Chapter 1

Descriptive statistics

Chapter outline

Objective

Review the basic concepts of elementary statistics.

Sir Ronald Aylmer Fisher
(Source: http://www.stetson.edu/~efriedma/periodictable/jpg/Fisher.jpg).

Sir Ronald Fisher F.R.S. (1890−1962) was one of the leading scientists of the 20th century who laid the foundations for modern statistics. As a statistician working at the Rothamsted Agricultural Experiment Station, the oldest agricultural research institute in the United Kingdom, he also made major contributions to evolutionary biology and genetics. The concept of randomization and the analysis of variance procedures that he introduced are now used

Mathematical Statistics with Applications in R. https://doi.org/10.1016/B978-0-12-817815-7.00001-4

throughout the world. In 1922 he gave a new definition of statistics. Fisher identified three fundamental problems in statistics: (1) specification of the type of population that the data came from; (2) estimation; and (3) distribution. His book *Statistical Methods for Research Workers* (1925) was used as a handbook for the methods for the design and analysis of experiments. Fisher also published the books titled *The Design of Experiments* (1935) and *Statistical Tables* (1947). While at the Agricultural Experiment Station, he had conducted breeding experiments with mice, snails, and poultry, and the results he obtained led to theories about gene dominance and fitness that he published in *The Genetical Theory of Natural Selection* (1930).

1.1 Introduction

In today's society, decisions are made on the basis of data. Most scientific or industrial studies and experiments produce data, and the analysis of these data and drawing useful conclusions from them have become one of the central issues. Statistics is an integral part of the quantitative approach to knowledge. The field of statistics is concerned with the scientific study of collecting, organizing, analyzing, and drawing conclusions from data. Statistics benefits all of us because of its ability to predict the future based on data we have previously gathered. Statistical methods help us to transform data into information and knowledge. Statistical concepts enable us to solve problems in a diversity of contexts, add substance to decisions, and reduce guesswork. The discipline of statistics stemmed from the need to place knowledge management on a systematic evidence base. Earlier works on statistics dealt only with the collection, organization, and presentation of data in the form of tables and charts. In order to place statistical knowledge on a systematic evidence base, we require a study of the laws of probability. In mathematical statistics we create a probabilistic model and view the data as a set of random outcomes from that model. Advances in probability theory enable us to draw valid conclusions and to make reasonable decisions on the basis of data.

Statistical methods are used in almost every discipline, including agriculture, astronomy, biology, business, communications, economics, education, electronics, geology, health sciences, and many other fields of science and engineering, and can aid us in several ways. Modern applications of statistical techniques include statistical communication theory and signal processing, information theory, network security and denial-of-service problems, clinical trials, artificial and biological intelligence, quality control of manufactured items, software reliability, and survival analysis. The first of these is to assist us in designing experiments and surveys. We desire our experiment to yield adequate answers to the questions that prompted the experiment or survey. We would like the answers to have good precision without involving a lot of expenditure. Statistically designed experiments facilitate the development of robust products that are insensitive to changes in the environment and internal component variation. Another way that statistics assists us is in organizing, describing, summarizing, and displaying experimental data. This is termed *descriptive statistics*. Many of the descriptive statistics methods presented in this chapter are also part of the general area known as exploratory data analysis (EDA). A third use of statistics is in drawing inferences and making decisions based on data. For example, scientists may collect experimental data to prove or disprove an intuitive conjecture or hypothesis. Through the proper use of statistics, we can conclude whether the hypothesis is valid or not. In the process of solving a real-life problem using statistics, the following three basic steps may be identified. First, consistent with the objective of the problem, we identify the model using the appropriate statistical method. Then, we justify the applicability of the selected model to fulfill the aim of our problem. Last, we properly apply the related model to analyze the data and make the necessary decisions, which results in answering the question of our problem with minimum risk. Starting with Chapter 2, we will study the necessary background material to proceed with the development of statistical methods for solving real-world problems.

In this chapter we briefly review some of the basic concepts of descriptive statistics. Such concepts will give us a visual and descriptive presentation of the problem under investigation. Now, we proceed with some basic definitions and procedures.

1.1.1 Data collection

One of the first problems that a statistician faces is obtaining the data. The inferences that we make depend critically on the data that we collect and analyze. Data collection involves the following important steps.

General procedure for data collection

1. Define the objectives of the problem and proceed to develop the experiment or survey.
2. Define the variables or parameters of interest.
3. Define the procedures of data-collection and -measuring techniques. This includes sampling procedures, sample size, and data-measuring devices (questionnaires, telephone interviews, etc.).

EXAMPLE 1.1.1

We may be interested in estimating the average household income in a certain community. In this case, the parameter of interest is the average income of a typical household in the community. To acquire the data, we may send out a questionnaire or conduct a telephone interview. Once we have the data, we may first want to represent the data in graphical or tabular form to better understand its distributional behavior. Then we will use appropriate analytical techniques to estimate the parameter(s) of interest, in this case the average household income.

Very often a statistician is confined to the data that have already been collected, possibly even collected for other purposes. This makes it very difficult to determine the quality of the data. Planned collection of the data, using proper techniques, is much preferred.

1.2 Basic concepts

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. It also involves model building. Suppose we wish to study household incomes in a certain neighborhood. We may decide to randomly select, say, 50 families and examine their household incomes. As another example, suppose we wish to determine the diameter of a rod, and we take 10 measurements of the diameter. When we consider these two examples, we note that in the first case the population (the household incomes of all families in the neighborhood) really exists, whereas in the second, the population (set of all possible measurements of the diameter) is only conceptual. In either case we can visualize the totality of the population values, of which our sample data are only a small part. Thus, we define a population to be the set of all measurements or objects that are of interest and a sample to be a subset of that population. The population acts as the sampling frame from which a sample is selected. Now we introduce some basic notions commonly used in statistics.

Definition 1.2.1 *A* **population** *is the collection or set of all objects or measurements that are of interest to the collector.*

EXAMPLE 1.2.1

Suppose we wish to study the heights of all female students at a certain university. The population will be the set of the measured heights of all female students in the university. The population is not the set of all female students in the university.

In real-world problems it is usually not possible to obtain information on the entire population. The primary objective of statistics is to collect and study a subset of the population, called a sample, to acquire information on some specific characteristics of the population that are of interest.

Definition 1.2.2 *The* **sample** *is a subset of data selected from a population. The* **size** *of a sample is the number of elements in it.*

EXAMPLE 1.2.2

We wish to estimate the percentage of defective parts produced in a factory during a given week (5 days) by examining 20 parts produced per day. The parts will be examined each day at randomly chosen times. In this case "all parts produced during the week" is the population and the (100) selected parts for 5 days constitutes a sample.

Other common examples of sample and population are:

Political polls: The population will be all voters, whereas the sample will be the subset of voters we poll.
Laboratory experiment: The population will be all the data we could have collected if we were to repeat the experiment a large number of times (infinite number of times) under the same conditions, whereas the sample will be the data actually collected by the one experiment.
Quality control: The population will be the entire batch of items produced, say, by a machine or by a plant, whereas the sample will be the subset of items we tested.

Clinical studies: The population will be all the patients with the same disease, whereas the sample will be the subset of patients used in the study.
Finance: All common stock listed in stock exchanges such as the New York Stock Exchange, the American Stock Exchanges, and over-the-counter is the population. A collection of 20 randomly picked individual stocks from these exchanges will be a sample.

The methods consisting mainly of organizing, summarizing, and presenting data in the form of tables, graphs, and charts are called *descriptive statistics*. The methods of drawing inferences and making decisions about the population using the sample are called *inferential statistics*. Inferential statistics uses probability theory.

Definition 1.2.3 *A* **statistical inference** *is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample.*

For example, we may be interested in the average indoor radiation level in homes built on reclaimed phosphate mine lands (many of the homes in west-central Florida are built on such lands). In this case, we can collect indoor radiation levels for a random sample of homes selected from this area, and use the data to infer the average indoor radiation level for the entire region. In the Florida Keys, one of the concerns is that the coral reefs are declining because of the prevailing ecosystems. In order to test this, one can randomly select certain reef sites for study and, based on these data, infer whether there is a net increase or decrease in coral reefs in the region. Here the inferential problem could be finding an estimate, such as in the radiation problem, or making a decision, such as in the coral reef problem. We will see many other examples as we progress through the book.

1.2.1 Types of data

Data can be classified in several ways. We will give two different classifications, one based on whether the data are measured on a numerical scale or not, and the other on whether the data are collected in the same time period or collected at different time periods.

Definition 1.2.4 **Quantitative data** *are observations measured on a numerical scale. Nonnumerical data that can only be classified into one of the groups of categories are said to be* **qualitative** *or* **categorical data.**

EXAMPLE 1.2.3

Data on response to a particular therapy could be classified as no improvement, partial improvement, or complete improvement. These are qualitative data. The number of minority-owned businesses in Florida is quantitative data. The marital status of each person in a statistics class as married or not married is qualitative or categorical data. The number of car accidents in different U.S. cities is quantitative data. The blood group of each person in a community as O, A, B, AB is qualitative data.

Categorical data could be further classified as *nominal data* and *ordinal data*. Data characterized as nominal have data groups that do not have a specific order. An example of this could be state names, or names of the individuals, or courses by name. These do not need to be placed in any order. Data characterized as ordinal have groups that should be listed in a specific order. The order may be either increasing or decreasing. One example would be income levels. The data could have numeric values such as 1, 2, 3, or values such as high, medium, or low.

Definition 1.2.5 **Cross-sectional data** *are data collected on different elements or variables at the same point in time or for the same period of time.*

EXAMPLE 1.2.4

The data in Table 1.1 represent U.S. federal support for the mathematical sciences in 1996, in millions of dollars (source: *AMS Notices*). This is an example of cross-sectional data, as the data are collected in one time period, namely in 1996.

Definition 1.2.6 **Time series data** *are data collected on the same element or the same variable at different points in time or for different periods of time.*

TABLE 1.1 Federal Support for the Mathematical Sciences, 1996.

Federal agency	Amount
National Science Foundation	91.70
DMS	85.29
Other MPS	4.00
Department of Defense	77.30
AFOSR	16.70
ARO	15.00
DARPA	22.90
NSA	2.50
ONR	20.20
Department of Energy	16.00
University Support	5.50
National Laboratories	10.50
Total, all agencies	185.00

EXAMPLE 1.2.5

The data in Table 1.2 represent U.S. federal support for the mathematical sciences during the years 1995–97, in millions of dollars (source: *AMS Notices*). This is an example of time series data, because they have been collected at different time periods, 1995 through 1997.

TABLE 1.2 United States Federal Support for the Mathematical Sciences in Different Years.

Agency	1995	1996	1997
National Science Foundation	87.69	91.70	98.22
DMS	85.29	87.70	93.22
Other MPS	2.40	4.00	5.00
Department of Defense	77.40	77.30	67.80
AFOSR	17.40	16.70	17.10
ARO	15.00	15.00	13.00
DARPA	21.00	22.90	19.50
NSA	2.50	2.50	2.10
ONR	21.40	20.20	16.10
Department of Energy	15.70	16.00	16.00
University Support	6.20	5.50	5.00
National Laboratories	9.50	10.50	11.00
Total, all agencies	180.79	185.00	182.02

For an extensive collection of statistical terms and definitions, we can refer to many sources such as http://www.stats.gla.ac.uk/steps/glossary/index.html. We will give some other helpful Internet sources that may be useful for various aspects of statistics: http://www.amstat.org/(American Statistical Association), http://www.stat.ufl.edu (University of

Florida statistics department), http://www.statsoft.com/textbook/ (covers a wide range of topics, the emphasis is on techniques rather than concepts or mathematics), http://www.york.ac.uk/depts/maths/histstat/welcome.htm (some information about the history of statistics), http://www.isid.ac.in/ (Indian Statistical Institute), http://www.isi-web.org/30-statsoc/statsoc/282-nsslist (International Statistical Institute), http://www.rss.org.uk/ (Royal Statistical Society), and http://lib.stat.cmu.edu/ (an index of statistical software and routines). For energy-related statistics, refer to http://www.eia.doe.gov/. The Earth Observing System Data and Information System (https://earthdata.nasa.gov/about-eosdis) is one of the largest data sources for geological data. The Environmental Protection Agency (http://www.epa.gov/datafinder/) is another great source of data on environmental-related areas. If you want market data, YAHOO! Finance (http://finance.yahoo.com/) is a good source. There are various other useful sites that you could explore based on your particular needs.

Exercises 1.2

1.2.1. Give your own examples for qualitative and quantitative data. Also, give examples for cross-sectional and time series data.

1.2.2. Discuss how you will collect different types of data. What inferences do you want to derive from each of these types of data?

1.2.3. Refer to the data in Example 1.2.4. State a few questions that you can ask about the data. What inferences can you make by looking at these data?

1.2.4. Refer to the data in Example 1.2.5. Can you state a few questions that the data suggest? What inferences can you make by looking at these data?

1.3 Sampling schemes

In any statistical analysis, it is important that we clearly define the target population. The population should be defined in keeping with the objectives of the study. When the entire population is included in the study, it is called a *census* study because data are gathered on every member of the population. In general, it is usually not possible to obtain information on the entire population because the population is too large to attempt a survey of all of its members, or it may not be cost effective. A small but carefully chosen sample can be used to represent the population. A sample is obtained by collecting information from only some members of the population. A good sample must reflect all the characteristics (of importance) of the population. Samples can reflect the important characteristics of the populations from which they are drawn with differing degrees of precision. A sample that accurately reflects its population characteristics is called a *representative* sample. A sample that is not representative of the population characteristics is called a *biased* sample. The reliability or accuracy of conclusions drawn concerning a population depends on whether or not the sample is properly chosen so as to represent the population sufficiently well.

There are many sampling methods available. We mention a few commonly used simple sampling schemes. The choice between these sampling methods depends on (1) the nature of the problem or investigation, (2) the availability of good sampling frames (a list of all of the population members), (3) the budget or available financial resources, (4) the desired level of accuracy, and (5) the method by which data will be collected, such as questionnaires or interviews.

Definition 1.3.1 *A sample selected in such a way that every element of the population has an equal chance of being chosen is called a* **simple random sample**. *Equivalently, each possible sample of size* n *has the same chance of being selected as any other subset of sample of size* n.

EXAMPLE 1.3.1

For a state lottery, 52 identical ping-pong balls with a number from 1 to 52 painted on each ball are put in a clear plastic bin. A machine thoroughly mixes the balls and then six are selected. The six numbers on the chosen balls are the six lottery numbers that have been selected by a simple random sampling procedure.

Some advantages of simple random sampling

1. Selection of sampling observations at random ensures against possible investigator biases.
2. Analytic computations are relatively simple, and probabilistic bounds on errors can be computed in many cases.
3. It is frequently possible to estimate the sample size for a prescribed error level when designing the sampling procedure.

Simple random sampling may not be effective in all situations. For example, in a U.S. presidential election, it may be more appropriate to conduct sampling polls by state, rather than a nationwide random poll. It is quite possible for a candidate to get a majority of the popular vote nationwide and yet lose the election. We now describe a few other sampling methods that may be more appropriate in a given situation.

Definition 1.3.2 *A* **systematic sample** *is a sample in which every K^{th} element in the sampling frame is selected after a suitable random start for the first element. We list the population elements in some order (say alphabetical) and choose the desired sampling fraction.*

Steps for selecting a systematic sample

1. Number the elements of the population from 1 to N.
2. Decide on the sample size, say n, that we need.
3. Choose $K = N/n$.
4. Randomly select an integer between 1 and K.
5. Then take every K^{th} element.

EXAMPLE 1.3.2

If the population has 1000 elements arranged in some order and we decide to sample 10% (i.e., $N = 1000$ and $n = 100$), then $K = 1000/100 = 10$. Pick a number at random between 1 and $K = 10$ inclusive, say 3. Then select elements numbered 3, 13, 23, …, 993.

Systematic sampling is widely used because it is easy to implement. If the population elements are ordered, systematic sampling is a better sampling method. If the list of population elements is in random order to begin with, then the method is similar to simple random sampling. If, however, there is a correlation or association between successive elements, or if there is some periodic structure, then this sampling method may introduce biases. Systematic sampling is often used to select a specified number of records from a computer file.

Definition 1.3.3 A *sample obtained by stratifying (dividing into nonoverlapping groups) the sampling frame based on some factor or factors and then selecting some elements from each of the strata is called a* **stratified sample**. *Here, a population with* N *elements is divided into* s *subpopulations.* A *sample is drawn from each subpopulation independently. The size of each subpopulation and sample sizes in each subpopulation may vary.*

A stratified sample is a modification of simple random sampling and systematic sampling and is designed to obtain a more representative sample, but at the cost of a more complicated procedure. Compared to random sampling, stratified sampling reduces sampling error.

Steps for selecting a stratified sample

1. Decide on the relevant stratification factors (sex, age, race, income, etc.).
2. Divide the entire population into strata (subpopulations) based on the stratification criteria. Sizes of strata may vary.
3. Select the requisite number of units using simple random sampling or systematic sampling from each subpopulation. The requisite number may depend on the subpopulation sizes.

Examples of strata might be males and females, undergraduate students and graduate students, managers and nonmanagers, or populations of clients in different racial groups such as African Americans, Asians, whites, and Hispanics. Stratified sampling is often used when one or more of the strata in the population have a low incidence relative to the other strata. Through stratified random sampling adequate representation of all subgroups can be ensured.

EXAMPLE 1.3.3

In a population of 1000 children from an area school, there are 600 boys and 400 girls. We divide them into strata based on their parents' income as shown in Table 1.3.

TABLE 1.3 Classification of School Children.

	Boys	Girls
Poor	120	240
Middle class	150	100
Rich	330	60

This is stratified data.

EXAMPLE 1.3.4

Refer to Example 1.3.3. Suppose we decide to sample 100 children from the population of 1000 (that is, 10% of the population). We also choose to sample 10% from each of the categories. For example, we would choose 12 (10% of 120) poor boys; 6 (10% of 60 rich girls) and so forth. This yields Table 1.4. This particular sampling method is called a *proportional stratified sampling*.

TABLE 1.4 Proportional Stratification of School Children.

	Boys	Girls
Poor	12	24
Middle class	15	10
Rich	33	6

Some uses of stratified sampling

1. In addition to providing information about the whole population, this sampling scheme provides information about the subpopulations, the study of which may be of interest. For example, in a U.S. presidential election, opinion polls by state may be more important in deciding on the electoral college advantage than a national opinion poll.
2. Stratified sampling can be considerably more precise than a simple random sample, because the population is fairly homogeneous within each stratum but there is a sizable variation between the strata.

Definition 1.3.4 *In* **cluster sampling**, *the sampling unit contains groups of elements called clusters instead of individual elements of the population. A cluster is an intact group naturally available in the field. Unlike the stratified sample where the strata are created by the researcher based on stratification variables, the clusters naturally exist and are not formed by the researcher for data collection. Cluster sampling is also called* **area sampling.**

To obtain a cluster sample, first take a simple random sample of groups and then sample all elements within the selected clusters (groups). Cluster sampling is convenient to implement. When cost and time are important, cluster sampling may be used. However, because it is likely that units in a cluster will be relatively homogeneous, this method may be less precise than simple random sampling. The standard errors of estimates in cluster sampling are higher than other sampling designs.

EXAMPLE 1.3.5

Suppose we wish to select a sample of about 10% from all fifth-grade children of a county. We randomly select 10% of the elementary schools assumed to have approximately the same number of fifth-grade students and select all fifth-grade children from these schools. This is an example of cluster sampling, each cluster being an elementary school that was selected.

Definition 1.3.5 **Multiphase sampling** *involves collection of some information from the whole sample and additional information either at the same time or later from subsamples of the whole sample. The multiphase or multistage sampling is basically a combination of the techniques presented earlier.*

EXAMPLE 1.3.6

An investigator in a population census may ask basic questions such as sex, age, or marital status for the whole population, but only 10% of the population may be asked about their level of education or about how many years of mathematics and science education they had.

1.3.1 Errors in sample data

Irrespective of which sampling scheme is used, the sample observations are prone to various sources of error that may seriously affect the inferences about the population. Some sources of error can be controlled. However, others may be unavoidable because they are inherent in the nature of the sampling process. Consequently, it is necessary to understand the different types of errors for a proper interpretation and analysis of the sample data. The errors can be classified as *sampling errors* and *nonsampling errors*. Nonsampling errors occur in the collection, recording and processing of sample data. For example, such errors could occur as a result of bias in selection of elements of the sample, poorly designed survey questions, measurement and recording errors, incorrect responses, or no responses from individuals selected from the population. Sampling errors occur because the sample is not an exact representative of the population. Sampling error is due to the differences between the characteristics of the population and those of a sample from the population. For example, we are interested in the average test score in a large statistics class of size, say, 80. A sample of size 10 grades from this resulted in an average test score of 75. If the average test for the entire 80 students (the population) is 72, then the sampling error is $75-72=3$.

1.3.2 Sample size

In almost any sampling scheme designed by statisticians, one of the major issues is the determination of the sample size. In principle, this should depend on the variation in the population as well as on the population size, and on the required reliability of the results, that is, the amount of error that can be tolerated. For example, if we are taking a sample of school children from a neighborhood with a relatively homogeneous income level to study the effect of parents' affluence on the academic performance of the children, it is not necessary to have a large sample size. However, if the income level varies a great deal in the feeding area of the school, then we will need a larger sample size to achieve the same level of reliability. In practice, another influencing factor is the available resources such as money and time. In later chapters, we present some methods of determining sample size in statistical estimation problems.

The literature on sample survey methods is constantly changing, with new insights that demand dramatic revisions in the conventional thinking. We know that representative sampling methods are essential to permit confident generalizations of results to populations. However, there are many practical issues that can arise in real-life sampling methods. For example, in sampling related to social issues, whatever the sampling method we employ, a high response rate must be obtained. It has been observed that most telephone surveys have difficulty in achieving response rates higher than 60%, and most face-to-face surveys have difficulty in achieving response rates higher than 70%. Even a well-designed survey may stop short of the goal of a perfect response rate. This might induce bias in the conclusions based on the sample we obtained. A low response rate can be devastating to the reliability of a study. We can obtain series of publications on surveys, including guidelines on avoiding pitfalls from the American Statistical Association (www.amstat.org). In this book, we deal mainly with samples obtained using simple random sampling.

Exercises 1.3

1.3.1. Give your own examples for each of the sampling methods described in this section. Discuss the merits and limitations of each of these methods.

1.3.2. Using the information obtained from the publications of the American Statistical Association (www.amstat.org) or any other reference, write a short report on how to collect survey data, and what the potential sources of error are.

1.4 Graphical representation of data

The source of our statistical knowledge lies in the data. Once we obtain the sample data values, one way to become acquainted with them is through data visualization techniques such as to display them in tables or graphically. Charts and graphs are very important tools in statistics because they communicate information visually, and in a way, it is compression of knowledge. Remember, our interest in the data lies with the story it tells. These visual displays may reveal the patterns of behavior of the variables being studied. In this chapter, we will consider one-variable data. The most common graphical displays are the *frequency table*, *pie chart, bar graph*, *Pareto chart*, and *histogram*. For example, in the business world, graphical representations of data are used as statistical tools for everyday process management and improvements by decision makers (such as managers and frontline staff) to understand processes, problems, and solutions. The purpose of this section is to introduce several tabular and graphical procedures commonly used to summarize both qualitative and quantitative data. Tabular and graphical summaries of data can be found in reports, newspaper articles, websites, and research studies, among others.

Now we shall introduce some ways of graphically representing both qualitative and quantitative data. Bar graphs and Pareto charts are useful displays for qualitative data. With bar graphs, we can see how different things are distributed between separate categories. In practice, if there are too many categories, it may be helpful to compare only a limited number of categories, or combine categories with very short bars into say, others, and draw the bar graphs.

Definition 1.4.1 *A graph of bars whose heights represent the frequencies (or relative frequencies) of respective categories is called a* **bar graph.**

EXAMPLE 1.4.1

The data in Table 1.5 represent the percentages of price increases of some consumer goods and services for the period December 1990 to December 2000 in a certain city. Construct a bar chart for these data.

TABLE 1.5 Percentages of Price Increases of Some Consumer Goods and Services.

Medical care	83.3%
Electricity	22.1%
Residential rent	43.5%
Food	41.1%
Consumer price index	35.8%
Apparel and upkeep	21.2%

Solution

In the bar graph of Fig. 1.1*, we use the notations MC for medical care, El for electricity, RR for residential rent, Fd for food, CPI for consumer price index, and A & U for apparel and upkeep.*

Looking at Fig. 1.1, we can identify where the maximum and minimum responses are located, so that we can descriptively discuss the phenomenon whose behavior we want to understand.

For a graphical representation of the relative importance of different factors under study, one can use the *Pareto chart*. This is a bar graph with the height of the bars proportional to the contribution of each factor. The bars are displayed from

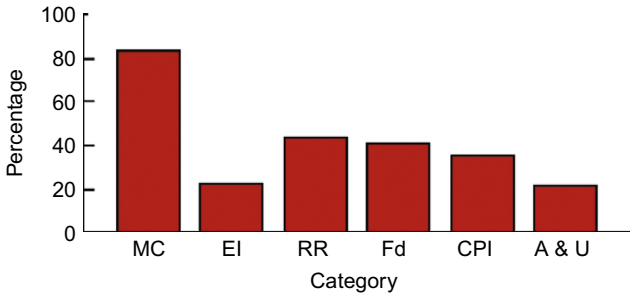


FIGURE 1.1 Percentage price increase of consumer goods.

the most numerous category to the least numerous category, as illustrated by the following example. A Pareto chart helps in separating significantly few factors that have larger influence from the trivial many.

EXAMPLE 1.4.2

For the data of Example 1.4.1, construct a Pareto chart.

Solution

First, rewrite the data in decreasing order. Then create a Pareto chart by displaying the bars from the most numerous category to the least numerous category.

Looking at Fig. 1.2, we can identify the relative importance of each category such as the maximum, the minimum, and the general behavior of the subject data.

Vilfredo Pareto (1848−1923), an Italian economist and sociologist, studied the distributions of wealth in different countries. He concluded that about 20% of people controlled about 80% of a society's wealth. This same distribution has been observed in other areas such as quality improvement: 80% of problems usually stem from 20% of the causes. This phenomenon has been termed the Pareto effect or 80/20 rule. Pareto charts are used to display the Pareto principle, arranging data so that the few vital factors that are causing most of the problems reveal themselves. Focusing improvement efforts on these few causes will have a larger impact and be more cost-effective than undirected efforts. Pareto charts are used in business decision-making as a problem-solving and statistical tool that ranks problem areas, or sources of variation, according to their contribution to cost or to total variation.

Definition 1.4.2 *A circle divided into sectors that represent the percentages of a population or a sample that belongs to different categories is called a* **pie chart.**

Pie charts are especially useful for presenting categorical data. The pie "slices" are drawn such that they have an area proportional to the frequency. The entire pie represents all the data, whereas each slice represents a different class or group within the whole. Thus, we can look at a pie chart and identify the various percentages of interest and how they compare among themselves. Most statistical software can create 3D charts. Such charts are attractive; however, they can make pieces at the front look larger than they really are. In general, a two-dimensional view of the pie is preferable.

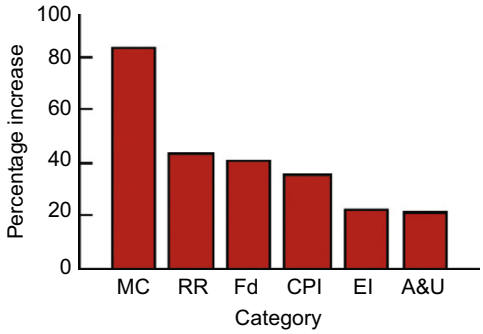


FIGURE 1.2 Pareto chart.

EXAMPLE 1.4.3

The combined percentages of carbon monoxide (CO) and ozone (O_3) emissions from different sources are listed in Table 1.6. Construct a pie chart.

TABLE 1.6 Combined Percentages of CO and O_3 Emissions.

Transportation (T)	Industrial process (I)	Fuel combustion (F)	Solid waste (S)	Miscellaneous (M)
63%	10%	14%	5%	8%

Solution

The pie chart is given in Fig. 1.3.

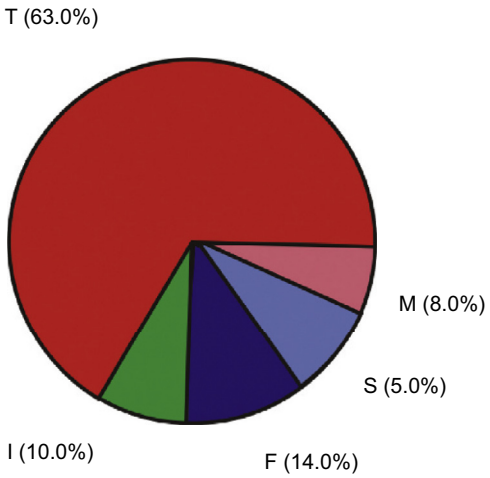


FIGURE 1.3 Pie chart for CO and O_3.

Definition 1.4.3 *A* **stem-and-leaf plot** *is a simple way of summarizing quantitative data and is well suited to computer applications. When data sets are relatively small, stem-and-leaf plots are particularly useful. In a stem-and-leaf plot, each data value is split into a "stem" and a "leaf." The "leaf" is usually the last digit of the number and the other digits to the left of the "leaf" form the "stem." Usually there is no need to sort the leaves, although computer packages typically do. For more details, we refer the student to elementary statistics books. We illustrate this technique with an example.*

EXAMPLE 1.4.4

Construct a stem-and-leaf plot for the 20 test scores given below.

78	74	82	66	94	71	64	88	55	80
91	74	82	75	96	78	84	79	71	83

Solution

At a glance, we see that the scores are distributed from the 50s through the 90s. We use the first digit of the score as the stem and the second digit as the leaf. The plot in Table 1.7 *is constructed with stems in the vertical position.*

TABLE 1.7 Stem-and-Leaf Display of 20 Exam Scores.

Stem	Leaves							
5	5							
6	6	4						
7	8	4	1	4	5	8	9	1
8	2	8	0	2	4	3		
9	4	1	6					

The stem-and-leaf plot condenses the data values into a useful display from which we can identify the shape and distribution of data such as the symmetry, where the maximum and minimum are located with respect to the frequencies, and whether they are bell shaped. This fact that the frequencies are bell shaped will be of paramount importance as we proceed to study inferential statistics. Also, note that the stem-and-leaf plot retains the entire data set and can be used only with quantitative data. Examples 1.8.1 and 1.8.6 explain how to obtain a stem-and-leaf plot using Minitab and SPSS, respectively. Refer to Section 1.8.3 for SAS commands to generate graphical representations of the data.

A *frequency table* is a table that divides a data set into a suitable number of categories (classes). Rather than retaining the entire set of data in a display, a frequency table essentially provides only a count of those observations that are associated with each class. Once the data are summarized in the form of a frequency table, a graphical representation can be given through bar graphs, pie charts, and histograms. Data presented in the form of a frequency table are called *grouped data*. A frequency table is created by choosing a specific number of classes in which the data will be placed. Generally, the classes will be intervals of equal length. The center of each class is called a *class mark*. The end points of each class interval are called class boundaries. Usually, there are two ways of choosing class boundaries. One way is to choose nonoverlapping class boundaries so that none of the data points will simultaneously fall in two classes. Another way is that for each class, except the last, the upper boundary is equal to the lower boundary of the subsequent class. When forming a frequency table this way, one or more data values may fall on a class boundary. One way to handle such a problem is to arbitrarily assign it one of the classes or to flip a coin to determine the class into which to place the observation at hand.

Definition 1.4.4 *Let* f_i *denote the frequency of the class* i *and let* n *be sum of all frequencies. Then the* **relative frequency** *for the class* i *is defined as the ratio* f_i/n. The **cumulative relative frequency** *for the class* i *is defined by* $\sum_{k=1}^{i} f_k/n$.

The following example illustrates the foregoing discussion.

EXAMPLE 1.4.5

The following data give the lifetime of 30 incandescent light bulbs (rounded to the nearest hour) of a particular type.

872	931	1146	1079	915	879	863	1112	979	1120
1150	987	958	1149	1057	1082	1053	1048	1118	1088
868	996	1102	1130	1002	990	1052	1116	1119	1028

Construct a frequency, relative frequency, and cumulative relative frequency table.

Solution

Note that there are n = *30 observations and that the largest observation is 1150 and the smallest one is 865 with a range of 285. We will choose six classes each with a length of 50.*

Class	Frequency f_i	Relative frequency $\frac{f_i}{\sum f_i}$	Cumulative relative frequency $\sum_{k=1}^{i} \frac{f_k}{n}$
50–900	4	4/30	4/30
900–950	2	2/30	6/30
950–1000	5	5/30	11/30
1000–1050	3	3/30	14/30
1050–1100	6	6/30	20/30
1100–1150	10	10/30	30/30

When data are quantitative in nature and the number of observations is relatively large, and there are no natural separate categories or classes, we can use a histogram to simplify and organize the data. Since the classes are listed in order, histograms are great to identify range and skew of quantitative data.

Definition 1.4.5 *A* **histogram** *is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. In a histogram, the bars are drawn adjacent to each other without any gaps.*

Histograms can be used only for quantitative data. A histogram compresses a data set into a compact picture that shows the location of the mean and modes of the data and the variation in the data, especially the range. It identifies patterns in the data. This is a good aggregate graph of one variable. In order to obtain the variability in the data, it is always a good practice to start with a histogram of the data. The following steps can be used as a general guideline to construct a frequency table and produce a histogram.

Guidelines for the construction of a frequency table and histogram

1. Determine the maximum and minimum values of the observations. The range, $R = \textit{maximum value} - \textit{minimum value}$.
2. Select from 5 to 20 classes that in general are nonoverlapping intervals of equal length, so as to cover the entire range of the data. The goal is to use enough classes to show the variation in the data, but not so many that there are only a few data points in many of the classes. The class width should be slightly larger than the ratio

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}.$$

3. The first interval should begin a little below the minimum value, and the last interval should end a little above the maximum value. The intervals are called class intervals and the boundaries are called class boundaries. The class limits are the smallest and the largest data values in the class. The class mark is the midpoint of a class.
4. None of the data values should fall on the boundaries of the classes.
5. Construct a table (frequency table) that lists the class intervals, a tabulation of the number of measurements in each class (tally), the frequency f_i of each class, and, if needed, a column with relative frequency, f_i/n, where n is the total number of observations.
6. Draw bars over each interval with heights being the frequencies (or relative frequencies).

Let us illustrate implementing these steps in the development of a histogram for the data given in the following example.

EXAMPLE 1.4.6

The following data refer to a certain type of chemical impurity measured in parts per million in 25 drinking-water samples randomly collected from different areas of a county.

11	19	24	30	12	20	25	29	15	21
24	31	16	23	25	26	32	17	22	26
35	18	24	18	27					

(a) Make a frequency table displaying class intervals, frequencies, relative frequencies, and percentages.
(b) Construct a frequency histogram.

Solution

(a) *We will use five classes. The maximum and minimum values in the data set are 35 and 11. Hence the class width is* $(35-11)/5 = 4.8 \simeq 5$. *Hence, we shall take the class width to be 5. The lower boundary of the first class interval will be chosen to be 10.5. With five classes, each of width 5, the upper boundary of the fifth class becomes 35.5. We can now construct the frequency table for the data.*

Class	Class interval	f_i = frequency	Relative frequency	Percentage
1	10.5–15.5	3	3/25 = 0.12	12
2	15.5–20.5	6	6/25 = 0.24	24
3	20.5–25.5	8	8/25 = 0.32	32
4	25.5–30.5	5	5/25 = 0.20	20
5	30.5–35.5	3	3/25 = 0.12	12

(b) *We can generate a histogram as in Fig. 1.4.*

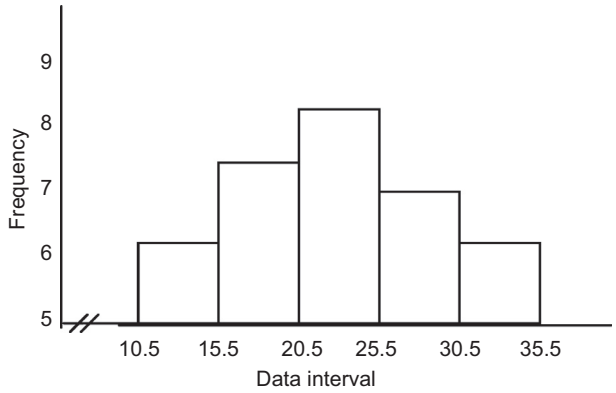


FIGURE 1.4 Frequency histogram of impurity data.

From the histogram we should be able to identify the center (i.e., the location) of the data, spread of the data, skewness of the data, presence of outliers, presence of multiple modes in the data, and whether the data can be capped with a bell-shaped curve. These properties provide indications of the proper distributional model for the data. Examples 1.8.2 and 1.8.7 explain how to obtain histograms using Minitab and SPSS, respectively.

Exercises 1.4

1.4.1. According to the recent U.S. Federal Highway Administration *Highway Statistics*, the percentages of freeways and expressways in various road mileage-related highway pavement conditions are as follows:
Poor 10%, Mediocre 32%, Fair 22%, Good 21%, and Very good 15%.
(a) Construct a bar graph.
(b) Construct a pie chart.

1.4.2. More than 75% of all species that have been described by biologists are insects. Of the approximately two million known species, only about 30,000 are aquatic in any life stage. The data in Table 1.8 give the proportion of total species by insect order that can survive exposure to salt (source: http://entomology.unl.edu/).
(a) Construct a bar graph.
(b) Construct a Pareto chart.
(c) Construct a pie chart.

1.4.3. The data in Table 1.9 are presented to illustrate the role of renewable energy consumption in the U.S. energy supply in 2007 (source: http://www.eia.doe.gov/fuelrenewable.html). Renewable energy consists of biomass, geothermal energy, hydroelectric energy, solar energy, and wind energy.
(a) Construct a bar graph.
(b) Construct a Pareto chart.
(c) Construct a pie chart.

1.4.4. A litter is a group of babies born from the same mother at the same time. Table 1.10 gives some examples of different mammals and their average litter size (source: http://www.saburchill.com/chapters/chap0032.html).
(a) Construct a bar graph.
(b) Construct a Pareto chart.

1.4.5. The following data give the letter grades of 20 students enrolled in a statistics course.

A C	B D	F B	A A	C B	C A	D F	A B	B C	F A

(a) Construct a bar graph.
(b) Construct a pie chart.

1.4.6. According to the U.S. Bureau of Labor Statistics (BLS), the median weekly earnings of fulltime wage and salary workers by age for the third quarter of 1998 is given in Table 1.11.

Construct a pie chart and bar graph for these data and interpret. Also, construct a Pareto chart.

1.4.7. The data in Table 1.12 are a breakdown of 18,930 workers in a town according to the type of work. Construct a pie chart and bar graph for these data and interpret.

1.4.8. The data in Table 1.13 represent the number (in millions) of adults and children living with HIV/AIDS by the end of 2000 according to their region of the world (source: http://w3.whosea.org/hivaids/factsheet.htm).

Construct a bar graph for these data. Also, construct a Pareto chart and interpret.

TABLE 1.8 Percentage of Species by Insect Order.

Species	Percentage	Species	Percentage
Coleoptera	26%	Odonata	3%
Diptera	35%	Thysanoptera	3%
Hemiptera	15%	Lepidoptera	1%
Orthoptera	6%	Other	6%
Collembola	5%		

TABLE 1.9 Renewable Energy Consumption.

Source	Percentage
Coal	22%
Natural gas	23%
Nuclear electric power	8%
Petroleum	40%
Renewable energy	7%

TABLE 1.10 Litter Size of Mammals.

Species	Litter size
Bat	1
Dolphin	1
Chimpanzee	1
Lion	3
Hedgehog	5
Red fox	6
Rabbit	6
Black rat	11

TABLE 1.11 Weakly Wages & Salary Distribution by Age.

16–19 years	$260
20–24 years	$334
25–34 years	$498
35–44 years	$600
45–54 years	$628
55–64 years	$605
65 years and over	$393

TABLE 1.12 Distribution of Workers by Type of Work.

Mining	58
Construction	1161
Manufacturing	2188
Transportation and public utilities	821
Wholesale trade	657
Retail trade	7377
Finance, insurance, and real estate	890
Services	5778
Total	18,930

TABLE 1.13 Number of People Living With HIV/AIDS.

Region of the world	Adults and children living with HIV/AIDS (in millions)
Sub-Saharan Africa	25.30
North Africa and Middle East	0.40
South and Southeast Asia	5.80
East Asia and Pacific	0.64
Latin America	1.40
Caribbean	0.39
Eastern Europe and Central Asia	0.70
Western Europe	0.54
North America	0.92
Australia and New Zealand	0.15

1.4.9. The data in Table 1.14 give the life expectancy at birth, in years, from 1900 through 2000 (source: National Center for Health Statistics). Construct a bar graph for these data.

1.4.10. Dolphins are usually identified by the shape and pattern of notches and nicks on their dorsal fin. Individual dolphins are cataloged by classifying the fin based on the location(s) of distinguishing marks. When a dolphin is sighted its picture can then be compared to the catalog of dolphins in the area, and if a match is found, the dolphin can be recorded as resighted. These methods of mark-resight are for developing databases regarding the life history of individual dolphins. From these databases we can calculate the levels of association between dolphins,

TABLE 1.14 Life Expectancy at Birth.

Year	Life expectancy
1900	47.3
1960	69.7
1980	73.7
1990	75.4
2000	77.0

population estimates, and general life history parameters such as birth and survival rates. The data in Table 1.15 represent frequently resighted individuals (as of January 2000) at a particular location (source: http://www.eckerd.edu/dolphinproject/biologypr.html).

Construct a bar graph for these data.

TABLE 1.15 Number of Dolphin Resights by Type.

Hammer (adult female)	59
Mid Button Flag (adult female)	41
Luseal (adult female)	31
84 Lookalike (adult female)	20

1.4.11. The data in Table 1.16 give death rates (per 100,000 population) for 10 leading causes in 1998 (source: National Center for Health Statistics, U.S. Department of Health and Human Services).

(a) Construct a bar graph.

(b) Construct a Pareto chart.

TABLE 1.16 Death Rate by Cause.

Cause	Death rate
Accidents and adverse effects	34.5
Chronic liver disease and cirrhosis	9.7
Chronic obstructive lung diseases and allied conditions	42.3
Cancer	199.4
Diabetes mellitus	23.9
Heart disease	268.0
Kidney disease	9.7
Pneumonia and influenza	35.1
Stroke	58.5
Suicide	10.8

1.4.12. In a fiscal year, a city collected $32.3 million in revenues. City spending for that year is expected to be nearly the same, with no tax increase projected.

Expenditure: Reserves 0.7%, capital outlay 29.7%, operating expenses 28.9%, debt service 3.2%, transfers 5.1%, personal services 32.4%.

Revenues: Property taxes 10.2%, utility and franchise taxes 11.3%, licenses and permits 1%, intergovernmental revenue 10.1%, charges for services 28.2%, fines and forfeits 0.5%, interest and miscellaneous 2.7%, transfers and cash carryovers 36%.

(a) Construct bar graphs for expenditures and revenues, and interpret.

(b) Construct pie charts for expenditures and revenues, and interpret.

1.4.13. Construct a histogram for the 24 examination scores given below:

78	74	82	66	94	71	64	88	55	80	73	86
91	74	82	75	96	78	84	79	71	83	78	79

1.4.14. The following table gives radon concentrations in pCi/liter (picocurie per liter) obtained from 40 houses in a certain area.

2.9	0.6	13.5	17.1	2.8	3.8	16.0	2.1	6.4	17.2
7.9	0.5	13.7	11.5	2.9	3.6	6.1	8.8	2.2	9.4
15.9	8.8	9.8	11.5	12.3	3.7	8.9	13.0	7.9	11.7
6.2	6.9	12.8	13.7	2.7	3.5	8.3	15.9	5.1	6.0

(a) Construct a stem-and-leaf display.
(b) Construct a frequency histogram and interpret.
(c) Construct a pie chart and interpret.

1.4.15. The following data give the mean of SAT mathematics scores by state for 1999 for a randomly selected 20 states (source: *The World Almanac and Book of Facts, 2000*).

558	503	565	572	546	517	542	605	493	499
568	553	510	525	595	502	526	475	506	568

(a) Construct a stem-and-leaf display and interpret.
(b) Construct a frequency histogram and interpret.
(c) Construct a pie chart and interpret.

1.4.16. A sample of 25 measurements is given here:

9	28	14	29	21	27	15	23	23	10
31	23	16	26	22	17	19	24	21	20
26	20	16	14	21					

(a) Make a frequency table displaying class intervals, frequencies, relative frequencies, and percentages.
(b) Construct a frequency histogram and interpret.

1.4.17. We may be interested in changing demographics of the U.S. population. The following table gives the demographics in 2010 (Overview of Race and Hispanic Origin: 2010, http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf). The Table 1.17 gives a pretty good summary understanding.

TABLE 1.17 US Population Demographics.

Race/Ethnicity	Number	% of population
White or European American	223,553,265	24.14
Black or African American	38,929,319	4.20
Asian American	14,674,252	1.58
American Indian or Alaska Native	2,932,248	0.32
Native Hawaiian or other Pacific Islander	540,013	0.06
Some other race	19,107,368	2.06
Two or more races	9,009,073	0.97
Not Hispanic nor Latino	258,267,944	27.88
Non-Hispanic white or European American	196,817,552	21.25
Non-Hispanic black or African American	37,685,848	4.07
Non-Hispanic Asian	14,465,124	1.56
Non-Hispanic American Indian or Alaska Native	2,247,098	0.24
Non-Hispanic Native Hawaiian or other Pacific Islander	481,576	0.05
Non-Hispanic some other race	604,265	0.07
Non-Hispanic two or more races	5,966,481	0.64
Hispanic or Latino	50,477,594	5.45
White or European American Hispanic	26,735,713	2.89

Continued

TABLE 1.17 US Population Demographics.—cont'd

Race/Ethnicity	Number	% of population
Black or African American Hispanic	1,243,471	0.13
American Indian or Alaska Native Hispanic	685,150	0.07
Asian Hispanic	209,128	0.02
Native Hawaiian or other Pacific Islander Hispanic	58,437	0.01
Some other race Hispanic	18,503,103	2
Two or more races Hispanic	3,042,592	0.33
Total	926,236,614	100%

Draw a pie chart.

1.5 Numerical description of data

In the previous section we looked at some graphical and tabular techniques for describing a data set. We shall now consider some numerical characteristics of a set of measurements. Suppose that we have a sample with values $x_1, x_2, \ldots, x_n$. There are many characteristics associated with this data set, for example, the central tendency and variability. A measure of the central tendency is given by the sample mean, median, or mode, and the measure of dispersion or variability is usually given by the sample variance or sample standard deviation or interquartile range.

Definition 1.5.1 *Let $x_1, x_2, \ldots, x_n$ be a set of sample values. Then the* **sample mean** *(or* **empirical mean***) $\bar{x}$ is defined by*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The **sample variance** *is defined by*

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The **sample standard deviation** *is*

$$s = \sqrt{s^2}$$

The sample variance s^2 and the sample standard deviation s both are measures of the variability or "scatteredness" of data values around the sample mean $\bar{x}$. The larger the variance, the greater is the spread. We note that s^2 and s are both nonnegative. One question we may ask is "why not just take the sum of the differences as a measure of variation?" The answer lies in the following result that shows that if we add up all deviations about the sample mean, we always get a zero value.

Theorem 1.5.1 *For a given set of measurements $x_1, x_2, \ldots, x_n$, let $\bar{x}$ be the sample mean. Then*

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Proof. Since $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, we have $\sum_{i=1}^{n} x_i = n\bar{x}$. Now

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}$$

$$= n\bar{x} - n\bar{x} = 0.$$

Thus, although there may be a large variation in the data values, $\sum_{i=1}^{n}(x_i - \bar{x})$ as a measure of spread would always be zero, implying no variability. So, it is not useful as a measure of variability.

Sometimes we can simplify the calculation of the sample variance s^2 by using the following computational formula:

$$s^2 = \frac{\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}{(n-1)}.$$

If the data set has a large variation with some extreme values (called outliers), the mean may not be a very good measure of the center. For example, average salary may not be a good indicator of the financial well-being of the employees of a company if there is a huge difference in pay between support personnel and management personnel. In that case, one could use the median as a measure of the center, roughly 50% of data fall below and 50% above. The median is less sensitive to extreme data values.

Definition 1.5.2 *For a data set, the* **median** *is the middle number of the ordered data set. If the data set has an even number of elements, then the median is the average of the middle two numbers. The* **lower quartile** *is the middle number of the half of the data below the median, and the* **upper quartile** *is the middle number of the half of the data above the median. We will denote*

$$Q_1 = \textit{lower quartile}$$

$$Q_2 = M = \textit{middle quartile (median)}$$

$$Q_3 = \textit{upper quartile}.$$

The difference between the quartiles is called the **interquartile range** *(IQR).*

$$IQR = Q_3 - Q_1.$$

A possible outlier (mild outlier) will be any data point that lies below

$$Q_1 - 1.5(IQR) \textit{ or above } Q_3 + 1.5(IQR).$$

Thus, about 25% of the data lie below Q_1, and about 75% of the data lie below Q_3. Note that the IQR is unaffected by the positions of those observations in the smallest 25% or the largest 25% of the data.

Mode is another commonly used measure of central tendency. A mode indicates where the data tend to concentrate most.

Definition 1.5.3 **Mode** *is the most frequently occurring member of the data set. If all the data values are different, then by definition, the data set has no mode.*

EXAMPLE 1.5.1

The following data give the time in months from hire to promotion to manager for a random sample of 25 software engineers from all software engineers employed by a large telecommunications firm.

5	7	229	453	12	14	18	14	14	483
22	21	25	23	24	34	37	34	49	64
47	67	69	192	125					

Calculate the mean, median, mode, variance, and standard deviation for this sample.

Solution

The sample mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 83.28 \textit{ months}.$$

To obtain the median, first arrange the data in ascending order:

5	7	12	14	14	14	18	21	22	23
24	25	34	34	37	47	49	64	67	69
125	192	229	453	483					

Now the median is the thirteenth number, which is 34 months.

Since 14 occurs most often (thrice), the mode is 14 months.

The sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$= \frac{1}{24}\left[(5 - 83.28)^2 + \cdots + (125 - 83.28)^2\right]$$

$$= 16,478.$$

and the sample standard deviation is, $s = \sqrt{s^2} = 128.36$ *months. Thus, we have sample mean* $\bar{x} = 83.28$ *months, median* $= 34$ *months, and mode* $= 14$ *months. Note that the mean is much different from the other two measures of the center because of a few large data values. Also, the sample variance* $s^2 = 16{,}478$ *months, and the sample standard deviation* $s = 128.36$ *months.*

EXAMPLE 1.5.2

For the data of Example 1.5.1, find lower and upper quartiles, median, and interquartile range (IQR). Check for any outliers.

Solution

Arrange the data in an ascending order.

5	7	12	14	14	14	18	21	22	23
24	25	34	34	37	47	49	64	67	69
125	192	229	453	483					

Then the median M *is the middle (13th) data value,* $M = Q_2 = 34$. *The lower quartile is the middle number below the median,* $Q_1 = [(14 + 18)/2] = 16$. *The upper quartile,* $Q_3 = [(67 + 69)/2] = 68$.

The interquartile range (IQR) $= Q_3 - Q_1 = 68 - 16 = 52$.

To test for outliers, compute

$$Q_1 - 1.5(IQR) = 16 - 1.5(52) = -62$$

and

$$Q_3 + 1.5(IQR) = 68 + 1.5(52) = 146.$$

Then all the data that fall above 146 are possible outliers. None is below −62. Therefore, the outliers are 192, 229, 453, and 483.

We have remarked earlier that the mean as a measure of central location is greatly affected by the extreme values or outliers. A robust measure of central location (a measure that is relatively unaffected by outliers) is the *trimmed mean*. For $0 \leq \alpha \leq 1$, a $100\alpha\%$ trimmed mean is found as follows: Order the data, and then discard the lowest $100\alpha\%$ and the highest $100\alpha\%$ of the data values. Find the mean of the rest of the data values. We denote the $100\alpha\%$ trimmed mean by $\bar{x}_\alpha$. We illustrate the trimmed mean concept in the following example.

EXAMPLE 1.5.3

For the data set representing the number of children in a random sample of 10 families in a neighborhood, find the 10% trimmed mean ($\alpha = 0.1$).

$$1 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 9 \quad 1 \quad 6 \quad 2.$$

Solution

Arrange the data in ascending order.

$$1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 6 \quad 9.$$

The data set has 10 elements. Discarding the lowest 10% (10% of 10 is 1) and discarding the highest 10% of the data values, we obtain the trimmed data set as

$$1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 6.$$

The 10% trimmed mean is

$$\bar{x}_{0.1} = \frac{1 + 2 + 2 + 2 + 2 + 3 + 3 + 6}{8} = 2.6.$$

Note that the mean for the data in the previous example without removing any observations is 3.1, which is different from the trimmed mean.

Although standard deviation is a more popular method, there are other measures of dispersion such as average deviation or interquartile range. We have already seen the definition of interquartile range. The average deviation for a sample $x_1, \ldots, x_n$ is defined by

$$\text{Average deviation} = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}.$$

Calculation of average deviation is simple and straightforward.

1.5.1 Numerical measures for grouped data

When we encounter situations where the data are grouped in the form of a frequency table (see Section 1.4), we no longer have individual data values. Hence, we cannot use the formulas in Definition 1.7.1. The following formulas will give approximate values for $\bar{x}$ and s^2. Let the grouped data have l classes, with m_i being the midpoint and f_i being the frequency of class i, $i = 1, 2, \ldots, l$. Let $n = \sum_{i=1}^{l} f_i$.

Definition 1.5.4 *The* **mean** *for a sample of size* n,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{l} f_i m_i,$$

where m_i *is the midpoint of the class* i *and* f_i *is the frequency of the class* i.

Similarly, the **sample variance**,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i(m_i - \bar{x})^2 = \frac{\sum m_i^2 f_i - \frac{\left(\sum_i f_i m_i\right)^2}{n}}{n-1}.$$

The following example illustrates how we calculate the sample mean for a grouped data.

EXAMPLE 1.5.4

The grouped data in Table 1.18 represent the number of children from birth through the end of the teenage years in a large apartment complex. Find the mean, variance, and standard deviation for these data.

Here we use the usual convention of until the child attains the next age, the age will be the previous year, for instance until a child is 4 years old, we will say the child is 3 years old.

Solution

Note that even though the classes are given as disjoint, in actuality these are adjacent age intervals, like [0, 4), [4, 8), etc. When we take the class midpoint, we have to take this into account. For simplicity of calculation we create Table 1.19.

The sample mean is

$$\bar{x} = \frac{1}{n}\sum_i f_i m_i = \frac{540}{50} = 10.80.$$

The sample variance is

$$s^2 = \frac{\sum m_i^2 f_i - \frac{\left(\sum_i f_i m_i\right)^2}{n}}{n-1} = \frac{7016 - \frac{(540)^2}{50}}{49} = 24.1632650 \approx 24.16.$$

The sample standard deviation is $s = \sqrt{s^2} = 4.9156144 \approx 4.92$.

TABLE 1.18 Number of Children and Their Age Group.

Class	0–3	4–7	8–11	12–15	16–19
Frequency	7	4	19	12	8

TABLE 1.19 Summary Statistics for Number of Children.

Class	*Interval*	f_i	m_i	$m_i f_i$	$m_i^2 f_i$
0–3	[0, 4)	7	2	14	28
4–7	[4, 8)	4	6	24	144
8–11	[8, 12)	19	10	190	1900
12–15	[12, 16)	12	14	168	2352
16–19	[16, 20)	8	18	144	2592
	$n = 50$		$\sum m_i f_i = 540$	$\sum m_i^2 f_i = 7016$	

Using the following calculations, we can also find the *median for grouped data.* We only know that the median occurs in a particular class interval, but we do not know the exact location of the median. We will assume that the measures are spread evenly throughout this interval. Let

$L =$ lower class limit of the interval that contains the median.
$n =$ total frequency.
$F_b =$ cumulative frequencies for all classes before the median class.
$f_m =$ frequency of the class interval containing the median.
$w =$ interval width of the interval that contains the median.
Then the median for the grouped data is given by

$$M = L + \frac{w}{f_m}(0.5n - F_b).$$

We proceed to illustrate with an example.

EXAMPLE 1.5.5

For the data in Example 1.5.4, find the median.

Solution
First, we develop Table 1.20.

TABLE 1.20 Frequency Distribution for Number of Children.

Class	f_i	Cumulative f_i	Cumulative f_i/n
0–3	7	7	0.14
4–7	4	11	0.22
8–11	19	30	0.6
12–15	12	42	0.84
16–19	8	50	1.00

The first interval for which the cumulative relative frequency exceeds 0.5 is the interval that contains the median. Hence, the interval 8 to 11 contains the median. Therefore, $L = 8$, $f_m = 19$, $n = 50$, $w = 3$, and $F_b = 11$. Then, the median is

$$M = L + \frac{w}{f_m}(0.5n - F_b) = 8 + \frac{3}{19}((0.5)(50) - 11) = 10.211.$$

It is important to note that all the numerical measures we calculate for grouped data are only approximations to the actual values of the ungrouped data if they are available.

One of the uses of the sample standard deviation will be clear from the following result, which is based on the data following a bell-shaped curve. Such an indication can be obtained from the histogram or stem-and-leaf display.

Empirical rule

When the histogram of a data set is "bell-shaped" or "mound-shaped," and symmetric, the *empirical rule* states:

1. Approximately 68% of the data are in the interval $(\overline{x} - s, \overline{x} + s)$.
2. Approximately 95% of the data are in the interval $(\overline{x} - 2s, \overline{x} + 2s)$.
3. Approximately 99.7% of the data are in the interval $(\overline{x} - 3s, \overline{x} + 3s)$.

The bell-shaped curve is called a normal curve and is discussed later in Chapter 3. A typical symmetric bell-shaped curve is given by Fig. 1.5.

1.5.2 Box plots

The sample mean or the sample standard deviation focuses on a single aspect of the data set, whereas histograms and stem-and-leaf displays express rather general ideas about the data. A pictorial summary called a *box plot* (also called *box-and-whisker plots*) can be used to describe several prominent features of a data set such as the center, the spread, the extent, and nature of any departure from symmetry, and identification of outliers. Box plots are a simple diagrammatic representation of the five number summary: minimum, lower quartile, median, upper quartile, maximum. Example 1.8.4 illustrates the method of obtaining box plots using Minitab.

Procedure to construct a box plot

1. Draw a vertical measurement axis and mark Q_1, Q_2 (median), and Q_3 on this axis as shown in Fig. 1.6, below. Let $IQR = Q_1 - Q_3$.
2. Construct a rectangular box whose bottom edge lies at the lower quartile, Q_1, and whose upper edge lies at the upper quartile, Q_3.
3. Draw a horizontal line segment inside the box through the median.
4. Extend the lines from each end of the box out to the farthest observation that is still within 1.5(*IQR*) of the corresponding edge. These lines are called *whiskers*.
5. Draw an open circle (or asterisks *) to identify each observation that falls between 1.5(*IQR*) and 3(*IQR*) from the edge to which it is closest; these are called *mild outliers*.
6. Draw a solid circle to identify each observation that falls more than 3(*IQR*) from the closest edge; these are called *extreme outliers*.

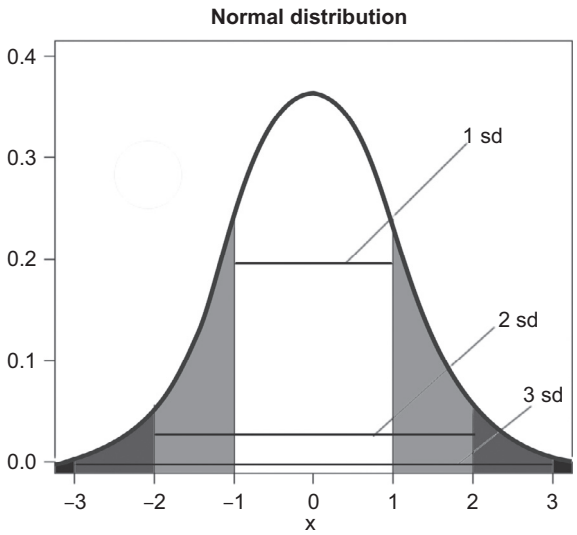


FIGURE 1.5 Bell-shaped curve.

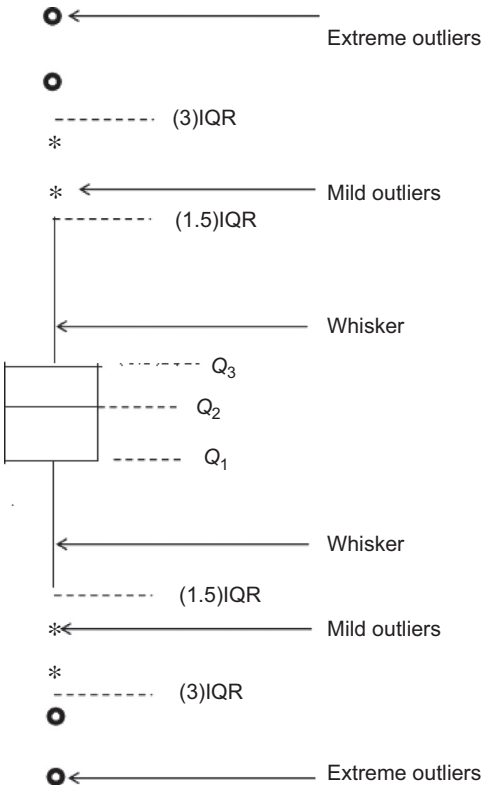


FIGURE 1.6 A typical box-and-whiskers plot.

We illustrate the procedure with the following example.

EXAMPLE 1.5.6

The following data identify the time in months from hire to promotion to chief pharmacist for a random sample of 25 employees from a certain group of employees in a large corporation of drugstores.

5	7	229	453	12	14	18	14	14	483
22	21	25	23	24	34	37	34	49	64
47	67	69	192	125					

Construct a box plot. Do the data appear to be symmetrically distributed along the measurement axis?

Solution

Referring to Example 1.5.2, we find that the median, $Q_2 = 34$.

The lower quartile is $Q_1 = \frac{14+18}{2} = 16$.

The upper quartile is $Q_3 = \frac{67+69}{2} = 68$.

The interquartile range is $IQR = 68 - 16 = 52$.

To find the outliers, compute

$$Q_1 - 1.5(IQR) = 16 - 1.5(52) = -62$$

and

$$Q_3 + 1.5(IQR) = 68 + 1.5(52) = 146.$$

*Using these numbers, we follow the procedure outlined earlier to construct the box plot shown by Fig. 1.7. The * in the box plot represents an outlier. The first horizontal line is the first quartile, the second is the median, and the third is the third quartile.*

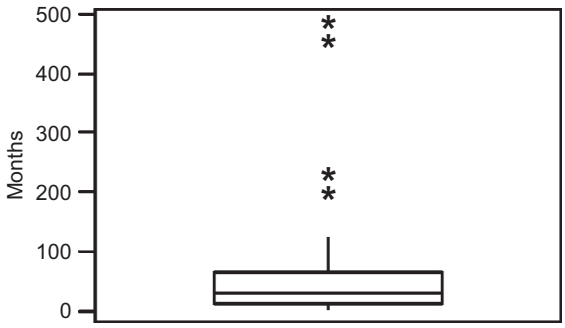


FIGURE 1.7 Box plot for months to promotion.

By examining the relative position of the median line (the middle line in Fig. 1.7), we can test the symmetry of the data. For example, in Fig. 1.7, the median line is closer to the lower quartile than the upper line, which suggests that the distribution is slightly nonsymmetrical. Also, a look at this box plot shows the presence of two mild outliers and two extreme outliers.

A box plot is an effective tool to visualize an entire range of data. Box plots can tell us if the data are uniform or diverse, and gives us a broad overview of the data at hand that will help us asking more questions in a practical application as well as selection of analytical methods.

Exercises 1.5

1.5.1. The prices of 12 randomly chosen homes in dollars (approximated to the nearest 1000) in a growing region of Tampa in the summer of 2002 are given below (data is given in 1000s).

176 105 133 140 305 215 207 210 173 150 78 96.

Find the mean and standard deviation of the sampled home prices from this area.

1.5.2. The following is a sample of nine mortgage companies' interest rates for 30-year home mortgages, assuming 5% down.

7.625 7.500 6.625 7.625 6.625 6.875 7.375 5.375 7.500

(a) Find the mean and standard deviation, and interpret.

(b) Find lower and upper quartiles, median, and interquartile range. Check for any outliers and interpret.

1.5.3. For four observations, it is given that mean is 6, median is 4, and mode is 3. Find the standard deviation of this sample.

1.5.4. The data given below pertain to a random sample of disbursements of state highway funds (in millions of dollars), to different states.

1188 537	1050 519	2882 2523	2802 316	780 1117	1171 1578	685 261

(a) Find the mean, variance, and range for these data and interpret.

(b) Find lower and upper quartiles, median and interquartile range. Check for any outliers and interpret.

(c) Construct a box plot and interpret.

1.5.5. Maximal static inspiratory pressure (PImax) is an index of respiratory muscle strength. The following data show the measure of PImax (cm H_2O) for 15 cystic fibrosis patients.

105	80	115	95	100	85	90	70
135	105	45	115	40	115	95	

(a) Find the lower and upper quartiles, median, and interquartile range. Check for any outliers and interpret.
(b) Construct a box plot and interpret.
(c) Are there any outliers?

1.5.6. Compute the mean, variance, and standard deviation for the data in Table 1.21 (assume that the data belong to a sample).

1.5.7. **(a)** For any grouped data with l classes with group frequencies f_i, and class midpoints m_i, show that

$$\sum_{i=1}^{l} f_i(m_i - \bar{x}) = 0.$$

(b) Verify this result for the data given in Exercise 1.5.6.

1.5.8. **(a)** Given the sample values $x_1, x_2, \ldots, x_n$, show that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

(b) Verify the result of part (a) for the data of Exercise 1.5.6.

1.5.9. The following are the closing prices of some securities that a mutual fund holds on a certain day:

10.25	5.31	11.25	13.13	18.00	32.56	37.06	39.00
43.25	45.00	40.06	28.56	22.75	51.50	47.00	53.50
32.00	25.44	22.50	30.00	24.75	53.37	51.38	26.00
53.50	29.87	32.00	28.87	42.19	37.50	30.44	41.37

(a) Find the mean, variance, and range for these data and interpret.
(b) Find lower and upper quartiles, median, and interquartile range. Check for any outliers.
(c) Construct a box plot and interpret.
(d) Construct a histogram.
(e) Locate on your histogram $\bar{x}$, $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. Count the data points in each of the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$ and compare this with the empirical rule.

1.5.10. The radon concentration (in pCi/liter) data obtained from 40 houses in a certain area are given below.

2.9	0.6	13.5	17.1	2.8	3.8	16.0	2.1	6.4	17.2
7.9	0.5	13.7	11.5	2.9	3.6	6.1	8.8	2.2	9.4
15.9	8.8	9.8	11.5	12.3	3.7	8.9	13.0	7.9	11.7
6.2	6.9	12.8	13.7	2.7	3.5	8.3	15.9	5.1	6.0

(a) Find the mean, variance, and range for these data.
(b) Find lower and upper quartiles, median, and interquartile range. Check for any outliers.
(c) Construct a box plot.
(d) Construct a histogram and interpret.
(e) Locate on your histogram $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. Count the data points in each of the intervals $\bar{x}$, $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. How do these counts compare with the empirical rule?

1.5.11. A random sample of 100 households' weekly food expenditure represented by x from a particular city gave the following statistics:

TABLE 1.21 Class and Frequency.

Class	0–4	5–9	10–14	15–19	20–24
Frequency	5	14	15	10	6

$$\sum x_i = 11,000, \quad \text{and} \quad \sum x_i^2 = 1,900,000.$$

(a) Find the mean and standard deviation for these data.
(b) Assuming that the food expenditure of the households of an entire city of 400,000 will have a bell-shaped distribution, how many households of this city would you expect to fall in each of the intervals, $\overline{x} \pm s$, $\overline{x} \pm 2s$, and $\overline{x} \pm 3s$?

1.5.12. The following numbers are the hours put in by 10 employees of a company in a randomly selected week:

40 46 40 54 18 45 34 60 39 42

(a) Calculate the values of the three quartiles and the interquartile range. Also, calculate the mean and standard deviation and interpret.
(b) Verify from this data set that $\sum_{i=1}^{10}(x_i - \overline{x}) = 0$.
(c) Construct a box plot.
(d) Does this data set contain any outliers?

1.5.13. For the following data:

6.3	2.9	4.5	1.1	1.8	4.0	1.2	3.1	2.0	4.0
7.0	2.8	4.3	5.3	2.9	8.3	4.4	2.8	3.1	5.6
4.5	4.5	5.7	0.5	6.2	3.7	0.9	2.4	3.0	3.5

(a) Find the mean, variance, and standard deviation.
(b) Construct a frequency table with five classes.
(c) Using the grouped data formula, find the mean, variance, and standard deviation for the frequency table constructed in part (b) and compare it to the results in part (a).

1.5.14. In order to assess the protective immunizing activity of various whooping cough vaccines, suppose that 30 batches of different vaccines are tested on groups of children. Suppose that the following data give immunity percentage in home exposure values (IPHE values).

85	51	41	90	91	40	39	69	45	47
42	12	70	38	97	34	94	77	88	91
79	90	43	40	89	85	71	30	25	21

(a) Find the mean, variance, and standard deviation and interpret.
(b) Construct a frequency table with five classes.
(c) Using the grouped data formula, find the mean, variance, and standard deviation for the table in part (b) and compare it to the results in part (a).

1.5.15. The grouped data in Table 1.22 give the number of births by age group of mothers between ages 10 and 39 in a certain state in 2000.

Find the median for this grouped data and interpret.

TABLE 1.22 Number of Births by Mother's Age Group.

Age of mother	Number of births
10–14	895
15–19	55,373
20–24	122,591
25–29	139,615
30–34	127,502
35–39	68,685

TABLE 1.23 Distribution of Salmon Mass.

Weight	155–164	165–174	175–184	185–194	195–204
Frequency	8	11	18	9	4

TABLE 1.24 Length of Dead Fish.

Length of fish (mm)	1–19	20–39	40–59	60–79	80–99
Frequency	38	31	59	45	7

1.5.16. Table 1.23 gives the distribution of the masses (in grams) of 50 salmon from a single young cohort.
(a) Using the grouped data formula, find the mean, variance, and standard deviation.
(b) Find the median for this grouped data.

1.5.17. After a pollution accident, 180 dead fish were recovered from a stream. Table 1.24 gives their lengths measured to the nearest millimeter.
(a) Using the grouped data formula, find the mean, variance, and standard deviation.
(b) Find the median for this grouped data and interpret.

1.6 Computers and statistics

With present-day technology, we can automate most statistical calculations. For small sets of data, many basic calculations such as finding means and standard deviations and creating simple charts, graphing calculators are sufficient. Students should learn how to perform statistical analysis using their handheld calculators. For deeper analysis and for large data sets, statistical software is necessary. Software also provides easier data entry and editing and much better graphics in comparison to calculators. There are many statistical packages available. Many such analyses can be performed with spreadsheet application programs such as Microsoft Excel, but a more thorough data analysis requires the use of more sophisticated software such as Minitab and SPSS. For students with programming abilities, packages such as MATLAB may be more appealing. For very large data sets and for complicated data analysis, one could use SAS. SAS is one of the most frequently used statistical packages. Many other statistical packages (such as Splus, and StatXact) are available; the utilities and advantages of each are based on the specific application and personal taste. The software R is free software that is being increasingly used by statisticians and can be downloaded from http://www.r-project.org/, and many statistical tutorials for R are freely available on the worldwide web. For a good introduction to doing statistics with R, refer to the book by Peter Dalgaard, *Introductory Statistics*, *with R,* Springer, 2002 or its newer edition.

In this book, we will give some representative R, Minitab, SPSS, and SAS commands at the end of each chapter just to get students started on the technology. These examples are by no means a tutorial for the respective software. For a more thorough understanding and use of technology, students should look at the users' manual that comes with the software or at references given at the end of the book. The computer commands are designed to be illustrative, rather than completely efficient. In dealing with data analysis for real-world problems, we need to know which statistical procedure to use, how to prepare the data sets suitable for use in the particular statistical package, and finally how to interpret the results obtained. A good knowledge of theory supplemented with a good working knowledge of statistical software will enable students to perform sophisticated statistical analysis, while understanding the underlying assumptions and the limitations of results obtained. This will prevent us from misleading conclusions when using computer-generated statistical outputs.

1.7 Chapter summary

In this chapter, we dealt with some basic aspects of descriptive statistics. First we gave basic definitions of terms such as *population* and *sample*. Some sampling techniques were discussed. We learned about some graphical presentations in Section 1.4. In Section 1.5 we dealt with descriptive statistics, in which we learned how to find mean, median, and variance and how to identify outliers. A brief discussion of the technology and statistics was given in Section 1.6. All the examples given in this chapter are for a univariate population, in which each measurement consists of a single value. Many populations are *multivariate,* where measurements consist of more than one value. For example, we may be interested in finding a relationship between blood sugar level and age, or between body height and weight. These types of problems will be discussed in Chapter 8.

In practice, it is always better to run descriptive statistics as a check on one's data. The graphical and numerical descriptive measures can be used to verify that the measurements are sound and that there are no obvious errors due to collection or coding.

We now list some of the key definitions introduced in this chapter.

- Population
- Sample
- Statistical inference
- Quantitative data
- Qualitative or categorical data
- Cross-sectional data
- Time series data
- Simple random sample
- Systematic sample
- Stratified sample
- Proportional stratified sampling
- Cluster sampling
- Multiphase sampling
- Relative frequency
- Cumulative relative frequency
- Bar graph
- Pie chart
- Histogram
- Sample mean
- Sample variance
- Sample standard deviation
- Median
- Interquartile range
- Mode
- Mean
- Empirical rule
- Box plots

In this chapter, we have also introduced the following important concepts and procedures:

- General procedure for data collection
- Some advantages of simple random sampling
- Steps for selecting a stratified sample
- Procedures to construct frequency and relative frequency tables and graphical representations such as stem-and-leaf displays, bar graphs, pie charts, histograms, and box plots
- Procedures to calculate measures of central tendency, such as mean and median, as well as measures of dispersion such as the variance and standard deviation for both ungrouped and grouped data
- Guidelines for the construction of frequency tables and histograms
- Procedures to construct a box plot

1.8 Computer examples

In this section, we give some examples of how to use Minitab, SPSS, and SAS for creating graphical representations of the data as well as methods for the computation of basic statistics. Sometimes, the outputs obtained using a particular software package may not be exactly as explained in the book; they vary from one package to another, and also depend on the particular software version. In fact, most of the outputs will not be shown in this book. It is important to obtain the explanation of outputs from the help menu of the particular software package for complete understanding. The "Computer Examples" sections of this book are not designed as manuals for the software, nor are they written in the most efficient way. The idea is only to introduce some basic procedures, so that the students can get started with applying the theoretical material they have seen in each of the chapters.

1.8.1 R introduction and examples

R is a free software for statistical computing and graphics that you can download from http://www.r-project.org/. Detailed help manuals are available from this site. In addition, you can get R help from numerous sources. One such book can be obtained at http://www.ecostat.unical.it/tarsitano/Didattica/LabStat2/Everitt.pdf. In this book, we are only introducing the reader to basic R-programming as a starting point. The R-commands are not optimal, nor is it comprehensive. If you don't have experience with R-program, we suggest that you start working with R-studio (https://www.rstudio.com/), which is much easier to use with its windows interface.

R you ready to start programming?

Introduction to R, imputing and importing data from the examples:

How to input data?

Using the following data:

66 74 79 80 69 77 78 65 79 81

we will make a single variable data set or vector named *x*. First manually, and second using the *scan()* function for convenience.

R code

```
x=c(66,74,79,80,69,77,78,65,79,81);
```

Typing the commas can be time consuming

OR

```
x=scan();
1: 66
2: 74
3: 79
4: 80
5: 69
6: 77
7: 78
8: 65
9: 79
10: 81
11:
```

This method allows you to type each number pressing enter between each entry designed with the number pad in mind. Notice the last entry is blank which ends the scan function.

Results: Both methods obtain the same output, which can be seen simply by typing *x* or *cat(x)* or *print(x)*, however, the scan method allows you to rapidly type your numbers into the variable using a numpad and enter key.

Importing a CSV file

It is common to import comma separated value (CSV) files into R; this imports Example 7.7.1 data into variable *x*.

This example assumes your file is located on a D:\ drive, you may need to modify the path preceding the file name for the CSV files you wish to import.

R code

```
x = read.csv("D:\ ch7_1.csv");
```

Results:

You should have obtained a variable containing the data from the CSV file, these files can be opened with notepad to see their contents.

Exporting a CSV

It is common to export a CSV file of data you wish to save, back up, or share.

Using R we will export the following data:

Sample 1 (x): 1 2 3 4 5 6 7 8 9 10.

This example is writing to the path C:\Users\Admin\Documents; please modify the path to work on your computer.

R code

```
x = c(1:10);
write.csv(x,"C:\Users\Admin\ Documents\myfile.csv");
```

Results: This should have created the specified file in the specified location; you can open this file with notepad and should see the exported data.

Example 1.8.1 **(Stem-and-leaf plot)** Using the following data construct a stem-and-leaf plot.

Sample X: 78 74 82 66 94 71 64 88 55 80 91 74 82 75 96 78 84 79 71 83

This assumes you've stored the data under variable *x*; please modify your code appropriately.

R code

```
stem(x);
```

Output:

```
The decimal point is 1 digit(s) to the right of the |
5 | 5
6 | 46
7 | 11445889
8 | 022348
9 | 146
```

Example 1.8.2 **(Histogram)** Using the following data construct a histogram.

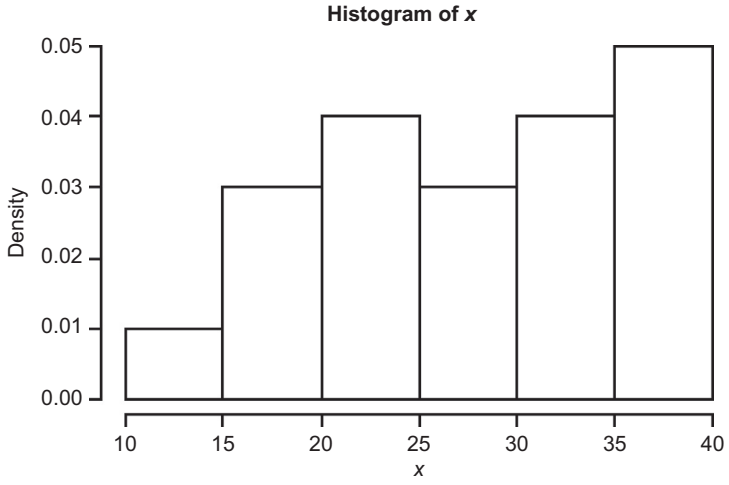
Sample X: 25 37 20 31 31 21 12 25 36 27 38 16 40 32 33 24 39 26 27 19

This assumes you've stored the data into variable *x*; please modify your code appropriately.

R code

```
hist(x);
```

Output:



Example 1.8.3 **(Descriptive Statistics)** Using the following data generate descriptive statistics.

Sample X: 5 7 229 453 12 14 18 14 18 14 14 483 22 21 25 23 24 34 37 34 49 64 47 67 69 192 125

This assumes you've stored the data into variable *x*; please modify your code appropriately.

R code

```
summary(x);
sd(x);
```

Standard deviation

```
length(x);
```

Length of variable

Output:

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
5.00	18.00	34.00	83.28	67.00	483.00

128.3649 ← Standard deviation

25 ← Length of variable

Example 1.8.4 **(Box Plot)** Using the following data create a box plot.

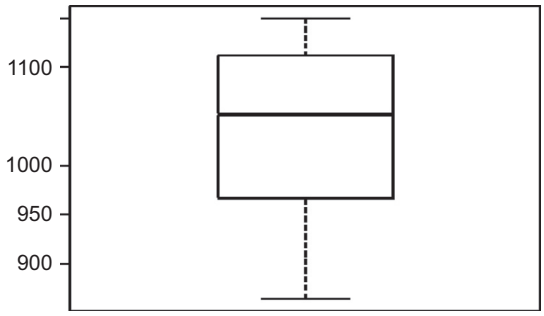
Sample X: 870 922 1146 1120 1079 905 888 865 1112 966 1150 977 958 1088 1139 1055 1082 1053 1048 1118 866 996 1102 1028 1130 1002 990 1052 1116 1109

This assumes you've stored the data into variable *x*; please adjust your code appropriately.

R code

```
boxplot(x);
```

Output:



Example 1.8.5 **(Test of Randomness)** Using the following data, test whether or not the sample is random (details of this test are left undisclosed):

Sample X: 24 31 28 43 28 56 48 39 52 32 38 49 51 49 62 33 41 58 63 56

This test is known as "Runs test" and assumes you've stored the data into variable *x*; please modify your code appropriately. Additionally you will need to install the "lawstat" package to use this test.

R code

```
install.packages('lawstat');
library('lawstat');
runs.test(x);
```

install.packages('lawstat'); ← Installs and loads the required package

Output:

Runs Test - Two sided

data: x

Standardized Runs Statistic = -1.3784, p-value = 0.1681

Using the methods of Chapter 6, we will see that since the *P*-value is not small, we cannot reject the hypothesis that the sample is random.

1.8.2 Minitab examples

A good place to get help on Minitab is http://www.minitab.com/resources/. There are many helpful sites available on Minitab procedures; for example, Minitab student tutorials can be obtained from http://www.minitab.com/resources/tutorials/. Here we illustrate only some of the basic uses of Minitab. In Minitab, we can enter the data in the spreadsheet and use the Windows pull-down menus, or we can directly enter the data and commands. We will mostly give procedures for the pull-down menus only. It is up to the user's taste to choose among these procedures. It should be noted that with different versions of Minitab, there will be some differences in the pull-down menu options. It is better to consult the Help menu for the actual procedure. Most of the time, we will not give the output.

EXAMPLE 1.8.1 (Stem-and-Leaf)

For the following data, construct a stem-and-leaf display using Minitab:

78	74	82	66	94	71	64	88	55	80
91	74	82	75	96	78	84	79	71	83

Solution

For the pull-down menu, first enter the data in column 1. Then follow the following sequence. The boldface represents the actions.

Graph > ***Character Graphs*** > ***Stem*-and-Leaf**.

In ***Variables*:** type ***C1*** and click ***OK***.

EXAMPLE 1.8.2 (Histogram)

For the following data, construct a histogram:

25	37	20	31	31	21	12	25	36	27
38	16	40	32	33	24	39	26	27	19

Solution

Enter the data in ***C1***, then use the following sequence.

Graph > ***Histogram*** … > *in* ***Graph variables:*** *type* ***C1*** > ***OK***.

If we want to change the number of intervals, after entering ***Graph variables****, click* ***Options*** *… and click* ***Number of intervals*** *and enter the* ***desired number****, then* ***OK***.

EXAMPLE 1.8.3 (Descriptive Statistics)

In this example, we will describe how to obtain basic statistics such as mean, median, and standard deviation for the following data:

5	7	229	453	12	14	18	14	14	483
22	21	25	23	24	34	37	34	49	64
47	67	69	192	125					

Solution

Enter the data in ***C1****. Then use*

Stat > ***Basic Statistics*** > ***Display Descriptive Statistics*** … > in ***Variables*:** type ***C1*** > click ***OK***.

EXAMPLE 1.8.4 (Sorting and Box Plot)

For the following data, first sort in the increasing order and then construct a box plot to check for outliers.

870	922	1146	1120	1079	905	888	865	1112	966
1150	977	958	1088	1139	1055	1082	1053	1048	1118
866	996	1102	1028	1130	1002	990	1052	1116	1109

Solution

After entering the data in C1, we can sort the data in increasing order as follows:

Manip > ***Sort*** … > in ***Sort column(s):*** type ***C1*** > in ***Store sorted column(s) in:*** type ***C2*** > in ***Sorted by column:*** type ***C1*** > ***OK***.

If we want to draw a box plot for the data, do the following:

Graph > ***Box plot*** … > in ***Graph variables:*** under ***Y,*** type ***C1*** > ***OK***.

EXAMPLE 1.8.5 (Test of Randomness)

Almost all of the analyses in this book assume that the sample is random. How can we verify whether the sample is really random? Project 12B explains a procedure called *run test*. Without going into details, this test is simple with Minitab. All we have to do is enter the data in C1. Then click.

Stat > ***Nonparametric*** > ***Runs Test*** … > *in* ***variables:*** *enter* ***C1*** > ***OK***.

For instance, if we have the following data:

24	31	28	43	28	56	48	39	52	32
38	49	51	49	62	33	41	58	63	56

we will get following output:

Run Test

C1

K = 44.0500

The observed number of runs = 14.

The expected number of runs = 11.0000.

10 Observations above K 10 below.

*N Small – The following approximation may be invalid.

The test is significant at 0.1681.

Cannot reject at alpha = 0.05.

"***Cannot reject***" in the output means that it is reasonable to assume that the sample is random. For any data, it is always desirable to do a run test to determine the randomness.

1.8.3 SPSS examples

For SPSS, we will give only Windows commands. For all the pull-down menus, the sequence will be separated by the > symbol.

EXAMPLE 1.8.6

Redo Example 1.8.1 with SPSS.

Solution

After entering the data in ***C1:***

Analyze > ***Descriptive Statistics*** > ***Explore*** … >

At the ***Explore*** window select the variable and move to ***Dependent List***; then click ***Plots*** …, select ***Stem-and-Leaf***, click ***Continue***, and click ***OK*** at the ***Explore*** Window.

We will get the output with a few other things, including box plots along with the stem-and-leaf display, which we will not show here.

EXAMPLE 1.8.7

Redo Example 1.8.2 with SPSS.

Solution

After entering the data:

Graphs > ***Histogram*** … >

At the ***Histogram*** window select the variable and move to ***Variable***, and click ***OK***.

We will get the histogram, which we will not display here.

EXAMPLE 1.8.8

Redo Example 1.8.3 with SPSS.

Solution

Enter the data, then:

Analyze > **Descriptive Statistics** > **Frequencies** … >

At the ***Frequencies*** window select the variable(s); then open the ***Statistics*** window and check whichever boxes you desire under ***Percentile***, ***Dispersion, Central Tendency***, and ***Distribution*** > ***continue*** > ***OK***.

For example, if you select Mean, Median, Mode, Standard Deviation, and Variance, we will get the following output and more:

Statistics

		VAR00001
N	**Valid**	**25**
	Missing	**0**
Mean		83.2800
Median		34.0000
Mode		14.00
Std. Deviation		128.36488
Variance		16,477.54333

1.8.4 SAS examples

We will now give some SAS procedures describing the numerical measures of a single variable. **PROC UNIVARIATE** will give mean, median, mode, standard deviation, skewness, kurtosis, etc. If we do not need median, mode, and so on, we could just as well use **PROC MEANS** in lieu of **PROC UNIVARIATE**. We can use the following general format in writing SAS programs with appropriate problem-specific modifications. There are many good online references as well as books available for SAS procedures. To get support on SAS, including many example codes, refer to the SAS support website: http://support.sas.com/. Another helpful site can be found at http://www.ats.ucla.edu/stat/sas/. There are many other sites that may suit your particular application.

General format of an SAS program

DATA give a name to the data set;
INPUT here we put variable names and column locations, if there are more than one variable;
CARDS; (also we can use DATALINES)
Enter the data here;
TITLE "here we include the title of our analysis";
PROC PRINT;
PROC name of procedure (such as PROC UNIVARIATE) goes here;
Options that we may want to include (such as the variables to be used) go here;
RUN;

After writing an SAS program, to execute it we can go to the menu bar and select run > submit, or click the "running man" icon. On execution, SAS will output the results to the Output window. All the steps used including time of execution and any error messages will be given in the Log window.

In order to make the SAS outputs more manageable, we can use the following SAS command at the beginning of an SAS program:

options ls = 80 ps = 50;

ls stands for line size, and this sets each line to be 80 characters wide. ps stands for page size and allows 50 lines on each page. This reduces the number of unnecessary page breaks. In order to avoid date and number, we can use the option commands:

Options *nodate nonumber*;

EXAMPLE 1.8.9

For the data of Example 1.8.3, use **PROC UNIVARIATE** to summarize the data.

Solution

*In the program editor window, type the following if you are entering the data directly. If you are using the data stored in a file, the comment line (with *) should be used instead of the input and data lines.*

```
Options nodate nonumber;
DATA e×9;
INPUT e×9 @@;
DATALINES;
5 7 229 453 12 14 18 14 14 483.
22 21 25 23 24 34 37 34 49 64.
47 67 69 192 125;
PROC UNIVARIATE;
TITLE;
RUN;
```

In this case we will get the following output:

The UNIVARIATE Procedure
Variable: ex9
Moments

N	25	Sum Weights	25
Mean	83.28	Sum Observations	2082
Std Deviation	128.364884	Variance	16,477.5433
Skewness	2.45719194	Kurtosis	5.47138396
Uncorrected SS	568,850	Corrected SS	395,461.04
Coeff Variation	154.136508	Std Error Mean	25.6729767

Basic Statistical Measures

Location		Variability	
Mean	83.28000	Std Deviation	128.36488
Median	34.00000	Variance	16,478
Mode	14.00000	Range	478.00000
Interquartile Range		49.00000	

Tests for Location: Mu0=0

Test	-Statistic-		-p Value-
Student's t t	3.243878	Pr>|t|	0.0035
Sign	M 12.5	Pr>=|M|	<0.0001
Signed Rank S	162.5	Pr>=|S|	<0.0001

Quartiles (Definition 5)

Quartile	Estimate
100% Max	483
99%	483
95%	453
90%	229
75% Q3	67
50% Median	34
25% Q1	18
10%	12
5%	7
1%	5
0% Min	5

The UNIVARIATE Procedure
Variable: ex9
Extreme Observations

-Lowest-		-Highest-	
Value	Obs	Value	Obs
5	1	125	25
7	2	192	24
12	5	229	3
14	9	453	4
14	8	483	10

We can observe from the previous output that ***PROC UNIVARIATE*** *gives much information about the data, such as mean, standard deviation, and quartiles. If we do not want all these details, we could use the* ***PROC MEANS*** *command. In the previous code, if we replace* ***PROC UNIVARIATE*** *by the* ***PROC MEANS*** *statement, we will get the following:*

The MEANS Procedure

Analysis Variable : ex9

N	Mean	Std Dev	Minimum	Maximum
25 83	2800000	128.3648836	5.0000000	483.0000000

The output is greatly simplified.

If we use ***PROC UNIVARIATE PLOT NORMAL***, this option will produce three plots: stem-and-leaf, box plot, and normal probability plot (this will be discussed later in the text). In order to obtain bar graphs at the midpoints of the class intervals, use the following commands:

PROC CHART DATA = e×9;
VBAR e×9;
If we want to create a frequency table, use the following:
PROC FREQ;
table ex9;
title "Frequency tabulation";

Every PROC or procedure has its own name and options. We will use different PROCs as we need them. Always remember to enclose titles in single quotes. There are various other actions that we can perform for the data analysis using SAS. It is beyond the scope of this book to explain general and efficient SAS codes. For details, we refer to books dedicated to SAS, such as the book by Ronald P. Cody and Jeffrey K. Smith, *Applied Statistics and the SAS Programming Language, Fifth Edition*, Prentice Hall, 2006. There are many websites that give SAS codes. One example with references for many aspects of SAS, including many codes, can be found at http://www.sas.com/service/library/onlinedoc/code.samples.html.

Exercises 1.8

1.8.1. The following data represent the lengths (to the nearest whole millimeter) of 80 shoots from seeds of a certain type planted at the same time.

75	72	76	76	72	74	71	75	77	72
74	71	76	76	76	72	71	73	73	71
72	72	75	70	74	74	78	74	76	79
75	76	73	73	71	72	79	74	77	72
76	70	72	75	78	72	69	75	72	71
77	79	76	73	75	73	72	75	74	78
73	77	73	77	70	74	66	74	73	77
75	79	75	70	72	73	80	73	78	75

Using one of the software packages (R, Minitab, SPSS, or SAS):
(a) Represent the data in a histogram.
(b) Find the summary statistics such as mean, median, variance, and standard deviation.
(c) Draw box plots and identify any outliers.

1.8.2. On a particular day, when asked, "How many minutes did you exercise today?" the following were the responses of 30 randomly selected people:

15	30	25	10	30	15	10	45	20	22
18	0	45	12	15	10	17	30	30	15
10	30	20	8	18	30	27	33	15	0

Using one of the software packages (R, Minitab, SPSS, or SAS):
(a) Represent the data in a histogram.
(b) Find the summary statistics such as mean, median, variance, and standard deviation.
(c) Draw box plots and identify any outliers.

Projects for chapter 1

1A World Wide Web and data collection

Statistical Abstracts of the United States is a rich source of statistical data (http://www.census.gov/prod/www/statistical-abstract-us.html). Pick any category of interest to you and obtain data (say, Income, Expenditures, and Wealth). Represent a section of the data graphically. Find mean, median, and standard deviation. Identify any outliers. There are many other sites, such as http://lib.stat.cmu.edu/datasets/ and http://it.stlawu.edu/∼rlock/datasurf.html, that we can use for obtaining real data sets.

1B Preparing a list of useful Internet sites

Prepare a list of Internet references for various aspects of statistical study.

1C Dot plots and descriptive statistics

From the local advertisements of apartments for rent, randomly pick 50 monthly rents for two-bedroom apartments. For these data, first draw a dot plot and then obtain descriptive statistics (use R, or any other statistical software).

1D Importance of statistics in our society

Write a short report on the importance of statistics in our modern-day society. Give different examples to illustrate your points. One interesting project will be to study the role of the Internet of Things (IoT), a vast network of smart objects that work together in collecting and analyzing data and autonomously performing actions.

1E Uses and misuses of statistics

"There are three types of lies—lies, damn lies, and statistics"—Benjamin Disraeli.

Write a short report on uses and misuses of statistics.