

Describing Data: Graphical

- 1.1 Decision Making in an Uncertain Environment
 - Random and Systematic Sampling
 - Sampling and Nonsampling Errors
- 1.2 Classification of Variables
 - Categorical and Numerical Variables
 - Measurement Levels
- 1.3 Graphs to Describe Categorical Variables
 - Tables and Charts
 - Cross Tables
 - Pie Charts
 - Pareto Diagrams
- 1.4 Graphs to Describe Time-Series Data
- 1.5 Graphs to Describe Numerical Variables
 - Frequency Distributions
 - Histograms and Ogives
 - Shape of a Distribution
 - Stem-and-Leaf Displays
 - Scatter Plots
- 1.6 Data Presentation Errors
 - Misleading Histograms
 - Misleading Time-Series Plots

Introduction

What are the projected sales of a new product? Will the cost of Google shares continue to increase? Who will win the 2020 UEFA Champions League? How satisfied were you with your last purchase at Starbucks, on alibaba.com, or at IKEA? If you were hired by the National Nutrition Council of your country, how would you determine if the Council's guidelines on consumption of fruit, vegetables, snack foods, and soft drinks are being met? Do people who are physically active have healthier diets than people who are not physically active? What factors (perhaps disposable income or grants) are significant in forecasting the aggregate consumption of durable goods? What effect will a 2% increase in interest rates have on residential investment? Do

credit scores, current balance, or outstanding maintenance balance contribute to an increase in the percentage of a mortgage company's delinquent accounts increasing? Answers to questions such as these come from an understanding of statistics, fluctuations in the market, consumer preferences, trends, and so on.

Statistics are used to predict or forecast sales of a new product, construction costs, customer-satisfaction levels, the weather, election results, university enrollment figures, grade point averages, interest rates, currency-exchange rates, and many other variables that affect our daily lives. We need to absorb and interpret substantial amounts of data. Governments, businesses, and scientific researchers spend billions of dollars collecting data. But once data are collected, what do we do with them? How do data impact decision making?

In our study of *statistics* we learn many tools to help us process, summarize, analyze, and interpret data for the purpose of making better decisions in an uncertain environment. Basically, an understanding of statistics will permit us to make sense of all the data.

In this chapter we introduce tables and graphs that help us gain a better understanding of data and that provide visual support for improved decision making. Reports are enhanced by the inclusion of appropriate tables and graphs, such as frequency distributions, bar charts, pie charts, Pareto diagrams, line charts, histograms, stem-and-leaf displays, or ogives. Visualization of data is important. We should always ask the following questions: What does the graph suggest about the data? What is it that we see?

1.1 DECISION MAKING IN AN UNCERTAIN ENVIRONMENT

Decisions are often made based on limited information. Accountants may need to select a portion of records for auditing purposes. Financial investors need to understand the market's fluctuations, and they need to choose between various portfolio investments. Managers may use surveys to find out if customers are satisfied with their company's products or services. Perhaps a marketing executive wants information concerning customers' taste preferences, their shopping habits, or the demographics of Internet shoppers. An investor does not know with certainty whether financial markets will be buoyant, steady, or depressed. Nevertheless, the investor must decide how to balance a portfolio among stocks, bonds, and money market instruments while future market movements are unknown.

For each of these situations, we must carefully define the problem, determine what data are needed, collect the data, and use statistics to summarize the data and make inferences and decisions based on the data obtained. Statistical thinking is essential from initial problem definition to final decision, which may lead to reduced costs, increased profits, improved processes, and increased customer satisfaction.

Random and Systematic Sampling

Before bringing a new product to market, a manufacturer wants to arrive at some assessment of the likely level of demand and may undertake a market research survey. The manufacturer is, in fact, interested in *all* potential buyers (the population). However, populations are often so large that they are unwieldy to analyze; collecting complete information for a population could be impossible or prohibitively expensive. Even in circumstances where sufficient resources seem to be available, time constraints make the examination of a subset (sample) necessary.

Population and Sample

A **population** is the complete set of all items that interest an investigator. Population size, N , can be very large or even infinite. A **sample** is an observed subset (or portion) of a population with sample size given by n .

Examples of populations include the following:

- All potential buyers of a new product
- All stocks traded on the London Stock Exchange (LSE)
- All registered voters in a particular city or country
- All accounts receivable for a corporation

Our eventual aim is to make statements based on sample data that have some validity about the population at large. We need a sample, then, that is representative of the population. How can we achieve that? One important principle that we must follow in the sample selection process is randomness.

Random Sampling

Simple random sampling is a procedure used to select a sample of n objects from a population in such a way that each member of the population is chosen strictly by chance, the selection of one member does not influence the selection of any other member, each member of the population is equally likely to be chosen, and every possible sample of a given size, n , has the same chance of selection. This method is so common that the adjective *simple* is generally dropped, and the resulting sample is called a **random sample**.

Another sampling procedure is systematic sampling (stratified sampling and cluster sampling are discussed in Chapter 17).

Systematic Sampling

Suppose that the population list is arranged in some fashion unconnected with the subject of interest. **Systematic sampling** involves the selection of every j th item in the population, where j is the ratio of the population size N to the desired sample size, n ; that is, $j = N/n$. Randomly select a number from 1 to j to obtain the first item to be included in your systematic sample.

Suppose that a sample size of 100 is desired and that the population consists of 5,000 names in alphabetical order. Then $j = 50$. Randomly select a number from 1 to 50. If your number is 20, select it and every 50th number, giving the systematic sample of elements numbered 20, 70, 120, 170, and so forth, until all 100 items are selected. A systematic sample is analyzed in the same fashion as a simple random sample on the grounds that, relative to the subject of inquiry, the population listing is already in random order. The danger is that there could be some subtle, unsuspected link between the ordering of the population and the subject under study. If this were so, bias would be induced if systematic sampling was employed. Systematic samples provide a good representation of the population if there is no cyclical variation in the population.

Sampling and Nonsampling Errors

Suppose that we want to know the average age of registered voters in the United States. Clearly, the population size is so large that we might take only a random sample, perhaps 500 registered voters, and calculate their average age. Because this average is based on sample data, it is called a *statistic*. If we were able to calculate the average age of the entire population, then the resulting average would be called a *parameter*.

Parameter and Statistic

A **parameter** is a numerical measure that describes a specific characteristic of a population. A **statistic** is a numerical measure that describes a specific characteristic of a sample.

Throughout this book we will study ways to make decisions about a population parameter, based on a sample statistic. We must realize that some element of uncertainty will always remain, as we do not know the exact value of the parameter. That is, when a sample is taken from a population, the value of any population parameter will not be able to be known *precisely*. One source of error, called **sampling error**, results from the fact that information is available on only a subset of all the population members. In Chapters 6, 7, and 8 we develop statistical theory that allows us to characterize the nature of the sampling error and to make certain statements about population parameters.

In practical analyses there is the possibility of an error unconnected with the kind of sampling procedure used. Indeed, such errors could just as well arise if a complete census of the population were taken. These are referred to as **nonsampling errors**. Examples of nonsampling errors include the following:

1. **The population actually sampled is not the relevant one.** A celebrated instance of this sort occurred in 1936, when *Literary Digest* magazine confidently predicted that Alfred Landon would win the presidential election over Franklin Roosevelt. However, Roosevelt won by a very comfortable margin. This erroneous forecast resulted from the fact that the members of the *Digest's* sample had been taken from telephone directories and other listings, such as magazine subscription lists and automobile registrations. These sources considerably underrepresented the poor, who were predominantly Democrats. To make an inference about a population (in this case the U.S. electorate), it is important to sample that population and not some subgroup of it, however convenient the latter course might appear to be.
2. **Survey subjects may give inaccurate or dishonest answers.** This could happen because questions are phrased in a manner that is difficult to understand or in a way that appears to make a particular answer seem more palatable or more desirable. Also, many questions that one might want to ask are so sensitive that it would be foolhardy to expect uniformly honest responses. Suppose, for example, that a plant manager wants to assess the annual losses to the company caused by employee thefts. In principle, a random sample of employees could be selected and sample members asked, What have you stolen from this plant in the past 12 months? This is clearly not the most reliable means of obtaining the required information!
3. **There may be no response to survey questions.** Survey subjects may not respond at all, or they may not respond to certain questions. If this is substantial, it can induce additional sampling and nonsampling errors. The sampling error arises because the achieved sample size will be smaller than that intended. Nonsampling error possibly occurs because, in effect, the population being sampled is not the population of interest. The results obtained can be regarded as a random sample *from the population that is willing to respond*. These people may differ in important ways from the larger population. If this is so, a bias will be induced in the resulting estimates.

There is no general procedure for identifying and analyzing nonsampling errors. But nonsampling errors could be important. The investigator must take care in such matters as identifying the relevant population, designing the questionnaire, and dealing with non-response in order to minimize the significance of nonsampling errors. In the remainder of this book it is assumed that such care has been taken, and our discussion centers on the treatment of sampling errors.

To think statistically begins with problem definition: (1) What information is required? (2) What is the relevant population? (3) How should sample members be selected? (4) How should information be obtained from the sample members? Next we will want to know how to use sample information to make decisions about our population of interest. Finally, we will want to know what conclusions can be drawn about the population.

After we identify and define a problem, we collect data produced by various processes according to a design, and then we analyze that data using one or more statistical procedures. From this analysis, we obtain information. Information is, in turn, converted into knowledge, using understanding based on specific experience, theory, literature, and additional statistical procedures. Both descriptive and inferential statistics are used to change data into knowledge that leads to better decision making. To do this, we use descriptive statistics and inferential statistics.

Descriptive and Inferential Statistics

Descriptive statistics focus on graphical and numerical procedures that are used to summarize and process data. **Inferential statistics** focus on using the data to make predictions, forecasts, and estimates to make better decisions.

1.2 CLASSIFICATION OF VARIABLES

A variable is a specific characteristic (such as age or weight) of an individual or object. Variables can be classified in several ways. One method of classification refers to the type and amount of information contained in the data. Data are either categorical or numerical. Another method, introduced in 1946 by American psychologist Stanley Smith Stevens is to classify data by levels of measurement, giving either qualitative or quantitative variables. Correctly classifying data is an important first step to selecting the correct statistical procedures needed to analyze and interpret data.

Categorical and Numerical Variables

Categorical variables produce responses that belong to groups or categories. For example, responses to yes/no questions are categorical. Are you a business major? and Do you own a car? are limited to yes or no answers. A health care insurance company may classify incorrect claims according to the type of errors, such as procedural and diagnostic errors, patient information errors, and contractual errors. Other examples of categorical variables include questions on gender or marital status. Sometimes categorical variables include a range of choices, such as “strongly disagree” to “strongly agree.” For example, consider a faculty-evaluation form where students are to respond to statements such as the following: The instructor in this course was an effective teacher (1: strongly disagree; 2: slightly disagree; 3: neither agree nor disagree; 4: slightly agree; 5: strongly agree).

Numerical variables include both discrete and continuous variables. A **discrete numerical variable** may (but does not necessarily) have a finite number of values. However, the most common type of discrete numerical variable produces a response that comes from a counting process. Examples of discrete numerical variables include the number of students enrolled in a class, the number of university credits earned by a student at the end of a particular semester, and the number of Microsoft stocks in an investor’s portfolio.

A **continuous numerical variable** may take on any value within a given range of real numbers and usually arises from a measurement (not a counting) process. Someone might say that he is 6 feet (or 72 inches) tall, but his height could actually be 72.1 inches, 71.8 inches, or some other similar number, depending on the accuracy of the instrument used to measure height. Other examples of continuous numerical variables include the weight of a cereal box, the time to run a race, the distance between two cities, or the temperature. In each case the value could deviate within a certain amount, depending on the precision of the measurement instrument used. We tend to truncate continuous variables in daily conversation and treat them as though they were the same as discrete variables without even giving it a second thought.

Measurement Levels

We can also describe data as either *qualitative* or *quantitative*. With **qualitative data** there is no measurable meaning to the “difference” in numbers. For example, one football player is assigned the number 7 and another player has the number 10. We cannot conclude that the first player plays twice as well as the second player. However, with **quantitative data** there is a measurable meaning to the difference in numbers. When one student scores 90 on an exam and another student scores 45, the difference is measurable and meaningful.

Qualitative data include nominal and ordinal levels of measurement. Quantitative data include interval and ratio levels of measurement.

Nominal and ordinal levels of measurement refer to data obtained from categorical questions. Responses to questions on gender, country of citizenship, political affiliation, and ownership of a mobile phone are nominal. **Nominal data** are considered the lowest or weakest type of data, since numerical identification is chosen strictly for convenience and does not imply ranking of responses.

The values of nominal variables are words that describe the categories or classes of responses. The values of the gender variable are male and female; the values of Do you own a car? are yes and no. We arbitrarily assign a code or number to each response. However, this number has no meaning other than for categorizing. For example, we could code gender responses or yes/no responses as follows:

1 = Male; 2 = Female
1 = Yes; 2 = No

Ordinal data indicate the rank ordering of items, and similar to nominal data the values are words that describe responses. Some examples of ordinal data and possible codes are as follows:

1. Product quality rating (1: poor; 2: average; 3: good)
2. Satisfaction rating with your current Internet provider (1: very dissatisfied; 2: moderately dissatisfied; 3: no opinion; 4: moderately satisfied; 5: very satisfied)
3. Consumer preference among three different types of soft drink (1: most preferred; 2: second choice; 3: third choice)

In these examples the responses are ordinal, or put into a rank order, but there is no measurable meaning to the “difference” between responses. That is, the difference between your first and second choices may not be the same as the difference between your second and third choices.

Interval and ratio levels of measurement refer to data obtained from numerical variables, and meaning is given to the *difference* between measurements. An interval scale indicates rank and distance from an arbitrary zero measured in unit intervals. That is, data are provided relative to an arbitrarily determined benchmark. Temperature is a classic example of this level of measurement, with arbitrarily determined benchmarks generally based on either Celsius degrees or Fahrenheit. Suppose that in March 2019, it is 30°C in Pune, India, and only 10°C in Tokyo, Japan. We can conclude that the difference in temperature is 20°, but we cannot say that it is three times as warm in Pune as it is in Tokyo. The year is another example of an interval level of measurement, with benchmarks based most commonly on the Gregorian calendar.

Ratio data indicate both rank and distance from a natural zero, with ratios of two measures having meaning. A person who weighs 200 pounds is twice the weight of a person who weighs 100 pounds; a person who is 40 years old is twice the age of someone who is 20 years old.

After collecting data, we first need to classify responses as categorical or numerical or by measurement scale. Next, we assign an arbitrary ID or code number to each response. Some graphs are appropriate for categorical variables, and others are used for numerical variables.

Note that data files usually contain “missing values.” For example, respondents to a questionnaire may choose not to answer certain questions about gender, age, income, or some other sensitive topic. Missing values require a special code in the data entry stage. Unless missing values are properly handled, it is possible to obtain erroneous output. Statistical software packages handle missing values in different ways.

EXERCISES




Visit www.MyStatLab.com or www.pearsonglobaleditions.com to access the data files.


Basic Exercises

- 1.1 State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?
 - a. Size of a vanilla chai (small to extra-large)
 - b. The number of shares of a stock purchased by a broker
 - c. The weight (in pounds, ounces, etc.) of a newborn baby
 - d. The nationality of a state’s incumbent governor
- 1.2 Upon visiting a newly opened Starbucks store, customers were given a brief survey. Is the answer to each of the following questions categorical or numerical? If categorical, give the level of measurement. If numerical, is it discrete or continuous?
 - a. Is this your first visit to this Starbucks store?
 - b. On a scale from 1 (very dissatisfied) to 5 (very satisfied), rate your level of satisfaction with today’s purchase?
 - c. What was the actual cost of your purchase today?
- 1.3 The Budapest Airport managers circulated a form to find out passengers’ level of satisfaction with the lounges and VIP services. The passengers who frequented the lounges and used the services were asked to indicate how much they spent on such services in a year. They were also asked to indicate their level of satisfaction on a scale from 1 (very satisfied) to 5 (very dissatisfied). Is a passenger’s response to each question numerical or categorical? If numerical, is it discrete or continuous? If categorical, give the level of measurement.
- 1.4 Faculty at one university were asked a series of questions in a recent survey. State the type of data for each question.
 - a. Indicate your level of satisfaction with your teaching load (very satisfied, moderately satisfied, neutral, moderately dissatisfied, or very dissatisfied).
 - b. How many of your research articles were published in refereed journals during the last 5 years?
 - c. Did you attend the last university faculty meeting?
 - d. Do you think that the teaching evaluation process needs to be revised?

- 1.5 Tourists visiting Croatia are asked to participate in a survey, consisting of various questions regarding their experience during their trip, which have been provided below. For each question, describe the type of data obtained.
 - a. Which of the following areas did you visit?
 - Coast
 - Islands
 - Mountains
 - Zagreb (Croatia’s capital)
 - b. Did you rent the sailing boat?
 - Yes
 - No
 - c. What was the average amount you spent on food per day?
 - d. What is the optimal number of days you would recommend a tourist spends in Croatia?
 - e. How often would you recommend visiting Croatia?
 - a. every year
 - b. once in a five years
 - c. once in a lifetime
 - d. never
- 1.6 Residents in one housing development were asked a series of questions by their homeowners’ association. Identify the type of data for each question.
 - a. Did you play golf during the last month on the development’s new golf course?
 - b. How many times have you eaten at the country club restaurant during the last month?
 - c. Do you own a camper?
 - d. Rate the new security system for the development (very good, good, poor, or very poor).

Application Exercises

- 1.7  A survey of students at one college was conducted to provide information to address various concerns about the college’s library. This information and other data about the students are stored in the data file **Library Survey**.
 - a. Give an example of a categorical variable with ordinal responses.
 - b. Give an example of a categorical variable with nominal responses.
 - c. Give an example of a numerical variable with discrete responses.

- 1.8  The Programme for International Student Assessment (PISA) is a global study by the Organization for Economic Co-operation and Development (OECD). It measures 15-year-old students' ability to use their reading, mathematics, and science knowledge and skills to meet real-life challenges. PISA data, available from the OECD website, is used for research into equity and inclusion in countries' education

worldwide. Using the data file **PISA Sample**, which provides a small sample for some variables, give an example of the following:

- A categorical variable with ordinal responses
- A categorical variable with nominal responses
- A numerical variable with continuous responses
- A numerical variable with discrete responses

1.3 GRAPHS TO DESCRIBE CATEGORICAL VARIABLES

We can describe categorical variables using frequency distribution tables and graphs such as bar charts, pie charts, and Pareto diagrams. These graphs are commonly used by managers and marketing researchers to describe data collected from surveys and questionnaires.

Frequency Distribution

A **frequency distribution** is a table used to organize data. The left column (called classes or groups) includes all possible responses on a variable being studied. The right column is a list of the frequencies, or number of observations, for each class. A **relative frequency distribution** is obtained by dividing each frequency by the number of observations and multiplying the resulting proportion by 100%.

Tables and Charts

The classes that we use to construct frequency distribution tables of a categorical variable are simply the possible responses to the categorical variable. Bar charts and pie charts are commonly used to describe categorical data. If our intent is to draw attention to the *frequency* of each category, then we will most likely draw a **bar chart**. In a bar chart the height of a rectangle represents each frequency. There is no need for the bars to touch.

Example 1.1 Healthy Eating Index 2005 (HEI-2005): Activity Level (Frequency Distribution and Bar Chart)

The U.S. Department of Agriculture (USDA) Center for Nutrition Policy and Promotion (CNPP) and the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), conduct surveys to assess the health and nutrition of the U.S. population. The CNPP conducts the Healthy Eating Index (Guenther et al. 2007) and the NCHS conducts the National Health and Nutrition Examination Survey (CDC 2003–2004). The Healthy Eating Index (HEI) monitors the diet quality of the U.S. population, particularly how well it conforms to dietary guidance. The HEI–2005 measures how well the population follows the recommendations of the 2005 *Dietary Guidelines for Americans* (Guenther et al.). In particular it measures, on a 100-point scale, the adequacy of consumption of vegetables, fruits, grains, milk, meat and beans, and liquid oils.

The data file **HEI Cost Data Variable Subset** contains considerable information on randomly selected individuals who participated in two extended interviews and medical examinations. Data for the first interview are identified by `daycode = 1`; data for the second interview are identified by `daycode = 2`. Other variables in the data file are described in the data dictionary in the Chapter 10 Appendix.

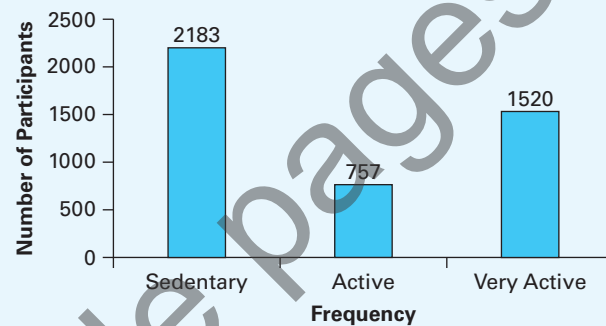
One variable in the HEI–2005 study is a participant’s activity level coded as 1 = sedentary, 2 = active, and 3 = very active. Set up a frequency distribution and relative frequency distribution and construct a simple bar chart of activity level for the HEI–2005 participants during their first interview.

Solution Table 1.1 is a frequency distribution and a relative frequency distribution of the categorical variable “activity level.” Figure 1.1 is a bar chart of this data.

Table 1.1 HEI–2005 Participants’ Activity Level: First Interview

	<i>PARTICIPANTS</i>	<i>PERCENT</i>
Sedentary	2,183	48.9
Active	757	17.0
Very active	1,520	34.1
Total	4,460	100.0

Figure 1.1 HEI–2005 Participants’ Activity Level: First Interview (Simple Bar Chart)



Cross Tables

There are situations in which we need to describe relationships between categorical or ordinal variables. Market-research organizations describe attitudes toward products, measured on an ordinal scale, as a function of educational levels, social status measures, geographic areas, and other ordinal or categorical variables. Personnel departments study employee evaluation levels versus job classifications, educational levels, and other employee variables. Production analysts study relationships between departments or production lines and performance measures to determine reasons for product change, reasons for interruption of production, and quality of output. These situations are usually described by cross tables and pictured by component or cluster bar charts. These bar charts are useful extensions of the simple bar chart in Figure 1.1.

Cross Table

A **cross table**, sometimes called a crosstab or a contingency table, lists the number of observations for every combination of values for two categorical or ordinal variables. The combination of all possible intervals for the two variables defines the cells in a table. A cross table with r rows and c columns is referred to as an $r \times c$ cross table.

Example 1.2 illustrates the use of cross tables, component bar charts, and cluster bar charts to describe graphically two categorical variables from the HEI–2005 study.

Example 1.2 HEI-2005: Activity Level and Gender (Component and Cluster Bar Charts)

Consider again the data in Table 1.1. Sometimes a comparison of one variable (activity level) with another variable (such as gender) is of interest. Construct component and cluster bar charts that compare activity level and gender. Use the data coded daycode = 1 in the data file **HEI Cost Data Variable Subset**.

Solution Table 1.2 is a cross table of activity levels (1 = sedentary; 2 = active; and 3 = very active) and gender (0 = male; 1 = female) obtained from the first interview for HEI-2005 participants.

Table 1.2 HEI-2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

	MALES	FEMALES	TOTAL
Sedentary	957	1,226	2,183
Active	340	417	757
Very active	842	678	1,520
Total	2,139	2,321	4,460

Figure 1.2 displays this information in a *component* or *stacked bar chart*. Figure 1.3 is a *cluster*, or *side-by-side*, bar chart of the same data.

Figure 1.2 HEI-2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

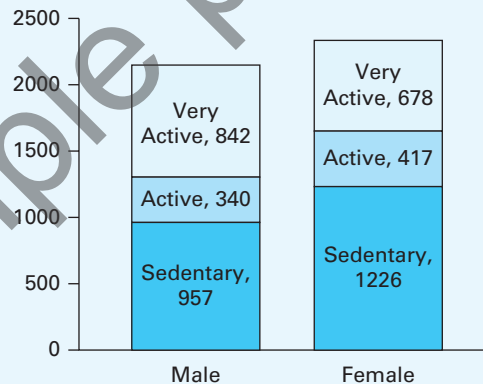
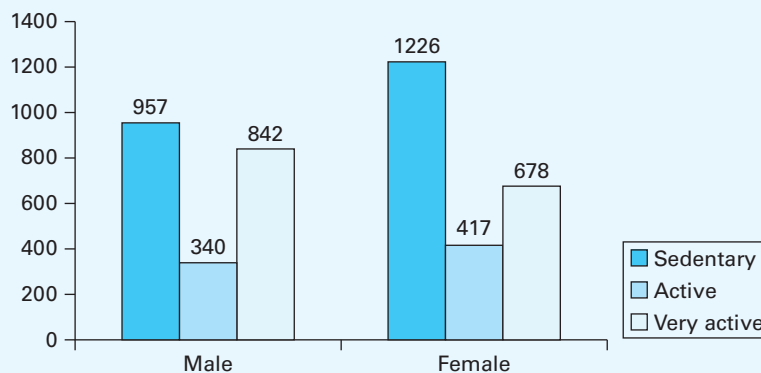


Figure 1.3 HEI-2005 Participants' Activity Level (First Interview) by Gender (Cluster Bar Chart)



Pie Charts

If we want to draw attention to the *proportion* of frequencies in each category, then we will probably use a **pie chart** to depict the division of a whole into its constituent parts. The circle (or “pie”) represents the total, and the segments (or “pieces of the pie”) cut from its center depict shares of that total. The pie chart is constructed so that the area of each segment is proportional to the corresponding frequency.

Example 1.3 Windows Wars: Market Shares (Pie Chart)

In the competition for market share by desktop Windows versions, StatCounter Global Stats, the research arm of StatCounter Stats reported that in January 2019, for the first time Windows 7 was not the lead operating system for Microsoft. However, we note that in January 2017 Windows 7's market share of 41.86% does not appear to be significantly different from Windows 10's market share of 42.78%. The data file **Windows Wars** contains market-share data for Win7, Win10, WinXP, WinVista, Win2003, and others for a 13-month period from January 2017 through January 2018 (StatCounter Global Stats Desktop Windows Version Market Share Worldwide, Jan 2017 - Jan 2018). Construct pie charts of the market shares for January 2017 and January 2018. In Section 1.4 we develop a graphical procedure to show the trend in market share over a period of time.

Solution Table 1.3 lists the market shares for Microsoft's various operating systems during January 2017 and January 2018. Figure 1.4 is a pie chart of the January 2017 market shares, and Figure 1.5 is a pie chart of the January 2018 market shares.

Table 1.3 Market Shares (Pie Chart)

	JANUARY 2017	JANUARY 2018
Win7	47.46	41.86
Win10	32.84	42.78
WinXP	5.72	3.36
WinVista	1.20	0.74
Win2003	0.08	0.07
Others	0.03	0.02

SOURCE: <http://gs.statcounter.com>

Figure 1.4 Windows Wars: January 2017 Market Share (Pie Chart)

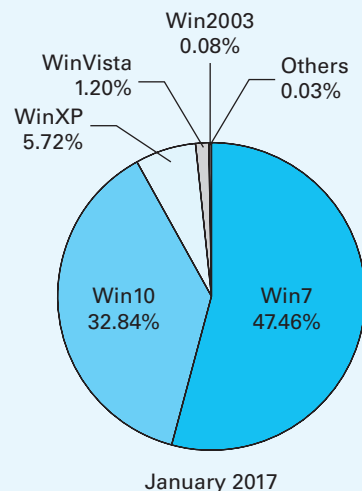
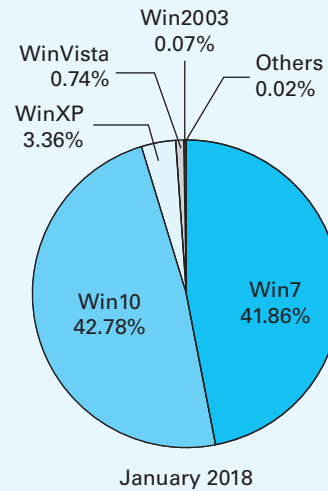


Figure 1.5 Windows Wars: January 2018 Market Share (Pie Chart)



Pareto Diagrams

Managers who need to identify major causes of problems and attempt to correct them quickly with a minimum cost frequently use a special bar chart known as a *Pareto diagram*. The Italian economist Vilfredo Pareto (1848–1923) noted that in most cases a small number of factors are responsible for most of the problems. We arrange the bars in a Pareto diagram from left to right to emphasize the most frequent causes of defects.

Pareto Diagram

A **Pareto diagram** is a bar chart that displays the frequency of defect causes. The bar at the left indicates the most frequent cause and the bars to the right indicate causes with decreasing frequencies. A Pareto diagram is used to separate the “vital few” from the “trivial many.”

Pareto’s result is applied to a wide variety of behavior over many systems. It is sometimes referred to as the 80–20 rule. A cereal manufacturer may find that most of the packaging errors are due to only a few causes. A student might think that 80% of the work on a group project was done by only 20% of the team members. The use of a Pareto diagram can also improve communication with employees or management and within production teams.

Example 1.4 illustrates the Pareto principle applied to a problem in a health insurance company.

Example 1.4 Insurance Claims Processing Errors (Pareto Diagram)

Analysis and payment of health care insurance claims is a complex process that can result in a number of incorrectly processed claims leading to an increase in staff time to obtain the correct information, an increase in costs, or a negative effect on customer relationships. A major health insurance company set a goal to reduce errors by 50%. Show how we would use Pareto analysis to help the company determine the most significant factors contributing to processing errors. The data are stored in the data file **Insurance**.

Solution The health insurance company conducted an intensive investigation of the entire claims’ submission and payment process. A team of key company personnel was selected from the claims processing, provider relations and marketing, internal auditing, data processing, and medical review departments. Based on their experience

and a review of the process, the team members finally agreed on a list of possible errors. Three of these errors (procedural and diagnostic, provider information, and patient information) are related to the submission process and must be checked by reviewing patient medical records in clinics and hospitals. Three possible errors (pricing schedules, contractual applications, and provider adjustments) are related to the processing of claims for payment within the insurance company office. The team also identified program and system errors.

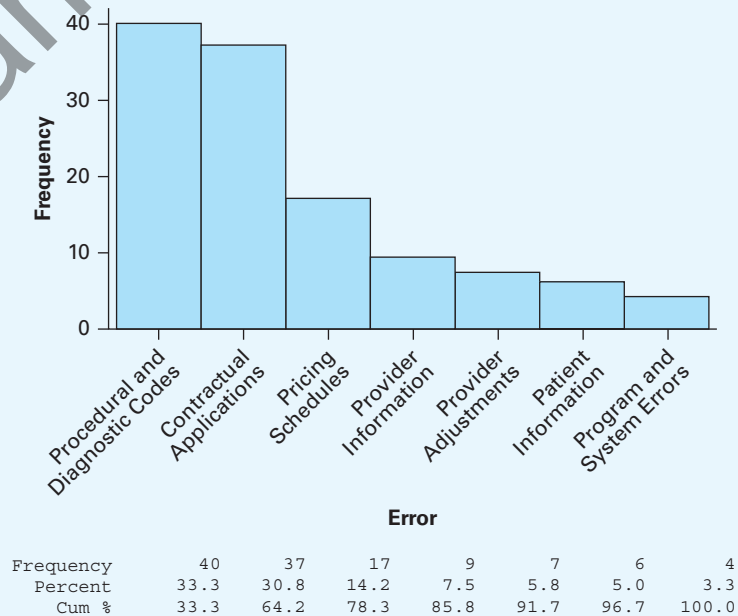
A complete audit of a random sample of 1,000 claims began with checking each claim against medical records in clinics and hospitals and then proceeded through the final payment stage. Claims with errors were separated, and the total number of errors of each type was recorded. If a claim had multiple errors, then each error was recorded. In this process many decisions were made concerning error definition. If a child were coded for a procedure typically used for adults and the computer processing system did not detect this, then this error was recorded as error 7 (Program and System Errors) and also as error 3 (Patient Information). If treatment for a sprain were coded as a fracture, this was recorded as error 1 (Procedural and Diagnostic Codes). Table 1.4 is a frequency distribution of the categories and the number of errors in each category.

Next, the team constructed the Pareto diagram in Figure 1.6.

Table 1.4 Errors in Health Care Claims Processing

CATEGORY	ERROR TYPE	FREQUENCY
1	Procedural and Diagnostic Codes	40
2	Provider Information	9
3	Patient Information	6
4	Pricing Schedules	17
5	Contractual Applications	37
6	Provider Adjustments	7
7	Program and System Errors	4

Figure 1.6 Errors in Health Care Claims Processing (Pareto Diagram)



From the Pareto diagram the analysts saw that error 1 (Procedural and Diagnostic Codes) and error 5 (Contractual Applications) were the major causes of error. The combination of errors 1, 5, and 4 (Pricing Schedules) resulted in nearly 80% of the errors. By examining the Pareto diagram in Figure 1.6, the analysts could quickly determine which causes should receive most of the problem correction effort. Pareto analysis separated the vital few causes from the trivial many.

Armed with this information, the team made a number of recommendations to reduce errors.

EXERCISES



Visit www.MyStatLab.com or www.pearsonglobaleditions.com to access the data files.

Basic Exercises

- 1.9 A university administrator requested a breakdown of travel expenses for faculty to attend various professional meetings. It was found that 41% of the travel expenses was spent for transportation costs, 20% was spent for lodging, 15% was spent for food, 8% for conference fees, and 16% was spent for conference registration fees; the remainder was spent for miscellaneous costs.
- Construct a pie chart.
 - Construct a bar chart.
- 1.10 A company has determined that there are seven possible defects for one of its product lines. Construct a Pareto diagram for the following defect frequencies:

Defect Code	A	B	C	D	E	F	G
Frequency	10	70	15	90	8	4	3

- 1.11 Bank clients were asked to indicate their level of satisfaction with the service provided by the bank's tellers. Responses from a random sample of customers were as follows: 67 were very satisfied, 53 were moderately satisfied, 8 had no opinion, 5 were moderately dissatisfied, and 3 were very dissatisfied.
- Construct a bar chart.
 - Construct a pie chart.
- 1.12 The supervisor of a factory conducted a survey on how long it takes employees to get to work based on the mode of transportation they use. The following table contains data from a random sample of 230 employees:

Mode of Transportation	Time		
	Less Than 15 Minutes	15 to Less Than 30 Minutes	30 to 45 Minutes
Bus	20	14	7
Train	11	8	27
Car	31	45	19
Walk	17	26	5

Graph the data with a component bar chart.

Application Exercises


- 1.13 Suppose that an estimate of a certain country's federal spending showed that 45% was for entitlements,

19% was for defense, 15% was for grants to various regions, 13% was for interest on debt, 5% was for other federal operations, and 3% was for deposit insurance. Construct a pie chart to show this information.


- 1.14 The European Central Bank (ECB) published a reliable and complete statistics of annual structural financial indicators for the banking sector in the European Union (EU). It comprises statistics on the number of employees of domestic credit institutions. The 2014 data highlighted that bank employees in the region declined by about 74,000. The following table gives a partial list of the number of employees of domestic credit institutions in certain countries (Table 1, *EU Structural Financial Indicators*, 2014):

Country	Number of Employees of Domestic Credit Institutions	
	2012	2013
Belgium	60,068	58,237
Bulgaria	33,527	32,756
Czech Republic	40,147	39,742
Denmark	44,900	36,367
Germany	659,100	655,600


SOURCE: Based on data from European Central Bank website, "EU Structural Financial Indicators: 2014," SSI Table, July 1, 2015

- Construct a bar chart on the number of employees of domestic credit institutions in 2012.
 - Construct a bar chart on the number of employees of domestic credit institutions in 2013.
 - Construct a bar chart to compare the number of employees of domestic credit institutions in 2012 to those in 2013.
- 1.15  A tennis coach kept a record of the most serious type of errors made by each player during a 1-week training camp. The data are stored in the data file **Tennis**.
- Construct a Pareto diagram of total errors committed by all players.
 - Construct a Pareto diagram of total errors committed by male players.
 - Construct a Pareto diagram of total errors committed by female players.
 - Construct a component bar chart showing type of error and gender of the player.

1.16 On what social media platform do you spend the most time? The responses from a random sample of 1,200 Internet users were Instagram, 382; Facebook, 226; LinkedIn, 350; Twitter, 85; Tumblr, 56; and Google+, 101. Describe the data graphically.

1.17  A random sample of 130 firms was asked whether they have been involved in merger and acquisition activities during the last two years. The researcher also noted the main sector each firm operates in. These data are contained in the file **M&A Survey**.

- Construct a cluster bar chart of the firms' sector and recent merger and acquisitions activities.
- Construct a pie chart of their sectors.

1.18  Part of the PISA project is a questionnaire on factors that influence study success. Enjoyment of reading is one such factor. The PISA questionnaire includes several questions about this, including the statement: For me, reading is a waste of time. The following table is a frequency distribution of responses from males and females to this statement in a sample of 500 PISA participants.

	Female	Male	Total
Strongly disagree	90	71	161
Agree	27	45	72
Disagree	96	107	203
Strongly agree	13	20	33
Total	226	243	469

- Use this data or the data file **PISA Sample** to construct a pie chart of the percent of males in each answer category.
- Use this data or the data file **PISA Sample** to construct a pie chart of the percent of females in each answer category.

1.19  Internet Explorer (IE) dropped below 50% of the worldwide market for the first time in September 2010 (StatCounter Global Stats Microsoft 2010). IE's worldwide market share continued to decrease over the next several months. Worldwide market share data from January 2010 through February 2011 for IE, Firefox, Chrome, Safari, and Opera are contained in the data file **Browser Wars**.

- Depict the worldwide market shares for February 2011 for the data contained in the data file **Browser Wars** using a pie chart.
- Use a pie chart to depict the current market shares for these Internet browsers (Source: gs.statcounter.com).
- Select a country or region from the list provided by StatCounter Global Stats and depict the market shares for the current time period with a pie chart (Source: gs.statcounter.com).

1.4 GRAPHS TO DESCRIBE TIME-SERIES DATA

Suppose that we take a random sample of 100 boxes of a new variety of cereal. If we collect our sample at one point in time and weigh each box, then the measurements obtained are known as *cross-sectional* data. However, we could collect and measure a random sample of 5 boxes every 15 minutes or 10 boxes every 20 minutes. Data measured at successive points in time are called *time-series* data. A graph of time-series data is called a *line chart* or *time-series plot*.

Line Chart (Time-Series Plot)

A **time series** is a set of measurements, ordered over time, on a particular quantity of interest. In a time series the sequence of the observations is important. A **line chart**, also called a **time-series plot**, is a series of data plotted at various time intervals. Measuring time along the horizontal axis and the numerical quantity of interest along the vertical axis yields a point on the graph for each observation. Joining points adjacent in time by straight lines produces a time-series plot.

Examples of time-series data include annual university enrollment, annual interest rates, the gross domestic product over a period of years (Example 1.5), daily closing prices for shares of common stock, daily exchange rates between various world currencies (Example 1.6), government receipts and expenditures over a period of years (Example 1.7), monthly product sales, quarterly corporate earnings, and social network weekly traffic (such as weekly number of new visitors) to a company's Web site (Example 1.8). In Chapter 16 we consider four components (trend, cyclical, seasonal, and irregular) that may affect the behavior of time-series data, and we present descriptive procedures for analyzing time-series data.