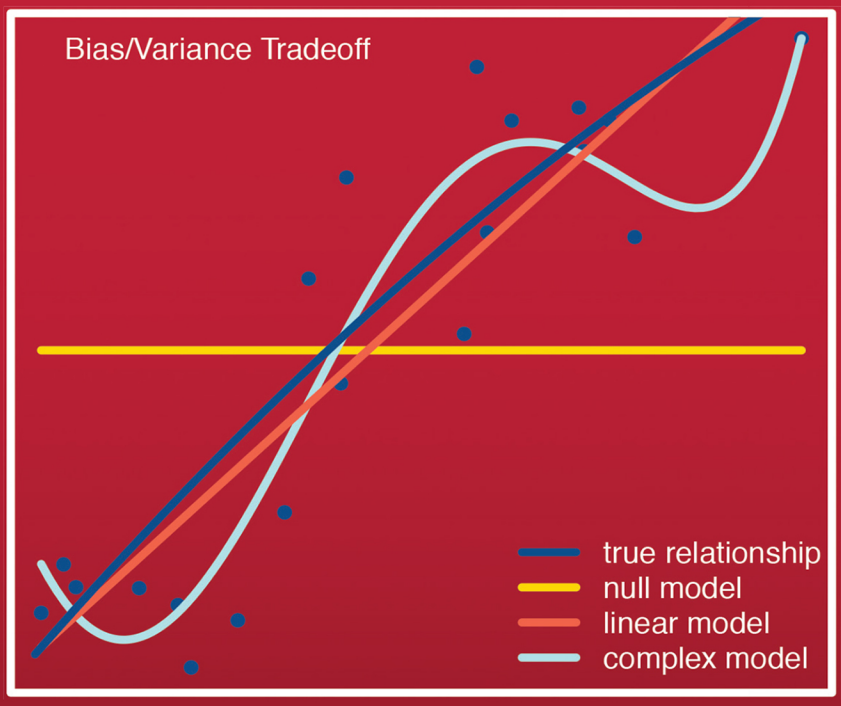
Texts in Statistical Science

Foundations of Statistics for Data Scientists

With R and Python



Alan Agresti
Maria Kateri

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Foundations of Statistics for Data Scientists

CHAPMAN & HALL/CRC
Texts in Statistical Science Series
Joseph K. Blitzstein, *Harvard University, USA*
Julian J. Faraway, *University of Bath, UK*
Martin Tanner, *Northwestern University, USA*
Jim Zidek, *University of British Columbia, Canada*

Recently Published Titles

Randomization, Bootstrap and Monte Carlo Methods in Biology
Fourth Edition
Bryan F. J. Manly, Jorje A. Navarro Alberto

Principles of Uncertainty, Second Edition
Joseph B. Kadane

Beyond Multiple Linear Regression
Applied Generalized Linear Models and Multilevel Models in R
Paul Roback, Julie Legler

Bayesian Thinking in Biostatistics
Gary L. Rosner, Purushottam W. Laud, and Wesley O. Johnson

Linear Models with Python
Julian J. Faraway

Modern Data Science with R, Second Edition
Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton

Probability and Statistical Inference
From Basic Principles to Advanced Models
Miltiadis Mavrakakis and Jeremy Penzer

Bayesian Networks
With Examples in R, Second Edition
Marco Scutari and Jean-Baptiste Denis

Time Series
Modeling, Computation, and Inference, Second Edition
Raquel Prado, Marco A. R. Ferreira and Mike West

A First Course in Linear Model Theory
Second Edition
Nalini Ravishanker, Zhiyi Chi, Dipak K. Dey

Foundations of Statistics for Data Scientists
With R and Python
Alan Agresti and Maria Kateri

For more information about this series, please visit: https://www.crcpress.com/Chapman--Hall/CRC-Texts-in-Statistical-Science/book-series/CHTEXSTASCI

Foundations of Statistics for Data Scientists

With R and Python

Alan Agresti and Maria Kateri

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-0-367-74845-6 (hbk)
ISBN: 978-1-003-15983-4 (ebk)

DOI: 10.1201/9781003159834

Typeset in LM Roman
by KnowledgeWorks Global Ltd.

Contents

Preface

This book presents an overview of the *foundations*—the key concepts and results—of statistical science. The primary intended audience is undergraduate students who are training to become data scientists. This is not a book, however, about how to become a data scientist or about the newest methods being used by data scientists or about how to analyze "big data" and the wide variety of types of data with which data scientists deal. It is a book that has the purpose of teaching potential data scientists the foundations of one of the core tenets of data science—statistical science.

Statistical science is by now a large subject, having many distinct specialties. This book highlights the topics with which we believe that any data scientist should be familiar: descriptive statistical methods, probability distributions, the inferential statistical methods of confidence intervals and significance testing, and linear and generalized linear modeling. When combined with courses that a student majoring in Data Science would take in computer science and mathematics, this book provides the background needed to follow this introduction to statistical science by studying specialized areas of it.[1]

This book assumes that students have knowledge of calculus, so we can focus on *why a statistical analysis works* as well as *how to do it*. University statistical science courses that require a calculus background often have the name *Mathematical Statistics*. We avoid this term in the title of our book, as we do not want students to think that statistical science is a subfield of mathematics or that complex mathematics is necessary to be proficient in understanding and applying statistical science. In fact, we mainly use only basic calculus tools of differentiation and integration, and then only for some topics. Compared to the content of traditional mathematical statistics textbooks, our book has less emphasis on probability theory, derivations of probability distributions of transformations of random variables, decision theory, and statement and formal proof of theorems. It introduces some modern topics that do not normally appear in such texts but are especially relevant for data scientists, such as generalized linear models for non-normal responses, Bayesian and regularized model-fitting, and classification and clustering. The greatest difference from a traditional mathematical statistics book, however, is that this book shows how to implement statistical methods with modern software and illustrates statistical concepts and theory using simulations.

To use and properly interpret methods of modern statistical science, computational skills are as important as mathematical skills. Besides using mathematics to show "why it works," we use computational simulations and Internet apps to help provide intuition about foundational results such as behavior of sampling distributions and error rates for statistical inferences. Throughout the book, examples with real data show how to use the free statistical software `R` to implement statistical methods. The book also contains software appendices that present greater detail about `R` as well as introducing `Python` for statistical analyses. The `Python` appendix shows analyses for the examples analyzed in the chapters with `R`, so an instructor can easily use the text in a course that has `Python` as its primary software. Since

[1] Such as multivariate analysis, nonparametrics, categorical data analysis, design and sample survey methods, time series, longitudinal data analysis, survival analysis, decision theory, Bayesian statistics, stochastic modeling, computational methods of statistics, and smoothing and nonlinear modeling

the book focuses on the foundations of statistical science, it has less emphasis on some practical issues of data analysis, such as preparing and cleaning data files. However, the software appendices also introduce additional analyses that supplement the examples presented in the chapters. A regularly-updated website `http://stat4ds.rwth-aachen.de` for the book has all data files analyzed and longer versions of the software appendices as well as an appendix about the use of `Matlab` for statistical analyses. The data files are also available at `www.stat.ufl.edu/~aa/ds` and at the GitHub site `https://github.com/stat4DS/data`.

Use of this book as a course textbook

Chapters 1–6 of this book are designed as a textbook for an introductory course on statistical science for undergraduate students majoring in Data Science or Statistics or Mathematics. Some instructors may prefer to skip some of the less central or more technical material, such as Sections 3.4, 4.9, 5.7, 5.8, and 6.7. (These and other such sections and subsections have an * next to their titles.) With all nine chapters and the extra material presented in the `R` and `Python` appendices, it is also appropriate for a two-term sequence of courses. The book also can serve programs that have a heavy focus on statistical science, such as econometrics and operations research. It also should be useful to graduate students in the social, biological, and environmental sciences who choose Statistics as their minor area of concentration, so they can learn about the foundations that underlie statistical methods that they use. An instructor can use either `R` or `Python` as the main software for the course, as the examples in the main part of the text use `R` but the same examples are shown with `Python` in its appendix.

Each chapter contains many exercises for students to practice and extend the theory and methods. The exercises are grouped into two parts: Exercises in *Data Analysis and Applications* request that students perform data analyses similar to the ones presented in that chapter. Exercises in *Methods and Concepts* relate directly to the foundations aspect of the book. They ask questions about properties of statistical methods, conceptual questions about their bases, as well as extend that chapter's results. An appendix contains outlines of solutions for the odd-numbered exercises.

This book is by no means a complete overview of statistical science. The field is large and grows more every year, with areas being developed now that did not even exist in the twentieth century. However, we do believe it provides a solid introduction to the core material with which we believe any data scientist should be familiar.

In preparing this book together, Agresti (`agresti@ufl.edu`) has taken main responsibility for the chapter material and Kateri (`maria.kateri@rwth-aachen.de`) has taken main responsibility for the appendices about `R` and `Python` statistical software and the expanded appendices about `R`, `Python`, and `Matlab` at the book's website. We welcome any comments or suggestions that you care to send either of us that we can take into account in future editions of this book.

Acknowledgments

Thanks to several friends and colleagues who provided comments on various versions of this manuscript or who provided data sets or other help, including Alessandra Brazzale, Jane Brockmann, Brian Caffo, Sir David Cox, Bianca De Stavola, Cristina Cuesta, Travis Gerke, Sabrina Giordano, Anna Gottard, Ralitza Gueorguieva, Bernhard Klingenberg, Bhramar Mukherjee, Ranjini Natarajan, Madan Oli, Euijung Ryu, Alessandra Salvan, Nicola Sartori, Elena Stanghellini, Stephen Stigler, Gerhard Tutz, Roberta Varriale, Larry Winner, and Daniela Witten. Thanks to Hassan Satvat for help in setting up the book's website and to Bernhard Klingenberg for developing the excellent apps at `www.artofstat.com` that are often cited in the book. Many thanks to Joyce Robbins, Mintaek Lee, Jason M. Graham, Christopher Gaffney, Tumulesh Solanky, and Steve Chung for providing helpful reviews to

CRC Press of our manuscript. Finally, special thanks to John Kimmel, Executive Editor of Statistics for Chapman & Hall/CRC Press, for his encouragement and support in this book project.

Alan Agresti and Maria Kateri
Gainesville Florida and Brookline Massachusetts, USA;
Aachen, Germany

April 2021

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Introduction to Statistical Science

Compared to mathematics and the physical and natural sciences, statistical science is quite young. The statistical methods that you'll learn about in this book were mainly developed within the past century. Modern computing power is causing a revolution in the sorts of data analyses that are possible, so new methods are continually being developed. In recent years, new statistical methods have resulted from challenges in analyzing data in diverse fields, such as medicine (e.g., genetic data relating to disease presence, data for personalized medical decisions) and business (e.g., data on consumer buying behavior, data from experiments comparing advertising strategies). This book presents the foundations underlying the methods of statistical science, explaining when and why these methods work, and shows how to use statistical software to apply them.

Statistical software also has become increasingly powerful and easily available. This has had benefits in the data analyses that are now possible, but a danger is that prospective data scientists might think that statistical science is merely a computational toolbox consisting of a variety of algorithms. A goal of this book, by contrast, is to show that *the methods of statistical science result from a unified theory*, although that theory itself has slight variations in the way it is implemented or interpreted. Another danger of the ubiquity of statistical software is that prospective data scientists might expect that software can automatically perform good data analyses without input from the user. We'll see, however, that careful thought is needed to decide which statistical methods are appropriate for any particular situation, as they all make certain assumptions, and some methods work poorly when the assumptions are violated. Moreover, a data scientist needs to be able to interpret and explain the results that software yields.

In this chapter, we introduce statistical science as a field that deals with describing data and using them to make inferences. We define types of *variables* that represent how measured characteristics can vary from observation to observation. We also introduce *graphical and numerical methods* for describing the data. When a study can use *randomization* in collecting the data or conducting an experiment, data analysts can exploit the random variation to make reliable estimations and predictions.

1.1 Statistical Science: Description and Inference

You already have a sense of what the word ***statistics*** means. You regularly hear statistics quoted about sports events, the economy, medical research, and opinions, beliefs, and behaviors of people. In this sense, a statistic is merely a number calculated from ***data*** —the observations that provide information about the subject matter. But the field of ***statistical science*** has a much broader sense—as a field that gives us a way of gathering and analyzing the data in an objective manner.

DOI: 10.1201/9781003159834-1

Statistical science

Statistical science is the science of developing and applying methods for collecting, analyzing, and interpreting data.

Many methods of statistical science incorporate reasoning using tools of *probability*. The methods enable us to deal with uncertainty and variability in virtually all scientific fields. With statistical methods, we learn from the data while measuring, controlling, and communicating uncertainty.

1.1.1 Design, Descriptive Statistics, and Inferential Statistics

Statistical science has three aspects:

1. ***Design***: Planning how to gather relevant data for the subject matter of interest.
2. ***Description***: Summarizing the data.
3. ***Inference***: Making evaluations, such as estimations and predictions, based on the data.

Design refers to planning a study so that it yields useful data. For example, for a poll taken to determine public opinion on some issue, the design specifies how to select the people to interview and constructs the questionnaire for interviews. For a research study to compare an experimental diet with a standard diet to address obesity, the design specifies how to obtain people for the study, how to determine which people use each diet, and specifies the characteristics to measure to compare the diets.

Description refers to summarizing the data, to mine the information that the data provide. For any study, the raw data are a complete listing of observations that can be overwhelming for comprehension. To present the results, we reduce the data to simpler and more understandable form without distorting or losing much information. Graphs, tables, and numerical summaries such as averages and percentages are called ***descriptive statistics***.

Inference refers to using the data to make estimations and other sorts of evaluations, such as predictions. These evaluations take into account random variability that occurs with the characteristics measured and the resulting uncertainty in decision-making. For instance, suppose that in the study comparing two diets, the people on the experimental diet had an average weight loss of 7.0 kilograms. What can we say about the average weight change if hypothetically *all* obese people used this diet? An inferential statistical method provides an interval of numbers within which we can predict that the average weight change would fall. The analysis might enable us to conclude, with small probability of being wrong, that the average weight change for all obese people would fall between 5.6 and 8.4 kilograms. Another inferential statistical method would enable us to decide whether that average weight change is greater than would be obtained with a standard diet or no special diet. Other inferential statistical methods evaluate whether weight change is associated with characteristics other than the diet, such as a person's gender, race, age, attained education, and amount of weekly exercise. Data-based evaluations such as estimations and predictions are called ***statistical inferences***.

Descriptive statistics and *inferential statistics* are the two main types of methods for analyzing data. Researchers use them to answer questions such as, "Does the experimental diet have a beneficial effect in reducing obesity, and is it more effective than a standard diet?" "How does the sales of a product compare if we place an advertisement for it at websites, or in mailings, or in newspapers, or on TV programs?" "Do states in the U.S.

that have stronger gun control laws tend to have lower murder rates, taking into account socioeconomic factors?" "Is student performance in Canada associated with the amount of money spent per student, the size of the classes, or the teachers' salaries?" "Do a majority of all New Zealanders favor legalization of marijuana?"

1.1.2 Populations and Samples

The entities on which a study makes observations are called the sample ***subjects***. Usually the subjects are individual people, such as in a survey, but they need not be. For example, an agricultural experiment might have cows as subjects if its goal is to compare milk yields for different diets. An ecological survey might have different forest areas as subjects in a study of species diversity. Subjects in social surveys might be people, families, schools, or counties.

Although we obtain data for the sample subjects, our ultimate interest is on the *population* from which the sample is taken.

Population and sample

The ***population*** is the total set of subjects of interest. A ***sample*** is the set of subjects from the population for which data are available.

The goal of most data analyses is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, polling organizations such as the Gallup poll (`www.gallup.com`) and the Pew Research Center (`www.pewresearch.org`) usually sample about 1000–2000 Americans to gather information about opinions and beliefs of the population of *all* adult Americans.

Inferential statistics provide evaluations about a population, based on data from a sample. For example, a survey taken in the U.S. in 2018 asked, "Do you believe in heaven?" The population of interest was all adults in the United States. Of the 1141 sampled subjects, 81% answered *yes*. We would be interested, however, not only in those 1141 people but in the *population* of more than 250 million adults in the U.S. An inferential method presented in Chapter 4 estimates that the population percentage believing in heaven almost certainly falls between 78% and 84%. That is, the sample value of 81% has a "margin of error" of 3%. Inferential statistical analyses can predict characteristics of entire populations quite well by selecting samples that are very small relative to the population size. In this book, we'll learn why this works.

1.1.3 Parameters: Numerical Summaries of the Population

To distinguish between a descriptive statistic calculated for a sample and the corresponding characteristic of the population, we use the term ***parameter*** for the population characteristic.

Parameter

A ***parameter*** is a numerical summary of a population.

In practice, our primary interest is in the values of parameters rather than sample descriptive statistics. For example, in viewing the results of a poll before an election, we would be more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed. However, parameter values are almost

always unknown. The sample and statistics describing it help us to make inferences about the unknown parameter values.

A key aspect of statistical inference involves reporting the *precision* of the sample statistic that estimates the population parameter. For the example on belief in heaven, the reported margin of error of 3% predicted how close the *sample* value of 81% was to the unknown *population* percentage. The other key aspect relates to the *probability* with which we obtain that precision. For instance, a statistical inference might state that we can be 95% sure that the sample value of 81% differs from the population value by no more than 3%.

1.1.4 Defining Populations: Actual and Conceptual

Usually the population to which inferences apply is an actual set of subjects, such as all adult residents of a nation. Sometimes, though, the inferences refer to a *conceptual* population—one that does not actually exist but is hypothetical.

For example, suppose a medical research team investigates a newly proposed drug for treating a virus by conducting a study at several medical centers. Such a medical study is called a *clinical trial*. The study would compare virus patients who are given the new drug to other virus patients who instead receive a standard treatment or a placebo, using descriptive statistics such as the percentages who respond positively. In applying inferential statistical methods, the researchers would like their inferences to apply to the conceptual population of *all* people suffering from the virus now or at some time in the future.

1.2 Types of Data and Variables

The observations gathered on the characteristics of interest are the *data*. For example, a survey of 1000 people to analyze opinions about the legalization of same-sex marriage might also observe characteristics such as political party affiliation, frequency of attending religious services, number of years of education, annual income, marital status, race, and gender. The data for a particular subject would consist of observations such as (opinion = do not favor legalization, political party = Republican, religiosity = attend services once a week, education = 12 years, annual income in the interval 40–60 thousand dollars, marital status = married, race = White, gender = male).

1.2.1 Data Files

Statistical software analyzes data organized in the spreadsheet form of a ***data file***:

- Any one row of a data file contains the observations for a particular subject (e.g., person) in the sample.
- Any one column of a data file contains the observations for a particular characteristic (e.g., opinion about legalized same-sex marriage).

The *number* of subjects in the data file, called the ***sample size***, is denoted by n. If we observe 10 characteristics for a sample of $n = 2000$ people, then the data file has 2000 rows and 10 columns.

Throughout this book, we use the statistical software `R` to illustrate statistical analyses. This software package is available to download for free at www.r-project.org. In `R`, we

save a data file as a *data frame*,[1] which is the fundamental data structure required by many R functions. Figure 1.1 shows the use of R to read a data file called `Survey.dat` (containing data from a student survey mentioned in Exercise 1.2) from a directory on a computer, save it as a data frame called *Opinions*, and display it.

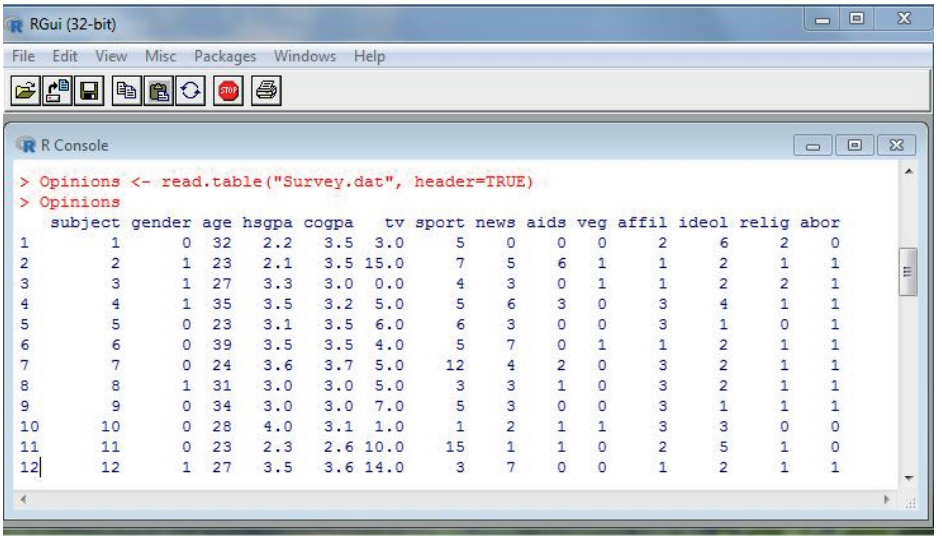


FIGURE 1.1 Part of a R session for loading and displaying a data file.

Existing archived collections of data are called ***databases***. Many databases result from a survey or study of some type, but some are existing records of data that result from other purposes. An example is a database of patients' electronic medical records. With the increasing variety of data that can be recorded electronically, not all data files have the format of a traditional data file with entries that are numbers or characters. For example, in medical records, some observations for some subjects may be images, such as a mammogram, a chest x-ray, or a brain scan,[2] or a continuous streaming of data over time, such as monitoring of heart-rate, respiratory-rate, blood pressure, and temperature.

1.2.2 Example: The General Social Survey (GSS)

Some databases are freely available on the Internet. An important database in the U.S. contains results since 1972 of the *General Social Survey* (GSS), conducted every other year by the National Opinion Research Center at the University of Chicago. It gathers information using personal interviews of a sample of about $n = 2000$ subjects from the U.S. adult population to provide a snapshot of opinions and behaviors. Researchers use it to investigate how adult Americans answer a wide diversity of questions, such as, "Do you believe in life after death?" and "Would you be willing to pay higher prices in order to protect the environment?" Similar social surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database:

[1] For details, see Section A.0.3 of the R Appendix.

[2] See https://aimi.stanford.edu/research/public-datasets for examples of data files of this type.

- Go to the website https://sda.berkeley.edu/archive.htm at the Survey Documentation and Analysis site at the University of California, Berkeley.
- Click on the most recently available *General Social Survey (GSS) Cumulative Datafile*. You will then see a "variable selection" listing in the left margin of characteristics measured over the years, and a menu on the right for selecting particular characteristics of interest.
- Type the name of a characteristic of interest in the *Row* box, and click on *Run the table*. The GSS site will then generate a table that shows the possible values for the characteristic and the number of people and the percentage who made each possible response.

For example, in one survey the GSS asked "About how many good friends do you have?" The GSS name for this characteristic is NUMFREND. The table that the GSS provides shows that the responses of 1, 2, 3, 4, 5, and 6 good friends had the percentages 6.1, 16.2, 15.7. 14.2, 11.3, and 8.8, respectively, with the remaining 27.7% spread around the other possible responses.

1.2.3 Variables

For the characteristics we measure in a study, *variability* occurs naturally among subjects in a sample or population. For instance, variation occurs from student to student in their college grade point average (GPA). A study to investigate the factors mainly responsible for that variability might also observe other characteristics that vary among students, such as high school GPA, college board score, time per day spent studying, time per day watching TV or browsing the Internet, and whether at least one parent attended college. Any characteristic that we can measure for the subjects is called a ***variable***. The term reflects that values of the characteristic *vary* among subjects.

Variable

A ***variable*** is a characteristic that can vary in value among subjects in a sample or population.

The values the variable can take form a *measurement scale*. The valid statistical methods for a variable depend on its measurement scale. We treat a numerical-valued variable such as number of good friends differently than a variable measured with categories, such as (*yes, no*) for whether employed. We next present two ways to classify variables. The first type refers to whether the measurement scale consists of numbers or categories. The second type refers to the fineness of measurement—the number of values in the measurement scale.

1.2.4 Quantitative Variables and Categorical Variables

A variable is called ***quantitative*** when the measurement scale has numerical values that represent different magnitudes of the variable. Examples of quantitative variables are number of good friends, annual income, college GPA, age, and weight.

A variable is called ***categorical*** when the measurement scale is a set of categories. Examples of categorical variables are marital status (with categories such as *single, married, divorced, widowed*), primary mode of transportation to work (*automobile, bicycle, bus, subway, walk*), preferred destination for clothes shopping (*downtown, Internet, mall, other*), and favorite type of music (*classical, country, folk, jazz, rap/hip-hop, rock*). Categorical variables having only two categories, such as whether employed (*yes, no*), are called ***binary***.

For categorical variables, distinct categories differ in quality, not in numerical magnitude. Categorical variables are often called ***qualitative***.

Categorical variables have two types of measurement scales. For some categorical variables, such as the ones just mentioned, the categories are unordered. The scale does not have a "high" or "low" end. The categories are then said to form a ***nominal scale***. By contrast, some categorical scales have a natural *ordering* of values. The categories form an ***ordinal scale***. Examples are perceived happiness (*not too happy, pretty happy, very happy*), headache pain (*none, slight, moderate, severe*), and political philosophy (*very liberal, slightly liberal, moderate, slightly conservative, very conservative*).

1.2.5 Discrete Variables and Continuous Variables

Another classification refers to the *number* of values in the measurement scale.

Discrete and continuous variables

A quantitative variable is ***discrete*** if it can take a set of distinct, separate values, such as the nonnegative integers (0, 1, 2, 3, ...). It is ***continuous*** if it can take an infinite continuum of possible real number values.

Examples of discrete variables are one's number of good friends, number of computers in household, and number of days playing a sport in the past week. Any variable phrased as "the number of ..." is discrete, because we can list its possible values (*0, 1, 2, 3, ...*). Examples of continuous variables are height, weight, age, distance a person walks in a day, winning time in a marathon race, and how long a cell phone works before it needs recharging. It is impossible to write down all the distinct potential values, because they form an interval of infinitely many real-number values. A person's age, for example, could take the value 20.6294473... years.

In practice, we round continuous variables when measuring them, so the actual measurement is discrete. We say that an individual is 20 years old whenever that person's age is somewhere between 20 and 21. On the other hand, some variables, although discrete, have a very large number of possible values. In measuring annual income in dollars, the potential values are 0, 1, 2, 3, ..., up to some very large value in many millions. Statistical methods for continuous variables are used for quantitative variables that can take a very large number of values, regardless of whether they are theoretically continuous or discrete.

1.2.6 Associations: Response Variables and Explanatory Variables

Most studies have more than one variable. With multivariable analyses, we say that an ***association*** occurs between two variables if certain values of one variable tend to go with certain values of the other. For example, consider religious affiliation, with categories (*Catholic, Protestant, Muslim, Jewish, Other*), and ethnic group, with categories (*African-American, Anglo-American, Hispanic, Other*). In the United States, Anglo-Americans are more likely to be Protestant than are Hispanics, who are overwhelmingly Catholic. African-Americans are even more likely to be Protestant. An association exists between religious affiliation and ethnic group, because the percentage of people having a particular religious affiliation changes as ethnic group changes.

When we study the association between two variables, usually one is an outcome variable on which comparisons are made at levels of the other variable. The outcome variable is called the ***response variable***. The variable that defines the groups is called the ***explanatory variable***. The analysis studies how the outcome on the response variable *depends on* or

is *explained by* the value of the explanatory variable. For example, when we describe how religious affiliation depends on ethnic group, religious affiliation is the response variable. In a comparison of men and women on annual income, annual income is the response variable and gender is the explanatory variable.

Most studies have one response variable and multiple explanatory variables. Sometimes the response variable is called the ***dependent variable*** and the explanatory variables are called the ***independent variables***. We prefer not to use these terms, because the terms *independent* and *dependent* are used in statistical science for many other things, and the terminology suggests causal interpretations that are usually inappropriate.

1.3 Data Collection and Randomization

When we apply inferential methods of statistical science to parameters for some population, the quality of the inferences depends on how well the *sample* represents the *population*.

1.3.1 Randomization

Randomization is a mechanism for achieving good sample representation of a population in a survey or an experiment.

Simple random sample

A ***simple random sample*** of n subjects from a population is one in which each possible sample of size n has the same probability (chance) of being selected.

With simple random sampling, everyone has the same chance of inclusion in the sample, so it is fair. It tends to yield a sample that resembles the population. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population.

Suppose that a researcher in a medical center plans to compare two drugs for some adverse condition. She has four patients with this condition, and she wants to randomly select two to use each drug. Denote the four patients by P_1, P_2, P_3, and P_4. In selecting $n = 2$ subjects to use the first drug, the six possible samples are

$$(P_1, P_2), (P_1, P_3), (P_1, P_4), (P_2, P_3), (P_2, P_4), (P_3, P_4).$$

More generally, let N denote the population size. The population has $\binom{N}{n}$ possible samples of size n. For example, a population of size $N = 4$ has $\binom{4}{2} = 4!/[2!(4-2)!] = 6$ possible samples of size $n = 2$. You could select the simple random sample by placing the four people's names on four identical ballots and selecting two blindly from a hat. This is unwieldy with the larger values for N and n usual in practice, and these days software can easily select the sample from a list of the population members using a *random number generator*.

We illustrate for random selection of $n = 5$ students out of a class of size $N = 60$. We assign the numbers $01, 02, \ldots, 60$ to the class members and generate five random numbers between 01 and 60. With the statistical software `R`, the `sample` function performs simple random sampling from a numbered population list:[3]

[3] We suggest that you read the Basics and Chapter 1 sections in this book's `R` Appendix to learn more about `R` and its use for descriptive statistical analysis.

```
> sample(1:60, 5)    # Comments about R commands follow the # symbol
[1] 11 55 48 59 29   # output line [1] shows the five integers randomly generated
```

The sample of size 5 selects the students numbered 11, 55, 48, 59, 29.

The *simple* adjective in "simple random sample" distinguishes this type of sampling from more complex sampling schemes that also have elements of randomization. For instance, *stratified random sampling* schemes divide the population into separate groups ("strata") and then select a simple random sample from each stratum. This is useful for comparing groups on some variable when a particular group is relatively small and may not be adequately represented in a simple random sample. *Cluster random sampling* schemes divide the population into a large number of clusters, such as city blocks, and select a simple random sample of the clusters. This is useful when a complete listing of the population is not available.

1.3.2 Collecting Data with a Sample Survey

Some studies sample people from a population and interview them to collect data. This method of data collection is called a ***sample survey***. The interview could be a personal interview, telephone interview, or self-administered questionnaire. Implementing a sample survey requires finding or constructing a valid *population list* for selecting the sample. A major challenge with collecting data with sample surveys is *nonresponse*—a significant percentage of those sampled refuse to be interviewed or don't respond on some items.

The *General Social Survey*, introduced in Section 1.2.2, is a sample survey. The GSS uses a random sampling design that is more complex than simple random sampling, incorporating multiple stages and clustering to make the survey easier to implement, but it ensures adequate coverage by giving each family the same chance of inclusion.

1.3.3 Collecting Data with an Experiment

Some studies use a planned ***experiment*** to generate data. An experiment compares subjects on a response variable under different conditions. Those conditions, which are levels of an explanatory variable, are called *treatments*. For instance, the treatments might be different drugs for treating some illness, compared in a clinical trial. The researcher specifies a plan for how to assign subjects to the treatments, called the *experimental design*. Good experimental designs use randomization to determine which treatment a subject receives. This reduces bias and allows us to use statistical inference.

For example, the Physicians' Health Study Research Group at Harvard Medical School designed an experiment to analyze whether regular intake of aspirin reduces mortality from heart disease. Of about 22,000 physicians, half were randomly chosen to take an aspirin every other day. The remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared. By using randomization to determine who received which treatment, the researchers knew the groups would roughly balance on all variables that could affect heart attack rates, such as age and quality of health. If the physicians could decide on their own which treatment to take, the groups might have been out of balance on some important factor. For instance, if younger physicians were more likely to select aspirin, then a lower heart attack rate among the aspirin group could occur merely because younger subjects are less likely to suffer heart attacks.

In medical research, *randomized clinical trials* are experiments using randomization that have been the gold standard for many years. But experiments are now used to address questions in an increasing variety of areas. For instance, the economists at Harvard and Massachusetts Institute of Technology (MIT) who won the Nobel Prize in 2019 pioneered

the use of experiments to determine the policies that best improve the lives of the poor.[4] They randomly selected participants for anti-poverty programs, such as to evaluate if access to textbooks or access to remedial instruction improves education results. Randomized experiments have addressed diverse topics such as how to improve child nutrition, protect forests, and reduce gender discrimination.

1.3.4 Collecting Data with an Observational Study

In many application areas, it is not possible to conduct experiments to answer the questions of interest. We cannot randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income or usage of guns. Many studies merely *observe* the outcomes for available subjects on the variables of interest, without any experimental control of the subjects. Such studies are called ***observational studies***. Sample surveys are examples of observational studies.

Sometimes we can envision an experiment that would help us answer some question, but the experiment would be *unethical* to conduct, so observational studies are used instead. For example, in the mid-20th century some researchers decided to investigate whether an association exists between lung cancer and smoking. The researchers could have answered the question by taking a group of youngsters, randomly splitting them into two groups, and instructing one group to smoke a pack of cigarettes each day and the other group not to smoke at all. Then, after 50 years, the study would have compared the percentages of smokers and non-smokers who got lung cancer. Such an experiment would not have been ethical or feasible to conduct, and the answer was needed at that time itself and not 50 years later, so it was necessary to use observational studies to address this issue. One useful approach sampled hospital patients, matching each adult suffering from lung cancer to an adult control of similar age who did not have it,[5] and compared the lung-cancer group with the control group in terms of how much they had smoked in the past.

1.3.5 Establishing Cause and Effect: Observational versus Experimental Studies

With observational studies, making causal conclusions based on comparing groups on a response variable is dangerous because the groups may be imbalanced on other variables that affect the response outcome. This is true even with random sampling. For instance, suppose we plan to compare performance on some standardized exam for Black, Hispanic, and White students. If White students have a higher average score, a variety of variables might account for that difference. Perhaps, on the average, White students have higher parents' attained education or higher parents' income or better quality of school attended. Those or other key variables may not even have been measured in the study.

Establishing *cause and effect* definitively is not possible with an observational study.[6] Unmeasured variables, referred to as *lurking variables*, could be responsible for associations observed in the data. By contrast, with an experiment that randomly assigns subjects to treatments, those treatments should balance on any unmeasured variables, at least approximately. For example, in the Harvard Medical School study of the association between heart attack prevalence and taking aspirin or placebo, those taking aspirin would not tend to be

[4] See www.nytimes.com/2019/11/29/business/economics-nobel.html

[5] Such a study is called a *matched case–control study*. It is a *retrospective* type of observational study, which compares groups, such as those with and without some medical condition, by "looking into the past" to measure relevant variables.

[6] Section 6.2.4 discusses this issue in more detail.

younger or of better health than those taking placebo. Because a randomized experiment balances the groups being compared on lurking variables, one can better establish cause and effect with it than with an observational study.

1.4 Descriptive Statistics: Summarizing Data

Descriptive statistics summarize the information that the data contain. Before calculating and analyzing descriptive statistics, you need to be cautious about complications such as missing data or improperly recorded data that can cause software to fail to work or to give invalid output.[7] Furthermore, the data readily accessible to us may be unstructured and messy. The process of organizing and cleaning the data and bringing them into an appropriate form of a data file for further statistical analysis is called *data wrangling*. This is an important preliminary stage of any statistical analysis and may take substantial time.

This section presents descriptive statistics for quantitative variables in clean data files. Tables and graphs describe the data by showing the number of times various outcomes occurred. The two key features to describe numerically are the *center* of the data and the *variability* of the data around the center.

1.4.1 Example: Carbon Dioxide Emissions in European Nations

Environmental scientists study how the increasing levels of carbon dioxide (CO2) emissions around the world over time are associated with climate change reflected by the rising temperatures of "global warming." To illustrate methods of this section, we analyze recent UN data on carbon dioxide emissions per capita, in metric tons, for 31 nations in Europe. The emissions range between 2.0 for Albania and 9.9 for the Netherlands. We do not consider trends over time here, but it has been estimated that CO2 emissions in the U.S. have doubled since the 1950s and are about 150 times higher than in 1850.

We can use `R` to read and view the `Carbon` data file from the book's website:

```
> Carbon <- read.table("http://stat4ds.rwth-aachen.de/data/Carbon.dat", header=TRUE)
                       # header=TRUE if variable names are at top of data file
> head(Carbon, 3)      # head(Carbon, n) shows n observations at top of data file
    Nation  CO2        # tail(Carbon, n) shows n observations at end of data file
1  Albania  2.0
2  Austria  6.9
3  Belgium  8.3
```

1.4.2 Frequency Distribution and Histogram Graphic

A ***frequency distribution*** is a listing of the possible values for a variable, together with the number, proportion, or percentage of observations at each value. For a discrete variable with relatively few values, the distribution lists each possible value. For a continuous variable or a discrete variable with many possible values, those values are divided into intervals.

The next output uses `R` to construct a frequency distribution for the CO2 values in the `Carbon` data file, using intervals of width 1.0 metric ton. The most common interval was 5.0 up to but not including 6.0, with a percentage occurrence of 25.8%.

[7] Section A.1.4 in the `R` Appendix shows the use of `R` software for identifying missing data and conducting descriptive statistical analyses without them.

```
> breaks <- seq(2.0, 10.0, by=1.0) # frequency dist. intervals of width 1 between 2 and 10
> freq <- table(cut(Carbon$CO2, breaks, right=FALSE))
> freq              # with right=FALSE, right-most value not included in interval
 [2,3)  [3,4)  [4,5)  [5,6)  [6,7)  [7,8)  [8,9) [9,10)
     1      3      7      8      5      1      3      3
> cbind(freq, freq/nrow(Carbon)) # Frequency distribution of CO2 values, showing
          freq                  # intervals of CO2 values, followed by frequency
[2,3)        1  0.03225806      # (freq) and proportion, which divides freq by
[3,4)        3  0.09677419      # nrow(Carbon) = number of rows in data frame = 31
[4,5)        7  0.22580645
[5,6)        8  0.25806452
[6,7)        5  0.16129032
[7,8)        1  0.03225806
[8,9)        3  0.09677419
[9,10)       3  0.09677419
> hist(Carbon$CO2, xlab="CO2", ylab="Proportion", freq=FALSE) # histogram
> plot(density(Carbon$CO2))      # smooth-curve approximation of histogram (not shown)
```

A graph of the frequency distribution of a continuous variable or a discrete variable with intervals of values is called a ***histogram***.[8] Each interval of possible values has a bar over it, with height representing its number, proportion or percentage of observations. For the data on European CO2 values, Figure 1.2 shows the histogram constructed using the `hist` function in `R`. For highly discrete variables, such as categorical variables, a graph that has a separate bar over each distinct value is called a ***bar graph***.

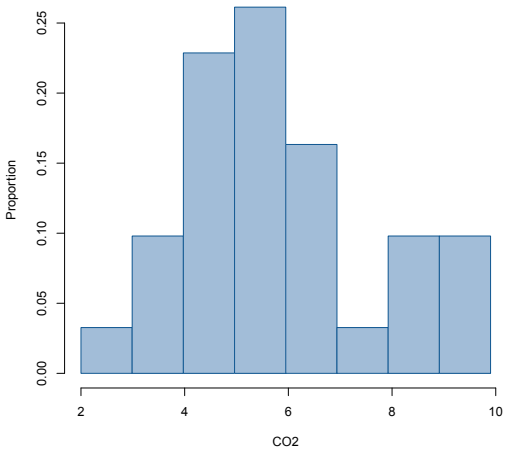


FIGURE 1.2 Histogram for frequency distribution of European CO2 values.

The *shape* of a histogram is informative. For a continuous variable, if we could increase n indefinitely with the number of intervals simultaneously increasing, so their width narrows, the shape would gradually approach a smooth curve.[9] For a smooth-curve approximation for any n, a bell-shaped appearance indicates that most subjects tend to fall near a central value. See Figure 1.3. The parts of the curve for the lowest values and the highest values are called the *tails* of the distribution. Often, one tail is much longer than the other. A distribution is then said to be ***skewed***: *skewed to the right* or *skewed to the left* according to which tail is longer.

[8] Section A.1.2 of the `R` Appendix presents options for constructing histograms.

[9] This curve can be approximated with the `density` function in `R`. See the code above (plot not shown) and Exercise 1.18.

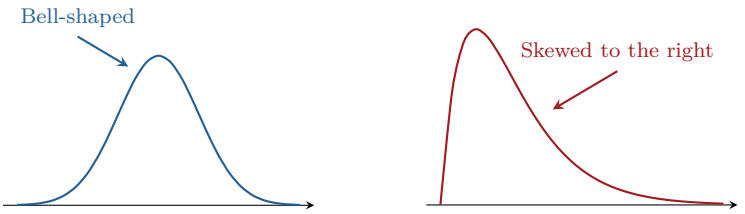


FIGURE 1.3 Smooth curve versions of bell-shaped and skewed frequency distributions: The longer tail indicates the direction of skew.

The `Carbon_West` data file at the book's website adds four non-European Western nations to the `Carbon` data file for Europe, with CO2 values of 15.4 for Australia, 15.1 for Canada, 7.7 for New Zealand, and 16.5 for the U.S. As an exercise, load that data file and form the histogram of CO2 values. The relatively large values for 3 of the 35 nations yields an extended right-tail reflecting skewness to the right.

1.4.3 Describing the Center of the Data: Mean and Median

For a sample, we use subscripts to identify particular observations in a data file. For example, we express the first three observations of a variable denoted by y as y_1, y_2, y_3.

How can we describe a typical observation for a quantitative variable, for example, describing the *center* of the data? A commonly used measure is the ***mean***.

Mean

The ***mean*** is the sum of the observations divided by the number of them. For a variable y with n observations y_1, y_2, ..., y_n in a sample from some population, the mean $\bar{y}$ is

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

The next section explains how the mean is the *center of gravity* of the data. Because of this, with small n, the mean can be highly influenced by an observation that falls well above or well below the rest of the data. Such an observation is called an ***outlier***. Here is an example: The owner of Leonardo's Pizza Joint reports that the mean annual income of his seven employees is $56,600. In fact, the annual incomes are $15,400, $15,600, $15,900, $16,400, $16,400, $16,600, and $299,900. The $299,900 income, which is the owner's income, is an outlier. The mean computed for the other six observations alone equals $16,050, quite different from the overall mean of $56,600.

The other commonly used measure of center for a quantitative variable is the ***median***. It is the middle value,[10] being larger than 50% of the other observations and smaller than 50%. For highly skewed distributions, this is a more representative summary of the center than is the mean. To illustrate, for the ordered income observations for the seven employees of Leonardo's Pizza, the median is $16,400. In particular, the median is *resistant* to outliers. When the highest observation of $299,900 is increased to $1,000,000, the median is still $16,400.

A measure sometimes reported with discrete data (quantitative or categorical) is the ***mode***, which is the most common outcome. For the GSS data on number of good friends,

[10] When n is even, the median is the midpoint between the two middle observations.

the most common response was 2, which is the mode. A frequency distribution that has two distinct peaks is called *bimodal*.

1.4.4 Describing Data Variability: Standard Deviation and Variance

To describe quantitative data more fully, we describe not only the center but also the *variability* about that center. The difference between the largest and smallest observations, called the ***range***, is a simple way to do this. The range incorporates only the two most extreme observations, however, so it is sensitive to outliers and does not reflect how far observations other than those two fall from the center.

More useful measures of variability use the distances of *all* the observations from the center. For observation i, the ***deviation*** of y_i from the mean $\bar{y}$ is $(y_i - \bar{y})$. The deviation is *positive* when y_i falls *above* the mean and *negative* when y_i falls *below* the mean. Calling the mean the *center of gravity* of the data reflects that the sum of the positive deviations equals the negative of the sum of negative deviations. That is, the sum of all the deviations,

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

Because of this, measures of variability use the squares or the absolute values of the deviations.

Standard deviation and variance

For a variable y with n observations $y_1, y_2, \ldots, y_n$ in a sample from some population, the ***standard deviation*** s is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}.$$

The standard deviation is the positive square root of the ***variance*** s^2,

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

The variance is approximately an average of the squared deviations. The units of measurement are the squares of those for the original data, so the standard deviation is simpler to interpret: s is a sort of *typical distance* of an observation from the mean. Also, *the larger the standard deviation, the greater the spread of the data*. The value of s is nonnegative, with $s = 0$ only when all observations have the same value. The denominator of s^2 uses $(n-1)$, rather than n, because this version of the measure naturally arises in inferential statistical methods.[11]

The magnitude of s partly reflects the shape of the frequency distribution. If the distribution is approximately bell-shaped, then:[12]

1. About 68% of the observations fall between $\bar{y} - s$ and $\bar{y} + s$.
2. About 95% of the observations fall between $\bar{y} - 2s$ and $\bar{y} + 2s$.

[11] With data for an entire population, we replace $(n-1)$ by n; the variance is then precisely the mean squared deviation. In using a sample to *estimate* variability around the population mean, whose value is unknown, Section 4.4.6 shows that the bias due to the numerator of s^2 having the *sample* mean instead of the *population* mean is eliminated when we use $(n-1)$ rather than n in the denominator.

[12] Section 2.5.1 will show where these percentages come from.

3. All or nearly all observations fall between $\bar{y} - 3s$ and $\bar{y} + 3s$.

Figure 1.4 is a graphical portrayal. To illustrate, suppose that grades on an exam are bell-shaped around $\bar{y} = 70$, with $s = 10$. Then, about 68% of the exam scores fall between 60 and 80, about 95% fall between 50 and 90, and all or nearly all fall between 40 and 100.

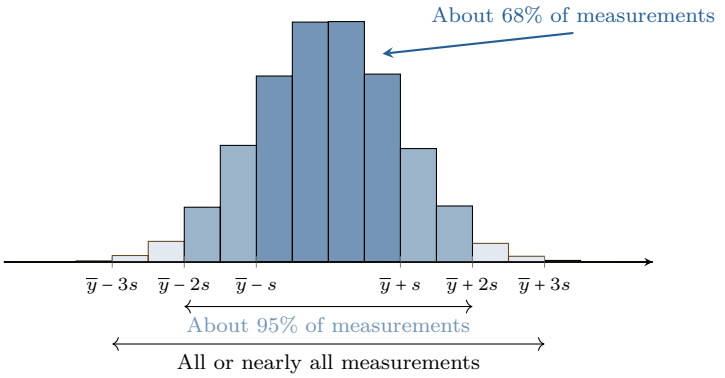


FIGURE 1.4 The standard deviation and mean determine approximate percentages in certain ranges for bell-shaped distributions.

Regardless of the shape of a frequency distribution, it is rare for observations to fall many standard deviations from the mean. The Russian mathematician Pafnuty Chebyshev proved in 1867 that *for any $k \geq 1$, the proportion of observations that fall at least k standard deviations from the mean can be no greater than $1/k^2$*. The result is called ***Chebyshev's inequality***.[13] For example, no more than 4% of the observations can fall at least five standard deviations from the mean. Percentages for most data sets, however, are much closer to the bell-shaped percentages than to the Chebyshev upper bounds.

1.4.5 Describing Position: Percentiles, Quartiles, and Box Plots

Besides center and variability, another way to describe a distribution is with a measure of *position*. The ***pth percentile*** is the point such that p% of the observations fall below or at that point and $(100 - p)$% fall above it.[14] For example, for $p = 95$, the 95th percentile falls above 95% of the observations and below 5% of them. The 50th percentile is the *median*. ***Quantiles*** are percentiles expressed in proportion form. For example, the 95th percentile is also called the 0.95 quantile.

Two especially useful percentiles are the the 25th percentile, called the ***lower quartile***, and the 75th percentile, called the ***upper quartile***. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the observations. The middle half of the data falls between the lower and upper quartiles. The ***interquartile range***, denoted by IQR, is the difference between the upper quartile and the lower quartile. Unlike the ordinary range, the IQR is not affected by outliers and takes into account variability by observations other than the most extreme ones.

[13] Sometimes it is referred to as the *Bienaymé–Chebyshev inequality*, since essentially the same result was shown by the French mathematician I. J. Bienaymé in 1853.

[14] This definition is imprecise, because no value or an infinite number of real numbers may have *exactly* p% falling below or at it. Software uses more elaborate definitions that make adjustments to yield precise values. The adjustments are tiny for large n.

The median, the upper and lower quartiles, and the maximum and minimum values provide a ***five-number summary*** of positions. Software can easily find these values as well as other percentiles and summary measures. For instance, with R, here is the five-number summary and the mean and standard deviation for the variable labeled as CO2 in the Carbon data file:

```
> summary(Carbon$CO2)           # 1st Qu = lower quartile, 3rd Qu = upper quartile
   Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
  2.000    4.350    5.400   5.819    6.700   9.900
> c(mean(Carbon$CO2), sd(Carbon$CO2), quantile(Carbon$CO2, 0.90))
                      90%
5.819355 1.964929 8.900000 # mean, standard deviation, 0.90 quantile
> boxplot(Carbon$CO2, xlab="CO2 values", horizontal=TRUE)
```

The five-number summary is the basis of a graphical display called the ***box plot***. The *box* contains the central 50% of the data, from the lower quartile to the upper quartile. The median is marked by a line drawn within the box. The dashed lines extending from the box, called *whiskers*, contain the outer 50% of the data except for outliers, which are marked separately.[15] For example, Figure 1.5 shows the box plot for the European CO2 values. The shape indicates that the right-tail of the distribution, which corresponds to the relatively large values, is slightly longer than the left tail. The plot reflects the slight skewness to the right of the observations.

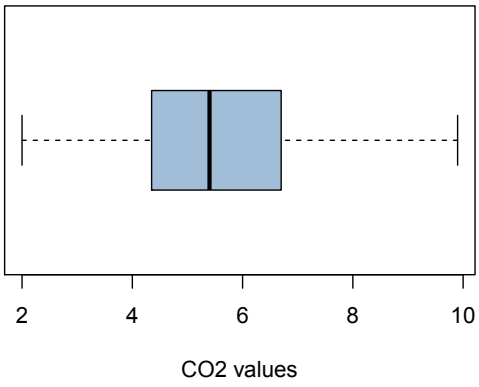


FIGURE 1.5 Box plot of CO2 values for European nations.

Many studies compare different groups on some variable. Side-by-side box plots are useful for making comparisons. To illustrate, Figure 1.6 shows side-by-side box plots of the number of murders per 100,000 population in 2018 for the 50 states and the District of Columbia (D.C.) in the U.S. and for the 10 provinces of Canada. The figure shows that the murder rates in the U.S. tended to be much larger, have much greater variability, and have an extremely large outlier (the murder rate of 24.2 in D.C.). Here is R code to construct this plot and organize summary statistics by group:[16]

```
> Crime <- read.table("http://stat4ds.rwth-aachen.de/data/Murder2.dat", header=TRUE)
> boxplot(Crime$murder ~ Crime$nation, xlab="Murder rate", horizontal=TRUE)
```

[15] An observation is identified as an *outlier* if it falls more than 1.5(IQR) below the lower quartile or above the upper quartile and as an *extreme outlier* if more than 3(IQR) away. The box plot is one of many methods of ***exploratory data analysis*** proposed in a landmark book by John Tukey (1977).

[16] Section A.1.3 shows other ways to present statistics by the level of a second variable.

```
> tapply(Crime$murder, Crime$nation, summary)  # applies summary to murder, by nation
$Canada
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.030   1.735   1.673   1.875   4.070
$US
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.650   5.000   5.253   6.450  24.200
```

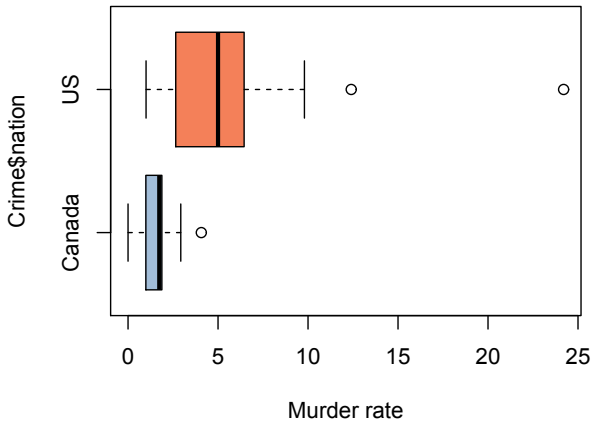


FIGURE 1.6 Side-by-side box plots for U.S. and Canadian murder rates.

1.5 Descriptive Statistics: Summarizing Multivariate Data

For statistical analyses with two or more variables, descriptive methods investigate *associations* between the response variable and the explanatory variables. For a quantitative response variable and categorical explanatory variable, we can compare means or medians on the quantitative variable for the groups that are formed by the categories of the explanatory variable and display results in a side-by-side box plot, such as the R output above and Figure 1.6 show for murder rate and nation. We next present descriptive methods for pairs of quantitative variables and for pairs of categorical variables.

1.5.1 Bivariate Quantitative Data: The Scatterplot, Correlation, and Regression

For a pair of quantitative variables, we can plot values for the explanatory variable on the horizontal (x) axis and for the response variable on the vertical (y) axis. The values of the two variables for any particular observation form a point relative to these axes. A ***scatterplot*** portrays the n observations as n points.

For example, Figure 1.7 plots the relation in the U.S. for the 50 states and D.C. between x = percent of people in state who own guns and y = suicide rate, measured as the annual number of suicides per 100,000 people in the state. The scatterplot for the 51 observations shows that relatively high values of x tend to occur with relatively high values of y, partly reflecting that guns are used in slightly more than half of all suicides in the U.S.

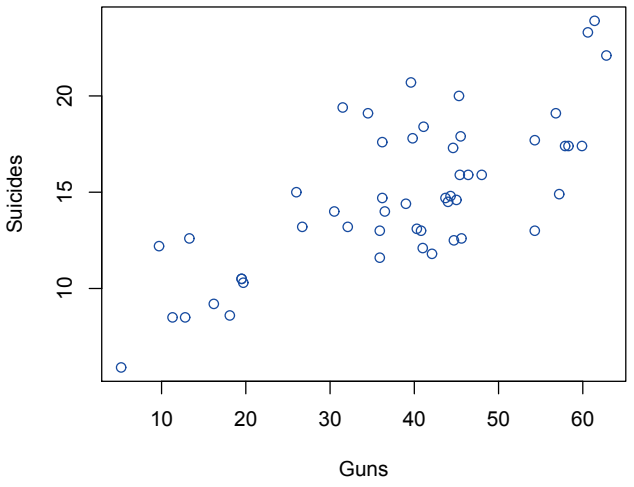


FIGURE 1.7 Scatterplot relating state-level data in the U.S. on percent gun ownership and suicide rate.

Chapter 6 presents two ways to describe such a trend. The ***correlation*** describes the strength of the association, in terms of how closely the data follow a *straight line trend*. For Figure 1.7, the correlation is 0.74. The positive value means that the suicide rate tends to go *up* as gun ownership goes *up*. The correlation takes values between −1 and +1. The larger it is in absolute value, the stronger the association. A ***regression analysis*** provides a straight-line formula for predicting the response variable from the explanatory variable. For the data in Figure 1.7, this equation is

$$\text{Predicted suicide rate} = 7.390 \ + \ 0.1936(\text{gun ownership}).$$

For a state with $x = 5.2$ (the lowest value in this sample, which is for D.C.), the predicted suicide rate is $7.390 + 0.1936(5.2) = 8.40$. For a state with $x = 62.8$ (the highest value, which is for Wyoming), the predicted suicide rate is $7.390 + 0.1936(62.8) = 19.55$.

Chapter 6 derives formulas for the correlation and the regression line. It is simple to implement them with software, as shown next with R.

```
> GS <- read.table("http://stat4ds.rwth-aachen.de/data/Guns_Suicide.dat", header=TRUE)
> Guns <- GS$guns; Suicides <- GS$suicide
> plot(Guns, Suicides)          # scatterplot with arguments x, y
> cor(Guns, Suicides)           # correlation
[1] 0.7386667
> summary(lm(Suicides ~ Guns)) # lm(y ~ x) is "linear model" for
Coefficients:                   # response variable y and explanatory variable x
             Estimate
(Intercept)  7.39008            # intercept of line fitted to data points
Guns         0.19356            # slope of line
```

1.5.2 Bivariate Categorical Data: Contingency Tables

For two categorical variables, a ***contingency table*** is a rectangular table that cross-classifies the variables. The *cells* show the combinations of categories and their counts. For instance,

TABLE 1.1 Contingency table cross-classifying political party identification (ID) and race.

	Political Party ID		
Race	Democrat	Independent	Republican
Black	281	66	30
Other	124	77	52
White	633	272	704

Table 1.1 is a contingency table for political party identification (ID = Democrat, Independent, or Republican) and race, using data from the 2018 General Subject Survey.

Treating political party ID as the response variable, we can summarize by finding percentages in each ID category, by race. For instance, 44% of Whites and 8% of Blacks identified as Republicans. The following code uses `R` with the `PartyID` data file at the book's website to construct the contingency table and find the ID proportions separately for each category of race:

```
> PID <- read.table("http://stat4ds.rwth-aachen.de/data/PartyID.dat", header=TRUE)
> PID
       race          id
1     white    Democrat
...
2238  other  Republican       # 2238 subjects in data file
> table(PID$race, PID$id)     # forms contingency table (not shown here; see Table 1.1)
> options(digits=2)
> prop.table(table(PID$race, PID$id), margin=1)  # For margin=1, proportions
       Democrat Independent Republican           # sum to 1.0 within rows
  black     0.75        0.17       0.08
  other     0.49        0.30       0.21
  white     0.39        0.17       0.44
> mosaicplot(table(PartyID$race, PartyID$id))    # graphical portrayal of cell sizes
```

As an exercise, check what the `mosaicplot` function portrays. Contingency tables extend to multi-dimensional tables to handle several variables at once.

1.5.3 Descriptive Statistics for Samples and for Populations

This chapter has introduced commonly-used graphical and numerical descriptive statistics. We introduce others in later chapters and in the book and website appendices. Of descriptive statistics, the mean $\bar{y}$ and the standard deviation s are the most commonly reported measures of center and variability. The correlation and the regression slope are commonly reported measures of association. Since the values of these statistics depend on the sample selected, they vary in value from sample to sample. In this sense, they are also variables.

For example, for a population, Stanford–Binet IQ scores have a bell shape and are scaled to have a mean of 100 and standard deviation of 16. The following `R` code randomly samples 30 people from a bell-shaped population with this center and spread and finds $\bar{y}$ and s and constructs a histogram. Then it repeats with another sample of size 30, to show that $\bar{y}$ and s and the histogram shape vary from sample to sample.

```
> y1 <- rnorm(30, 100, 16)        # randomly sample normal distribution (Sec. 2.5.1)
> mean(y1); sd(y1); hist(y1)      # histogram (not shown)
[1] 102.7339                      # mean
[1] 11.64643                      # standard deviation
> y2 <- rnorm(30, 100, 16)        # another random sample of size n=30
> mean(y2); sd(y2); hist(y2)
[1] 99.17068   # mean and standard deviation change for each sample of size n=30
[1] 16.52736
```

Do this several times to investigate how $\bar{y}$ and s vary from sample to sample around the population values. You will observe that the histogram may portray the population poorly, sometimes even showing multiple modes. Now do this several times with $n = 1000$ instead of 30. The statistics will vary less from sample to sample, and the histogram will better portray the population. Chapter 3 shows that, for a particular value of n, we can predict how much $\bar{y}$ varies among samples.

Inferential statistical methods use sample descriptive statistics to make predictions about corresponding parameters for the population. In this text, *lower-case Greek letters denote population parameters and Roman letters denote sample statistics.* For example, $\bar{y}$ and s denote a sample mean and standard deviation, and μ and σ denote the population mean and standard deviation. For the IQ sampling just shown, $\mu = 100$ and $\sigma = 16$, whereas $\bar{y} = 102.73$ and $s = 11.65$ for the first sample of 30 observations and $\bar{y} = 99.17$ and $s = 16.53$ for the second sample. We use π for a population proportion.

Before learning about inferential statistical methods, however, you need some basic tools of *probability*, which serves as the language for expressing uncertainty about inferences. Probability is the subject of Chapter 2.

1.6 Chapter Summary

The field of statistical science includes methods for

- designing research studies,
- summarizing the data (***descriptive statistics***),
- making estimations and predictions using the data (***inferential statistics***).

Statistical methods apply to observations in a ***sample*** taken from a ***population***. With randomization, samples are likely to be representative of the population. For a ***simple random sample***, every possible sample has the same chance of selection. ***Statistics*** summarize sample data, while ***parameters*** summarize entire populations. Inferential statistics use sample data to make predictions about population parameters.

A ***data file*** has a separate row of data for each subject and a separate column for each characteristic. Statistical methods analyze data on ***variables***, which are characteristics that vary among subjects.

- Numerically measured variables, such as family income and number of children in a family, are ***quantitative***.
- Variables taking values in a set of categories, such as race and gender, are ***categorical***.
- Variables are also classified as ***discrete***, such as categorical variables and quantitative variables that take values in a set of separate numbers (e.g., 0, 1, 2, ...), or ***continuous***, having a continuous, infinite set of possible values.

Descriptive statistics summarize key characteristics of the data. A ***frequency distribution*** lists numbers of observations for possible values or intervals of values of a variable. For a quantitative variable, a ***histogram*** uses bars over possible values or intervals of values to portray a frequency distribution. It shows whether the distribution is approximately bell-shaped or skewed to the right (longer tail pointing to the right) or to the left. A ***box***

plot portrays the quartiles (25th and 75th percentiles), the extreme values, and any outliers. Table 1.2 summarizes the most important numerical measures that describe the *center* of n observations and their *variability* (spread).

TABLE 1.2 Summary of descriptive statistical measures of center and variability.

Measure	Definition	Interpretation
Center		
Mean	$\overline{y} = (\sum_i y_i)/n$	Center of gravity
Median	Middle observation of ordered sample	50th percentile, splits sample into two equal parts
Variability		
Standard deviation	$s = \sqrt{\sum(y_i - \overline{y})^2/(n-1)}$	If bell-shaped, 68%, 95%, nearly all within $s, 2s, 3s$ of $\overline{y}$

Bivariate statistics summarize the ***association*** between two variables and how the outcome on a ***response variable*** depends on the value of an ***explanatory variable***.

- For quantitative variables, a ***scatterplot*** graphs the observations as points with axes for the variables. The ***correlation*** describes the strength of straight-line association and the ***regression line*** can predict the response variable using the explanatory variable.
- For categorical variables, a ***contingency table*** shows the number of observations at the combinations of possible category outcomes for the two variables.

Exercises

Data Analysis and Applications

1.1 In the 2018 election for Senate in California, a CNN exit poll of 1882 voters stated that 52.5% voted for the Democratic candidate, Diane Feinstein. Of all 11.1 million voters, 54.2% voted for Feinstein.

(a) What was the (i) subject, (ii) sample, (iii) population?

(b) Identify a relevant statistic and corresponding parameter.

1.2 The `Students` data file at `http://stat4ds.rwth-aachen.de/data` shows responses of a class of 60 social science graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), *cogpa* = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every

week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* = belief in life after death (1 = yes, 2 = no, 3 = undecided). You will use this data file for some exercises in this book.

(a) Practice accessing a data file for statistical analysis with your software by going to the book's website and copying and then displaying this data file.

(b) Using responses on *abor*, state a question that could be addressed with (i) descriptive statistics, (ii) inferential statistics.

1.3 Identify each of the following variables as categorical or quantitative: **(a)** Number of smartphones that you own; **(b)** County of residence; **(c)** Choice of diet (vegetarian, nonvegetarian); **(d)** Distance, in kilometers, commute to work

1.4 Give an example of a variable that is **(a)** categorical; **(b)** quantitative; **(c)** discrete; **(d)** continuous.

1.5 In analyzing data about patients who developed Covid-19 from coronavirus, many research studies used the scale (1. Death; 2. Hospitalized with invasive ventilation; 3. Hospitalized with non-invasive ventilation; 4. Hospitalized with supplemental oxygen; 5. Hospitalized, not requiring supplemental oxygen but requiring ongoing medical care; 6. Hospitalized, not requiring ongoing medical care (quarantine or awaiting rehab); 7. Not hospitalized, limitation on activities; 8. Not hospitalized, no limitations on activities). Is this categorical scale *nominal* or *ordinal*? Why?

1.6 Give an example of a variable that is (**a**) technically discrete but essentially continuous for purposes of data analysis; (**b**) potentially continuous but highly discrete in the way it is measured in practice.

1.7 The student directory for a large university has 400 pages with 130 names per page, a total of 52,000 names. Using software, show how to select a simple random sample of 10 names.

1.8 Explain whether an experiment or an observational study would be more appropriate to investigate the following:

(a) Whether cities with higher unemployment rates tend to have higher crime rates.

(b) Whether a Honda Accord hybrid or a Toyota Prius gets better gas mileage.

(c) Whether higher college grade point averages tend to occur for students who had higher scores on college entrance exams.

(d) Whether design A or design B for an Internet page makes a person more likely to buy the product advertised.

1.9 For the GSS data on number of good friends (search at `https://sda.berkeley.edu/sdaweb/analysis/?dataset=gss18` for variable `NUMFREND`), the responses (1, 2, 3, 4, 5, 6, 7, 8, 9, ... , 96+) had percentages of (6.1, 16.2, 15.7. 14.2, 11.3, 8.8, 1.2, 2.4, 0.7, ..., 0.8). Report the median and the mode. Would you expect the mean to be smaller, the same, or larger than the median? Why?

1.10 Analyze the `Carbon_West` data file at the book's website by (**a**) constructing a frequency distribution and a histogram, (**b**) finding the mean, median, and standard deviation. Interpret each.

1.11 According to Statistics Canada, for the Canadian population having income in 2019, annual income had a median of $35,000 and mean of $46,700. What would you predict about the shape of the distribution? Why?

1.12 Give an example of a variable that is nonnegative but has the majority of its sample values at 0, so the median is not especially informative despite the skew.

1.13 A report indicates that public school teacher's annual salaries in New York city have an approximate mean of $69,000 and standard deviation of $6,000. If the distribution has approximately a bell shape, report intervals that contain about (**a**) 68%, (**b**) 95%, (**c**) all or nearly all salaries. Would a salary of $100,000 be unusual? Why?

1.14 In 2017, according to the Kaiser Family Foundation (www.kff.org), the five-number summary for the U.S. statewide percentage of people without health insurance had minimum = 3% (Massachusetts), lower quartile = 6%, median = 8%, upper quartile = 9.5%, and maximum = 17% (Texas). Interpret the quartiles and the interquartile range, and sketch a box plot.

1.15 According to www.salary.com, the mean salary (in dollars) of secondary school teachers in the United States in 2019 varied among states with a five-number summary of maximum = 67,600 (California), upper quartile = 64,700, median = 55,500, lower quartile = 53,100, and minimum = 51,600 (South Dakota). Sketch a box plot, and indicate whether the distribution seems to be symmetric, skewed to the right, or skewed to the left.

1.16 Access the most recent General Social Survey at https://sda.berkeley.edu/archive.htm. Entering TVHOURS for the row variable and year(2018) in the selection filter, you obtain data on hours per day of TV watching in the U.S. in 2018.

(a) Construct the frequency distribution for the values 0, 1, 2, 3, 4, 5, 6, 7 or more. How would you describe its shape?

(b) Find the median and the mode.

(c) Check *Summary statistics* in the output options, and report the mean and standard deviation. From these, explain why you would not expect the distribution to be bell-shaped.

1.17 From the `Murder` data file at the book's website, use the variable *murder*, which is the murder rate (per 100,000 population) for each state in the U.S. in 2017 according to the FBI Uniform Crime Reports. At first, do not use the observation for D.C. (`DC`). Using software:

(a) Find the mean and standard deviation and interpret their values.

(b) Find the five-number summary, and construct the corresponding box plot. Interpret.

(c) Now include the observation for D.C. What is affected more by this outlier: The mean or the median? The range or the inter-quartile range?

1.18 The `Income` data file at the book's website reports annual income values in the U.S., in thousands of dollars.

(a) Using software, construct a histogram. Describe its shape.

(b) Find descriptive statistics to summarize the data. Interpret them.

(c) The *kernel density estimation* method finds a smooth-curve approximation for a histogram. At each value, it takes into account how many observations are nearby and their distance, with more weight given those closer. Increasing the *bandwidth* increases the influence of observations further away. Plot a smooth-curve approximation for the histogram of income values, using the `density` function in `R`. Summarize the impact of increasing and of decreasing the bandwidth (option `bw` in the `density` function) substantially from the default value.[17]

(d) Construct and interpret side-by-side box plots of income by race (B = Black, H = Hispanic, W = White). Compare the incomes using numerical descriptive statistics.

1.19 The `Houses` data file at the book's website lists the selling price (thousands of dollars), size (square feet), tax bill (dollars), number of bathrooms, number of bedrooms, and whether the house is new (1 = yes, 0 = no) for 100 home sales in Gainesville, Florida. Let's analyze the selling prices.

(a) Construct a frequency distribution and a histogram. Describe the shape.

(b) Find the percentage of observations that fall within one standard deviation of the mean. Why is this not close to 68%?

(c) Construct a box plot, and interpret.

(d) Use descriptive statistics to compare selling prices according to whether the house is new.

1.20 Refer to the previous exercise. Let y = selling price and x = size of home.

(a) Construct a scatterplot. Interpret. Identify any observation that seems to fall apart from the others.

(b) Find the correlation. Interpret.

(c) Find the regression line. Interpret the slope, and find the predicted selling price for a home of (i) 1000 square feet, (ii) 4000 square feet.

1.21 For the `Students` data file introduced in Exercise 1.2, summarize the relationship between *hsgpa* and *cogpa* using correlation and regression. Find the predicted college GPA of a student who had a high school GPA of 4.0.

1.22 Using the `Happy` data file, construct the contingency table relating marital status and happiness. Which variable is the natural response variable? Report the proportions in its categories, separately at each category of the explanatory variable, and interpret.

1.23 For the `Students` data file introduced in Exercise 1.2, construct and summarize a contingency table relating religiosity and opinion about legalized abortion.

1.24 The `UN` data file at the book's website has United Nations data for 42 nations on per capita gross domestic product (GDP, in thousands of dollars), a human development index (HDI, which has components referring to life expectancy at birth, educational attainment, and income per capita), a gender inequality index (GII, a composite measure reflecting inequality in achievement between women and men in reproductive health, empowerment, and the labor market), fertility rate (number of births per woman), carbon dioxide emissions per capita (CO2, in metric tons), a homicide rate (number of homicides per 100,000 people), prison population (per 100,000 people), and percent using the Internet.

[17] The default is the so-called *Silverman's rule of thumb*, option `bw = "nrd0"`.

(a) Conduct a descriptive statistical analysis of the prison rates. Summarize your conclusions, highlighting any unusual observations.

(b) Using a command like `cor(cbind(GDP, HDI, GII, …, Internet))` in `R`, construct a *correlation matrix* showing the correlation for each pair of variables. Which pair has the strongest association?

(c) Conduct correlation and regression analyses to study the association between CO2 use and GDP. Show how the predicted CO2 varies as GDP goes from its minimum to its maximum value in this sample.

1.25 The `ScotsRaces` data file at the book's website shows the record times for men and for women for several hill races in Scotland. Use graphical and numerical descriptive statistics to compare the men's winning times with the women's winning times. Summarize your analyses in a short report, with software output as an appendix.

1.26 Refer to the previous exercise. Explanatory variables listed in the data file are the distance of the race and the climb in elevation. Use graphical and numerical descriptive statistics to summarize the men's winning times and their relationship with the race distance and climb. Summarize your analyses in a one-page report.

1.27 A study of sheep[18] analyzed whether the sheep survived for a year from the original observation time (1 = yes, 0 = no) as a function of their weight (kg) at the original observation. Using `Sheep` data file at the text website, use graphical and numerical methods of this chapter to compare weights of the sheep that survived to weights of the sheep that did not survive. Summarize the results of your analysis in a few sentences.

Methods and Concepts

1.28 The beginning of Section 1.2 mentioned a potential survey to observe characteristics such as opinion about the legalization of same-sex marriage, political party affiliation, frequency of attending religious services, number of years of education, annual income, marital status, race, and gender. Describe two ways you could select one of these variables as a response variable and the others as explanatory variables, and explain the reasoning for these choices.

1.29 A *systematic random sample* of n subjects from a population of size N selects a subject at random from the first $k = N/n$ in the population list and then selects every kth subject listed after that one. Explain why this is not a simple random sample.

For the following two multiple-choice questions, select the best response.

1.30 A simple random sample of size n is one in which:

(a) Every nth member is selected from the population.

(b) Each possible sample of size n has the same chance of being selected.

(c) There must be exactly the same proportion of women in the sample as is in the population.

(d) You keep sampling until you have a fixed number of people having various characteristics (e.g., males, females).

[18] Summarized in article by T. Coulson, *Oikos*, **121**: 1337–1350 (2012); thanks to Prof. M. K. Oli for the data from this study.

(e) A particular minority group member of the population is less likely to be chosen than a particular majority group member.

(f) All of the above.

1.31 If we use random numbers to take a simple random sample of 50 students from the 6500 undergraduate students at the University of Rochester:

(a) We would never get the random number 1111, because it is not a random sequence.

(b) The draw 1234 is no more or less likely than the draw 1111.

(c) Since the sample is random, it is *impossible* that it will be non-representative, such as having only females in the sample.

(d) Since the sample is random, it is impossible to get the sequence of random numbers 0001, 0002, 0003, ..., 0049, 0050.

1.32 With an Internet search, find a study that used an (**a**) experiment; (**b**) observational study. In each case, describe how the sample was obtained and summarize results.

1.33 An article[19] in the *New England Journal of Medicine* (October 12, 2012) observed a correlation of 0.79 for 23 countries between per capita annual chocolate consumption and the number of Nobel laureates per 10 million population. Was this study an experiment or an observational study? Can we conclude that increasing chocolate consumption increases the chance of a Nobel prize? Why or why not?

1.34 A research study funded by Wobegon Springs Mineral Water, Inc., discovers that children with dental problems are less common in families that regularly buy bottled water than in families that do not. Explain why this association need not reflect a causal link between drinking bottled water and having fewer dental problems. Identify lurking variables that could be responsible for the result, and explain how.

1.35 Suppose that grade-point averages at your university are bell-shaped with mean 3.0 and standard deviation 0.3. Randomly sample n students several times for $n = 20$ (using a function such as `rnorm` in `R`), and then repeat several times for $n = 1000$, each time constructing a histogram. What does this suggest about the difficulty of determining the shape of a population distribution when n is small?

1.36 Construct a set of data for which the mean and median are identical.

1.37 Suppose you estimate the mean number of friends that members of Facebook have by randomly sampling 100 members of Facebook and (i) averaging the numbers of friends that they have, (ii) averaging how many friends the friends of those members have. Which estimate do you think would be larger? Why? (*Hint*: See article by J. A. Paulos, *Scientific American*, Feb. 1, 2011.)

1.38 To measure center, why is the **(a)** median sometimes preferred over the mean? **(b)** mean sometimes preferred over the median? (*Hint*: A wide variety of highly discrete frequency distributions can have the same median.) To illustrate, in your answers use the variables annual income for (a) and number of times you played a sport in the past week for (b).

1.39 To measure variability, why is the **(a)** standard deviation s usually preferred over the range? **(b)** interquartile range often preferred over the range?

[19] See www.nejm.org/doi/full/10.1056/NEJMon1211064

1.40 The largest value in a sample is moved upwards so that it is an extreme outlier. Explain how, if at all, this affects the mean, median, range, and interquartile range.

1.41 To investigate how $\bar{y}$ can vary from sample to sample of size n, for the simulation from a bell-shaped population shown at the end of Section 1.5.3, take **(a)** 10,000 random samples of size $n = 30$ each; **(b)** 10,000 random samples of size $n = 1000$ each. In each case, form a histogram of the 10,000 $\bar{y}$ values and find their standard deviation. Compare results and explain what this simulation reveals about the impact of sample size on how study results can vary. (Chapter 3 shows that in sampling from a population with standard deviation 16, the theoretical standard deviation of $\bar{y}$ values is $16/\sqrt{n}$.)

1.42 The Internet site `www.artofstat.com/web-apps` has useful apps [20] for illustrating data analyses and properties of statistical methods.

(a) Using the *Explore Quantitative Data* app, construct a sample of 20 observations on y = number of hours of physical exercise in the past week having $\bar{y} < s$. What aspect of the distribution causes this to happen?

(b) Using the *Explore Linear Regression* app with the *Draw Own* option, create 20 data points that are plausible for x = number of hours of exercise last week and y = number of hours of exercise this week. Describe your data by the correlation and by the linear regression line, and interpret them.

1.43 For a sample with mean $\bar{y}$, show that adding a constant c to each observation changes the mean to $\bar{y} + c$, and the standard deviation s is unchanged. Show that multiplying each observation by c changes the mean to $c\bar{y}$ and the standard deviation to $|c|s$.

1.44 Suppose the sample data distribution of $\{y_i\} = \{y_1, \ldots, y_n\}$ is very highly skewed to the right, and we take logs and analyze $\{x_i = \log(y_i)\}$.

(a) Is $\bar{x} = \log(\bar{y})$? Why or why not?

(b) Is $\text{median}(\{x_i\}) = \log[\text{median}(\{y_i\})]$? Why or why not?

(c) To summarize $\{y_i\}$, we find $\bar{x}$ and then use $\exp(\bar{x})$. Show that $\exp(\bar{x}) = (\prod_i y_i)^{1/n}$, called the ***geometric mean*** of $\{y_i\}$.

1.45 Find the Chebyshev inequality upper bound for the proportion of observations falling at least (**a**) 1, (**b**) 2, (**c**) 3 standard deviations from the mean. Compare this to the approximate proportions for a bell-shaped distribution. Why are the differences so large?

1.46 The `R` output in Section 1.5.2 shows a `mosaicplot` function. Implement it for the `PartyID` data. Do an Internet search and look at the `R` Appendix to this book to learn about *mosaic plots*, and describe what the plot shows you. (The `mosaic` function in the `vcd` package constructs more sophisticated mosaic plots.)

1.47 The *least squares* property of the mean states that the data fall closer to $\bar{y}$ than to any other number c, in the sense that

$$\sum_i (y_i - \bar{y})^2 < \sum_i (y_i - c)^2.$$

Prove this property by treating $f(c) = \sum_i (y_i - c)^2$ as a function of c and deriving the value of c that minimizes it.

[20] The apps at that website were developed by Dr. Bernhard Klingenberg.

1.48 For a sample $\{y_i\}$ of size n, $\sum_i |y_i - c|$ is minimized at c = median. Explain why this property holds. (Hint: Starting at c = median, what happens to $\sum_i |y_i - c|$ as you move away from it in either direction?)

1.49 The Delphi group at Carnegie-Mellon University has tracked statistics about coronavirus. Figure 1.8 is a scatterplot of U.S. statewide data compiled between x = percentage wearing masks and y = percentage knowing someone with Covid-19 symptoms. Which do you think best describes the scatterplot?

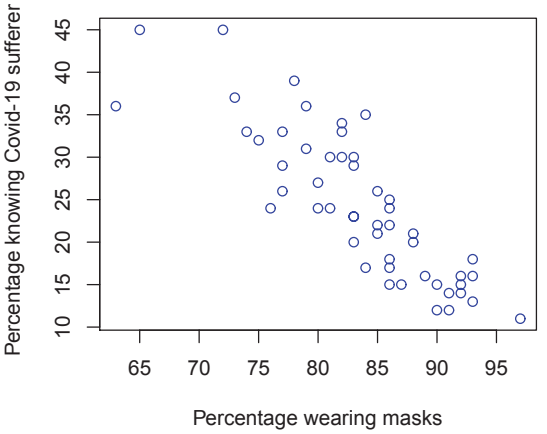


FIGURE 1.8 Scatterplot for data file `CovidMasks` at book's website, showing U.S. statewide data in October 2020 of percentage of people wearing masks in public all or most of the time and percentage of people knowing someone who has had Covid-19 symptoms

(a) Correlation = −0.85 and predicted $y = 112.9 - 1.06\,x$

(b) Correlation = −0.85 and predicted $y = 112.9 + 1.06\,x$

(c) Correlation = 0.85 and predicted $y = 112.9 - 1.06\,x$

(d) Correlation = 0.85 and predicted $y = 112.9 + 1.06\,x$

2

Probability Distributions

This chapter introduces *probability*, the language used to describe uncertainty, such as in inferential statistical analyses. We first define probability and give basic rules for calculating probabilities. *Random variables* specify possible values of variables in experiments or observational studies that incorporate randomization. A *probability distribution* summarizes random variation in the values of a random variable by specifying probabilities for all the possible outcomes. Probability distributions are themselves summarized by measures of center and variability, such as the mean and standard deviation. These measures are *expected values*, describing what we expect, on the average, with observations taken at random from a probability distribution.

Like ordinary variables, random variables can be *discrete* or *continuous*. The most important probability distributions for discrete random variables are the *binomial* for binary outcomes and the *Poisson* for count outcomes. The most important probability distributions for continuous random variables are the *normal*, which has a bell shape over the real line, and the *gamma*, which is skewed to the right over positive real number values.

Joint probability distributions describe how two or more random variables co-vary. The *correlation* describes their strength of association, with *independence* being the absence of any association.

2.1 Introduction to Probability

In everyday life, the term *probability* is used in an informal manner to mean the *chance* of an outcome. But this definition is rather vague. For studies that can employ *randomization* for gathering data, a more precise definition relates to how often that outcome occurs.

2.1.1 Probabilities and Long-Run Relative Frequencies

For random phenomena, such as observations in a randomized experiment, the probability of an outcome refers to the proportion of times the outcome would occur in a very long sequence of like observations.

Probability

For an observation of a random phenomenon, the ***probability*** of a particular outcome is the proportion of times that outcome would occur in an indefinitely long sequence of like observations, under the same conditions.

Probabilities are sometimes expressed as percentages. For this definition, a weather forecast that the probability of rain tomorrow is 0.20, or 20%, means that in a long run of repeated observations of days with atmospheric conditions like those expected tomorrow, rain would occur on 20% of the days. We could simulate the outcome for tomorrow by

DOI: 10.1201/9781003159834-2

generating a random digit from $(0, 1, 2, \ldots, 9)$, where 2 of the 10 possible outcomes (say, 0 and 1) denote *rain* and the other 8 possible outcomes denote *no rain*; for example, using R:

```
> sample(0:9, 1)      # randomly generate 1 integer between 0 and 9
[1] 8
```

The outcome of the simulation was an 8, representing no rain. Let's simulate for a week under the same conditions:

```
> sample(0:9, 7, replace=TRUE)    # with replace=TRUE, numbers replaced,
[1] 5 8 1 9 8 9 5                 # each selection has same choice set
```

In this simulation, it rains on day 3 but not on any of the other six days.

The outcome on a particular day corresponds to flipping an unbalanced coin that has probability 0.20 of a head, where we identify the *head* outcome with *rain*. More generally, we can simulate how often rain occurs in n days by flipping a coin n times, when each flip has probability 0.20 of a head and probability 0.80 of a tail. We can do this using the R function **rbinom**, which provides simulations with binary outcomes for each observation.[1] Here is the result for 1 simulation of $n = 7$ days:

```
> rbinom(1, 7, 0.20) # 1 simulation of 7 coin flips, probability 0.20 of head
[1] 3                # outcome is 3 heads in 7 flips (i.e., rain on 3 days)
> rbinom(7, 1, 0.20) # Or, 7 simulations of 1 coin flip, probability 0.20 of head
[1] 0 1 0 0 0 0 1    # 1 = head, 0 = tail; simulated outcome is rain on days 2 and 7
```

For the first simulation, it rained on 3 of the 7 days. The proportion $3/7 = 0.43$ is quite far from the stated probability of 0.20. But a probability is a *long-run* relative frequency, in theory letting the number of days $n \to \infty$. Let's see what happens with much larger n for a simulation:

```
> rbinom(1, 100, 0.20)
[1] 18                    # proportion 0.180; 18 heads in 100 coin flips
> rbinom(1, 1000, 0.20)
[1] 204                   # proportion 0.204
> rbinom(1, 10000, 0.20)
[1] 2010                  # proportion 0.2010
> rbinom(1, 100000, 0.20)
[1] 20032                 # proportion 0.2003
> rbinom(1, 1000000, 0.20)
[1] 199859                # proportion 0.1999; 199859 heads in 1000000 coin flips
```

According to this simulation, the proportions of days of rain out of n = (100, 1000, 10000, 100000, 1000000) days are (0.180, 0.204, 0.201, 0.2003, 0.1999). The proportion converges to 0.2000 as n increases indefinitely. Figure 2.1 illustrates how the proportion stabilizes as n increases.

The long-run sequence of like observations for a definition of probability is not always appropriate. For example, it is not meaningful for the probability that intelligent life exists elsewhere in the universe or for the probability that a new business is successful, because no long-run sequence of observations is available. We must then rely on *subjective* beliefs rather than *objective* data. In fact, an alternative definition of probability is subjective:

Subjective definition of probability: The probability of an outcome is the degree of belief that the outcome will occur, based on all the available information.

[1] Besides assuming the same probability for each coin flip, this function assumes that observations are "independent events" in a sense to be introduced in Section 2.1.6; for instance, whether it rains one day is not influenced by whether it rained the previous day.