STATISTICS WITH R
Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company’s continued independence.
STATISTICS WITH R
A BEGINNER'S GUIDE

ROBERT STINEROCK

SAGE
## contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Terminology</td>
<td>2</td>
</tr>
<tr>
<td>Data: Qualitative or Quantitative</td>
<td>5</td>
</tr>
<tr>
<td>Data: Cross-Sectional or Longitudinal</td>
<td>7</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>9</td>
</tr>
<tr>
<td>Probability</td>
<td>9</td>
</tr>
<tr>
<td>Statistics: Estimation and Inference</td>
<td>10</td>
</tr>
<tr>
<td>Summary</td>
<td>11</td>
</tr>
<tr>
<td>Appendix: An Introduction to R</td>
<td>14</td>
</tr>
<tr>
<td>Our Approach to Writing This Book</td>
<td>26</td>
</tr>
</tbody>
</table>
learning objectives

1. Understand the meaning of the following terms: data, elements, variables, observations, dataset, population, sample, population parameters, and sample statistics.
2. Obtain an appreciation of what can be done with data that are quantitatively scaled versus those that are qualitatively scaled.
3. Learn about the difference between data that are cross-sectional and data that are longitudinal.
4. Grasp the connection between descriptive statistics, probability, and statistics.
5. Be able to distinguish between a population and a sample.
6. Appreciate the role a sample plays in statistical estimation and statistical inferences about a population parameter.
7. Learn a bit about R, how to download it, how to install it, and how to operate it in some basic ways.

BASIC TERMINOLOGY

Statistics can be seen as a methodological discipline, and, like all areas of methodology, statistics has its own basic set of terms for describing the components and methods that form its conceptual foundation. In this vein, we begin with nine definitions of key components upon which the most widely used statistical methods are based. Throughout the book, we build upon and extend these terms by introducing additional definitions.

Definition 1.1. Data. Data are the facts or measurements that are collected, analyzed, presented, and interpreted. As such, data comprise the raw material of all statistical analyses. Data are further categorized as either quantitative or qualitative—and either longitudinal or cross-sectional—distinctions which are developed in more detail in Sections 1.2 and 1.3. Methods of summarizing and displaying data are provided in Chapters 2 and 3.

As an example of data that might be found in a typical marketing study, consider Table 1.1. In this case, a market researcher contacts all individuals in a statistics class with the purpose of learning something about banking habits, and asks four questions. Do you bank online? What is your age? How many years of formal education do you have? What is your family status (single and never married, married without children, married with children, separated, divorced, or widowed)? The answers to these questions are entered in each of the columns; the column on the far left has the surname of each of the 10 students as well as the title of the instructor. The information about each of the 11 individuals is displayed across each row.

The next eight definitions are described in terms of the different aspects of Table 1.1.

Definition 1.2. Elements. An element is a unit of data which is represented as a set of attributes or measurements. Elements are the entities on which the data are collected.

Examples of elements include households, cities, companies, products, transactions, and persons. In Table 1.1, the elements consist of the 10 students and instructor comprising the statistics class.
Definition 1.3. Variables. A variable is an attribute of an element that may assume different values. Examples of variables are income, age, weight, occupation, industry classification, presence or absence of a disease, gender, and marital status. Variables are the measurements or the characteristics of interest of the elements, and they are what is usually analyzed using statistical methods. In Table 1.1, there are four variables, one for each question: (1) Do you bank online? (2) What is your age? (3) How many years of formal education do you have? (4) What is your family status?

Definition 1.4. Observations. An observation is the set of values on the variables for a single element. In Table 1.1, the observation corresponding to the instructor consists of the following set of values: yes, 41, 20, married with children. For the student named Adams, the observation is composed of these values: yes, 28, 16, single never married. These four pieces of information completely characterize the person who is the element in question for this data set.

Definition 1.5. Data Set. A “data set” is all the data, across all observations and all variables, collected for any given study. A single data item is referred to as a “datum.” Table 1.1 is an example of a data set.

To find the total number of data values in a data set, it is normally necessary only to multiply the number of elements by the number of variables. Exceptions to this rule of thumb occur, however, when there is missing information about any variable measurement for any elements. In this case, it is said that the data set contains missing values.

In Table 1.1 all 11 persons are represented by four variables each, and there are no missing values. Thus, by multiplying 11 times 4, it can be seen that this data set consists of 44 data values.

---

Table 1.1 A set of data: 11 observations on four variables for a study of banking habits

<table>
<thead>
<tr>
<th>Name</th>
<th>Bank online</th>
<th>Age</th>
<th>Years of education</th>
<th>Family status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>yes</td>
<td>28</td>
<td>16</td>
<td>single never married</td>
</tr>
<tr>
<td>Butler</td>
<td>yes</td>
<td>31</td>
<td>19</td>
<td>married with children</td>
</tr>
<tr>
<td>Danielson</td>
<td>no</td>
<td>22</td>
<td>16</td>
<td>single never married</td>
</tr>
<tr>
<td>Fitzgerald</td>
<td>yes</td>
<td>26</td>
<td>17</td>
<td>single never married</td>
</tr>
<tr>
<td>Johnson</td>
<td>no</td>
<td>23</td>
<td>16</td>
<td>single never married</td>
</tr>
<tr>
<td>Martinez</td>
<td>yes</td>
<td>27</td>
<td>18</td>
<td>married with children</td>
</tr>
<tr>
<td>Park</td>
<td>yes</td>
<td>24</td>
<td>16</td>
<td>single never married</td>
</tr>
<tr>
<td>Shah</td>
<td>yes</td>
<td>28</td>
<td>17</td>
<td>married with children</td>
</tr>
<tr>
<td>Smith</td>
<td>no</td>
<td>21</td>
<td>16</td>
<td>single never married</td>
</tr>
<tr>
<td>Taylor</td>
<td>yes</td>
<td>30</td>
<td>19</td>
<td>married without children</td>
</tr>
<tr>
<td>Instructor</td>
<td>yes</td>
<td>41</td>
<td>20</td>
<td>married with children</td>
</tr>
</tbody>
</table>
Definition 1.6. Population. A population consists of all the elements of interest in a study that have some quality (or qualities) in common. In the Table 1.1 data set, the population is the set of all persons enrolled in this particular statistics class: Adams, Butler, Danielson, Fitzgerald, Johnson, Martinez, Park, Shah, Smith, Taylor, and the Instructor. As another example, consider an opinion poll which seeks to predict the outcome of the election of the mayor of London. In this case, the population would consist of all elements having a characteristic in common: they are registered voters in London. Residents of Manchester or Hong Kong are not members of this population since they do not possess the quality of being registered voters in London. Similarly, in a study concerning the effectiveness of a drug intended to help patients control their level of cholesterol, the population would be made up of all persons known to be suffering from high levels of cholesterol.

Definition 1.7. Sample. A sample is a subset of the population that is selected for the purposes of the study. There are many possible unique samples that can be drawn from any given population. For example, for the population specified in Table 1.1, there are 165 samples consisting of 3 elements that can be selected from the population of 11 elements. One such sample would contain Butler, Taylor, and the Instructor; another would consist of Danielson, Johnson, and Smith. In Chapter 7, several methods of selecting samples from a population are discussed.

Often it is not practical or even possible to work with the entire population. Accordingly, most statistical analyses involve working with a sample rather than a population. The choice to use a sample instead of a population is a trade-off decision, however, and it involves sacrificing some richness and accuracy in the findings for the advantage of working with a more manageable, more readily available, less expensive set of data.

Definition 1.8. Population Parameters. Population parameters are the characteristics of interest in a study. But unless the entire population is available, the values of the population parameters are rarely (if ever) known with certainty. Population parameters are typically true but unknown; while they represent the unchanging truth about some characteristic of the population at a given moment in time, they are usually unobservable. Population parameters are often referred to simply as “parameters.” Since Table 1.1 displays the entire population of interest in this particular study, the parameters can be derived easily. For example, one parameter, the mean age, can be found by summing the age of all 11 individuals of the population and dividing the result by 11:

\[
\mu = \frac{28 + 31 + 22 + 26 + 23 + 27 + 24 + 28 + 21 + 30 + 41}{11} = \frac{301}{11} \approx 27.4.
\]

Thus, the mean age for the population, denoted by the Greek letter \( \mu \), is approximately 27.4 years. Chapter 3 includes a more complete discussion of population parameters such as the population mean.

Definition 1.9. Sample Statistics. Sample statistics are the characteristics of interest derived from a sample rather than a population. The values of these sample statistics vary from sample to sample. That is, when a sample is drawn, the sample statistic assumes one value; when another sample is selected, the sample statistic usually takes on a different value. Sample statistics are often referred to simply as “statistics.”
Returning to Table 1.1, the sample mean age of the sample consisting of Butler, Taylor, and the Instructor (see Definition 1.7 above) is

\[ \bar{x} = \frac{31 + 30 + 41}{3} = \frac{102}{3} = 34, \]

while the sample mean age for the other sample (made up of Danielson, Johnson, and Smith) is

\[ \bar{x} = \frac{22 + 23 + 21}{3} = \frac{66}{3} = 22. \]

Note that the values of the statistics from the two samples differ from one another as well as from the population mean of 27.4 years. This is not surprising since the population and the samples consist of entirely different elements. In later chapters, various methods for measuring and controlling this difference or error are introduced and developed in detail.

**DATA: QUALITATIVE OR QUANTITATIVE**

Now that we have defined the terms that form the basis of the statistical methods we encounter in subsequent chapters, we are in a position to draw the distinction between qualitative and quantitative data, and between data that are cross-sectional and longitudinal.

**Definition 1.10. Qualitative Data.** Qualitative data are measurements that can be categorized into one of several classifications. That is, qualitative data are typically labels or names that are used to identify a quality of each element in a data set. They are characteristics but not numerical measurements of anything. In general, qualitative data are either nominal-scaled or ordinal-scaled.

**Nominal-Scaled Data.** As an example of nominal-scaled data, consider the case of the Stevens Institute of Technology, which has three schools in which undergraduate students may enroll: Sciences, Engineering, and Business Technology. If the elements in a data set consist of the undergraduate students at Stevens, it is clear that each student can be classified in terms of the undergraduate school in which she is enrolled. For this reason, nominal-scaled data are sometimes referred to as “categorical” data: they indicate into which category the element should be placed. Note that this representation can be a non-numeric label such as Sciences, Engineering, and Business Technology. Alternatively, it is possible to employ a numeric representation such as Sciences = 1, Engineering = 2, and Business Technology = 3.

It is important to note, however, that since numerically represented, nominal-scaled data are not inherently quantitative—that is, they are not actually numbers but rather labels—we cannot perform quantitative operations on them in the usual ways. (There are some quantitative methods of a more advanced nature that make use of nominal-scaled data, but they are beyond the coverage of this book.) Just as it would make no sense to calculate the average of a list of telephone numbers, or the average of the numbers on the jerseys of a soccer team’s players, it is pointless to find the average of the schools in which students are enrolled. Mathematically, it can be done; the result, nevertheless, would have no meaning. Chapter 2 gives several descriptive methods of presenting and summarizing qualitative data.
Ordinal-Scaled Data. As an example of data that are ordinal-scaled, each element (or student) from the above data set can be further classified in terms of his class standing: freshman, sophomore, junior, or senior. This variable is ordinal-scaled because it has all the properties of a nominal-scaled variable—that is, each undergraduate year constitutes a category—and yet it contains some additional information beyond the simple classification. In this case, the rank of the data is meaningful: seniors normally have rank over juniors, juniors over sophomores, and sophomores over freshmen. As with nominal-scaled data, this information can be represented using the non-numeric label such as senior, junior, sophomore or freshman; it also can be represented using a number, for example senior = 4, junior = 3, and so on. However, we must remember that since numerically represented, ordinal-scaled data are not strictly quantitative or metric, we typically cannot perform quantitative operations on them as if they were. Even so, some statisticians consider ordinal-scaled data to be somewhat quantitative. In fact, ordinal-scaled data can be summarized by the use of statistical measures known as “nonparametric methods.”

Consider the rank order of the population of the world’s five most populous countries. Those countries are:

1. China
2. India
3. US
4. Indonesia
5. Brazil.

The difference between any two elements on an ordinal-scaled variable is not necessarily meaningful, however. If one considers only the ordinal-scaled data (the ranking) above, one might conclude that the US and India have comparable populations when in fact they do not (see Table 1.2).

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>1,350,000,000</td>
</tr>
<tr>
<td>India</td>
<td>1,200,000,000</td>
</tr>
<tr>
<td>US</td>
<td>314,000,000</td>
</tr>
<tr>
<td>Indonesia</td>
<td>246,000,000</td>
</tr>
<tr>
<td>Brazil</td>
<td>204,000,000</td>
</tr>
</tbody>
</table>

The population of India is almost four times as great as that of the US. The difference between these two data elements, India and the US, is not equal to the difference between the US and Indonesia, the fourth-ranked country. Indonesia’s population is clearly much closer to that of the US than is India’s. Ordinal-scaled data often mask these differences while the original data do not.

Definition 1.11. Quantitative Data. Quantitative data are those that can be characterized as metric (or quantitative); they report how many or how much. In general, quantitative data are one of two types: interval-scaled or ratio-scaled.
**Interval-Scaled Data.** A variable is interval-scaled if the data have all the above-mentioned properties of ordinal data, and if the interval between the values is expressed in terms of fixed units of measurement. An example of interval-scaled data is temperature. The difference between 50 and 51 degrees Celsius is the same as the difference between 75 and 76 degrees. Zero degrees Celsius, however, has no special meaning in that it does not imply the absence of the underlying construct being measured. Zero degrees Celsius is simply one degree warmer than –1°C, and one degree colder than +1°C. Returning to the example concerning the students, it is also possible to measure them in terms of their SAT admission test scores, an interval-scaled variable. Since interval-scaled data are quantitative, many quantitative operations can be performed on them. For example, it is possible to report meaningfully the average SAT score of engineering undergraduates.

**Ratio-Scaled Data.** A variable is ratio-scaled if the data have all the properties of interval data, and (additionally) if the ratio of two values is meaningful. Examples of ratio-scaled data abound: distance, height, weight, time, money, and so on. Note also that in the case of ratio-scaled data, zero now indicates the absence of the construct being measured: zero distance, zero height, zero weight, zero money, and zero time are universally understood concepts. To understand the difference between ratio-scaled and interval-scaled data, consider the matter of meaningful ratios: 50 kilometers is twice as great a distance as 25 kilometers, but 50°C is not twice as warm as 25°C. Finally, data that are ratio-scaled allow the widest range of quantitative operations because ratio-scaled data contain more information than any other kind of data.

The practical reason why it is important to draw the distinction between quantitative and qualitative data is that the statistical analysis that is appropriate depends on whether the data for the variable are quantitative or qualitative. In general, there are more alternatives for statistical analysis when the data are quantitative. Moreover, quantitative data indicate either how many, in which case the data are described as discrete, or how much, in which case the data are characterized as continuous. We return to develop these properties further in Chapters 5 and 6.

### DATA: CROSS-SECTIONAL OR LONGITUDINAL

A final important characteristic to consider is whether the data are cross-sectional or longitudinal.

**Definition 1.12. Cross-Sectional Data.** Cross-sectional data are collected at the same (or approximately the same) point in time. For example, Table 1.3 is a cross-sectional data set consisting of the population of each of the five counties comprising New York City on July 1, 2012.

<table>
<thead>
<tr>
<th>County</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronx</td>
<td>1,408,473</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>2,565,635</td>
</tr>
<tr>
<td>Manhattan</td>
<td>1,619,090</td>
</tr>
<tr>
<td>Queens</td>
<td>2,272,771</td>
</tr>
<tr>
<td>Staten Island</td>
<td>470,728</td>
</tr>
</tbody>
</table>

**Table 1.3** Population of the five counties of New York City: July 1, 2012
Note that while the population of each of the five counties is reported on July 1, 2012, it is clear that the exercise of counting such a large number of people no doubt took considerable time. Put another way, it is misleading for anyone to think that the population of Brooklyn was exactly 2,565,635 on July 1, 2012. All that number really represents is the best estimate by the Census Bureau of what the actual population figure is likely to be. The real population is bound to be some other number, either larger or smaller.

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1900</td>
<td>1,166,820</td>
</tr>
<tr>
<td>1910</td>
<td>1,634,510</td>
</tr>
<tr>
<td>1920</td>
<td>2,018,560</td>
</tr>
<tr>
<td>1930</td>
<td>2,560,010</td>
</tr>
<tr>
<td>1940</td>
<td>2,698,285</td>
</tr>
<tr>
<td>1950</td>
<td>2,738,175</td>
</tr>
<tr>
<td>1960</td>
<td>2,627,319</td>
</tr>
<tr>
<td>1970</td>
<td>2,602,012</td>
</tr>
<tr>
<td>1980</td>
<td>2,231,028</td>
</tr>
<tr>
<td>1990</td>
<td>2,300,664</td>
</tr>
<tr>
<td>2000</td>
<td>2,465,326</td>
</tr>
<tr>
<td>2010</td>
<td>2,504,700</td>
</tr>
</tbody>
</table>

**Table 1.4** Population of Kings County (Brooklyn), New York: 1900–2010

**Definition 1.13. Longitudinal Data.** Longitudinal data, also known as time-series data, are collected over several time periods. As an example of longitudinal data, consider the population of Brooklyn, New York, for the years 1900–2010 (Table 1.4). In this case, we are measuring the same quantity repeatedly, over time.

Some data sets are both cross-sectional and longitudinal. For example, consumer products companies and market research firms often test-market new-to-the-world, frequently purchased, packaged goods, such as salty snacks or personal care items, before readying the product for a nationwide roll-out. They might do this by monitoring the monthly purchases of 10,000 continuously reporting households in a selected test market city, like Minneapolis–St. Paul, Minnesota, who have agreed to participate in a consumer panel over a 12-month period. The data set then would be structured something like the arrangement depicted in Table 1.5 where each cell would contain the number of purchases made by each household of the product in question over each of the 12 months of the year.
Clearly, this data set is both cross-sectional and longitudinal in nature. Note that there are 10,000 elements or households, and that each element is represented 12 times over the course of a year. As indicated above, the measurement would be the number of units purchased by the household of the new product over the period of one month. Thus, assuming no missing data items, the total number of data values in this data set is 10,000 times 12, or 120,000.

Having discussed some of the most important characteristics associated with data, data sets, and scales of measurement, we now consider the difference between descriptive statistics (covering the first three chapters of this book), probability (treated in Chapters 4–6), and statistical estimation and inference (our main focus in Chapter 7 and beyond).

### DESCRIPTIVE STATISTICS

**Definition 1.14. Descriptive Statistics.** Descriptive statistics consists of a wide variety of methods of organizing, summarizing, and reporting data. In general, these methods are classified as either tabular, graphical, or numerical methods, and are developed in detail in Chapters 2 and 3.

While descriptive statistics consists of the three aforementioned data summarization approaches, we first look at tabular methods that can be used to present data in tables (hence the name “tabular”). We then move on to the graphical methods that often can be thought of as pictures (such as pie charts, bar graphs, and histograms) of the information organized in a table. Tabular and graphical methods are covered in Chapter 2. Finally, in Chapter 3, we introduce and discuss a number of the most useful numerical methods that provide additional procedures for summarizing data.

Regardless of the particular method of descriptive statistics we employ, the objective in using them is to provide insights about the data that cannot be easily or quickly obtained by a cursory examination of the original data.

### PROBABILITY

**Definition 1.15. Probability.** We encounter probability problems whenever we draw a sample from a population, and then attempt to estimate the probability that the sample will have
certain characteristics. At least for the type of probability problems encountered in this book, the primary focus is on the sample rather than the population. Here are two examples of this type of probability problem.

1. If a single card is drawn from a deck of 52 playing cards, what is the probability the card will be an ace?
   In this instance, the characteristics of the population are well known: there are four aces in a normal deck of 52 playing cards. The probability that the sample (consisting of only one playing card) has a certain characteristic (it is an ace) is $\frac{4}{52}$.

2. The campaign manager for a certain political candidate knows that 75% of visitors to her website are men and 25% are women. In a sample of the next 100 visitors to the website, what is the probability that at least 80% will be men?
   Here the population is comprised of all visitors to the website, and the relevant population parameter is known at the outset: the percentage of male visitors to the website is 75%. The sample consists of the next 100 visitors to the website, and the issue concerns what the sample will most likely look like. In other words, in light of what is known about the population, what will the sample most probably be like? The emphasis here is on the characteristic of the sample, not the population.

---

**Definition 1.16. Statistics.** We encounter problems of statistics whenever we select a sample from a population and want to draw conclusions about the population’s properties based on what we learn from the sample. Unlike in the case of the probability problem, the population parameters are unknown. Here is an example of a statistics problem.

The leadership of a political party recently surveyed 1200 registered party members with the purpose of determining how receptive they would be to the party nominating a particular candidate for an open seat in the Senate. Once the survey results had been collected and analyzed, the leadership announced that because 624 of the 1200 survey respondents, or 52%, expressed their approval of the candidate, “The majority of our members favor our running this particular candidate for the Senate seat.” How confident can we be that the leadership is correct in this conclusion? Is it possible that they have drawn the wrong conclusion about the population from the sample data? How strong is the evidence supporting their conclusion?

In this example, the members of the political party comprise the population, and the 1200 people whose preferences were sought constitute the sample. We note, however, the crucial difference between probability and statistics: in probability, we draw a sample from a population with known parameters, and make statements about what the sample might look like. By contrast, in statistics, we draw a sample with the purpose of estimating or inferring what the population characteristics might be. In other words, in statistics problems, we do not know the characteristics of the population, but we would like to know them. Accordingly, we draw a sample from that population in order to estimate or infer statistically what those population characteristics might be. The methods of statistics comprise the lion’s share of material in this book.
There are several approaches we can take when using statistical methods, but in this book we will make abundant use of statistical estimation and statistical inference. As we will see in Chapters 7 and 8, estimation consists of both point estimation and confidence interval estimation, and the motivation of each is to puzzle out how small or how large something may be.

As an example of a research question for which we might employ a statistical estimation method, consider the following: in the 2012 US Presidential election, what percentage of women and men voted for President Obama?

Note that the research issue in question involves developing an understanding of how large or small something might be. In this example, that something is the percentage vote received by a political candidate for public office.

In Chapter 9, we learn that statistical inference involves testing formal hypotheses for the purpose of drawing conclusions or inferences. As an example of a research hypothesis which we might wish to test using the methods of statistical inference, consider the following statement: in the 2012 US Presidential election, a higher percentage of women voted for President Obama than of men.

A research hypothesis is a formal statement that can be tested for the purpose of determining if it is true or false. Because the statement above (concerning the relatively greater support that President Obama enjoyed among women over men) is a claim we could confront with real data, statistical inference would be a research methodology we might use when considering whether the statement is true.

On the other hand, a statement such as “Bill Clinton was a better president than Barack Obama has been” is not a claim that can be tested using empirically grounded data: there are none. We sometimes hear that people try to make these types of comparisons—historians and journalists among them—but in doing this they inevitably resort to subjectively chosen criteria on which the presidents might be measured. Naturally, because there are no universally agreed-upon standards for what criteria are used, and what are not used, different people select different criteria for inclusion. Many people will have different views on the question, but there is no objective evidence pointing either way. Therefore, we would say that this type of hypothesis is not one that can be tested for the purpose of establishing its truth.

Clearly, both statistical approaches—estimation and inference—can be used to get at the same research question. Which we use depends on what specifically we want to know.

**SUMMARY**

There is a reason why descriptive statistics is covered first, probability second, and statistical estimation and inference last. Broadly speaking, the three areas are sequenced this way because they build on one another. The methods used to solve probability problems could not be effectively applied by someone unfamiliar with some of the concepts from descriptive statistics, such as the mean and the standard deviation (discussed in Chapter 3). Furthermore, problems involving statistical estimation and inference (beginning with Chapter 8) cannot be solved, or perhaps even understood, by someone who is unaware of how to use probability distributions, a topic at the very heart of probability (introduced in Chapters 5 and 6).
These three areas form the content and sequencing of the course material. The first area, descriptive statistics, introduces and develops methodologies for organizing, summarizing, and presenting data that many students and readers use in their work lives. The second area, probability, is approached from several vantage points. While the central ideas of probability are inherently interesting to many students, they are often found to be challenging and abstract. We do our level best to reduce the ambiguity surrounding some of the more difficult ideas and to make almost everything accessible to everyone in a language that is conversational and user-friendly. The third area, statistical inference, is also rather challenging for many students. Like probability, however, it can be seen as formalized common sense. Its usefulness and power, however, cannot be overstated. Indeed, it comprises numerous tools and methods that many students and readers find of immediate relevance in both their coursework and workplace.

**definitions**

**Cross-Sectional Data** Cross-sectional data are collected at the same (or approximately the same) point in time.

**Data** Data consist of facts or measurements that we collect, analyze, and interpret; they comprise the raw material of statistical analysis, and may be either quantitative or qualitative.

**Data Set** A data set refers to all the data, across observations and variables, collected for any given investigative purpose. A single data item is referred to as a “datum.”

**Descriptive Statistics** Descriptive statistics consists of a wide variety of methods of organizing, summarizing, reporting, and interpreting data characteristics, both qualitative and quantitative. In general, these methods are classified as tabular, graphical, or numerical methods.

(a) **Tabular Methods** A collection of data summarization procedures designed to display data in a table format.

(b) **Graphical Methods** A group of pictorial methods for presenting in a pictorial format the information reported in a table.

(c) **Numerical Methods** A collection of numerical measures which summarize various characteristics of a set of data such as the measures of central tendency, location, dispersion, and association.

**Elements** An element is a unit of data which is represented as a set of attributes or measurements. Elements are the entities (e.g., the households, cities, companies, products, transactions, and persons) on which the data are collected.

**Longitudinal Data** Longitudinal data, also known as time-series data, are collected over several time periods.

**Observations** An observation is the set of values of the variables for an element. Observations represent the set of measurements for a single data entity.

**Population** A population consists of all the entities of interest in an investigative study.

**Population Parameters** These are the characteristics or qualities of the population in terms of the variables of interest.

**Probability** We encounter probability problems whenever we draw a sample from a population with known parameters, and then attempt to estimate the probability that the sample will have certain characteristics.
Qualitative Data Qualitative data are those data items that can be classified into categories or classes. That is, qualitative data are typically labels or names that are used to identify a characteristic of each element in a data set. In this book, we will consider qualitative data as one of two types, nominal-scaled or ordinal-scaled.

(a) **Nominal-Scaled Data** Nominal-scaled data are sometimes referred to as "categorical" data: they indicate into which category the data element should be placed. Such data may be represented either numerically or non-numerically.

(b) **Ordinal-Scaled Data** Ordinal-scaled data have all the properties of nominal-scaled data but, in addition, have the quality that order is meaningful. Such data may be represented either numerically or non-numerically.

Quantitative Data Quantitative data are those data items which can be characterized as metric (or quantitative), and which report how many or how much. In this book, we will consider quantitative data as one of two types, interval-scaled or ratio-scaled.

(a) **Interval-Scaled Data** Interval-scaled data have all the above-mentioned properties of ordinal-scaled data but, in addition, have the additional quality that the interval between the values is expressed in terms of fixed units of measurement.

(b) **Ratio-Scaled Data** Ratio-scaled data have all the above-mentioned properties of interval-scaled data, plus the additional quality that the ratio of two values is meaningful.

Sample The portion or subset of the population we select for the purposes of our study.

Sample Statistics These are the characteristics or qualities of the sample in terms of the variables of interest.

Statistics We encounter problems of statistics whenever we select a sample from a population and want to draw conclusions about the population's properties based on what we learn from the sample. Unlike in the case of a probability problem, the population parameters are unknown.

(a) **Statistical Estimation** Statistical estimation consists of both point estimation and confidence interval estimation, and the motivation of each is to puzzle out how small or how large something may be.

(b) **Statistical Inference** Statistical inference involves testing formal hypotheses for the purpose of drawing conclusions or inferences.

Variables A variable is an attribute of an element that may assume different values. Examples of variables are income, age, weight, occupation, industry, disease, gender, and marital status. Variables are the measurements or the characteristics of interest for the elements, and they are what we usually analyze using statistical methods.

---

**R functions**

- `c()` The concatenate function combines elements within the parentheses, called arguments, into a single entity, called a "vector."
- `data.frame()` Combines elements within the parentheses into a single entity, called a "data frame." The data frame is one of the most important structures used in connection with R statistical methods.
- `head()` Shows the first six lines (by default) of a data object. To see `n` lines instead of six lines, include `n` in the argument: `head(,n).
- `mean()` Reports the mean of the data object.

(Continued)
str() Displays the internal structure of the data object (e.g., vector, data frame.)
tail() Shows the last six lines (by default) of a data object. To see n lines instead of six lines, include n in the argument: tail(n).

**summary()** Reports the minimum and maximum values plus the median, first and third quartiles, as well as the mean.

---

**APPENDIX: AN INTRODUCTION TO R**

1. Where to Get It and How to Install It

Before we embark on our statistical and probabilistic journey, let us describe R (and RStudio), the statistical programming language we will be using. We describe how to download and install it, how to read data into it, how to write data and analytic results out of it, and how to perform some elementary computations.

Before we can work with R, we have to install it. R is provided free of charge, and can be downloaded from the following website:

http://www.r-project.org/

Once you get to this location, note the area in the left-hand margin named “Download.” Just below this heading, click on the CRAN link. When you do this, you are presented with the list of hosts called CRAN mirrors (arranged by geographic location) which contain identical R content. Click on the location closest to your own, because the download will typically be faster. For example, since we live in New York City, we used the link to the National Institutes of Health in Bethesda, Maryland: http://watson.nci.nih.gov/cran mirror/

Depending on whether you use Windows or Mac OS, the next steps differ slightly.

(a) **Windows.** If you are a Windows user:

   (i) Click on Download R for Windows.
   (ii) When the next screen appears, click on base.
   (iii) Once you see the next page, click on the link Download R 3.2.3 for Windows. (The number 3.2.3 is associated only with the current version of R; this number changes as updated versions of R are made available.)
   (iv) Save the file R-3.2.3-win.exe, double-click, and follow the onscreen instructions. Assuming the installation has been successful, there will be an entry in the All Programs area of the Start Menu. By right-clicking on R 3.2.3, you may place a shortcut R icon on the desktop. To start up R, simply click on the icon in the usual way.

(b) **Mac OS X.** If you are a Mac user:

   (i) Click on the Download R for (Mac) OS X link.
   (ii) Click on the latest pkg file. As of this writing, the latest file is R-3.2.3.pkg. (The number will change, however, as updated versions of R are made available.) Once downloaded, the file will appear in the Downloads area.
(iii) Move R-3.2.3.pkg from the Downloads area to the Applications area and double-click on its entry (icon). Follow the instructions for full installation.

(iv) The installation process should create an R entry in the Application area. Drag the R entry to the desktop for an R shortcut. To start up R, simply click on the icon in the usual way.

You will also want to install RStudio, an integrated development environment, once you have downloaded the R statistical programming package itself. RStudio can be thought of as a shell in which R operates. It simplifies many basic functions and facilitates the running of R itself, thus making R easier to interact with and more user-friendly. RStudio is provided free of charge and can be downloaded from:

http://rstudio.com

(a) When you get to this location, click Download RStudio.
(b) When the next screen appears, select Desktop.
(c) At the next screen, select Download RStudio Desktop.
(d) Depending on whether you use Windows or a Mac, follow the above steps (for installing R). The installation steps are the same for RStudio as they are for R.

Make sure that you have installed R before installing RStudio. RStudio by itself does not include the R program. Once both R and RStudio have been successfully downloaded and installed, a double-click on the RStudio icon launches the program (Figure 1.1) and you are ready to start using R.
There are four panes or windows in the RStudio interface.

(a) The **Source** window, located in the upper left-hand quadrant, is where we write code for our programs. It includes an intuitive, feature-rich text editor that makes it easy to run, save, and access programs.

(b) The **Console** is in the lower left-hand window and is the location where interactive work is done. The RStudio Console is the same as the Console in R.

(c) As the name implies, the tabbed **Environment/History** pane in the upper right-hand area is where the data sets and variables are listed and described (in Environment), and where the history of R commands is displayed (in History).

(d) The tabbed **Files/Plots/Packages/Help/Viewer** pane in the lower right-hand area is a repository of helpful tools which we discuss in subsequent chapters.

We point out that it is not absolutely necessary to run R through RStudio. While RStudio is a relatively new addition to R, the R package itself has been around for a couple of decades, and people ran (and still run) R very well without RStudio. However, since RStudio makes interacting with R more straightforwardly intuitive, we will use it throughout this book.

To exit RStudio, enter `q()` at the R prompt and hit the enter key. The R prompt can be seen on the command line of the Console, the lower left-hand pane. It is the greater-than symbol `>`, and it is the location where we write our R commands when working in an interactive mode.

## 2. Basic File Management and Working with Data

(a) **Installing the introstats package that accompanies this book**

This book comes with an R package called **introstats** which includes the data sets used with the examples and exercises for the chapters. Packages are extensions to R that implement various methods and provide package-specific data sets, and we will encounter more of them in later chapters. In order to access the data sets used in this book, you will want to install the package by entering the following two commands at the R prompt in the Console.

```r
install.packages("devtools")
devtools::install_github("ericnovik/introstats")
```

Once these two steps have been executed, it is necessary to enter one final command at the R prompt in the Console each time we begin an R session requiring one of the data sets:

```r
library(introstats)
```

This option for importing data into the R environment (referred to as the R Workspace) is used for examples and exercises throughout the book. But because this approach cannot be taken when we are using R in any other context, two additional methods are introduced below (see (b) and (c)): entering the data directly in the R Console, and importing the data from a spreadsheet.
How to get data into R: entering the data directly in the R Console

Even though it is an abstract characterization, R is often referred to as an \textit{object-based} programming language. In R, two of the most common data objects are vectors and data frames. One can think of vectors as a single column in a spreadsheet that contains elements of the same type. One example of a vector might consist of the following elements: 1, 3, 5, 7. Another vector could be made up of these items: “a,” “d,” “t.” However, 1, 5, “d,” “t” do not make up a valid vector since the elements are not of the same type.

How to use R to analyze a data set if it is organized in a vector

Analyzing data that are organized in vectors is convenient when we have a small set of data and want to analyze it quickly. Consider the data in Table 1.6, which consist of the US unemployment rates (percent) for the 12 months from January to December 2012.

\begin{table}[h]
\centering
\caption{Official monthly US unemployment rate (%) for 2012.}
\begin{tabular}{lrrrrrrrrrr}
\hline
\textbf{Rate} & 8.3 & 8.3 & 8.1 & 8.2 & 8.2 & 8.2 & 8.1 & 7.8 & 7.9 & 7.8 & 7.8 \\
\hline
\end{tabular}
\label{table:unemployment_rates}
\end{table}


In this book, we write out the R code in a shaded block similar to the one below. We refer to these as R blocks. When explanation is helpful, each line of code in the R block is accompanied by a comment. When there is more than one comment in an R block, the comments are numbered. Finally, whenever the code requires more detailed explanation, it is written just after the R block itself in numbered comments. This is done with an eye to keeping a more clutter-free R block environment.

As a naming convention, we identify an object by three elements: a letter indicating whether the object is associated with a chapter (C) or its exercises (E), and two numbers, the first indicating the chapter number, the second the object number. For example, the first object of Chapter 1 is C1_1; the first object in Chapter 2 is C2_1; and the second object in Chapter 3 is C3_2. Similarly, E1_1 names the first object of the Chapter 1 exercises, while E3_4 names the fourth object in the Chapter 3 exercises. When the data objects are named unambiguously, there is less chance of confusion.

Consider the following R block in which we create an object C1_1 in the R Workspace, check its accuracy, and find the mean of C1_1.

\begin{verbatim}
#Comment1. Read data into a vector named C1_1.
C1_1 <- c(8.3, 8.3, 8.2, 8.1, 8.2, 8.2, 8.2, 8.1, 7.8, 7.9, 7.8, 7.8)

#Comment2. To confirm that C1_1 contains what we want.
C1_1
\end{verbatim}
## Comment3. Find the mean of C1_1.
mean(C1_1)
## [1] 8.075

Some explanation of these R commands is helpful.

A To enter the Table 1.6 data into the R Workspace, place the cursor at the command line in the Console next to the R prompt >, write out the following line of code, and enter.

C1_1 <- c(8.3, 8.3, 8.2, 8.1, 8.2, 8.2, 8.2, 8.1, 7.8, 7.9, 7.8, 7.8)

The c() expression is referred to as the “concatenate” function and it combines all the elements in the parentheses into a vector. The symbol <- is known as the assignment operator, and it assigns whatever is on its right-hand side to whatever is on its left-hand side, in this case C1_1. Even though it is constructed from two symbols, the less-than sign and a hyphen, we should think of <- as a single symbol. Finally, the direction of the assignment operator -> can be reversed so that whatever is on its left-hand side is stored in whatever is on its right-hand side.

B To make certain that the vector C1_1 contains the data items we want, enter the vector name C1_1 at the R prompt >, hit enter, and visually inspect the result. In this case, we need only confirm that all the data values from Table 1.6 are indeed stored in the vector named C1_1.

C Once we are satisfied that C1_1 includes the desired data values, we calculate their mean with the function mean().

The R function c() is the first of many we will use; mean() is another. The values that are entered within the parentheses of the function are referred to as “arguments,” and each R function has rules governing how its arguments are to be specified. If a function has two or more arguments, they are separated by commas. A third function is data.frame(), and it can be used to assign a heading (or variable name) to a vector; it can also be used to bring together more than one vector to form a matrix-like object.

(ii) How to use R to analyze a data set if it is organized in a data frame

Working with data frames (rather than vectors) is more common when performing statistical analysis, particularly when one desires to incorporate variable names in the objects. (C1_1 did not include a variable name.) Once the set of data extends to more than one variable, it usually becomes necessary to include variable names. Below, the function data.frame() is used to include a variable name rate (for % unemployment rate).

#Comment1. Use function data.frame() to read C1_1 into C1_2.
C1_2 <- data.frame(rate = C1_1)

#Comment2. Examine contents of first 3 rows of data in C1_2.
head(C1_2,3)
INTRODUCTION AND R INSTRUCTIONS

## rate
## 1 8.3
## 2 8.3
## 3 8.2

#Comment3. Find the mean of the variable named rate.
mean(C1_2$rate)
## [1] 8.075

The three R commands above require additional explanation.

A To add the variable name rate to vector C1_1, use the function data.frame(rate = C1_1) and assign the result to a data frame C1_2.

B Use the function head(C1_2,3) to check that C1_2 includes the variable name rate. Note that the second argument of head(C1_2,3) (which is 3) specifies that only the first three rows C1_2 should be reported. Remember: if a function has two or more arguments, they are separated by commas.

C Using the function mean(C1_2$rate), find the mean value of rate. Note that C1_2 and the variable name rate are separated by the dollar sign $.

The function data.frame can also be used to create a single object with two or more variables. For example, suppose that six students are measured on three variables: grade point average (gpa); math Scholastic Aptitude Test (SAT) score (satm); and verbal SAT score (satv).

#Comment1. Read data into 3 vectors: gpa, satm, and satv.
gpa <- c(2.7, 3.5, 3.7, 3.3, 3.6, 3.0)
satm <- c(450, 560, 700, 620, 640, 570)
satv <- c(540, 650, 700, 720, 540, 750)

#Comment2. Use function data.frame() to read 3 vectors into
#C1_3. Name each vector GPA, SATM, and SATV, respectively.
C1_3 <- data.frame(GPA = gpa,SATM = satm,SATV = satv)

#Comment3. Examine the contents of C1_3.
C1_3
## GPA SATM SATV
## 1 2.7 450 540
## 2 3.5 560 650
## 3 3.7 700 700
## 4 3.3 620 720
## 5 3.6 640 540
## 6 3.0 570 750

#Comment4. Find the mean of the variable named GPA.
mean(C1_3$GPA)
## [1] 3.3
The data frame C1_3 now contains all three variables—GPA, SATM, and SATV. The mean of GPA is 3.3; the mean of the other variables is found in the same manner, taking care to separate C1_3 from the variable names (SATM or SATV) with a dollar sign $.

(c) How to get data into R: importing the data from a spreadsheet

While the method described above works well with small data sets, we do not recommend using it for larger data sets. There are many ways to import data into the R Workspace; one of the more widely used methods involves carrying out the following two steps: (1) convert a spreadsheet data file to a comma-separated values (csv) file; (2) import the csv file into the R Workspace. Spreadsheets are widely used, and most have the ability to convert data to csv files. Here is how this can be done if your spreadsheet software is Excel or Numbers.

1 Convert a spreadsheet data file to a csv file: first Excel, then Numbers.
   A Once the Excel file is loaded, select the File drop-down menu and click on the Save As... option.
   B Once the Save As dialog box appears, enter the desired filename in the Save As window; change from Excel Workbook to Comma Separated Value (.csv) in the Format window. Click on Save.

   If you use a Mac and your preferred spreadsheet software is Numbers:
   A Once the Numbers file is loaded, select the File drop-down menu and click on Export option.
   B Once the Export dialog box appears, select CSV and click on Next. When the Save As dialog box appears, enter the desired filename in the Save As window. Click on Export.

2 Import the csv file into the R Workspace.

Now that your spreadsheet data file has been converted to the csv format, the second (and final) step is to import it into the R Workspace. There are several ways of doing this but we believe that the best method exploits a feature of RStudio known as Project. RStudio Project helps us organize and automate the migration of our data files between the various folders that might reside in the Documents area of our hard drive and the R Workspace where the actual data analysis must be done.

To demonstrate how to use Project, we first recreate the data sets above, C1_2 and C1_3, but this time using a spreadsheet such as Excel or Numbers, making sure to save both in the csv format (see step 1 above). In this demonstration, we save C1_2.csv and C1_3.csv in folders named Economics and Advising, respectively, for course material for two university classes, Economics and Educational Psychology. The Economics and Advising folders are organized within a folder named Class Assignments. When successfully executed, the tabbed window of the lower right-hand area of the RStudio interface—Files/Plots/Packages/Help/Viewer—provides confirmation (see Figures 1.2 and 1.3). Figure 1.2 shows that C1_2.csv is located in the Economics folder which itself is organized inside another folder, Class Assignments. Similarly, Figure 1.3 reports the path address for the other data set, C1_3.csv.
Suppose we wish to create two projects, one each for our two courses, because we want to organize all of our work—data files, R programs, and graphical illustrations such as plots, histograms, and so on—in its own designated space. Here is how RStudio Project helps us do that. After
opening RStudio, click on File and select New Project... from the drop-down menu. When the New Project window opens, click on Existing Directory. (Note that “directory” and “folder” are two terms that refer to the same thing.) When the next window appears, identify the target directory in the Project working directory field by clicking on the Browse button and navigating to the desired directory. In this case, the desired directory is the Advising folder. Click on the Create Project button (see Figure 1.4).

![Figure 1.4 Final step of creating the Advising project](image)

That the Advising project has been successfully created is confirmed in two ways: Advising is now indicated in the upper right-hand corner, next to the RStudio icon; and Advising, as reported in the lower right-hand panel, now includes two new files, .Rhistory and Advising. Rproj, along with the data set we want to import into the R Workspace, C1_3.csv (see Figure 1.5).

![Figure 1.5 The Advising project is now created](image)
In this demonstration, the next step requires a simple repetition of the above procedure to create the Economics Project. Everything is executed in the same way except that at Figure 1.4, we specify Economics rather than Advising. Assuming the execution is successful, the upper right-hand corner of the RStudio interface will now show the two project names, Advising and Economics. When the contents of Economics is examined, we will once again find two new files, .Rhistory and Economics.Rproj, along with the data set we want to import into the R Workspace, C1_2.csv. Both Economics and Advising projects can be thought of as ordinary folders residing in the Class Assignments folder.

Remember that our purpose here is to demonstrate how to import, export, and organize our files. To this end, we now (1) import the data set C1_3.csv to the R Workspace, (2) perform a simple statistical summarization procedure, and (3) export the output back to the Advising folder.

We open the Advising project by clicking on the down arrow next to the RStudio icon in the upper right-hand corner of the RStudio interface and then navigating to Advising (see Figure 1.5). Once in the Advising project, we enter C1_3 <- read.csv("C1_3.csv") at the prompt in the Console to import C1_3.csv to the Workspace. Note that the argument of the function read.csv() does not specify the full path address (an advantage of using RStudio) but does require that the filename itself be expressed within quotes, either single or double: "C1_3.csv" (see Figure 1.6).

As with the first time we used C1_3 above, we only need to enter C1_3 at the prompt to examine the contents of the new object (see Figure 1.6). Now that C1_3 resides in the Workspace, we are in a position to perform our statistical analysis.

We use the function summary(C1_3) to provide the minimum, maximum, mean, median, and first and third quartiles for all three variables in C1_3. Once the results are assigned to an object named Descriptives, we review the contents by entering the object name at the prompt (see Figure 1.7).
The expression `write.csv()` directs R to export the object `Descriptives` from the R Workspace to the Advising project folder and name it `SummaryTable.csv`.

**Figure 1.7** Summary statistics for C1_3.csv assigned to Descriptives

The `summary()` function is only one among many that we might use in this instance. The important point to bear in mind is that our purpose here is to demonstrate how to organize, import, and analyze our files.

(d) How to export data and analysis results out of R

Finally, we show how to export an object from the R Workspace to an external folder that might reside in the Documents area on the hard drive, a capability that is useful when one wishes to share data or analysis files with colleagues. In this instance, we use the function `write.csv()` to send the object we created above, Descriptives, back to the Advising project folder (see Figure 1.8).

**Figure 1.8** Exporting the object Descriptives to the Advising project folder

```r
> Descriptives <- summary(C1_3)
>
> Descriptives

<table>
<thead>
<tr>
<th>GPA</th>
<th>SATM</th>
<th>SATV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>2.700</td>
<td>Min.</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>3.075</td>
<td>1st Qu.</td>
</tr>
<tr>
<td>Median</td>
<td>3.400</td>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
<td>3.300</td>
<td>Mean</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>3.575</td>
<td>3rd Qu.</td>
</tr>
<tr>
<td>Max.</td>
<td>3.700</td>
<td>Max.</td>
</tr>
</tbody>
</table>

```
Note that the `write.csv()` function does not specify the full path address to the external folder (this was done when we created the Project) but does require two arguments: the name of the object in the R Workspace, `Descriptives`, and the name of the data file once it has been exported to the folder, `SummaryTable.csv`. Note that quotes must enclose the file name.

A final inspection of the `Advising` project folder now lists the four files that reside there, including `SummaryTable.csv` (see Figure 1.9).

![Figure 1.9](image)

**Figure 1.9** The *Advising* project folder with four files

We have just summarized three methods that can be used to import data into the Workspace: (a) installing the `introstats` package that accompanies this book; (b) entering data directly into the R Console; and (c) most useful of all, importing the data from a spreadsheet. Despite the fact that we have just taken nearly six pages to cover method (c), it is not the approach we use in the examples and exercises for this book. Instead, we use method (a) and to a lesser extent method (b). The reason why we take so much care and so many pages to demonstrate method (c) is that it is the approach you will use most often when working with R in any context unconnected with the book, particularly in your other classes or in your work life.

### 3. Five Hints for Using R

1. Since R is case sensitive, remember that `x` and `X` are not the same thing.
2. If a line of code is incomplete, a `+` sign will appear, prompting us to finish the line.
3. Remember that comments (preceded with a `#`) describe what lines of R code are intended to accomplish, and can be very helpful when we cannot otherwise make out the purpose of the code.
4. We can avoid re-entering R commands by using the up-arrow key, an action that recovers the lines above for recycling. In fact, it is possible to go back through as many lines of code as necessary by repeatedly hitting this key.
5. We can confirm that R is connected to a certain external directory by entering `getwd()` at the prompt in the Console. Similarly, we can connect R to a desired directory by entering `setwd()`, where the argument is the path address enclosed in quotes. Even so, one of the advantages of working with RStudio Projects is that these steps are usually unnecessary.
OUR APPROACH TO WRITING THIS BOOK

There is much additional material we could have included but have not. Textbooks are normally defined as much by what they leave out as by what they include, and this one is no exception. This book is primarily about statistics and probability, and only secondarily about the R statistical programming language. Moreover, R provides many different ways to solve the problems we encounter in this book. We have chosen to include those concepts and methods that have worked best for the thousands of graduate and undergraduate students we have had when teaching statistics at four different universities in the US and Europe.

---

exercises

1.1 Using R, create a vector consisting of the following elements: 81, 17, 7, 55, 2, 98, 71, 47, 19, 8, 3, 10, 28, 65, 80. Name it \( E_{1.1} \).

(a) How many data values are in \( E_{1.1} \)? Use the function `length()`.
(b) What is the mean of \( E_{1.1} \)?
(c) What is the median of \( E_{1.1} \)? Use the function `median()`.
(d) Use the functions `min()` and `max()` to find the minimum and maximum values.
(e) What is the sum of the values in \( E_{1.1} \)? Use the function `sum()`.

1.2 Use the functions `sum()` and `length()` to find the mean of \( E_{1.1} \).

1.3 Using the built-in data set `LakeHuron`, please answer the following questions.

(a) What are the first five values in `LakeHuron`?
(b) How many data values are in `LakeHuron`?
(c) What is the lowest level (in feet) of Lake Huron during the 1875–1972 period?
(d) What is the highest level of Lake Huron?
(e) What is the mean level?
(f) What is the median level?
(g) What are the last four values in `LakeHuron`? Use the function `tail(,4)`.

1.4 Suppose we interview five individuals who are registering to vote in the 2016 US Presidential election, and learn the following about them in terms of their age (years) and annual income (US$): voter 1 is 25 years of age with an income of $24,000; voter 2 is 37 years with an income of $42,000; voter 3 is 45 years with an income of $39,000; voter 4 is 57 years with an income of $77,000; and voter 5 is 65 years with an income of $84,000. Use R to create a data frame consisting of these five individuals and two variables, named `Age` and `Income`. Name the data frame \( E_{1.2} \).

1.5 Use the function `summary()` to find the minimum, maximum, mean, median, and first and third quartiles of \( E_{1.2} \).

---

https://study.sagepub.com/stinerock