	50.72
4	34	49.28	69	100.00

Table 1.6 Distribution of Home Prices in Intervals of $50,000

Range	Frequency	THE FREQ PROCEDURE Percent	Cumulative Frequency	Cumulative Percent
less than 50k	4	5.80	4	5.80
50k to 100k	22	31.88	26	37.68
100k to 150k	23	33.33	49	71.01
150k to 200k	10	14.49	59	85.51
200k to 250k	2	2.90	61	88.41
250k to 300k	1	1.45	62	89.86
300k to 350k	4	5.80	66	95.65
350k to 400k	3	4.35	69	100.00

were to randomly draw an observation from the data in Table 1.2, there is an 18.84% chance that it will be from zip area 2. For this reason a relative frequency distribution is often referred to as an observed or **empirical probability distribution** (Chapter 2).

Constructing a frequency distribution of a numeric variable is a little more complicated. Defining individual values of the variable as categories will usually only produce a listing of the original observations since very few, if any, individual observations will normally have identical values. Therefore, it is customary to define categories as intervals of values, which are called **class** intervals. These intervals must be nonoverlapping and usually each class interval is of equal size with respect to the scale of measurement. A frequency distribution of the variable `price` is shown in Table 1.6. Clearly the preponderance of homes is in the 50- to 150-thousand-dollar range.

The column labeled `Cumulative Frequency` in Table 1.6 is the **cumulative frequency distribution**, which gives the frequency of observed values less than or equal to the upper limit of that class interval. Thus, for example, 59 of the homes are priced at less than $200,000. The column labeled `Cumulative Percent` is the cumulative relative frequency distribution, which gives the proportion (percentage) of observed values less than the upper limit of that class interval. Thus the 59 homes priced at less than $200,000 represent 85.51% of the number of homes offered. We will see later that cumulative relative frequencies — especially those near 0 and 100% — can be of considerable importance.

1.4.1 Graphical Representation of Distributions

Using the principle that a picture is worth a thousand words (or numbers), the information in a frequency distribution is more easily grasped if it is presented in graphical form. The most common graphical presentation of a frequency distribution for numerical data is a **histogram** while the most common presentation

for nominal, categorical, or discrete data is a **bar chart**. Both these graphs are constructed in the same way. Heights of vertical rectangles represent the frequency or the relative frequency. In a histogram, the width of each rectangle represents the size of the class and the rectangles are usually contiguous and of equal width so that the *areas* of the rectangles reflect the relative frequency. In a bar chart the width of the rectangle has no meaning; however, all the rectangles should be the same width to avoid distortion. Figure 1.1 shows a frequency bar chart for `exter` from Table 1.2 that shows the large proportion of brick homes clearly. Figure 1.2 shows a frequency histogram for `price`, clearly showing the preponderance of homes selling from 50 to 150 thousand dollars.

Another presentation of a distribution is provided by a **pie chart**, which is simply a circle (pie) divided into a number of slices whose sizes correspond to the frequency or relative frequency of each class. Figure 1.3 shows a pie chart for the variable `zip`. We have produced these graphs with different programs and options to show that, although there may be slight differences in appearances, the basic information remains the same.

The use of graphs and charts is pervasive in the news media, business and economic reports, and governmental reports and publications, mainly due to the

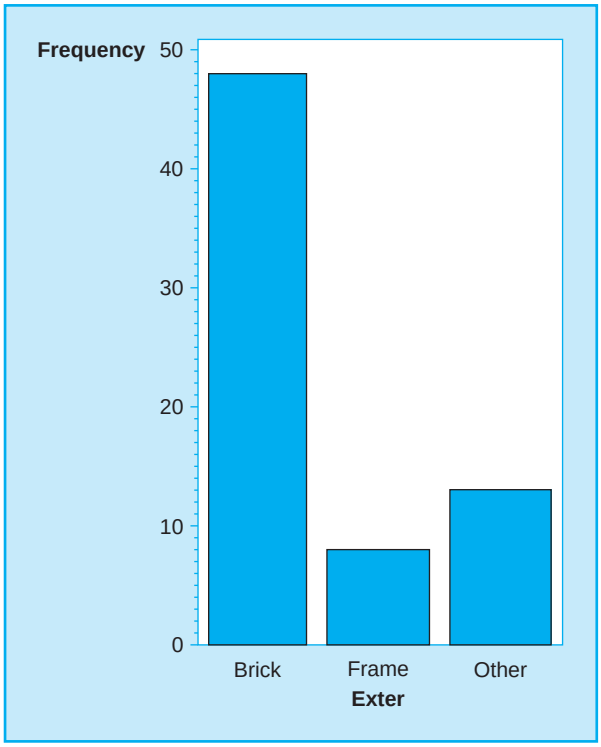


FIGURE 1.1
Bar Chart for `exter`.

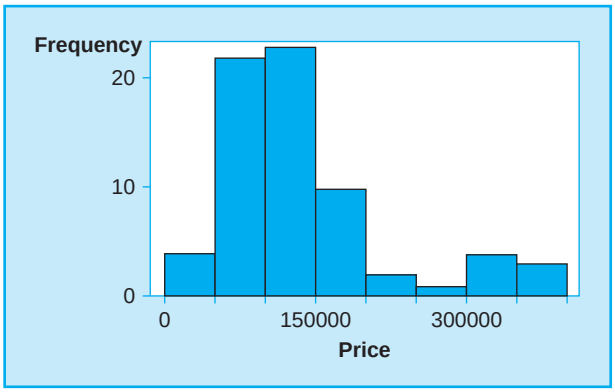


FIGURE 1.2
Histogram of `price`.

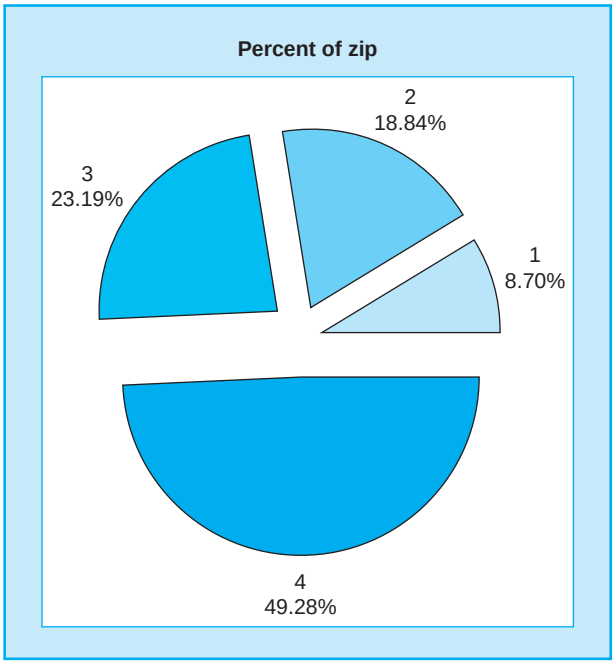


FIGURE 1.3
Pie Chart for the Relative Frequency Distribution of `zip`.

ease of storage, retrieval, manipulation, and summary of large sets of data using modern computers. Because of this, it is extremely important to be able to evaluate critically the information contained in a graph or chart. After all, a graphical presentation is simply a visual impression, which is quite easy to distort. In fact, distortion is so easy and commonplace that in 1992 the Canadian Institute of Chartered Accountants deemed it necessary to begin setting guidelines for financial graphics, after a study of hundreds of the annual reports of major corporations reported almost 10% of the reports contained at least one misleading graph that masked unfavorable data.

Whether intentional or by honest mistake, it is very easy to mislead with an incorrectly presented chart or graph. Darrell Huff, in a book entitled *How to Lie with Statistics* (1982) illustrates many such charts and graphs and discusses various issues concerning misleading graphs. In general, a correctly constructed chart or graph should have

1. all axes labeled correctly, with clearly identifiable scales,
2. be captioned correctly,
3. have bars and/or rectangles of equal width to avoid distortion,
4. have sizes of figures properly proportioned, and
5. contain only relevant information.

Histograms of numeric variables provide information on the **shape** of a distribution, a characteristic that we will later see to be of importance when performing statistical analyses. The shape is roughly defined by drawing a reasonably smooth line through the tops of the bars. In such a representation of a distribution, the region of highest frequency is known as the "peak" and the ends as "tails." If the tails are of approximately equal length, the distribution is said to be symmetric. If the distribution has an elongated tail on the right side, the distribution is skewed to the right and vice versa. Other features may consist of a sharp peak and long "fat" tails, or a broad peak and short tails. We can see that the distribution of `price` is slightly skewed to the right, which, in this case, is due to a few unusually high prices. We will see later that recognizing the shape of a distribution can be quite important.

We continue the study of shapes of distributions with another example.

■ Example 1.3

The discipline of forest science is a frequent user of statistics. An important activity is to gather data on the physical characteristics of a random sample of trees in a forest. The resulting data may be used to estimate the potential yield of the forest, to obtain information on the genetic composition of a particular species, or to investigate the effect of environmental conditions.

Table 1.7 is a listing of such a set of data. This set consists of measurements of three characteristics of 64 sample trees of a particular species. The researcher would like to summarize this set of data in graphic form to aid in its interpretation.

Solution

As we can see from Table 1.7, the data set consists of 64 observations of three ratio variables. The three variables are measurements characterizing each tree and are identified by brief mnemonic identifiers in the column headings as follows:

1. `DFOOT`, the diameter of the tree at one foot above ground level, measured in inches,

Table 1.7 Data on Tree Measurements

OBS	DFOOT	HCRN	HT	OBS	DFOOT	HCRN	HT	OBS	DFOOT	HCRN	HT
1	4.1	1.5	24.5	23	4.3	2.0	25.6	45	4.7	3.3	29.7
2	3.4	4.7	25.0	24	2.7	3.0	20.4	46	4.6	8.9	26.6
3	4.4	2.8	29.0	25	4.3	2.0	25.0	47	4.8	2.4	28.1
4	3.6	5.1	27.0	26	3.3	1.8	20.6	48	4.5	4.7	28.5
5	4.4	1.6	26.5	27	5.0	1.7	24.6	49	3.9	2.3	26.0
6	3.9	1.9	27.0	28	5.2	1.8	26.9	50	4.4	5.4	28.0
7	3.6	5.3	27.0	29	4.7	1.5	26.7	51	5.0	3.2	30.4
8	4.3	7.6	28.0	30	3.8	3.2	26.3	52	4.6	2.5	30.5
9	4.8	1.1	28.5	31	3.8	2.6	27.6	53	4.1	2.1	26.0
10	3.5	1.2	26.0	32	4.2	1.8	23.5	54	3.9	1.8	29.0
11	4.3	2.3	28.0	33	4.7	2.7	25.0	55	4.9	4.7	29.5
12	4.8	1.7	28.5	34	5.0	3.1	27.3	56	4.9	8.3	29.5
13	4.5	2.0	30.0	35	3.2	2.9	26.2	57	5.1	2.1	28.4
14	4.8	2.0	28.0	36	4.1	1.3	25.8	58	4.4	1.7	29.0
15	2.9	1.1	20.5	37	3.5	3.2	24.0	59	4.2	2.2	28.5
16	5.6	2.2	31.5	38	4.8	1.7	26.5	60	4.6	6.6	28.5
17	4.2	8.0	29.3	39	4.3	6.5	27.0	61	5.1	1.0	26.5
18	3.7	6.3	27.2	40	5.1	1.6	27.0	62	3.8	2.7	28.5
19	4.6	3.0	27.0	41	3.7	1.4	25.9	63	4.8	2.2	27.0
20	4.2	2.4	25.4	42	5.0	3.8	29.5	64	4.0	3.1	26.0
21	4.8	2.9	30.4	43	3.3	2.4	25.8				
22	4.3	1.4	24.5	44	4.3	3.0	25.2				

2. HCRN, the height to the base of the crown measured in feet, and
3. HT, the total height of the tree measured in feet.

A histogram for the heights (HT) of the 64 trees is shown in Fig. 1.4 as produced by PROC INSIGHT of the SAS System. Due to space limitations, not all boundaries of class intervals are shown, but we can deduce that the default option of PROC INSIGHT yielded a class interval width of 1.5 feet with the first interval being from 20.25 to 21.75 and the last from 30.75 to 32.25. In this program the user can adjust the size of class intervals by clicking on an arrow at the lower left (not shown in Fig 1.4) that causes a menu to pop up allowing such changes. For example, by changing the first "tick" to 20, the last to 32, and the "tick interval" to 2, the histogram will have 6 classes instead of the 8 shown. Many graphics programs allow this type of interactive modification. Of course, the basic shape of the distribution is not changed by such modifications. Also note that in these histograms, the legend gives the boundaries of the intervals; other graphic programs may give the midpoints.

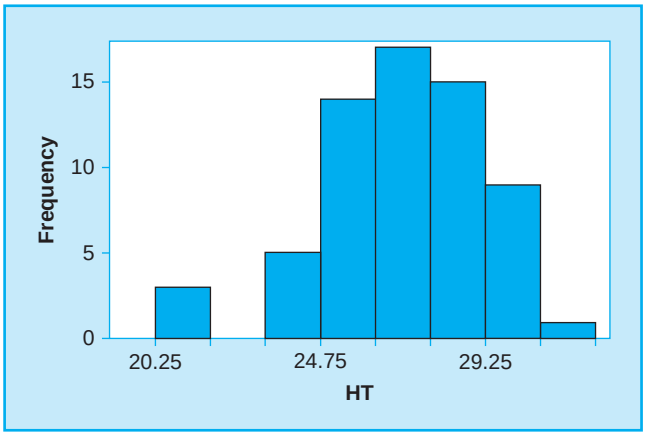


FIGURE 1.4
Histogram of Tree Height.

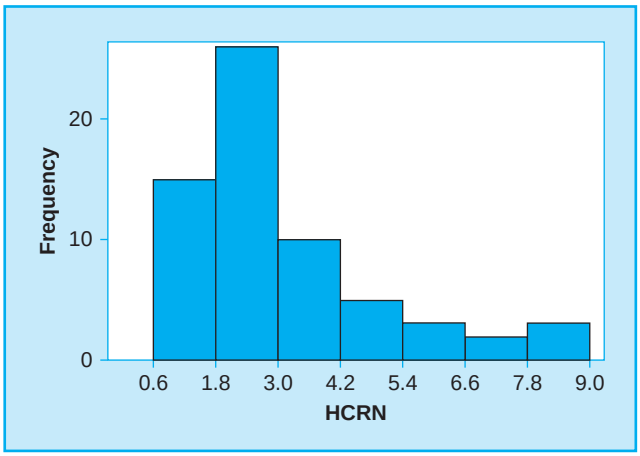


FIGURE 1.5
Histogram of HCRN Variable.

The histogram for the variable HCRN is shown in Fig. 1.5. We can now see that the distribution of HT is slightly skewed to the left while the distribution of HCRN is quite strongly skewed to the right. ■

1.5 NUMERICAL DESCRIPTIVE STATISTICS

Although distributions provide useful descriptions of data, they still contain too much detail for some purposes. Assume, for example, that we have collected data on tree dimensions from several forests for the purpose of detecting possible differences in the distribution of tree sizes among these forests. Side-by-side histograms of the distributions would certainly give some indication of such differences, but would not produce measures of the differences that could be used for quantitative comparisons. **Numerical** measures that provide descriptions of the characteristics of the

distributions, which can then be used to provide more readily interpretable information on such differences, are needed. Of course, since these are numerical measures, their use is largely restricted to numeric variables, that is, variables measured in the ratio or interval scales (see, however, Chapters 12 and 14).

Note that when we first started evaluating the tree measurement data (Table 1.7) we had 64 observations to contend with. As we attempted to summarize the data using a frequency distribution of heights and the accompanying histogram (Fig. 1.4) we represented these data with only eight entries (classes). We can use numerical descriptive statistics to reduce the number of entries describing a set of data even further, typically using only using two numbers. This action of reducing the number of items used to describe the distribution of a set of data is referred to as **data reduction**, which is unfortunately accompanied by a progressive loss of information. In order to minimize the loss of information, we need to determine the most important characteristics of the distribution and find measures to describe these characteristics. The two most important aspects are the **location** and the **dispersion** of the data. In other words, we need to find a number that indicates where the observations are on the measurement scale and another to indicate how widely the observations vary.

1.5.1 Location

The most useful single characteristic of a distribution is some typical, average, or representative value that describes the set of values. Such a value is referred to as a descriptor of **location** or **central tendency**. Several different measures are available to describe this concept. We present two in detail. Other measures not widely used are briefly noted.

The most frequently used measure of location is the arithmetic mean, usually referred to simply as the mean.

Definition 1.12 *The **mean** is the sum of all the observed values divided by the number of values.*

Denote by $y_i, i = 1, \ldots, n$, an observed value of the variable Y, then the sample mean[4] denoted by $\bar{y}$ is obtained by the formula

$$\bar{y} = \frac{\sum y_i}{n},$$

where the symbol $\sum$ stands for "the sum of." For example, the mean for `DFOOT` in Table 1.7 is 4.301, which is the mean diameter (at one foot above the ground) of the 64 trees measured. A quick glance at the observed values of `DFOOT` reveals that this value is indeed representative of the values of that variable.[5]

[4] It is also often called the **average**. However, this term is often used as a generic term for any unspecified measure of location and will therefore not be used in this context.

[5] Some small data sets suitable for practicing computations are available in the following pages as well as in exercises at the end of the chapter.

Another useful measure of location is the median.

Definition 1.13 *The **median** of a set of observed values is defined to be the middle value when the measurements are arranged from lowest to highest; that is, 50% of the measurements lie above it and 50% fall below it.*

The precise definition of the median depends on whether the number of observations is odd or even as follows:

1. If n is odd, the median is the middle observation; hence, exactly $(n-1)/2$ values are greater than and $(n-1)/2$ values are less than the median, respectively.
2. If n is even, there are two middle values and the median is the mean of the two middle values and $n/2$ values are greater than and $n/2$ values are less than the median, respectively.[6]

Although both mean and median are measures of central tendency, they do differ in interpretation. For example, consider the following data for two variables, X and Y, given in Table 1.8.

Table 1.8 Data for Comparing Mean and Median

X	Y
1	1
2	1
3	1
3	2
4	5
5	8

We first compute the means

$$\bar{x} = (1/6)(1+2+3+3+4+5) = 3.0$$

and

$$\bar{y} = (1/6)(1+1+1+2+5+8) = 3.0.$$

The means are the same for both variables.

Denoting the medians by m_x and m_y, respectively, and noting that there are an even number of observations, we find

$$m_x = (3+3)/2 = 3.0$$

and

$$m_y = (1+2)/2 = 1.5.$$

The medians are different. The reason for the difference is seen by examining the histograms of the two variables in Fig. 1.6.

The distribution of the variable X is symmetric, while the distribution of the variable Y is skewed to the right. For symmetric or nearly symmetric distributions, the mean and median will be the same or nearly the same, while for skewed distributions the value of the mean will tend to be "pulled" toward the long tail. This phenomenon can be explained by the fact that the mean can be interpreted as the center of gravity

[6] If there are some identical values of the variable, the phrase "or equal to" may need to be added to these statements.

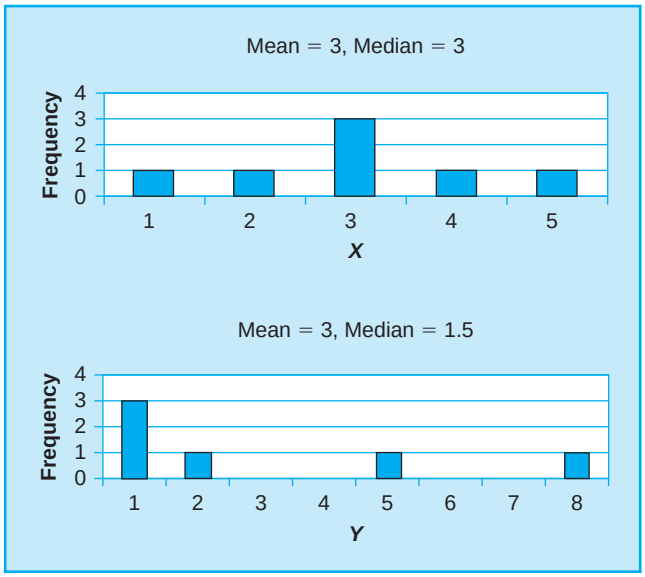


FIGURE 1.6
Data for Comparing Mean and Median.

of the distribution. That is, if the observations are viewed as weights placed on a plane, then the mean is the position at which the weights on each side balance. It is a well-known fact of physics that weights placed further from the center of gravity exert a larger degree of influence (also called leverage); hence the mean must shift toward those weights in order to achieve balance. However, the median assigns equal weights to all observations regardless of their actual values; hence the extreme values have no special leverage.

The difference between the mean and median is also illustrated by the tree data (Table 1.7). The heights variable (HT) was seen to have a reasonably symmetric distribution (Fig. 1.4). The mean diameter is 26.96 and its median is 27.0.[7] The variable HCRN has a highly right-skewed distribution (Fig. 1.5) and its mean is 3.04, which is quite a bit larger than its median of 2.4.

Now that we have two measures of location, it is logical to ask, which is better? Which one should we use? Note that the mean is calculated using the value of each observation, so all the information available from the data is utilized. This is not so for the median. For the median we only need to know where the "middle" of the data is. Therefore, the mean is the more useful measure and, in most cases, the mean will give a better measure of the location of the data. However, as we have seen, the value of the mean is heavily influenced by extreme values and tends to become a distorted

[7]It is customary to give a mean with one more decimal than the observed values. Computer programs usually give all decimal places that the space on the output allows. If a median corresponds to an observed value (n odd), the value is presented as is; if it is the mean of two observations (n even), the extra decimal may be used.

measure of location for a highly skewed distribution. In this case, the median may be more appropriate.

The choice of the measure to be used may depend on its ultimate interpretation and use. For example, monthly rainfall data often contain a few very large values corresponding to rare floods. For this variable, the mean does indicate the total amount of water derived from rain but hardly qualifies as a typical value for monthly rainfall. On the other hand, the median does qualify as a typical value, but certainly does not reflect the total amount of water.

In general, we will use the mean as the single measure of location unless the distribution of the variable is skewed. We will see later (Chapter 4) that variables with highly skewed distributions can be regarded as not fulfilling the assumptions required for methods of statistical analysis that are based on the mean. In Section 1.6 we present some techniques that may be useful for detecting characteristics of distributions that may make the mean an inappropriate measure of location.

Other occasionally used measures of location are as follows:

1. The **mode** is the most frequently occurring value. This measure may not be unique in that two (or more) values may occur with the same greatest frequency. Also, the mode may not be defined if all values occur only once, which usually happens with continuous numeric variables.
2. The **geometric** mean is the nth root of the product of the values of the n observations. This measure is related to the arithmetic mean of the logarithms of the observed values. The geometric mean cannot exist if there are any values less than or equal to 0.
3. The **midrange** is the mean of the smallest and largest observed values. This measure is not frequently used because it ignores most of the information in the data. (See the following discussion of the range and similar measures.)

1.5.2 Dispersion

Although location is generally considered to be the most important single characteristic of a distribution, the **variability** or **dispersion** of the values is also very important. For example, it is imperative that the diameters of $\frac{1}{4}$-in. nuts and bolts have virtually no variability, or else the nuts may not match the bolts. Thus the mean diameter provides an almost complete description of the size of a set of $\frac{1}{4}$-in. nuts and bolts. However, the mean or median incomes of families in a city provide a very inadequate description of the distribution of that variable since a listing of incomes would include a wide range of values.

Figure 1.7 shows histograms of two small data sets. Both have 10 observations, both have a mean of 5 and, since the distributions are symmetric, both have a median of 5. However, the two distributions are certainly quite different. Data set 2 may be described as having more variability since it has fewer observations near the mean and more observations at the extremes of the distribution.

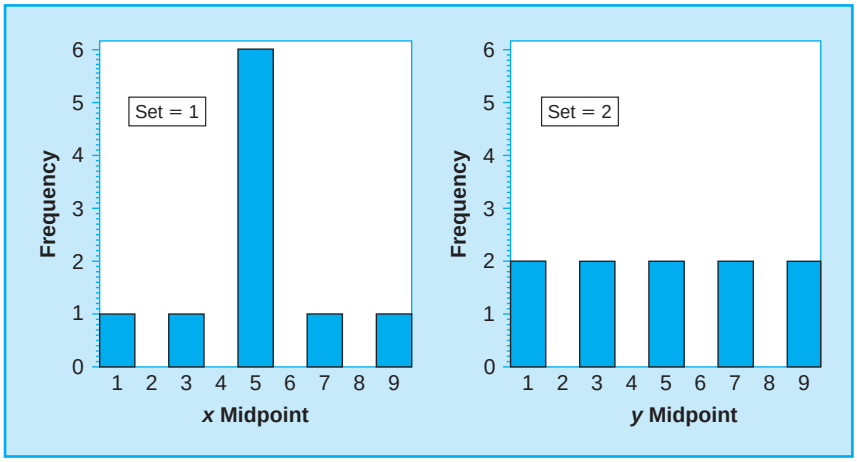


FIGURE 1.7
Illustration of Dispersion.

The simplest and intuitively most obvious measure of variability is the **range**, which is defined as the difference between the largest and smallest observed values. Although conceptually simple, the range has one very serious drawback: It completely ignores any information from all the other values in the data. This characteristic is also illustrated by the two data sets in Fig. 1.7. Both of these data sets exhibit the same range (eight), but data set 2 exhibits more variability.

Since greater dispersion means that observations are farther from the center of the distribution, it is logical to consider distances of observations from that center as indication of variability. The preferred measure of variation when the mean is used as the measure of center is based on the set of distances or differences of the observed values (y_i) from the mean ($\bar{y}$). These differences, $(y_i - \bar{y}), i = 1, 2, \ldots, n$, are called the **deviations** from the mean. Large magnitudes of deviation imply a high degree of variability, and small magnitudes of deviation imply a low degree of variability. If all deviations are zero, the data set exhibits no variability; that is, all values are identical.

The mean of these deviations would seem to provide a reasonable measure of dispersion. However, a relatively simple exercise in algebra shows that the sum of these deviations, that is, $\sum(y_i - \bar{y})$, is always zero. Therefore, this quantity is not useful. The mean absolute deviation (the mean of deviations ignoring their signs) will certainly be an indicator of variability and is sometimes used for that purpose. However, this measure turns out not to be very useful as the absolute values make theoretical development difficult.

Another way to neutralize the effect of opposite signs is to base the measure of variability on the *squared* deviations. Squaring each deviation gives a nonnegative value

and summing the squares of the deviations gives a positive measure of variability. This criterion is the basis for the most frequently used measure of dispersion, the **variance**.

Definition 1.14 *The **sample variance**, denoted by s^2, of a set of n observed values having a mean $\bar{y}$ is the sum of the squared deviations divided by $n-1$:*

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}.$$

Note that the variance is actually an average or mean of the squared deviations and is often referred to as a **mean square**, a term we will use quite often in later chapters. Note also that we have divided the sum by $(n-1)$ rather than n. While the reason for using $(n-1)$ may seem confusing at this time, there is a good reason for it. As we see later in the chapter, one of the uses of the sample variance is to estimate the population variance. Dividing by n tends to underestimate the population variance; therefore by dividing by $(n-1)$ we get, on average, a more accurate estimate. Recall that we have already noted that the sum of deviations $\sum(y_i - \bar{y}) = 0$; hence, if we know the values of any $(n-1)$ of these values, the last one must have that value that causes the sum of all deviations to be zero. Thus there are only $(n-1)$ "free" deviations. Therefore, the quantity $(n-1)$ is called the **degrees of freedom**.

An equivalent argument is to note that in order to compute s^2, we must first compute $\bar{y}$. Starting with the concept that a set of n observed values of a variable provides n units of information, when we compute s^2 we have already used one piece of information, leaving only $(n-1)$ "free" units or $(n-1)$ degrees of freedom.

Computing the variance using the above formula is straightforward but somewhat tedious. First we must compute $\bar{y}$, then the individual deviations $(y_i - \bar{y})$, square these, and then sum. For the two data sets represented by Fig. 1.7 we obtain

Data set 1:

$$\begin{aligned} s^2 &= (1/9)[(1-5)^2 + (3-5)^2 + \cdots + (9-5)^2] \\ &= (1/9) \cdot 40 = 4.44, \end{aligned}$$

Data set 2:

$$\begin{aligned} s^2 &= (1/9)[(1-5)^2 + (1-5)^2 + \cdots + (9-5)^2] \\ &= (1/9) \cdot 80 = 8.89, \end{aligned}$$

showing the expected larger variance for data set 2.

Calculations similar to that for the numerator of the variance are widely used in many statistical analyses and if done as shown in Definition 1.14 are quite tedious. This numerator, called the **sum of squares** and often denoted by SS, is more easily

calculated by using the equivalence

$$\text{SS} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 \Big/ n.$$

The first portion, $\sum y_i^2$, is simply the sum of squares of the original y values. The second part, $(\sum y_i)^2/n$, the square of the sum of the y values divided by the number of observations, is called the **correction factor**, since it "corrects" the sum of squared values to become the sum of squared deviations from the mean. The result, SS, is called the corrected, or centered, sum of squares, or often simply the sum of squares. This sum of squares is divided by the degrees of freedom to obtain the mean square, which is the variance. In general, then the variance

$$s^2 = \text{mean square} = (\text{sum of squares})/(\text{degrees of freedom}).$$

For the case of computing a variance from a single set of observed values, the sum of squares is the sum of squared deviations from the mean of those observations, and the degrees of freedom are $(n - 1)$. For more complex situations, which we will encounter in subsequent chapters, we will continue with this general definition of a variance; however, there will be different methods for computing sums of squares and degrees of freedom.

The computations are now quite straightforward, especially since many calculators have single-key operations for obtaining sums and sums of squares.[8] For the two data sets we have

Data set 1:

$$n = 10, \quad \sum y_i = 50, \quad \sum y_i^2 = 290,$$
$$\text{SS} = 290 - 50^2/10 = 40,$$
$$s^2 = 40/9 = 4.44,$$

Data set 2:

$$n = 10, \quad \sum y_i = 50, \quad \sum y_i^2 = 330,$$
$$\text{SS} = 330 - 50^2/10 = 80,$$
$$s^2 = 80/9 = 8.89.$$

For purposes of interpretation, the variance has one major drawback: It measures the dispersion in the square of the units of the observed values. In other words, the numeric value is not descriptive of the variability of the observed values. This flaw is remedied by using the square root of the variance, which is called the **standard deviation**.

[8] Many calculators also automatically obtain the variance (or standard deviation). Some even provide options for using either n or $(n - 1)$ for the denominator of the variance estimate! We suggest practice computing a few variances without using this feature.

Definition 1.15 *The **standard deviation** of a set of observed values is defined to be the positive square root of the variance.*

This measure is denoted by s and does have, as we will see shortly, a very useful interpretation as a measure of dispersion. For the two example data sets, the standard deviations are

$$\text{Data set 1:} \quad s = 2.11,$$
$$\text{Data set 2:} \quad s = 2.98.$$

Usefulness of the Mean and Standard Deviation

Although the mean and standard deviation (or variance) are only two descriptive measures, together the two actually provide a great deal of information about the distribution of an observed set of values. This is illustrated by the **empirical rule**: If the shape of the distribution is nearly bell shaped, the following statements hold:

1. The interval $(\bar{y} \pm s)$ contains approximately 68% of the observations.
2. The interval $(\bar{y} \pm 2s)$ contains approximately 95% of the observations.
3. The interval $(\bar{y} \pm 3s)$ contains virtually all of the observations.

Note that for each of these intervals the mean is used to describe the location and the standard deviation is used to describe the dispersion of a given portion of the data. We illustrate the empirical rule with the tree data (Table 1.7). The height (HT) was seen to have a nearly bell-shaped distribution, so the empirical rule should hold as a reasonable approximation. For this variable we compute

$$n = 64, \quad \bar{y} = 26.959, \quad s^2 = 5.163, \quad s = 2.272.$$

According to the empirical rule:

$(\bar{y} \pm s)$, which is 26.959 ± 2.272, defines the interval 24.687 to 29.231 and should include $(0.68)(64) = 43$ observations,
$(\bar{y} \pm 2s)$, which is 26.959 ± 4.544, defines the interval from 22.415 to 31.503 and should include $(0.95)(64) = 61$ observations, and
$(\bar{y} \pm 3s)$ defines the interval from 20.143 to 33.775 and should include all 64 observations.

The effectiveness of the empirical rule is verified using the actual data. This task may be made easier by obtaining an ordered listing of the observed values or using a stem and leaf plot (Section 1.6), which we do not reproduce here. For this variable, 46 values fall between 24.687 and 29.231, 61 fall between 22.415 and 31.503, and all observations fall between 20.143 and 33.775. Thus the empirical rule appears to work reasonably well for this variable.

The empirical rule furnishes us with a quick method of estimating the standard deviation of a bell-shaped distribution. Since at least 95% of the observations fall within 2 standard deviations of the mean in either direction, the range of the data covers

Table 1.9 The Empirical Rule Applied to a Nonsymmetric Distribution

Interval		Number of Obervations	
Specified	**Actual**	**Should Include**	**Does Include**
$\bar{y} \pm s$	1.146 to 4.926	43	51
$\bar{y} \pm 2s$	−0.744 to 6.816	61	60
$\bar{y} \pm 3s$	−2.634 to 8.706	64	63

about 4 standard deviations. Thus, we can estimate the standard deviation (a crude estimate by the way) by taking the range divided by 4. For example, the range of the data on the HT variable is $31.5 - 20.4 = 11.1$. Divided by 4 we get about 2.77. The actual standard deviation had a value of 2.272, which is approximately "in the ball park," so to speak.

The HCRN variable had a rather skewed distribution (Fig. 1.5); hence the empirical rule should not work as well. The mean is 3.036 and the standard deviation is 1.890. The expected and actual frequencies are given in Table 1.9. As expected, the empirical rule does not work as well. In other words, for a nonsymmetric distribution the mean and standard deviation (or variance) do not provide as complete a description of the distribution as they do for a more nearly bell-shaped one. We may want to include a histogram or general discussion of the shape of the distribution along with the mean and standard deviation when describing data with a highly skewed distribution.

Actually the mean and standard deviation provide useful information about a distribution no matter what the shape. A much more conservative relation between the distribution and its mean and standard deviation is given by Tchebysheff's theorem.

Definition 1.16 ***Tchebysheff's theorem*** *For any arbitrary constant k, the interval $(\bar{y} \pm ks)$ contains a proportion of the values of at least $[1 - (1/k^2)]$.*[9]

Note that Tchebysheff's theorem is more conservative than the empirical rule. This is because the empirical rule describes distributions that are approximately "bell" shaped, whereas Tchebysheff's theorem is applicable for any shaped distribution. For example, for $k = 2$, Tchebysheff's theorem states that the interval $(\bar{y} \pm 2s)$ will contain at least $[1 - (1/4)] = 0.75$ of the data. For the HCRN variable, this interval is from −0.744 to 6.816 (Table 1.9), which actually contains $60/64 = 0.9375$ of the values. Thus we can see that Tchebysheff's theorem provides a guarantee of a proportion in an interval but at the cost of a wider interval.

[9]Tchebysheff's theorem is usually described in terms of a theoretical distribution rather than for a set of data. This difference is of no concern at this point.

The empirical rule and Tchebysheff's theorem have been presented not because they are quoted in many statistical analyses but because they demonstrate the power of the mean and standard deviation to describe a set of data. The wider intervals specified by Tchebysheff's theorem also show that this power is diminished if the assumption of a bell-shaped curve is not made.

1.5.3 Other Measures

A measure of dispersion that has uses in some applications is the **coefficient of variation**.

Definition 1.17 *The **coefficient of variation** is the ratio of the standard deviation to the mean, expressed in percentage terms.*

Usually denoted by CV, it is

$$\text{CV} = \frac{s}{\bar{y}} \cdot 100.$$

That is, the CV gives the standard deviation as a proportion of the mean. For example, a standard deviation of 5 has little meaning unless we can compare it to something. If $\bar{y}$ has a value of 100, then this variation would probably be considered small. If, however, $\bar{y}$ has a value of 1, a standard deviation of 5 would be quite large relative to the mean. If we were evaluating the precision of a laboratory measuring device, the first case, CV = 5%, would probably be acceptable. The second case, CV = 500%, probably would not.

Additional useful descriptive measures are the **percentiles** of a distribution.

Definition 1.18 *The **pth percentile** is defined to be that value for which at most (p)% of the measurements are less and at most $(100 - p)$% of the measurements are greater.*[10]

For example, the 75th percentile of the diameter variable (`DFOOT`) corresponds to the 48th ($0.75 \cdot 64 = 48$) ordered observation, which is 4.8. This means that 75% of the trees have diameters of 4.8 in. or less. By definition, cumulative relative frequencies define percentiles.

To illustrate how a computer program calculates percentiles, the Frequency option of SPSS was instructed to find the 30th percentile for the same variable, `DFOOT`. The program returned the value 4.05. To find this value we note that $0.3 \times 64 = 19.2$. Therefore we want the value of `DFOOT` for which 19.2 of the observations are smaller and 60.8 are larger. This means that the 30th percentile falls between the 19th observation, 4.00, and the 20th observation, 4.10. The computer program simply took the midpoint between these two values and gave the 30th percentile the value of 4.05.

[10] Occasionally the percentile desired falls between two of the measurements in the data set. In that case interpolation may be used to obtain the value. To avoid becoming unnecessarily pedantic, most people simply choose the midpoint between the two values involved. Different computer programs may use different interpolation methods.

A special set of percentiles of interest are the **quartiles**, which are the 25th, 50th, and 75th percentiles. The 50th percentile is, of course, the median.

Definition 1.19 *The **interquartile range** is the length of the interval between the 25th and 75th percentiles and describes the range of the middle half of the distribution.*

For the tree diameters, the 25th and 75th percentiles correspond to 3.9 and 4.8 inches; hence the interquartile range is 0.9 inches. We will use this measure in Section 1.6 when we discuss the box plot. We will see later that we are often interested in the percentiles at the extremes or tails of a distribution, especially the 1, 2.5, 5, 95, 97.5, and 99th percentiles.

Certain measures may be used to describe other aspects of a distribution. For example, a measure of skewness is available to indicate the degree of skewness of a distribution. Similarly, a measure of kurtosis indicates whether a distribution has a narrow "peak" and fat "tails" or a flat peak and skinny tails. Generally, a "fat-tailed" distribution is characterized by having an excessive number of outliers or unusual observations, which is an undesirable characteristic. Although these measures have some theoretical interest, they are not often used in practice. For additional information, see Snedecor and Cochran (1980), Sections 5.13 and 5.14.

1.5.4 Computing the Mean and Standard Deviation from a Frequency Distribution

If a data set is presented as a frequency distribution, a good approximation of the mean and variance may be obtained directly from that distribution. Let y_i represent the midpoint and f_i the frequency of the ith class. Then

$$\bar{y} \approx \sum f_i y_i \Big/ \sum f_i$$

and

$$s^2 \approx \sum f_i (y_i - \bar{y})^2 \Big/ \sum f_i$$

or, using the computational form,

$$s^2 \approx \left[\sum f_i y_i^2 - \left(\sum f_i y_i \right)^2 \Big/ \sum f_i \right] \Big/ \sum f_i.$$

Note that these formulas use **weighted** sums of the observed values[11] or squared deviations. That is, each value is weighted by the number of observations it represents. If the y_i are the actual values (rather than midpoints of intervals) of a discrete distribution, these formulas provide exactly the same values as those using the formulas presented previously in this section.

[11]These formulas are primarily used for large data sets where $n \approx n - 1$; hence $\sum f_i = n$, rather than $(n - 1)$, is used as the denominator for computing the variance.

Equivalent formulas may be used for data represented as a relative frequency distribution. Let p_i be the relative frequency of the ith class. Then

$$\bar{y} \approx \sum p_i y_i \quad \text{and} \quad s^2 \approx \sum p_i (y_i - \bar{y})^2$$

or, using the computational form,

$$s^2 \approx \sum p_i y_i^2 - \left(\sum p_i y_i \right)^2.$$

Most data sets are available in their original form and since computers readily perform direct computation of mean and variance these formulas are not often used. We will, however, find these formulas useful in discussions of theoretical probability distributions in Chapter 2.

1.5.5 Change of Scale

Change of scale is often called **coding** or **linear transformation**. Most interval and ratio variables arise from measurements on a scale such as inches, grams, or degrees Celsius. The numerical values describing these distributions naturally reflect the scale used. In some circumstances it is useful to change the scale such as, for example, changing from imperial (inches, pounds, etc.) to metric units. Scale changes may take many forms, including a change from ratio to ordinal scales as mentioned in Section 1.3. Other scale changes may involve the use of functions such as logarithms or square roots (see Chapter 6).

A useful form of scaling is the use of a linear transformation. Let Y represent a variable in the observed scale, which is transformed to a rescaled or transformed variable X by the equation

$$X = a + bY,$$

where a and b are constants. The constant a represents a change in the **origin**, while the constant b represents a change in the unit of measurement, or **scale**, identified with a ratio or interval scale variable (Section 1.3). A well-known example of such a transformation is the change from degrees Celsius to degrees Fahrenheit. The formula for the transformation is

$$X = 32 + 1.8Y,$$

where X represents readings in degrees Fahrenheit and Y in degrees Celsius.

Many descriptive measures retain their interpretation through linear transformation. Specifically, for the mean and variance:

$$\bar{x} = a + b\bar{y} \quad \text{and} \quad s_x^2 = b^2 s_y^2.$$

A useful application of a linear transformation is that of reducing round-off errors. For example, consider the following values $y_i, i = 1, 2, \ldots, 6$:

$$10.004 \quad 10.002 \quad 9.997 \quad 10.000 \quad 9.996 \quad 10.001.$$

Using the linear transformation

$$x_i = -10{,}000 + 1000\, y_i$$

results in the values of x_i

$$4 \quad 2 \quad -3 \quad 0 \quad -4 \quad 1,$$

from which it is easy to calculate

$$\bar{x} = 0 \quad \text{and} \quad s_x^2 = 9.2.$$

Using the above relationships, we see that $\bar{y} = 10.000$ and $s_y^2 = 0.0000092$.

The use of the originally observed y_i may induce round-off error. Using the original data,

$$\sum y_i = 60.000, \quad \sum y_i^2 = 600.000046, \quad \text{and} \quad \left(\sum y_i\right)^2 \Big/ n = 600.000000.$$

Then

$$\text{SS} = 0.000046 \quad \text{and} \quad s^2 = 0.0000092.$$

If the calculator we are using has only eight digits of precision, then $\sum y^2$ would be truncated to 600.00004, and we would obtain $s^2 = 0.000008$. Admittedly this is a pathological example, but round-off errors in statistical calculations occur quite frequently, especially when the calculations involve many steps as will be required later. Therefore, scaling by a linear transformation is sometimes useful.

1.6 EXPLORATORY DATA ANALYSIS

We have seen that the mean and variance (or standard deviation) can do a very good job of describing the characteristics of a frequency distribution. However, we have also seen that these do not work as well when the distribution is skewed and/or includes some extreme or outlying observations. Because the vast majority of statistical analyses make use of the mean and standard deviation, the results of such analyses may prove misleading if the distribution has such features. Therefore, it is imperative that some preliminary checks of the data be performed to see if other methods (see Section 4.5 and Chapter 14) may be more appropriate.

Fortunately, the same computers that can so easily produce inappropriate analyses can just as easily be used to perform preliminary data screening to provide an overview of the nature of the data and thus provide information on unusual distributions and/or data anomalies. A variety of such procedures have been developed and many are available on most popularly used computer software. These procedures are called **exploratory data analysis** techniques or **EDA**, a concept first introduced

by Tukey (1977). We present here two of the most frequently used EDA tools: the **stem and leaf plot** and the **box plot**.

1.6.1 The Stem and Leaf Plot

The stem and leaf plot is a modification of a histogram for a ratio or interval variable that provides additional information about the distribution of the variable. The first one or two digits specify the class interval, called the "stem," and the next digit (rounded if necessary) is used to construct increments of the bar, which are called the "leaves." Usually in a stem and leaf plot, the bars are arranged horizontally and the leaf values are arranged in ascending order.

We illustrate the construction of a stem and leaf plot using the data on `size` for the 69 homes. To make construction easier, we first arrange the observations from low to high as shown in Table 1.10.

Normally the first or first two digits are used to define stem values, but in this case using one would result in an inadequate five stems, while using two would generate an overwhelming 40 stems. A compromise is to use the first two digits, in sets of two, a procedure automatically done by computer programs. In this example, the first stem value (the "." corresponds to the missing value) is 6, which identifies the range of 600 to 799 square feet. There is one observation in that range, 676, so the leaf value is 8 (76 rounded to 80). The second stem value has two observations, 951 and 994, producing leaf values of 5 and 9. When there are homes represented by both individual stem values, the leaf values for the first precede those for the second. For example, the stem value of 24 represents the range from 2400 to 2599. The first four leaf values 4, 6, 6, and 8, are in the range 2400 to 2499, while the values 1, 5, and 7 are in the range 2500 to 2599. The last stem value is 40 with a leaf value of 1. The

Table 1.10 Home Sizes Measured in Square Feet Arranged from Low to High

.	1344	1624	2016	2483	3055
676	1368	1636	2036	2510	3056
951	1387	1647	2038	2553	3253
994	1410	1750	2082	2572	3310
1036	1450	1752	2113	2670	3314
1064	1456	1770	2262	2805	3472
1152	1456	1770	2298	2809	3627
1176	1500	1800	2336	2921	3846
1186	1524	1852	2370	2949	4106
1216	1532	1920	2436	2992	
1312	1540	1972	2456	2993	
1344	1550	1980	2463	3045	

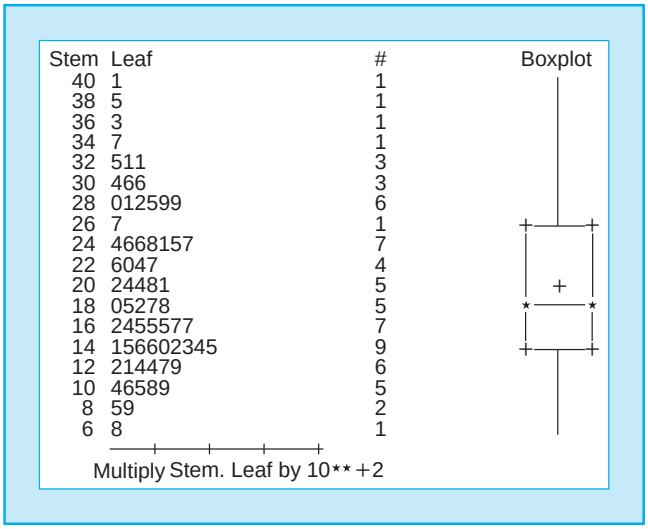


FIGURE 1.8
Stem and Leaf Plot and Box Plot for `size`.

resulting plot is shown in Fig. 1.8, produced by `PROC UNIVARIATE` of the SAS System, which automatically also provides the box plot discussed later in this section.[12]

At first glance, the stem and leaf plot looks like a histogram, which it is. However, the stem and leaf plot usually has a larger number of bars (or stems), 18 in this case, which provide greater detail about the nature of the distribution. In this case the stem and leaf chart does not provide any new information on this data set. The leaves provide rather little additional information here, but could, for example, provide evidence of rounding or imprecise measurements by showing an excessive number of zeros and fives. The leaves may also provide evidence of bunching of specific values within a stem by showing disproportionate frequencies of specific digits.

For some data sets minor modifications may be necessary to provide an informative plot. For example, the first digit of the `HCRN` variable in the tree data (Table 1.7) provides for only eight stems (classes) while using the first two digits creates too many stems. In such cases it is customary to use two lines for each digit, the first representing leaves with values from 0 through 4, and a second for values from 5 through 9. Most computer programs automatically adjust for such situations. This plot is given in Fig. 1.9 (also produced by `PROC UNIVARIATE`). The extreme skewness we have previously noted is quite obvious.

1.6.2 The Box Plot

The box plot[13] is used to show distributional shapes and to detect unusual observations. Figure 1.10 illustrates a typical box plot and the procedure is illustrated in

[12]This provides a good illustration of the fact that computer programs do not always provide only what is needed.
[13]Also referred to as a "box and whisker plot" by Tukey (1977).

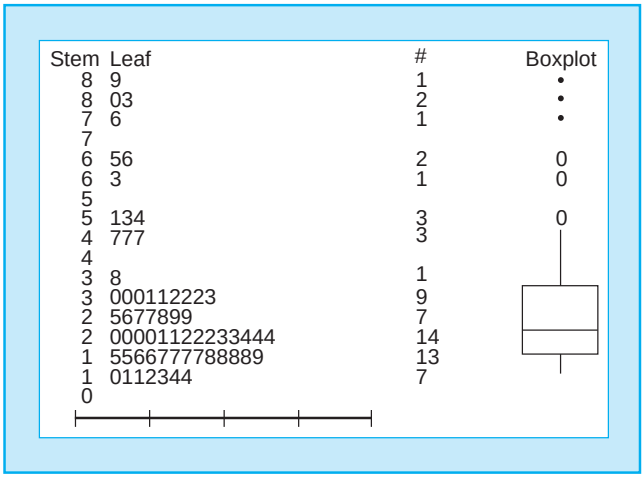


FIGURE 1.9
Stem and Leaf Plot and Box Plot for HCRN Variable.

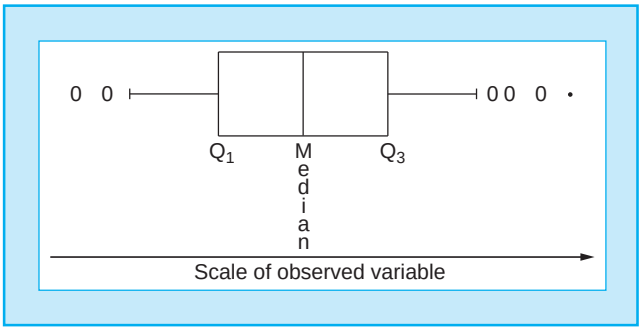


FIGURE 1.10
Typical Box Plot.

Fig. 1.8 for the size variable from the housing data set and in Fig. 1.9 for the HCRN variable from the trees data set.

The scale of the plot is that of the observed variable and may be presented horizontally as in Fig. 1.10 or vertically as produced by the SAS System in Figs. 1.8 and 1.9. The features of the plot are as follows:

1. The "box," representing the interquartile range, has a value we denote by R and the endpoints Q_1 and Q_3.
2. A vertical line inside the box indicates the median. If the median is in the center of the box, the middle portion of the distribution is symmetric.
3. Horizontal lines extending from the box represent the range of observed values inside the "inner fences," which are located 1.5 times the value of the interquartile range ($1.5R$) beyond Q_1 to the left and Q_3 on the right. The relative lengths of these lines are an indicator of the skewness of the distribution as a whole.

4. Individual symbols ∘ represent "mild" outliers, which are defined as values between the inner and outer fences that are located $3R$ units beyond Q_1 and Q_3.
5. Individual symbols • represent the location of extreme outliers, which are defined as being beyond the outer fences. Different computer programs may use different symbols for outliers and may provide options for different formats.

Symmetric distributions, which can be readily described by the mean and variance, should have the median line close to the middle of the box and reasonably equal length lines on both sides, a few mild outliers preferably equally distributed on both sides, and virtually no extreme outliers.

An ordered listing of the data or a stem and leaf plot can be used to construct the box plot. We illustrate the procedure for the HCRN variable for which the stem and leaf and box plots are shown in Fig. 1.9. Note that the box plot is arranged vertically in that plot. The scale is the same as the stem and leaf plot on the left. The details of the procedure are as follows:

1. The quartiles Q_1 and Q_3 are found by counting $(n/4) = 16$ leaf values from the top and bottom, respectively. The resulting values of 1.8 and 3.2 define the box. These values also provide the interquartile range: $R = Q_3 - Q_1 = 3.2 - 1.8 = 1.4$. The median of 2.4 defines the line in the box.
2. The inner fences are

$$f_1 = Q_1 - 1.5R = 1.8 - 2.1 = -0.3 \quad \text{and}$$
$$f_2 = Q_3 + 1.5R = 3.2 + 2.1 = 5.3.$$

   The lines extend on each side to the nearest actual values inside the inner fences. In this example the lines extend to 1.0 (the smallest value in the data set) and 5.3, respectively. The much longer line on the high side clearly indicates the skewness.
3. The outer fences are $F_1 = -2.4$ and $F_2 = 7.4$. The fact that the lower fence has a negative value that cannot occur is a clear indicator of a skewed distribution. The four mild outliers lying between the inner and outer fences are 5.4, 6.3, 6.5, and 6.6, and are indicated by the symbol ∘. Note that they are all on the high side, again indicating the skewness.
4. The extreme outliers are beyond the outer fences. They are 7.6, 8.0, 8.3, and 8.9, and are indicated by •. These are also all on the high side.

Thus we see that the box plot clearly shows the lack of symmetry for the distribution of the HCRN variable. On the other hand, the box plot for the house sizes (Fig. 1.8) shows little lack of symmetry and also has neither mild nor extreme outliers. Obviously the box plot provides a good bit of information on the distribution and outliers, but cannot be considered a complete replacement for the stem and leaf plot in terms of total information about the observations.

1.6.3 Comments

The presence of outliers in a set of data may cause problems in the analysis to be performed. For example, a single outlier (or several in the same direction) usually causes a distribution to be skewed, thereby affecting the mean of the distribution. In the box plot in Fig. 1.9 we see that there are several large values of the HCRN variable identified as outliers. If the mean is to be used for the analysis, it may be larger than is representative of the data due to the presence of these outliers. However, we cannot simply ignore or discard these observations as the trees do exist and to ignore them would be dishonest. A closer examination of the larger trees may reveal that they actually belong to an older grove that represents a different population from that being studied. In that case we could eliminate these observations from the analysis, but note that older trees that belonged to a population not included in the study were present in the data.

Descriptive statistical techniques, and in particular the EDA methods discussed here, are valuable in identifying outliers; however, the techniques very rarely furnish guidance as to what should be done with the outliers. In fact, the concern for "unrepresentative," "rogue," or "outlying" observations in sets of data has been voiced by many people for a long time. There is evidence that concern for outliers predates most of statistical methodology. Treatments of outliers are discussed in many texts, and in fact a book by Barnett and Lewis (1994), entitled *Outliers in Statistical Data*, is completely devoted to the topic. The sheer volume of literature addressing outliers points to the difficulty of adjusting the analysis when outliers are present.

All outliers are not deleterious to the analysis. For example, the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of some unexpectedly useful chemical reaction or surprisingly successful variety of corn. Often it is not necessary to take either of the extreme positions — reject the outlier or include the outlier — but instead to use some form of "robust" analysis that minimizes the effect of the outlier. One such example would be to use the median in the analysis of the variable HCRN in the tree data instead of the mean.

■ Example 1.4

A biochemical assay for a substance we will abbreviate to cytosol is supposed to be an indicator of breast cancer. Masood and Johnson (1987) report on the results of such an assay, which indicates the presence of this material in units per 5 mg of protein on 42 patients. Also reported are the results of another cancer detection method, which are simply reported as "yes" or "no." The data are given in Table 1.11. We would like to summarize the data on the variable CYTOSOL.

Solution

All the descriptive measures, the stem and leaf plot, and the box plot for these observations are given in Fig. 1.11 as provided by the Minitab DES-CRIBE, STEM-AND-LEAF, and BOXPLOT commands.

Table 1.11 Cytosol Levels in Cancer Patients

OBS	CYTOSOL	CANCER	OBS	CYTOSOL	CANCER
1	145.00	YES	22	1.00	NO
2	5.00	NO	23	3.00	NO
3	183.00	YES	24	1.00	NO
4	1075.00	YES	25	269.00	YES
5	5.00	NO	26	33.00	YES
6	3.00	NO	27	135.00	YES
7	245.00	YES	28	1.00	NO
8	22.00	YES	29	1.00	NO
9	208.00	YES	30	37.00	YES
10	49.00	YES	31	706.00	YES
11	686.00	YES	32	28.00	YES
12	143.00	YES	33	90.00	YES
13	892.00	YES	34	190.00	YES
14	123.00	YES	35	1.00	YES
15	1.00	NO	36	1.00	NO
16	23.00	YES	37	7.20	NO
17	1.00	NO	38	1.00	NO
18	18.00	NO	39	1.00	NO
19	150.00	YES	40	71.00	YES
20	3.00	NO	41	189.00	YES
21	3.20	YES	42	1.00	NO

The first portion gives the numerical descriptors. The mean is 136.9 and the standard deviation is 248.5. Note that the standard deviation is greater than the mean. Since the variable (CYTOSOL) cannot be negative, the empirical rule will not be applicable, implying that the distribution is skewed. This conclusion is reinforced by the large difference between the mean and the median. Finally, the first quartile is the same as the minimum value, indicating that at least 25% of the values occur at the minimum. The asymmetry is also evident from the positions of the quartiles, with values of 1.0 and 158.3 respectively. The output also gives the minimum and maximum values, along with two measures (TRMEAN and SEMEAN), which are not discussed in this chapter.

The stem and leaf and box plots reinforce the extremely skewed nature of this distribution. It is of interest to note that in this plot the mild outliers are denoted by * (there are none) and extreme outliers by 0.

A conclusion to be reached here is that the mean and standard deviation are not particularly useful measures for describing the distribution of this variable. Instead, the median should be used along with a brief description of the shape of the distribution. ■

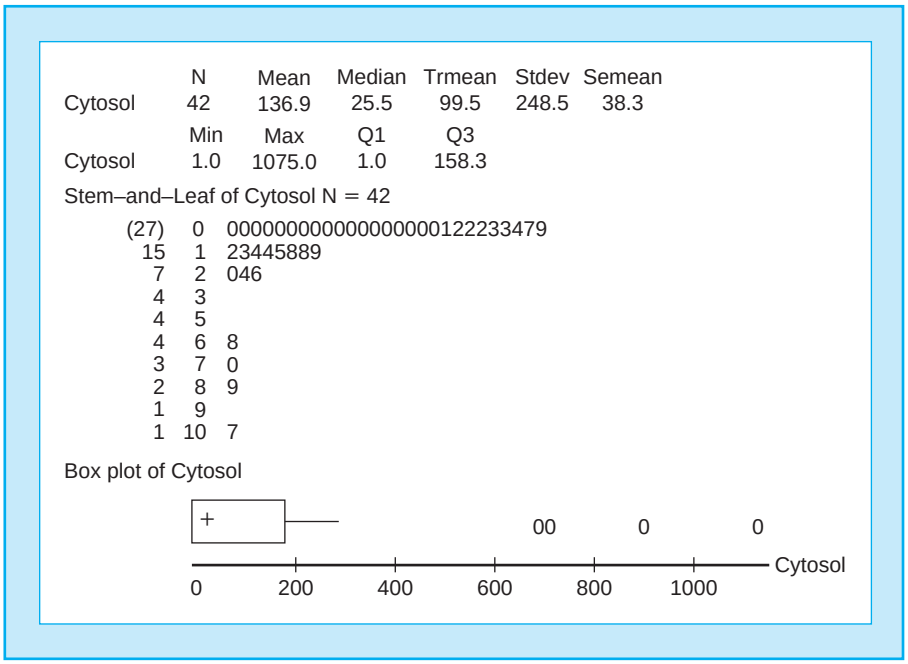


FIGURE 1.11
Descriptive Measures of CYTOSOL.

1.7 BIVARIATE DATA

So far we have presented methods for describing the distribution of observed values of a single variable. These methods can be used individually to describe distributions of each of several variables that may occur in a set of data. However, when there are several variables in one data set, we may also be interested in describing how these variables may be related to or associated with each other. We present in this section some graphic and tabular methods for describing the association between two variables. Numeric descriptors of association are presented in later chapters, especially Chapters 7 and 8.

Specific methods for describing association between two variables depend on whether the variables are measured in a nominal or numerical scale. (Association between variables measured in the ordinal scale is discussed in Chapter 14.) We illustrate these methods by using the variables on home sales given in Table 1.2.

1.7.1 Categorical Variables

Table 1.12 reproduces the home sales data for the two categorical variables sorted in order of zip and exter. Association between two variables measured in the nominal scale (categorical variables) can be described by a two-way frequency distribution,

Table 1.12 Home Sales Data for the Categorical Variables

zip	exter	zip	exter	zip	exter	zip	exter	zip	exter	zip	exter
1	Brick	2	Brick	3	Frame	4	Brick	4	Brick	4	Brick
1	Brick	2	Brick	3	Frame	4	Brick	4	Brick	4	Brick
1	Brick	2	Brick	3	Frame	4	Brick	4	Brick	4	Brick
1	Brick	2	Brick	3	Frame	4	Brick	4	Brick	4	Brick
1	Frame	2	Frame	3	Other	4	Brick	4	Brick	4	Brick
1	Other	2	Other	3	Other	4	Brick	4	Brick	4	Frame
2	Brick	2	Other	3	Other	4	Brick	4	Brick	4	Other
2	Brick	3	Brick	3	Other	4	Brick	4	Brick	4	Other
2	Brick	3	Brick	3	Other	4	Brick	4	Brick	4	Other
2	Brick	3	Brick	3	Other	4	Brick	4	Brick		
2	Brick	3	Brick	3	Other	4	Brick	4	Brick		
2	Brick	3	Frame	4	Brick	4	Brick	4	Brick		

Table 1.13 Association between `zip` and `exter`

The FREQ Procedure
Table of zip by exter

ZIP Frequency Row pct	EXTER			
	Brick	Frame	Other	Total
1	4 66.67	1 16.67	1 16.67	6
2	10 76.92	1 7.69	2 15.38	13
3	4 25.00	5 31.25	7 43.75	16
4	30 88.24	1 2.94	3 8.82	34
Total	48	8	13	69

which is a two-dimensional table showing the frequencies of combinations of the values of the two variables. Table 1.13 is such a table showing the association between the zip and exterior siding material of the houses. This table has been produced by `PROC FREQ` of the SAS System. The table shows the frequencies of the six combinations of the `zip` and `exter` variables. The headings at the top and left indicate the categories of the two variables. Each of the combinations of the two variables is referred to as a **cell**. The last row and column (each labeled `Total`) are the individual or marginal frequencies of the two variables. As indicated by the legend at the top left of the table, the first number in each cell is the frequency.

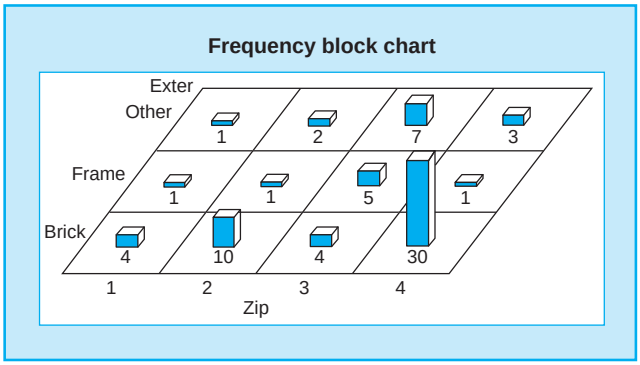


FIGURE 1.12
Block Chart for `exter` and `zip`.

The second number in each cell is the row percentage, that is, the percentage of each row (`zip`) that is brick, frame, or other. We can now see that brick homes predominate in all zip areas except 3, which has a mixture of all types.

The relationship between two categorical variables can also be illustrated with a block chart (a three-dimensional bar chart) with the height of the blocks being proportional to the frequencies. A block chart of the relationship between `zip` and `exter` is given in Fig. 1.12. Numeric descriptors for relationships between categorical variables are presented in Chapter 12.

1.7.2 Categorical and Interval Variables

The relationship between a categorical and interval (or ratio) variable is usually described by computing frequency distributions or numerical descriptors for the interval variables for each value of the nominal variable. For example, the mean and standard deviation of sales prices for the four zip areas are

$$\begin{aligned}
&\texttt{zip}\ \text{area 1},\ \bar{y} = 86,892, \quad s = 26,877\\
&\texttt{zip}\ \text{area 2},\ \bar{y} = 147,948, \quad s = 67,443\\
&\texttt{zip}\ \text{area 3},\ \bar{y} = 96,455, \quad s = 50,746\\
&\texttt{zip}\ \text{area 4},\ \bar{y} = 169,624, \quad s = 98,929.
\end{aligned}$$

We can now see that `zip` areas 2 and 4 have the higher priced homes. Side-by-side box plots can illustrate this information graphically as shown in Fig. 1.13 for `price` by `zip`. This plot reinforces the information provided by the means and standard deviations, but additionally shows that all of the very-high-priced homes are in `zip` area 4.

Box plots may also be used to illustrate differences among distributions. We illustrate this method with the cancer data, by showing the side-by-side box plots of `CYTOSOL` for the two groups of patients who were diagnosed for cancer by the other method. The results, produced this time with `PROC INSIGHT` of the SAS System in Fig. 1.14, shows that both the location and dispersion differ markedly between the two groups.

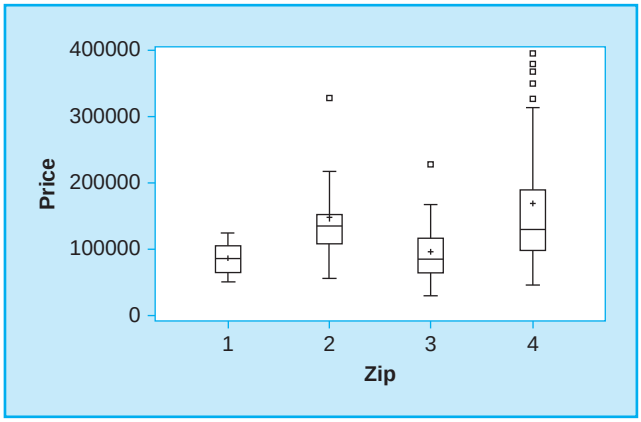


FIGURE 1.13
Side-by-Side Box Plots of Home Prices.

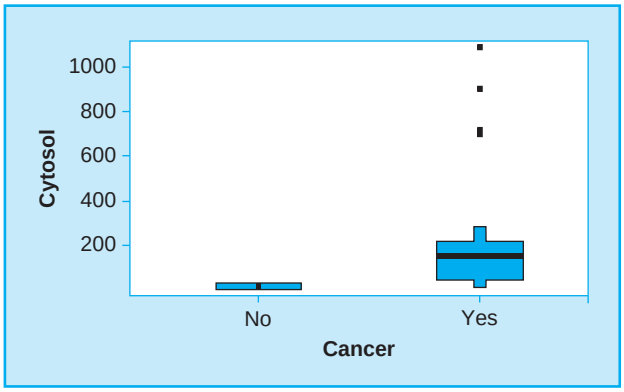


FIGURE 1.14
Side-by-Side Box Plots for Cancer Data.

Apparently both methods can detect cancer, although contradictory diagnoses occur for some patients.

1.7.3 Interval Variables

The relationship between two interval variables can be graphically illustrated with a **scatterplot**. A scatterplot has two axes representing the scales of the two variables. The choice of variables for the horizontal or vertical axes is immaterial, although if one variable is considered more important it will usually occupy the vertical axis. Also, if one variable is used to predict another variable, the variable being predicted always goes on the vertical axis. Each observation is plotted by a point representing the two variable values. Special symbols may be needed to show multiple points with identical values. The pattern of plotted points is an indicator of the nature of the relationship between the two variables. Figure 1.15 is a scatterplot showing the relationship between `price` and `size` for the data in Table 1.2.

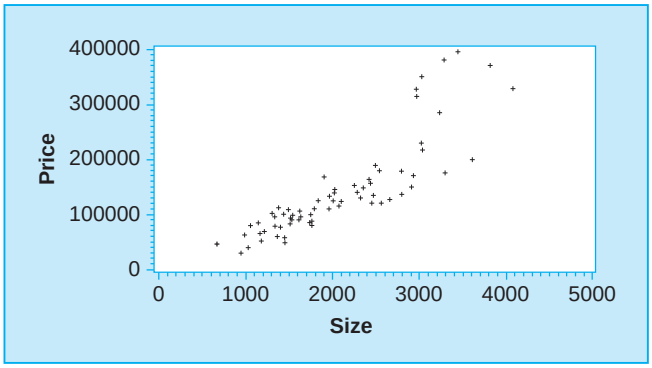


FIGURE 1.15
Scatterplot of `price` against `size`.

The pattern of the plotted data points shows a rather strong association between `price` and `size`, except for the higher price homes. Apparently these houses have a wider range of other amenities that affect the price. Numeric descriptors for this type of association are introduced in Chapter 7.

We should note at this point that the increased sophistication of computer graphics is rapidly leading to more informative graphs and plots. For example, some software packages provide a scatterplot with box plots on each axis describing the distribution of each of the individual variables.

1.8 POPULATIONS, SAMPLES, AND STATISTICAL INFERENCE — A PREVIEW

In the beginning of this chapter we noted that a set of data may represent either a population or a sample. Using the terminology developed in this chapter, we can now more precisely define a **population** as the set of values of one or more variables for the entire collection of units relevant to a particular study. Most researchers have at least a conceptual picture of the population for a given study. This population is usually called the **target** population. A target population may be well defined. For example, the trees in Table 1.7 are a sample from a population of trees in a specified forest. On the other hand, a population may be only conceptually defined. For example, an experiment measuring the decrease in blood pressure resulting from a new drug is a sample from a hypothetical population consisting of all sufferers of high blood pressure who are potential users of the drug. A population can, in fact, be infinite. For example, a laboratory experiment can hypothetically be reproduced an infinite number of times.

We are rarely afforded the opportunity of measuring all the elements of an entire population. For this reason, most data are normally some portion or **sample** of the target population. Obviously a sample provides only partial information on the

population. In other words, the characteristics of the population cannot be completely known from sample data.

We can, however, draw certain parallels between the sample and the population. Both population and sample may be described by measures such as those presented in this chapter (although we cannot usually calculate them for a population). To differentiate between a sample and the population from which it came, the descriptive measures for a sample are called **statistics** and are calculated and symbolized as presented in this chapter. Specifically, the sample mean is $\bar{y}$ and the sample variance is s^2. Descriptive measures for the population are called **parameters** and are denoted by Greek letters. Specifically, we denote the mean of a population by μ and the variance by σ^2. If the population consists of a finite number of values, $y_1, y_2, \ldots, y_N$, then the mean is calculated by

$$\mu = \sum y_i / N,$$

and the variance is found by

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N}.$$

It is logical to assume that the sample statistics provide some information on the values of the population parameters. In other words, the sample statistics may be considered to be **estimates** of the population parameters. However, the statistics from a sample cannot exactly reflect the values of the parameters of the population from which the sample is taken. In fact, two or more individual samples from the same population will invariably exhibit different values of sample estimates. The magnitude of variation among sample estimates is referred to as the **sampling error** of the estimates. Therefore, the magnitude of this sampling error provides an indication of how closely a sample estimate approximates the corresponding population parameter. In other words, if a sample estimate can be shown to have a small sampling error, that estimate is said to provide a good estimate for the corresponding population parameter.

We must emphasize that sampling error is not an error in the sense of making a mistake. It is simply a recognition of the fact that a sample statistic does not exactly represent the value of a population parameter. The recognition and measurement of this sampling error is the cornerstone of statistical inference.

1.9 DATA COLLECTION

Usually, our goal is to use the findings in our sample to make statements about the population from which the sample was drawn, that is, we want to make statistical inferences. But to do this, we have to be careful about the way the data was collected. If the process in some way, perhaps quite subtle, favored getting data that indicated a certain result, then we will have introduced a **bias** into the process. Bias produces

a systematic slanting of the results. Unlike sampling error, its size will not diminish even for very large samples. Worse, its nature cannot be guessed from information contained within the sample itself.

To avoid bias, we need to collect data using random sampling, or some more advanced probability sampling technique. All the statistical inferences discussed in this text assume the data came from random sampling, where "blind chance" dominates the selection of the units. A **simple random sample** is one where each possible sample of the specified size has an equal chance of occurring.

The process of drawing a simple random sample is conceptually simple, but difficult to implement in practice. Essentially, it is like drawing for prizes in a lottery: the population consists of all the lottery tickets and the sample of winners is drawn from a well-shaken drum containing all the tickets. The most straightforward method for drawing a random sample is to create a numbered list, called a **sampling frame**, of all the **sampling units** in the population. A random number generator from a computer program, or a table of random numbers, is used to select units from the list.

■ Example 1.5

Medicare has selected a particular medical provider for audit. The Medicare carrier begins by defining the target population—say all claims from Provider X to Medicare for office visits with dates of service between 1/1/2007 and 12/31/2007. The carrier then combs its electronic records for a list of all claims fitting this description, finding 521. This set of 521 claims, when sorted by beneficiary ID number and date of service, becomes the sampling frame. The sampling units are the individual claims. Units in the list are numbered from 1 to 521. The carrier decides that it has sufficient time and money to carry out an exploratory audit of 30 claims. To select the claims, the carrier uses a computer program to generate 30 integers with values between 1 and 521. Since it would be a waste to audit the same claim twice, these integers will be selected without replacement. The 30 claims in the sampling frame that correspond to these integers are the ones for which the carrier will request medical records and carry out a review.

This procedure can be used for relatively small finite populations but may be impractical for large finite populations, and is obviously impossible for infinite populations. Nevertheless, some blind, unbiased sampling mechanism is important, particularly for observational studies. Human populations are notoriously difficult to sample. Aside from the difficulty of constructing reasonably complete sampling frames for a target population such as "all American men between the ages of 50 and 59," people will frequently simply refuse to participate in a survey, poll, or experiment. This **nonresponse** problem often results in a sample that is drastically different from the target population in ways that cannot be readily assessed.

Convenience samples are another dangerous source of data. These samples consist of whatever data the researcher was most easily able to obtain, usually without

any random sampling. Often these samples allow people to self-select into the data set, as in polls in the media where viewers call in or click a choice on-line to give their opinion. These samples are often wildly biased, as the most extreme opinions will be over-represented in the data. You should never attempt to generalize convenience sample results to the population.

True random samples are difficult. Designed experiments partially circumvent these difficulties by introducing randomization in a different way. Convenience samples are indeed selected, usually with some effort at obtaining a representative group of individuals. This nonrandom sample is then randomly divided into subgroups one of which is often a placebo, control, or standard treatment group. The other subgroups are given alternative treatments. Participants are not allowed to select which treatment they will be given; rather, that is randomly determined. Suppose, for example, that we wanted to know whether adding nuts to a diet low in saturated fat would lead to a greater drop in cholesterol than would the diet alone. We could advertise for volunteers with high total cholesterol levels. We would then randomly divide them into two groups. One group would go on the low saturated-fat diet, the second group would go on the same diet but with the addition of nuts. At the end of three months, we would compare their changes in cholesterol levels. The assumption here is that even though the participants were not recruited randomly, the randomization makes it fair to generalize our results regarding the effect of the addition of the nuts.

For more information on selecting random samples, or for advanced sampling, see a text on sampling (for example, Scheaffer *et al.*, 2006 or Cochran, 1977). Designed experiments are covered in great detail in texts on experimental design (for example, Maxwell and Delaney, 2000). The overriding factor in all types of random sampling is that the actual selection of sample elements not be subject to personal or other bias.

In many cases experimental conditions are such that nonrestricted randomization is impossible; hence the sample is not a random sample. For example, much of the data available for economic research consists of measurements of economic variables over time. For such data the normal sequencing of the data cannot be altered and we cannot really claim to have a random sample of observations. In such situations, however, it is possible to define an appropriate model that contains a random element. Models that incorporate such random elements are introduced in Chapters 6 and 7. ■

1.10 CHAPTER SUMMARY

Solution to Example 1.1

We now know that the data listed in Table 1.1 consists of 50 observations on four variables from an observational study. Two of the variables (AGE and TVHOURS) are numerical and have the ratio level of measurement. The other two are categorical

(nominal) level variables. We will explore the nature of these variables and a few of the relationships between them.

We start by using SPSS to construct the frequency histograms of AGE and TVHOURS as shown in Fig. 1.16. From these it appears that the distribution of age is somewhat skewed positively while that of TVHOURS is extremely skewed positively.

To further explore the shape of the distributions of the two variables we construct the box plots shown in Fig. 1.17. Note the symmetry of the variable AGE while the obvious positive skewness of TVHOURS is highlighted by the long whisker on the positive side of the box plot. Also, note that there is one potential outlier identified in the TVHOURS box plot. This is the value 10 corresponding to the 20th respondent

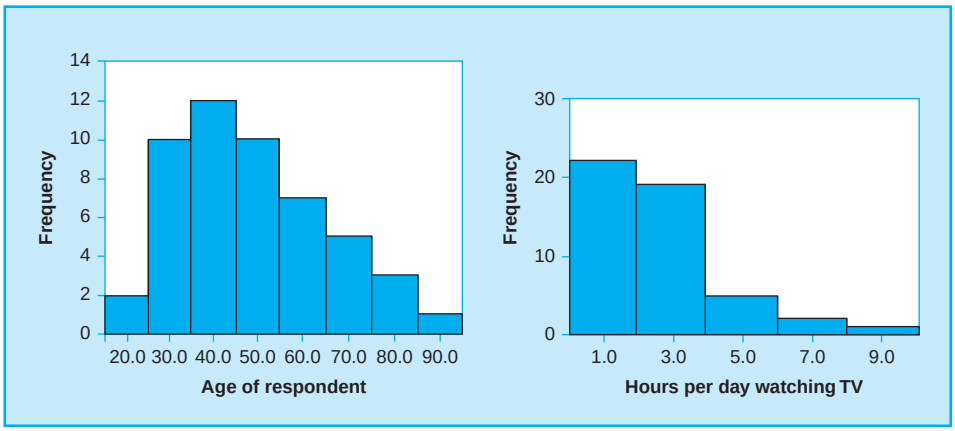


FIGURE 1.16
Histograms of AGE and TVHOURS.

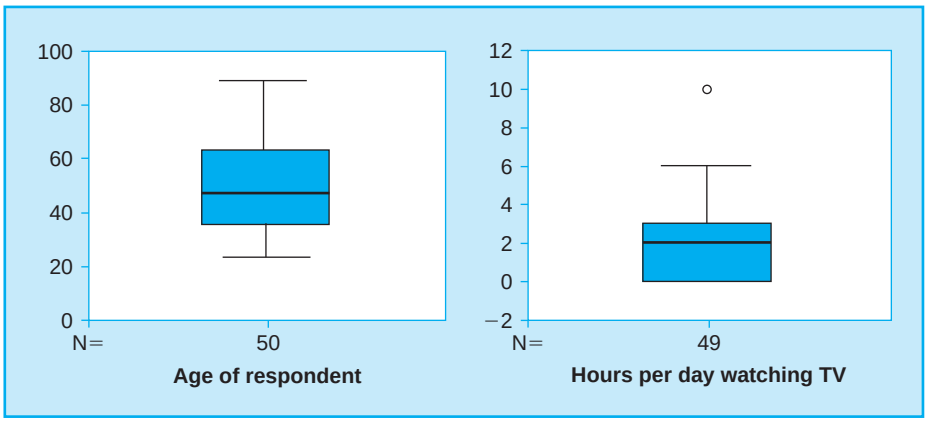


FIGURE 1.17
Box Plots of AGE and TVHOURS.

in the data set. It is also interesting to see that fully 25% of the respondents reported their average number of hours watching TV as 0 as indicated by the fact that the lower quartile (the lower edge of the box) is at the level "0."

We now examine some of the numerical descriptive statistics for these two measures as seen in Table 1.14.

The first two rows of Table 1.14 tell us that all 50 of our sample respondents answered the questions concerning age and number of hours per day watching TV. There were no missing values for these variables. The mean age is 48.26 and the age of the respondents ranges from 23 to 89. The mean number of hours per day watching TV is 1.88 and ranges from 0 to 10. Note that the standard deviation of the number of hours watching TV is actually larger than the mean. This is another indication of the extremely skewed distribution of these values.

Figure 1.18 shows a relative frequency (percent) bar chart of the variable HAPPY. From this we can see that only about 12% of the respondents considered themselves not happy with their lives. Figure 1.18 also shows a pie chart of the variable SEX. This indicates that 56% of the respondents were female vs. 44% male.

To see if there is any noticeable relationship between the variables AGE and TVHOURS, a scatter diagram is constructed. The graph is shown in Fig. 1.19. There does not seem to be a strong relationship between these two variables. There is one respondent who seems to be "separated" from the group, and that is the respondent who watches TV about 10 hours per day.

To examine the relationship between the two variables SEX and HAPPY, we will construct side-by-side relative frequency bar charts. These are given in Fig. 1.20. Note that the patterns of "happiness" seem to be opposite for the sexes. For example, of

Table 1.14 Numerical Statistics for AGE and TVHOURS

	Age of Respondent	Hours per Day Watching TV
N		
Valid	50	50
Missing	0	0
Mean	48.26	1.88
Median	46.00	2.00
Mode	53	0
Std. deviation	17.05	2.14
Variance	290.65	4.60
Minimum	23	0
Maximum	89	10

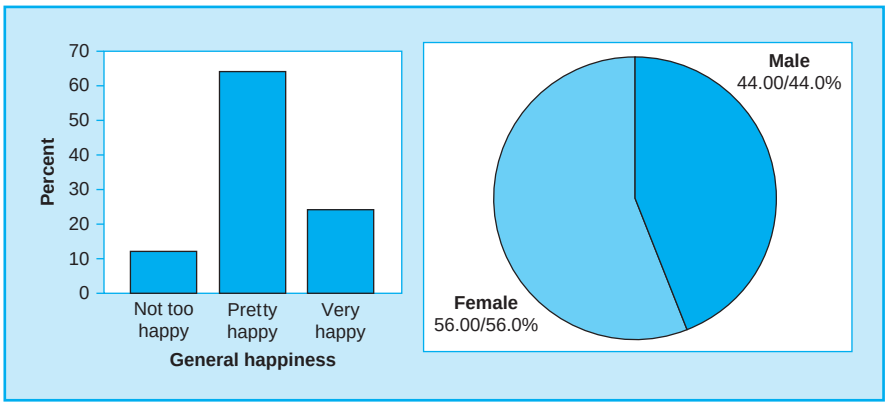


FIGURE 1.18
Bar Chart of HAPPY and Pie Chart of SEX.

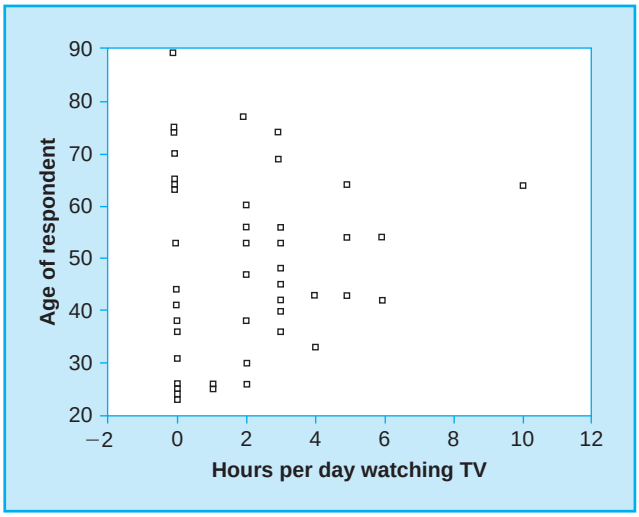


FIGURE 1.19
Scatter Diagram of AGE and TVHOURS.

those who identified themselves as being "Very Happy," 67% were female while only 33% were male.

Finally, to see if there is any difference in the relationship between AGE and TVHOURS when the respondents are identified by SEX, we construct a scatter diagram identifying points by SEX. This graph is given in Fig. 1.21.

The graph does not indicate any systematic difference in the relationship by sex. The respondent who watches TV about 10 hours per day is male, but other than that nothing can be concluded by examination of this graph. ■

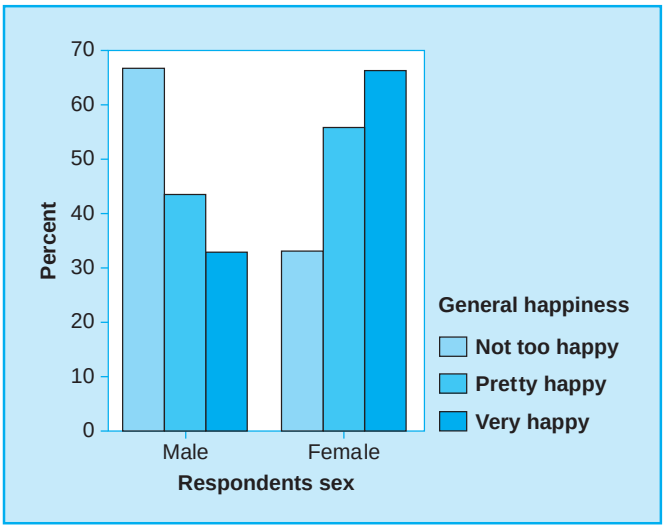


FIGURE 1.20
Side-by-Side Bar Charts for HAPPY by SEX.

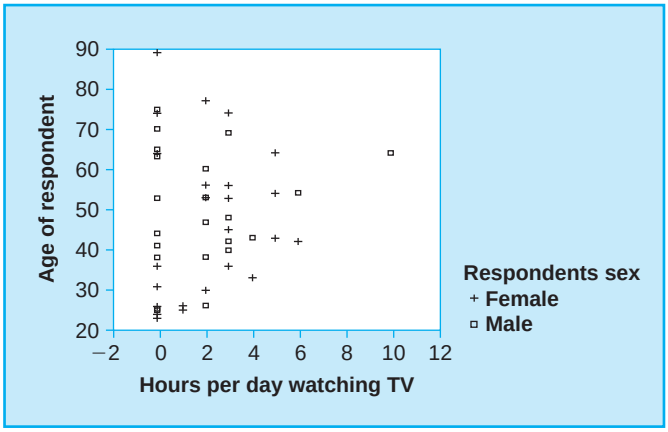


FIGURE 1.21
AGE vs. TVHOURS Identified by SEX.

Summary

Statistics is concerned with the analysis of data. A set of data is defined as a set of observations on one or more variables. Variables may be measured on a nominal, ordinal, interval, or ratio scale with the ratio scale providing the most information. Additionally, interval and ratio scale variables, also called numerical variables, may be discrete or continuous. The nature of a statistical analysis is largely dictated by the type of variable being analyzed.

A set of observations on a variable is described by a distribution, which is a listing of the frequencies with which different values of the variable occur. A relative frequency distribution shows the proportion of the total number of observations associated with each value or class of values and is related to a probability distribution, which is extensively used in statistics.

Graphical representation of distributions is extremely useful for investigating various characteristics of distributions, especially their shape and the existence of unusual values. Frequently used graphical representations include bar charts, stem and leaf plots, and box plots.

Numerical measures of various characteristics of distributions provide a manageable set of numeric values that can readily be used for descriptive and comparative purposes. The most frequently used measures are those that describe the location (center) and dispersion (variability) of a distribution. The most frequently used measure of location is the mean, which is the sum of observations divided by the number of observations. Also used is the median, which is the center value.

The most frequently used measure of dispersion is the variance, which is the average of the squared differences between the observations and the mean. The square root of the variance, called the standard deviation, describes dispersion in the original scale of measurement. Other measures of dispersion are the range, which is the difference between the largest and smallest observations, and the mean absolute deviation, which is the average of the absolute values of the differences between the observations and the mean.

Other numeric descriptors of the characteristics of a distribution include the percentiles, of which the quartile and interquartile ranges are special cases.

The importance of the mean and standard deviation is underscored by the empirical rule and Tchebysheff's theorem, which show that these two measures provide a very adequate description of data distributions.

The chapter concludes with brief sections on descriptions of relationships between two variables and a look ahead at the uses of descriptive measures for statistical inference. The chapter concludes with brief sections that describe certain relationships between two variables, look ahead at the uses of descriptive measures for statistical inference, and highlight some of the issues associated with data collection.

1.11 CHAPTER EXERCISES

Concept Questions

The following multiple choice questions are intended to provide practice in methods and reinforce some of the concepts presented in this chapter.

1. The scores of eight persons on the Stanford–Binet IQ test were:

   95 87 96 110 150 104 112 110

   The median is:

   (1) 107
   (2) 110
   (3) 112

(4) 104
(5) none of the above

2. The concentration of DDT, in milligrams per liter, is:
(1) a nominal variable
(2) an ordinal variable
(3) an interval variable
(4) a ratio variable

3. If the interquartile range is zero, you can conclude that:
(1) the range must also be zero
(2) the mean is also zero
(3) at least 50% of the observations have the same value
(4) all of the observations have the same value
(5) none of the above is correct

4. The species of each insect found in a plot of cropland is:
(1) a nominal variable
(2) an ordinal variable
(3) an interval variable
(4) a ratio variable

5. The "average" type of grass used in Texas lawns is best described by
(1) the mean
(2) the median
(3) the mode

6. A sample of 100 IQ scores produced the following statistics:

$$\begin{aligned} \text{mean} &= 95 & \text{lower quartile} &= 70 \\ \text{median} &= 100 & \text{upper quartile} &= 120 \\ \text{mode} &= 75 & \text{standard deviation} &= 30 \end{aligned}$$

Which statement(s) is (are) correct?
(1) Half of the scores are less than 95.
(2) The middle 50% of scores are between 100 and 120.
(3) One-quarter of the scores are greater than 120.
(4) The most common score is 95.

7. A sample of 100 IQ scores produced the following statistics:

$$\begin{aligned} \text{mean} &= 100 & \text{lower quartile} &= 70 \\ \text{median} &= 95 & \text{upper quartile} &= 120 \\ \text{mode} &= 75 & \text{standard deviation} &= 30 \end{aligned}$$

Which statement(s) is (are) correct?
(1) Half of the scores are less than 100.
(2) The middle 50% of scores are between 70 and 120.
(3) One-quarter of the scores are greater than 100.
(4) The most common score is 95.

8. Identify which of the following is a measure of dispersion:
- **(1)** median
- **(2)** 90th percentile
- **(3)** interquartile range
- **(4)** mean

9. A sample of pounds lost in a given week by individual members of a weight-reducing clinic produced the following statistics:

mean $= 5$ pounds	first quartile $= 2$ pounds
median $= 7$ pounds	third quartile $= 8.5$ pounds
mode $= 4$ pounds	standard deviation $= 2$ pounds

Identify the correct statement:
- **(1)** One-fourth of the members lost less than 2 pounds.
- **(2)** The middle 50% of the members lost between 2 and 8.5 pounds.
- **(3)** The most common weight loss was 4 pounds.
- **(4)** All of the above are correct.
- **(5)** None of the above is correct.

10. A measurable characteristic of a population is:
- **(1)** a parameter
- **(2)** a statistic
- **(3)** a sample
- **(4)** an experiment

11. What is the primary characteristic of a set of data for which the standard deviation is zero?
- **(1)** All values of the variable appear with equal frequency.
- **(2)** All values of the variable have the same value.
- **(3)** The mean of the values is also zero.
- **(4)** All of the above are correct.
- **(5)** None of the above is correct.

12. Let X be the distance in miles from their present homes to residences when in high school for individuals at a class reunion. Then X is:
- **(1)** a categorical (nominal) variable
- **(2)** a continuous variable
- **(3)** a discrete variable
- **(4)** a parameter
- **(5)** a statistic

13. A subset of a population is:
- **(1)** a parameter
- **(2)** a population
- **(3)** a statistic
- **(4)** a sample
- **(5)** none of the above

14. The median is a better measure of central tendency than the mean if:
- **(1)** the variable is discrete
- **(2)** the distribution is skewed
- **(3)** the variable is continuous
- **(4)** the distribution is symmetric
- **(5)** none of the above is correct

15. A small sample of automobile owners at Texas A & M University produced the following number of parking tickets during a particular year: 4, 0, 3, 2, 5, 1, 2, 1, 0. The mean number of tickets (rounded to the nearest tenth) is:
- **(1)** 1.7
- **(2)** 2.0
- **(3)** 2.5
- **(4)** 3.0
- **(5)** none of the above

16. In Problem 15, the implied sampling unit is:
- **(1)** an individual automobile
- **(2)** an individual automobile owner
- **(3)** an individual ticket

17. To judge the extent of damage from Hurricane Ivan, an Escambia County official randomly selects addresses of 30 homes from the county tax assessor's roll and then inspects these homes for damage.

Identify each of the following by writing the appropriate letter into the blank.

_____	Target population	(a)	The tax assessor's roll
_____	Sampling unit	(b)	The 30 homes inspected
_____	Sampling frame	(c)	An individual home
_____	Sample	(d)	All homes in Escambia County

Practice Exercises

Most of the exercises in this and subsequent chapters are based on data sets for which computations are most efficiently done with computers. However, manual computations, although admittedly tedious, provide a feel for how various results arise and what they may mean. For this reason, we have included a few exercises with small numbers of simple-valued observations that can be done manually. The solutions to all these exercises are given in the back of the text.

1. A university published the following distribution of students enrolled in the various colleges:

College	Enrollment	College	Enrollment
Agriculture	1250	Liberal arts	2140
Business	3675	Science	1550
Earth sciences	850	Social sciences	2100

Construct a bar chart of these data.

2. On ten days, a bank had 18, 15, 13, 12, 8, 3, 7, 14, 16, and 3 bad checks. Find the mean, median, variance, and standard deviation of the number of bad checks.

3. Calculate the mean and standard deviation of the following sample:

$$-1, \quad 4, \quad 5, \quad 0.$$

4. The following is the distribution of ages of students in a graduate course:

Age (years)	Frequency
20–24	11
25–29	24
30–34	30
35–39	18
40–44	11
45–49	5
50–54	1

   (a) Construct a bar chart of the data.
   (b) Calculate the mean and standard deviation of the data.

5. The percentage change in the consumer price index (CPI) is widely used as a gauge of inflation. The following numbers show the percentage change in the average CPI for the years 1993 through 2007:

3.0 2.6 2.8 3.0 2.3 1.6 2.2 3.4 2.8 1.6 2.3 2.7 3.4 3.2 2.8

   (a) Using time as the horizontal axis and CPI as the vertical axis, construct a trend graph showing how the CPI moved during this period. Comment on the trend.
   (b) Calculate the mean, standard deviation, and median of the CPI.
   (c) Calculate the inner and outer fences, and use this to say whether there are any outliers in this data.
   (d) Construct a box plot of the CPI values, and comment on the shape of the distribution.

Exercises

1. Most of the problems in this and other chapters deal with "real" data for which computations are most efficiently performed with computers. Since a little experience in manual computing is healthy, here are 15 observations of a variable having no particular meaning:

12 18 22 17 20 15 19 13 23 8 14 14 19 11 30.

   (a) Compute the mean, median, variance, range, and interquartile range for these observations.

(b) Produce a stem and leaf plot.
(c) Write a brief description of this data set.

2. Because waterfowl are an important economic resource, wildlife scientists study how waterfowl abundance is related to various environmental variables. In such a study, the variables shown in Table 1.15 were observed for a sample of 52 ponds.

Table 1.15 Waterfowl Data

OBS	WATER	VEG	FOWL	OBS	WATER	VEG	FOWL
1	1.00	0.00	0	27	0.25	0.00	0
2	0.25	0.00	10	28	1.50	0.00	240
3	1.00	0.00	125	29	2.00	1.50	2
4	15.00	3.00	30	30	31.00	0.00	0
5	1.00	0.00	0	31	149.00	9.00	1410
6	33.00	0.00	32	32	1.00	2.75	0
7	0.75	0.00	16	33	0.50	0.00	15
8	0.75	0.00	0	34	1.50	0.00	16
9	2.00	0.00	14	35	0.25	0.00	0
10	1.50	0.00	17	36	0.25	0.25	0
11	1.00	0.00	0	37	0.75	0.00	125
12	16.00	1.00	210	38	0.25	0.00	2
13	0.25	0.00	11	39	1.25	0.00	0
14	5.00	1.00	218	40	6.00	0.00	179
15	10.00	2.00	5	41	2.00	0.00	80
16	1.25	0.50	26	42	5.00	8.00	167
17	0.50	0.00	4	43	2.00	0.00	0
18	16.00	2.00	74	44	0.25	0.00	11
19	2.00	0.00	0	45	5.00	1.00	364
20	1.50	0.00	51	46	7.00	2.25	59
21	0.50	0.00	12	47	9.00	7.00	185
22	0.75	0.00	18	48	0.00	1.25	0
23	0.25	0.00	1	49	0.00	4.00	0
24	17.00	5.25	2	50	7.00	0.00	177
25	3.00	0.75	16	51	4.00	2.00	0
26	1.50	1.75	9	52	1.00	2.00	0

WATER: the amount of open water in the pond, in acres.
VEG: the amount of aquatic and wetland vegetation present at and around the pond, in acres.
FOWL: the number of waterfowl recorded at the pond during a (random) one-day visit to the pond in January.

The results of some intermediate computations:

$$\begin{aligned} \text{WATER: } \sum y &= 370.5 & \sum y^2 &= 25735.9 \\ \text{VEG: } \sum y &= 58.25 & \sum y^2 &= 285.938 \\ \text{FOWL: } \sum y &= 3933 & \sum y^2 &= 2449535 \end{aligned}$$

(a) Make a complete summary of one of these variables. (Compute mean, median, and variance, and construct a bar chart or stem and leaf and box plots.) Comment on the nature of the distribution.

(b) Construct a frequency distribution for FOWL, and use the frequency distribution formulas to compute the mean and variance.

(c) Make a scatterplot relating WATER or VEG to FOWL.

3. Someone wants to know whether the direction of price movements of the general stock market, as measured by the New York Stock Exchange (NYSE) Composite Index, can be predicted by directional price movements of the New York Futures Contract for the next month. Data on these variables have been collected for a 46-day period and are presented in Table 1.16. The variables are:

Table 1.16 Stock Prices

DAY	INDEX	FUTURE	DAY	INDEX	FUTURE
1	0.58	0.70	24	1.13	0.46
2	0.00	−0.79	25	2.96	1.54
3	0.43	0.85	26	−3.19	−1.08
4	−0.14	−0.16	27	1.04	−0.32
5	−1.15	−0.71	28	−1.51	−0.60
6	0.15	−0.02	29	−2.18	−1.13
7	−1.23	−1.10	30	−0.91	−0.36
8	−0.88	−0.77	31	1.83	−0.02
9	−1.26	−0.78	32	2.86	0.91
10	0.08	−0.35	33	2.22	1.56
11	−0.15	0.26	34	−1.48	−0.22
12	0.23	−0.14	35	−0.47	−0.63
13	−0.97	−0.33	36	2.14	0.91
14	−1.36	−1.17	37	−0.08	−0.02
15	−0.84	−0.46	38	−0.62	−0.41
16	−1.01	−0.52	39	−1.33	−0.81
17	−0.86	−0.28	40	−1.34	−2.43
18	0.87	0.28	41	1.12	−0.34
19	−0.78	−0.20	42	−0.16	−0.13
20	−2.36	−1.55	43	1.35	0.18
21	0.48	−0.09	44	1.33	1.18
22	−0.88	−0.44	45	−0.15	0.67
23	0.08	−0.63	46	−0.46	−0.10

INDEX: the percentage change in the NYSE composite index for a one-day period.
FUTURE: the percentage change in the NYSE futures contract for a one-day period.

(a) Make a complete summary of one of these variables.
(b) Construct a scatterplot relating these variables. Does the plot help to answer the question posed?

4. The data in Table 1.17 consist of 25 values for four computer-generated variables called Y1, Y2, Y3, and Y4. Each of these is intended to represent a particular distributional shape. Use a stem and leaf and a box plot to ascertain the nature of each distribution and then see whether the empirical rule works for each of these.

Table 1.17 Data for Recognizing Distributional Shapes

Y1	Y2	Y3	Y4	Y1	Y2	Y3	Y4
4.0	3.5	1.3	5.0	8.1	4.7	2.7	2.3
6.7	6.4	6.7	1.0	6.3	3.3	1.3	0.1
6.2	3.3	1.3	0.6	6.9	3.9	2.7	3.9
2.4	4.0	2.7	4.5	8.4	5.7	5.4	1.4
1.6	3.5	1.3	1.8	3.1	3.3	1.3	2.2
5.3	4.8	4.0	0.3	4.5	5.2	4.0	0.9
6.8	3.2	1.3	0.1	1.6	4.0	2.7	4.8
6.8	6.9	9.4	4.7	1.8	6.7	8.0	1.6
2.8	6.5	6.7	2.7	5.3	5.2	4.0	0.1
7.3	6.6	6.7	1.1	2.7	5.8	5.4	3.9
5.8	4.4	2.7	2.1	3.2	5.9	5.4	0.9
6.1	4.2	2.7	2.3	4.2	3.1	0.0	7.4
3.1	4.6	2.7	2.5				

5. Climatological records provide a rich source of data suitable for description by statistical methods. The data for this example (Table 1.18) are the number of January days in London, England, having rain (Days) and the average January temperature (Temp, in degrees Fahrenheit) for the years 1858 through 1939.

(a) Summarize these two variables.
(b) Draw a scatterplot to see whether the two variables are related.

6. Table 1.19 gives data on population (in thousands) and expenditures on criminal justice activities (in millions of dollars) for the 50 states and the District of Columbia as obtained from the 2005 Statistical Abstract of the United States.

Table 1.18 Rain Days and Temperatures, London Area, January

Year	Days	Temp	Year	Days	Temp	Year	Days	Temp
1858	6	40.5	1886	23	35.8	1914	12	39.7
1859	10	40.0	1887	13	37.9	1915	19	45.9
1860	21	34.0	1888	9	37.2	1916	14	35.5
1861	7	39.3	1889	10	43.6	1917	18	39.6
1862	19	42.2	1890	21	34.1	1918	18	37.8
1863	15	36.6	1891	14	36.6	1919	22	42.4
1864	8	36.5	1892	13	35.5	1920	21	46.1
1865	13	43.1	1893	17	38.5	1921	20	40.2
1866	23	34.6	1894	25	33.7	1922	20	41.5
1867	17	37.6	1895	16	40.5	1923	15	40.8
1868	19	41.4	1896	9	35.4	1924	18	41.7
1869	15	38.5	1897	21	43.7	1925	11	40.5
1870	17	33.4	1898	9	42.8	1926	18	41.0
1871	17	41.5	1899	19	40.4	1927	17	42.1
1872	22	42.3	1900	21	38.8	1928	21	34.8
1873	18	41.9	1901	12	42.0	1929	12	44.0
1874	17	43.6	1902	11	41.1	1930	17	39.0
1875	23	37.3	1903	17	39.5	1931	20	44.0
1876	11	42.9	1904	22	38.4	1932	13	37.4
1877	25	40.4	1905	8	42.4	1933	14	39.6
1878	15	31.8	1906	18	38.8	1934	18	40.7
1879	12	33.3	1907	8	36.8	1935	13	40.9
1880	5	31.7	1908	10	38.8	1936	21	41.9
1881	8	40.5	1909	13	40.0	1937	23	43.6
1882	7	41.4	1910	14	38.2	1938	21	41.7
1883	21	43.9	1911	12	40.2	1939	22	30.8
1884	16	36.6	1912	17	41.1			
1885	16	36.3	1913	17	38.4			

(a) Describe the distribution of states' criminal justice expenditures with whatever measures appear appropriate. Comment on the features and implications of these data.

(b) Compute the per capita expenditures (`EXPEND / POP`) for these data. Repeat part (a). Discuss any differences in the nature of the distribution you may have stated in part (a).

(c) Make a scatterplot of total and per capita expenditures on the vertical axis against population on the horizontal axis. Which of these plots is more useful?

Table 1.19 Criminal Justice Expenditures

STATE	POP	EXPEND	STATE	POP	EXPEND
AK	669	563	MT	936	441
AL	4540	1757	NC	8679	3706
AR	2772	1145	ND	636	224
AZ	5952	3589	NE	1754	726
CA	35990	29332	NH	1303	515
CO	4674	2519	NJ	8657	5982
CT	3486	1949	NM	1916	1156
DC	582	671	NV	2409	1730
DE	841	588	NY	19263	15449
FL	17736	11351	OH	11460	6028
GA	9108	4490	OK	3536	1485
HI	1268	678	OR	3630	2076
IA	2956	1126	PA	12367	6629
ID	1426	646	RI	1067	607
IL	12720	6500	SC	4255	1570
IN	6257	2390	SD	780	280
KS	2742	1215	TN	5989	2504
KY	4171	1706	TX	22844	10668
LA	4496	2491	UT	2505	1225
MA	6429	3465	VA	7558	3794
MD	5573	3578	VT	620	284
ME	1312	484	WA	6271	3292
MI	10108	5681	WI	5540	3092
MN	5114	2470	WV	1806	643
MO	5788	2425	WY	507	459
MS	2900	1050			

7. Make scatterplots for all pairwise combinations of the variables from the tree data (Table 1.7). Which pairs of variables have the strongest relationship? Is your conclusion consistent with prior knowledge?

8. The data set in Table 1.20 lists all cases of Down syndrome in Victoria, Australia, from 1942 through 1957, as well as the number of births classified by the age of the mother (Andrews and Herzberg, 1985).

(a) Construct a relative frequency histogram for total number of births by age group.

(b) Construct a relative frequency histogram for number of mothers of Down syndrome patients by age group.

(c) Compare the shape of the two histograms. Does the shape of the histogram for Down syndrome suggest that age alone accounts for number of Down syndrome patients born?

Table 1.20 Down Syndrome in Victoria, Australia[a]

Age Group, Years	Total Number of Births	Number of Mothers of Down Syndrome Patients
20 or less	35,555	15
20–24	207,931	128
25–29	253,450	208
30–34	170,970	194
35–39	86,046	297
40–44	24,498	240
45 or over	1,707	37

[a] *Reprinted with permission from Andrews and Herzberg (1985).*

(d) Construct a scatter diagram of total number of births versus number of mothers of Down syndrome. Does the scatter diagram support the conclusion in part (c)?

9. Table 1.21 shows the times in days from remission induction to relapse for 51 patients with acute nonlymphoblastic leukemia who were treated on a common protocol at university and private institutions in the Pacific Northwest. This is a portion of a larger study reported by Glucksberg *et al.* (1981).

Table 1.21 Ordered Remission Durations for 51 Patients with Acute Nonlymphoblastic Leukemia (in days)

24	46	57	57	64	65	82	89	90	90	111	117	128	143	148	152
166	171	186	191	197	209	223	230	247	249	254	258	264	269	270	273
284	294	304	304	332	341	393	395	487	510	516	518	518	534	608	642
697	955	1160													

Since data of this type are notoriously skewed, the distribution of the times can be examined using the following output from `PROC UNIVARIATE` in SAS as seen in Fig. 1.22.

(a) What is the relation between the mean and the median? What does this mean about the shape of the distribution? Do the stem and leaf plot and the box plot support this?

(b) Identify any outliers in this data set. Can you think of any reasons for these outliers? Can we just "throw them away"? Note that the mean time of remission is 292.39 days and the median time is 249.

(c) Approximately what percent of these patients were in remission for less than one year?

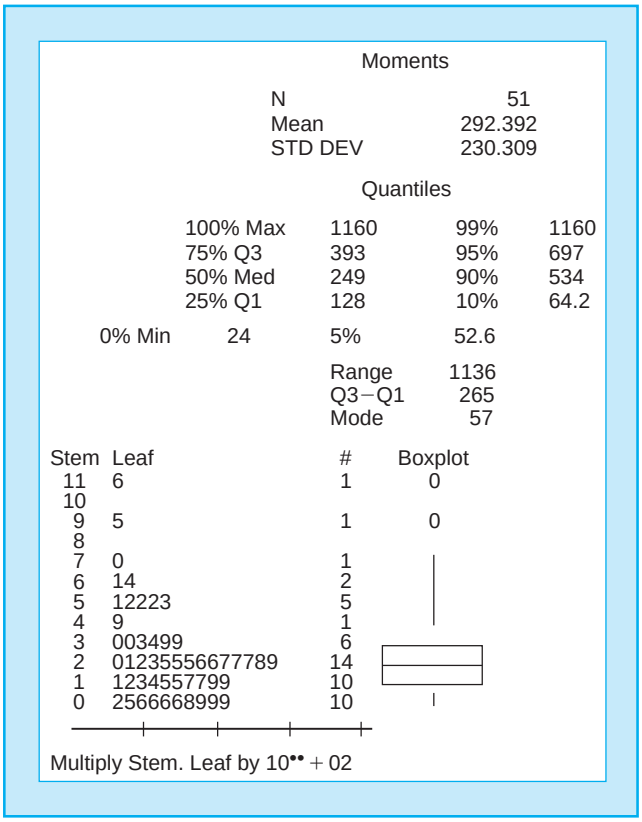


FIGURE 1.22
Summary Statistics for Remission Data.

10. The use of placement exams in elementary statistics courses has been a controversial topic in recent times. Some researchers think that the use of a placement exam can help determine whether a student will successfully complete a course (or program). A recent study in a large university resulted in the data listed in Table 1.22. The placement test administered was an in-house written general mathematics test. The course was Elementary Statistics. The students were told that the test would not affect their course grade. After the semester was over, students were classified according to their status. In Table 1.22 are the students' scores on the placement test (from 0 to 100), and the status of the student (coded as 0 = passed the course, 1 = failed the course, and 2 = dropped out before the semester was over) related?

(a) Construct a frequency histogram for Score. Describe the results.

(b) Construct a relative frequency histogram for Score for each value of Status. Describe the differences among these distributions. Are there some surprises?

11. The Energy Information Administration (http://www.eia.doe.gov) tabulates information on the average cost of electricity (in cents per kwh) for residential

Table 1.22 Placement Scores for Elementary Statistics

Student	Score	Status	Student	Score	Status	Student	Score	Status
1	90	2	36	85	0	71	97	2
2	65	2	37	99	1	72	90	0
3	30	1	38	45	0	73	30	0
4	55	0	39	90	0	74	1	0
5	1	0	40	10	1	75	1	0
6	5	1	41	56	0	76	70	0
7	95	0	42	55	2	77	90	0
8	99	0	43	50	0	78	70	0
9	40	0	44	1	1	79	75	0
10	95	0	45	45	0	80	75	2
11	1	0	46	50	0	81	70	2
12	55	0	47	85	2	82	85	0
13	85	0	48	95	2	83	45	0
14	95	0	49	15	0	84	50	0
15	15	2	50	35	0	85	55	0
16	95	0	51	85	0	86	15	0
17	15	0	52	85	0	87	55	0
18	65	0	53	50	0	88	20	1
19	55	0	54	10	1	89	1	1
20	75	0	55	60	0	90	75	0
21	15	0	56	45	1	91	45	2
22	35	2	57	90	0	92	70	0
23	90	0	58	1	1	93	70	0
24	10	0	59	80	2	94	45	0
25	10	1	60	45	0	95	90	0
26	20	0	61	90	0	96	65	2
27	25	0	62	45	0	97	75	2
28	15	1	63	20	0	98	70	0
29	40	0	64	35	1	99	65	0
30	15	0	65	40	2	100	55	0
31	50	0	66	40	0	101	55	0
32	80	0	67	60	0	102	40	0
33	50	1	68	15	0	103	56	0
34	50	2	69	45	0	104	85	0
35	97	0	70	45	0	105	80	0

customers in the United States. The data is shown in Table 1.23 for the years 1995 through 2008.

(a) Plot the cost versus the year.

(b) Comment on the trends, or general patterns, that are present in the data.

Table 1.23 Electricity Costs

Year	Cost	Year	Cost	Year	Cost
1995	8.40	2000	8.24	2005	9.45
1996	8.36	2001	8.58	2006	10.40
1997	8.43	2002	8.44	2007	10.65
1998	8.26	2003	8.72	2008	11.36
1999	8.16	2004	8.95		

12. A study of characteristics of successful salespersons in a certain industry included a questionnaire given to sales managers of companies in this industry. In this questionnaire the sales manager had to choose a trait that the manager thought was most important for salespersons to have. The results of 120 such responses are given in Table 1.24.

Table 1.24 Traits of Salespersons Considered Most Important by Sales Managers

Trait	Number of Responses
Reliability	44
Enthusiastic/energetic	30
Self-starter	20
Good grooming habits	18
Eloquent	6
Pushy	2

(a) Convert the number of responses to percents of total. What can be said about the first two traits?
(b) Draw a bar chart of the data.

13. A measure of the time a drug stays in the blood system is given by the half-life of the drug. This measure is dependent on the type of drug, the weight of the patient, and the dose administered. To study the half-life of aminoglyco sides in trauma patients, a pharmacy researcher recorded the data in Table 1.25 for patients in a critical care facility. The data consist of measurements of dosage per kilogram of weight of the patient, type of drug, either Amikacin or Gentamicin, and the half-life measured 1 hour after administration.
(a) Draw a scatter diagram of half-life versus dose per kilogram, indexed by drug type (use A's and G's). Does there appear to be a difference in the prescription of initial doses in types of drugs?
(b) Does there appear to be a relation between half-life and dosage? Explain.
(c) Find the mean and standard deviation for dose per kilogram for the two types of drugs. Does this seem to support the conclusion in part (a)?

Table 1.25 Half-Life of Aminoglycosides and Dosage by Drug Type

Patient	Drug	Half-Life	Dosage (mg drug/kg patient)	Patient	Drug	Half-Life	Dosage (mg drug/kg patient)
1	G	1.60	2.10	23	A	1.98	10.00
2	A	2.50	7.90	24	A	1.87	9.87
3	G	1.90	2.00	25	G	2.89	2.96
4	G	2.30	1.60	26	A	2.31	10.00
5	A	2.20	8.00	27	A	1.40	10.00
6	A	1.60	8.30	28	A	2.48	10.50
7	A	1.30	8.10	29	G	1.98	2.86
8	A	1.20	8.60	30	G	1.93	2.86
9	G	1.80	2.00	31	G	1.80	2.86
10	G	2.50	1.90	32	G	1.70	3.00
11	A	1.60	7.60	33	G	1.60	3.00
12	A	2.20	6.50	34	G	2.20	2.86
13	A	2.20	7.60	35	G	2.20	2.86
14	G	1.70	2.86	36	G	2.40	3.00
15	A	2.60	10.00	37	G	1.70	2.86
16	A	1.00	9.88	38	G	2.00	2.86
17	G	2.86	2.89	39	G	1.40	2.82
18	A	1.50	10.00	40	G	1.90	2.93
19	A	3.15	10.29	41	G	2.00	2.95
20	A	1.44	9.76	42	A	2.80	10.00
21	A	1.26	9.69	43	A	0.69	10.00
22	A	1.98	10.00				

Project

1. **Lake Data Set.** The Florida Lakes data set (Appendix C.1) shows total phosphorus levels for a sample of lakes in Northeast Florida. (Phosphorus is one of the nutrients that encourages algal growth in lake water.) For the lakes in the Hawthorne geologic formation (HAW), compare the winter total phosphorus levels (WTRTP) for those with clayey sand soil (CS) to those with quartzite sand soil (QS). Use both graphical and numerical statistical comparisons. Then transform the phosphorus levels by computing LOGWTRTP = ln(WTRTP). Are the transformed values easier to compare? Which variable is better described by means and standard deviations?