Cover Design: Wiley
Cover Image: © andresr/Getty Images

www.wiley.com/go/johnson/statistics8e

WILEY

ISBN 978-1-119-49711-0
90000
9 781119 497110

JOHNSON
BHATTACHARYYA

STATISTICS Principles and Methods

Eighth Edition

WILEY

RICHARD A. JOHNSON • GOURI K. BHATTACHARYYA

STATISTICS

Principles and Methods

Eighth Edition

25
10
9
2
4
3.0 3.5 4.0 4.5 5.0 5.5 6.0 pH
Histogram of acid rain data.
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ or $\frac{\sum x_i}{n}$

WILEY

8th Edition

Statistics

Principles and Methods

Richard A. Johnson
University of Wisconsin at Madison

Gouri K. Bhattacharyya

WILEY

V.P. & DIRECTOR	Laurie Rosatone
EDITORIAL DIRECTOR	Don Fowley
EDITOR	Jennifer Brady
EDITORIAL MANAGER	Judy Howarth
CONTENT MANAGEMENT DIRECTOR	Lisa Wojcik
CONTENT MANAGER	Nichole Urban
SENIOR CONTENT SPECIALIST	Nicole Repasky
PRODUCTION EDITOR	Loganathan Kandan
COVER PHOTO CREDIT	© andresr/iStock.com

This book was set in 10/12 pt BerlingLTStd-Roman by SPi Global and printed and bound by Quad Graphics.

ISBN: 978-1-119-49711-0 (PBK)
ISBN: 978-1-119-49710-3 (EVALC)

Library of Congress Cataloging in Publication Data:

Names: Johnson, Richard A. (Richard Arnold), 1937- author. | Bhattacharyya, Gouri K., 1940- author.
Title: Statistics : principles and methods / Richard A. Johnson (University of Wisconsin at Madison), Gouri K. Bhattacharyya.
Description: 8th edition. | Hoboken, NJ : Wiley, [2019] | Includes bibliographical references and index. |
Identifiers: LCCN 2018037971 (print) | LCCN 2018039352 (ebook) | ISBN 9781119497073 (Adobe PDF) | ISBN 9781119497059 (ePub) | ISBN 9781119497110 (pbk.)
Subjects: LCSH: Statistics–Textbooks.
Classification: LCC QA276.12 (ebook) | LCC QA276.12 .J63 2019 (print) | DDC 519.5–dc23

LC record available at https://lccn.loc.gov/2018037971

The inside back cover will contain printing identification and country of origin if omitted from this page. In addition, if the ISBN on the back cover differs from the ISBN on this page, the one on the back cover is correct.

Preface

THE NATURE OF THE BOOK

The first decades of the twenty-first century are earmarked by rapid advances in digital processing and the resulting digitalization of data. Extremely large numbers of data sets and very big data sets abound. The need to explore these data to uncover patterns and relationships has elevated the importance of statistics.

Statistics is the science of learning from data. It encompasses making measurements, reaching conclusions, and communicating the amount of related uncertainty. Its principles help determine what conclusions are reliable and, more generally, provide the navigational guidance for making advances in science and addressing societal issues. Statistics—the subject of data analysis and data-based reasoning—is necessarily playing a vital role in virtually all professions. Some familiarity with this subject is now an essential component of any college education. Yet, pressures to accommodate a growing list of academic requirements often necessitate that this exposure be brief. Keeping these conditions in mind, this book provides students with a first exposure to the powerful ideas of modern statistics. It presents the key statistical concepts and the most commonly applied methods of statistical analysis. It is accessible to freshmen and sophomores from a wide range of disciplines because we have avoided most mathematical derivations. They can pose a stumbling block to learning the essentials in a short period of time.

This book does not require students to have a strong background in mathematics. It is intended for those who seek to learn the basic ideas of statistics and their application in a variety of practical settings. The core material of this book is common to almost all first courses in statistics and is designed to be covered well within a one-semester course in introductory statistics for freshmen to seniors. It is supplemented with some additional special-topics chapters.

ORIENTATION

The topics treated in this text are, by and large, the ones typically covered in an introductory statistics course. They span three major areas: (i) descriptive statistics, which deals with summarization and description of data; (ii) ideas of probability and an understanding of the manner in which sample-to-sample variation influences our conclusions; and (iii) a collection of statistical methods for analyzing the types of data that are of common occurrence. However, it is the treatment of these topics that makes the text distinctive. Throughout, we have endeavored to give clear and concise explanations of the concepts and important statistical terminology and methods. By means of good motivation, sound explanations, and an abundance of illustrations given in real-world contexts, it emphasizes more than just a superficial understanding.

Each statistical concept or method is motivated by first presenting it in an interesting real-life setting and then focusing on an example to further elaborate important aspects and to illustrate its usefulness. The subsequent discussion is not only limited to showing how a method works but includes an explanation of the why. Even without recourse to mathematics, we are able to make the reader aware of possible pitfalls in the statistical analysis. Students can gain a proper appreciation of statistics only when they are provided with a careful explanation of the underlying logic. Without this understanding, a learning of elementary statistics is bound to be rote and transient.

When describing the various methods of statistical analysis, the reader is continually reminded that the validity of a statistical inference is contingent upon certain model assumptions. Misleading conclusions may result when these assumptions are violated. We think that the teaching of statistics, even at an introductory level, should not be limited to the prescription of methods. Students should be encouraged to develop a critical attitude in applying the methods and to be cautious when interpreting the results. This attitude is especially important in the study of relationship among variables, which is perhaps the most widely used (and also abused) area of statistics. In addition to discussing inference procedures in this context, we have particularly stressed critical examination of the model assumptions and careful interpretation of the conclusions.

To allow an instructor to avoid just teaching formulas and to concentrate on statistical reasoning, this text contains an abundance of data sets in the Data Bank, examples, and exercises. For instance, rather than just declare that a population is normal, the instructor can guide students to an understanding of the importance of checking the validity of the normal assumption and also identifying outliers.

Student projects are becoming an integral part of many first courses in statistics. Our larger data sets provide the basis for numerous projects while the examples help point students toward appropriate statistical analyses.

SPECIAL FEATURES

1. **Crucial elements are boxed** to highlight important concepts and methods. These boxes provide an ongoing summary of the important items essential for learning statistics. At the end of each chapter, all of its **key ideas and formulas** are summarized.
2. **Unique combination of understanding formulas and software**
   Important formulas are illustrated with examples based mostly on integer arithmetic. The student can then verify the result with simple software commands. The process is augmented by exercises that repeat the connection. This helps students comfortably rely on the software and also understand the calculation.

   The same approach is used for tables of the basic statistical distributions. The inclusion of commands for the free-ware R gives every student simple to use software.
3. **A rich collection of examples and exercises** is included. These are drawn from a large variety of real-life settings. In fact, many data sets stem from genuine experiments, surveys, or reports.
4. **Exercises** are provided at the end of each major section. They give the reader the opportunity to practice the ideas just learned. Occasionally, exercises supplement some points raised in the text. A larger collection of exercises appears at the end of a chapter. The starred problems are relatively difficult and suited to the more mathematically competent student.
5. **Using Statistics Wisely**, a feature at the end of each chapter, provides important guidelines for the appropriate use of the statistical procedures presented in the chapter.
6. **Statistics in Context** sections, in the first four chapters and Chapter 7, each describe an important statistical application where a statistical approach to understanding variation is vital. These extended examples reveal, early on in the course, the value of understanding the subject of statistics.
7. ***P*−values** are emphasized in examples concerning tests of hypotheses. Graphs giving the relevant normal or t density curve, rejection region, and P−value are presented.
8. **Regression analysis** is a primary statistical technique so we provide a more thorough coverage than is usual at this level. The basics of regression are introduced in Chapter 11, whereas Chapter 12 stretches the discussion to several issues of practical importance. These include methods of model checking, handling nonlinear relations, and multiple regression analysis. Complex formulas and calculations are judiciously replaced by computer output so the main ideas can be learned and appreciated with a minimum of stress.

9. **Integrated Technology**, at the end of most chapters, details the steps for using R, MINITAB, EXCEL, and TI-84/-83 calculator. With this presentation available, discussion can center on interpreting the computer output discussed in the text. Software packages remove much of the drudgery of hand calculation and they allow students to work with larger data sets where patterns are more pronounced. Some computer exercises are included in all chapters where relevant.
10. **Convenient Electronic Data Bank** at the end of the book contains a substantial collection of data. These data sets, together with numerous others throughout the book, allow for considerable flexibility in the choice between concept-orientated and applications-orientated exercises. The Data Bank and the other larger data sets are available to instructors.
11. **Technical Appendix A** presents a few statistical facts of a mathematical nature. These are separated from the main text so that the instructor can include or exclude them.

ABOUT THE EIGHTH EDITION

The Eighth edition of *STATISTICS—Principles and Methods* maintains the objectives and level of presentation of the earlier editions. The goals are the developing (i) of an understanding of the reasonings by which findings from sample data can be extended to general conclusions and (ii) a familiarity with some basic statistical methods. There are numerous data sets and computer outputs which give an appreciation of the role of the computer in modern data analysis.

Clear and concise explanations introduce the concepts and important statistical terminology and methods. Real-life settings motivate the statistical ideas and well organized discussions proceed to cover statistical methods with heavy emphasis on examples. The Eighth edition enhances these special features. The major improvements are:

New Examples A substantial number of new examples, almost all based on real data, are included. Being based on variety of interesting current activities and issues, they expose the student to the widening sphere of application of statistical methods. The majority are in the core chapters together with Chapter 11 on regression.

Unique Combination of Examples Involving Simple Calculations and Software Verification Key formulas are illustrated with examples based mostly on integer arithmetic. The software commands for R, MINITAB, EXCEL, and TI84 Plus are provided so that students can then verify the result. The process is augmented by exercises that aid the student in both understanding the key formulas and using software.

More Emphasis on *P*−values. There is more discussion and several new graphs illustrating *P*−values in examples that involve testing hypotheses. Several new graphs pertain to the t and χ^2 distributions.

Bayes' Theorem Several new examples and exercises give this subject more prominence.

More Data-Based Exercises Many new exercises are keyed to new data-based examples in the text. Several new data sets are also presented in the exercises. The abundance of real data sets allows an instructor to address such issues as the assumption of normality and outliers.

Poisson Distribution Several examples illustrate its application and the discussion covers rare events and the approximation to the binomial. Key parts can be covered in part of one lecture.

New Exercises Helping Understanding of Concepts Many new exercises provide practice on understanding the concepts while others address understanding related calculations. These new exercises, that augment the already rich collection of exercises, are placed in real-life settings to help promote a greater appreciation of the wide span of applicability of statistical methods.

This book presents the first steps towards understanding the basic statistical concepts and the techniques for making generalizations from a specific data set. In keeping with this century's gigantic advances both in the digitalization of data and the computing power available to students, the teaching of statistics is evolving to include more technology. In turn, this leads to major changes in the first course of statistics. In the eighth edition, we have placed more emphasis on interpretation and downplayed calculation formulas. Because most current students are well versed in keyboard activities, we approach the subject with statistical software playing a more central role.

Because some understanding of the formulas is necessary, we give examples and exercises involving only a few numbers and that require mostly integer arithmetic. We then turn to software calculations to verify the result. This aids in the understanding of calculations when the software is applied to larger more complicated data sets.

Tables of statistical distributions are approached in the same spirit. The full table gives some indication of how the entries vary as sample sizes change. It is desirable that students get an overall feel but are then able to turn to software calculations. This promotes the calculation of exact *P*−values.

In this edition, we have also included R software commands. R project http://CRAN.R-project.org is readily available free to all students.

ORGANIZATION

This book is organized into 15 chapters, an optional technical appendix (Appendix A), and a collection of tables (Appendix B). Although designed for a one-semester or a two-quarter course, it is enriched with ample additional material to allow the instructor some choices of topics. Beyond Chapter 1, which sets the theme of statistics and distinguishes population and sample, the subject matter could be classified as follows:

Topic	**Chapter**
Descriptive study of data	2, 3
Probability and distributions	4, 5, 6
Sampling variability	7
Core ideas and methods of statistical inference	8, 9, 10
Special topics of statistical inference	11, 12, 13, 14, 15

We regard Chapters 1 to 10 as constituting the core material of an introductory statistics course, with the exception of the starred sections in Chapter 6. Although this material is substantial enough for a one-semester course, many instructors may wish to eliminate some sections in order to cover the basics of regression analysis in Chapter 11. This is most conveniently done by initially skipping Chapter 3 and then taking up only those portions that are linked to Chapter 11. Also, instead of a thorough coverage of probability that is provided in Chapter 4, the later sections of that chapter may receive a lighter coverage.

SUPPLEMENTS

Instructor's Solution Manual. This manual contains complete solutions to all exercises. The ISM can be downloaded by instructors from the book companion Web site www.wiley.com/go/johnson/statistics8e.

Test Bank. (Available on the accompanying website: www.wiley.com/go/johnson/statistics8e) Contains a large number of additional questions for each chapter.

Electronic Data Bank. (Available on the accompanying website: www.wiley.com/go/johnson/statistics8e) Contains interesting data sets used in the text but that can be used to perform additional analyses with statistical software packages.

ACKNOWLEDGMENTS

We thank R (http://CRAN.R-project.org) and MINITAB (State College, Pa.) for permission to include commands and output from their software packages. A special thanks to K. T. Wu and Kam Tsui for many helpful suggestions and comments on earlier editions. We also thank all those

who have contributed the data sets which enrich the presentation and all those who reviewed the previous editions. The following people gave their careful attention to this edition:

Wei Lin, Ohio University
Jun Masamune, Penn State Altoona
Tatjana Miljkovic, North Dakota State University
Dan Ostrov, Santa Clara University
Rachel Rader, Ohio Northern University
Laurence D. Robinson, Ohio Northern University
Alla Sikorskii, Michigan State University
Robert L. Sims, George Mason University
James Stamey, Baylor University
Natalia Stepanova, Carleton University
Kenneth Strazzeri, George Mason University
Quoc-Phong Vu, Ohio University
George Petrov Yanev, The University of Texas-Pan American

Richard A. Johnson
Gouri K. Bhattacharyya

Contents

1

Introduction to Statistics

John W. McDonough/Sports Illustrated/Getty Images Inc.

Hometown fans attending today's game are but a sample of the population of all local football fans. A self-selected sample may not be entirely representative of the population on issues such as ticket price increases.

Surveys Provide Information About the Population

What is your favorite spectator sport?	
Football	43%
Baseball	15%
Basketball	9%
Other	33%

College and professional sports are combined in our summary.[1] Clearly, football is the most popular spectator sport. Actually, the National Football League by itself is more popular than baseball.

For many years, baseball was most popular according to similar surveys. Surveys, repeated at different times, can detect trends in opinion.

1. THE SUBJECT AND SCOPE OF STATISTICS

Epic advances in digital processing and storage of massive amounts of data have thrust statistics to the forefront of commercial, governmental, and research activities. Big data sets must be scrutinized for patterns and relationships between variables. Statistical reasoning and methods provide the basis for these analyses.

Statistical reasoning allows us to learn from observations that exhibit considerable variation and to make good decisions in the presence of uncertainty. These are, of course, the conditions that dominate our modern world. To introduce this subject, we first develop a definition of statistics as a subject and then give examples that are typically encountered in our everyday lives.

1.1 WHAT IS STATISTICS?

Historically, the word statistics originated from the Latin word "status," meaning "state." For a long time, it was identified solely with charts and displays of prevailing economic and demographic conditions. Governments require these statistics to underpin their setting of tax policies and the raising of armies. A major segment of today's population still thinks of statistics as synonymous with forbidding arrays of numbers and myriad graphs like those that predominate government reports. But this view is no longer valid. Gigantic advances during the twentieth century have propelled statistics to recognized importance as a field of data-based reasoning.

What, then, are the role and principal objectives of statistics as a scientific discipline? Stretching well beyond the confines of data display, statistics deals with collecting informative data, interpreting these data, and drawing conclusions about a phenomenon under study. The scope of this subject naturally extends to all processes of acquiring knowledge that involve fact finding by collecting and examining data. Opinion polls (surveys of households to study sociological, economic, or health-related issues), agricultural field experiments (with new seeds, pesticides, or farming equipment), clinical studies of vaccines, and cloud seeding for artificial rain production are just a few examples. The principles and methodology of statistics are useful in answering questions such as, What kind and how much data need to be collected? How should we organize and interpret the data? How can we analyze the data and draw conclusions? How do we assess the strength of the conclusions and gauge their uncertainty?

Statistics as a subject provides a body of principles and methodology for designing the process of data collection, summarizing and interpreting the data, and drawing conclusions or generalities.

1.2 STATISTICS IN OUR EVERYDAY LIFE

Fact finding by collecting and interpreting data is not confined to professional researchers. All of us must be able to review and interpret numerical facts and figures to better understand issues of environmental protection, the state of the economy, or the performance of competing football

[1] Obtained by combing professional and college sports from a Harris poll published in 2016.

teams. In our daily lives, we often learn by doing an implicit analysis of factual information. Statistical reasoning will provide a solid basis for improving this step in the learning process.

We are all familiar to some extent with reports in the news media on important statistics.

Employment. A key to providing timely unemployment numbers is choosing to use a sample rather than attempting a complete enumeration. Monthly, as part of the Current Population Survey, the Bureau of Census collects information about employment status from a sample of about 60,000 households. Households are contacted on a rotating basis with three-fourths of the sample remaining the same for any two consecutive months.

The survey data are analyzed by the Bureau of Labor Statistics, which reports monthly unemployment rates. □

Gallup Poll. This, the best known of the national polls, produces estimates of the percentage of popular vote for each candidate based on interviews with a minimum of 1,000 adults per day. Beginning several months before the presidential election, results are regularly published. These reports help predict winners and track changes in voter preferences. □

Our sources of factual information range from individual experience to reports in news media, government records, and articles in professional journals. As consumers of these reports, citizens need some idea of statistical reasoning to properly interpret the data and evaluate the conclusions. Statistical reasoning provides criteria for determining which conclusions are supported by the data and which are not. The credibility of conclusions also depends greatly on the use of statistical methods at the data collection stage. Statistics provides a key ingredient for any systematic approach to improve any type of process from manufacturing to service.

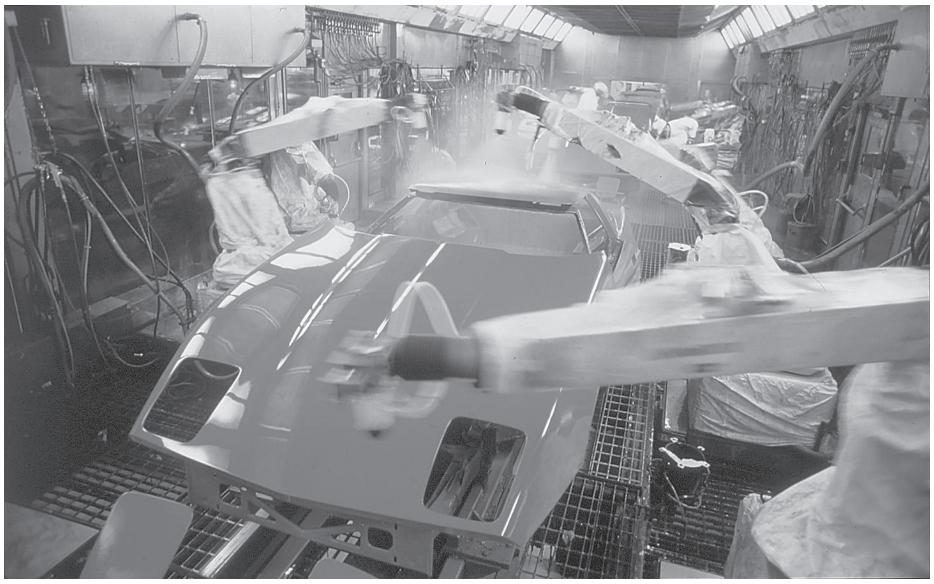
A/B Testing Approach to Online. Many e-commerce companies are currently implementing this strategy for improving their web sites to produce substantially better results. This strategy is an updated method that takes full advantage of the real-time collection of hundreds or more consumer actions each day for the purpose of comparing two versions of a web page or paths to a web page.

Today, many of the largest companies are using and refining this technique to improve their internet sales. For instance, one particular product is selected and then one element of its web page is changed. A picture may be added, an existing picture may be made larger or smaller, or the picture itself may be changed. For instance, a picture of a barbecue grill may be replaced by people grilling at a picnic.

Next, a fraction of the incoming traffic to the product page is directed to the modified page. Each day, for a week or so, the fraction of visitors to the product page who purchase is recorded both for the original page and for the modified page. The two cases give rise to the terminology A and B. If the new page is the more successful of the two, it becomes the new standard and an additional change is tested for improvement. Because of high traffic at the site, even seemingly marginal improvements in drop-off or purchase rates will create substantial additional sales.

This technique was employed, in its early stages of development, during the 2012 presidential campaign to substantially increase donations at one party's web site.

Quality and Productivity Improvement. In the past 35 years, the United States has faced increasing competition in the world marketplace. An international revolution in quality and productivity improvement has heightened the pressure on the U.S. economy. The ideas and teaching of W. Edwards Deming helped rejuvenate Japan's industry in the late 1940s and 1950s. In the 1980s and 1990s, Deming stressed to American executives that, in order to survive, they must mobilize their workforce to make a continuing commitment to quality improvement. His ideas have also been applied to government. The city of Madison, WI, implemented quality improvement projects in the police department and in bus repair and scheduling. In each case, the project goal was better service at less cost. Treating citizens as the customers of government services, the first step was to collect information from them in order to identify situations that needed improvement. One end result was the strategic placement of a new police substation and a subsequent increase in the number of foot patrol persons to interact with the community.



Andrew Sacks/Stone/Getty Images

Statistical reasoning can guide the purposeful collection and analysis of data toward the continuous improvement of any process.

Once a candidate project is selected for improvement, data are collected to assess the current status and then more data are collected after making changes. At this stage, statistical skills in the collection and presentation of summaries are not only valuable but necessary for all participants.

In an industrial setting, statistical training for all employees—production line and office workers, supervisors, and managers—is vital to the quality transformation of American industry. □

2. STATISTICS IN AID OF SCIENTIFIC INQUIRY

The phrase **scientific inquiry** refers to a systematic process of learning. A scientist sets the goal of an investigation, collects relevant factual information (or data), analyzes the data, draws conclusions, and decides further courses of action. We briefly outline a few illustrative scenarios.

Training Programs. Training or teaching programs designed for a specific type of clientele (college students, industrial workers, minority groups, physically handicapped people, developmentally challenged children, etc.) are continually monitored, evaluated, and modified to improve their usefulness to society. To learn about the comparative effectiveness of different programs, it is essential to collect data on the achievement or growth of skill of the trainees at the completion of each program. □

Monitoring Advertising Claims. The public is constantly bombarded with commercials that claim the superiority of one product brand in comparison to others. When such comparisons are founded on sound experimental evidence, they serve to educate the consumer. Not infrequently, however, misleading advertising claims are made due to insufficient experimentation, faulty analysis of data, or even blatant manipulation of experimental results. Government agencies and consumer groups must be prepared to verify the comparative quality of products by using adequate data collection procedures and proper methods of statistical analysis. □

Plant Breeding. To increase food production, agricultural scientists develop new hybrids by cross-fertilizing different plant species. Promising new strains need to be compared with the current best ones. Their relative productivity is assessed by planting some of each variety at a number of sites. Yields are recorded and then analyzed for apparent differences. Scientists are now identifying genes and manipulating them to create new crops. □

bjonesphotography/iStockphoto

Statistically designed experiments are needed to document the advantages of the new hybrid versus the old species.

Genomics. This century's most exciting scientific advances are occurring in biology and genetics. Scientists can now study the genome, or sum total of all of a living organism's genes. The human DNA sequence is now known along with the DNA sequences of hundreds of other organisms.

A primary goal of many studies is to identify the specific genes and related genetic states that give rise to complex traits (e.g., diabetes, heart disease, cancer). New instruments for measuring genes and their products are continually being developed to measure thousands of genes. Due to the impact of the disease and the availability of human tumor specimens, many early studies focused on human cancer. Significant advances have already improved cancer classification, knowledge of cancer biology, and prognostic prediction. A hallmark example of prognostic prediction is MammaPrint, the first genetic test approved by the FDA. This test classifies a breast cancer patient as low or high risk for recurrence.

This is clearly only the beginning. Typically, genomics experiments feature the simultaneous measurement of a great number of responses. As more and more data are collected, there is a growing need for novel statistical methods for analyzing data and thereby addressing critical scientific questions. Statisticians and other computational scientists are playing a major role in these efforts to better human health. □

Factual information is crucial to any investigation. The branch of statistics called **experimental design** can guide the investigator in planning the manner and extent of data collection.

After the data are collected, statistical methods are available that summarize and describe the prominent features of data. These are commonly known as **descriptive statistics**. Today, a major thrust of the subject is the evaluation of information present in data and the assessment of the new learning gained from this information. This is the area of **inferential statistics** and its associated methods are known as the methods of **statistical inference**.

It must be realized that a scientific investigation is typically a process of trial and error. Rarely, if ever, can a phenomenon be completely understood or a theory perfected by means of a single, definitive experiment. It is too much to expect to get it all right in one shot. Even after his first success with the light bulb, Thomas Edison had to continue to experiment with numerous materials for the filament before it was perfected.

Data obtained from an experiment provide new knowledge. This knowledge often suggests a revision of an existing theory, and this itself may require further investigation through more experiments and analysis of data. The box below captures the vital point that a scientific process of learning is essentially iterative in nature.

The Conjecture-Experiment-Analysis Learning Cycle

Learning to Swim

	1st cycle	**2nd cycle**	**3rd Cycle**
Conjecture:	Watch others swim	Try inflatable swim aid	Need instruction
Deduction:	Be comfortable in deep water	Keep head above water	Able to swim
Experiment:	Enter deep end of pool	Float with sport tube	Attend a two week class
Analysis:	Sink and swallow water	Better but really not swimming	Can now float and swim

3. TWO BASIC CONCEPTS—POPULATION AND SAMPLE

In the preceding sections, we cite a few examples of situations where evaluation of factual information is essential for acquiring new knowledge. These examples, drawn from widely differing fields, have limited descriptions of the scope and objectives of the studies. However, a few common characteristics are readily discernible.

First, in order to acquire new knowledge, relevant data must be collected. Second, some amount of variability in the data is unavoidable even though observations are made under the same or closely similar conditions. For instance, the treatment for an allergy may provide long-lasting relief for some individuals whereas it may bring only transient relief or even none at all to others. Likewise, it is unrealistic to expect that college freshmen whose high school records are alike would perform equally well in college. Nature does not follow such a rigid law.

A third notable feature is that access to a complete set of data is either physically impossible or from a practical standpoint not feasible. When data are obtained from laboratory experiments or field trials, no matter how much experimentation has been performed, more can always be done. In public opinion or consumer expenditure studies, a complete body of information would emerge only if data were gathered from every individual in the nation—undoubtedly a monumental if not impossible task. To collect an exhaustive set of data related to the damage sustained by all cars of a particular model under collision at a specified speed, every car of that model coming off the production lines would have to be subjected to a collision! Thus, the limitations of time, resources, and facilities, and sometimes the destructive nature of the testing, mean that we must work with incomplete information—the data that are actually collected in the course of an experimental study.

The preceding discussions highlight a distinction between the data set that is actually acquired through the process of observation and the vast collection of all potential observations that can be conceived in a given context. The statistical name for the former is **sample**; for the latter, it is **population**, or **statistical population**. To further elucidate these concepts, we observe that each measurement in a data set originates from a distinct source that may be a patient, tree, farm, household, or some other entity depending on the object of a study. The source of each measurement is called a **sampling unit**, or simply, a **unit**.

To emphasize population as the entire collection of units, we refer to it as the **population of units**.

A **unit** is a single entity, usually a person or an object, whose characteristics are of interest.

The **population of units** is the complete collection of units about which information is sought.

There is another aspect to any population and that is the value, for each unit, of a characteristic or variable of interest. There can be several characteristics of interest for a given population of units, as indicated in Table 1.

For a given variable or characteristic of interest, we call the collection of values, evaluated for every unit in the population, the **statistical population** or just the **population**. We refer to the collection of units as the **population of units** when there is a need to differentiate it from the collection of values.

A statistical **population** is the set of measurements (or record of some qualitative trait) corresponding to the entire collection of units about which information is sought.

The population represents the target of an investigation. We learn about the population by taking a sample from the population. A **sample** or **sample data set** then consists of measurements recorded for those units that are actually observed. It constitutes a part of a far larger collection about which we wish to make inferences—the set of measurements that would result if all the units in the population could be observed.

A **sample** from a statistical population is the subset of measurements that are actually collected in the course of an investigation.

TABLE 1 Populations, Units, and Variables

Population	Unit	Variables/Characteristics
Registered voters in your state	Voter	Political party Voted or not in last election Age Sex Conservative/liberal
All rental apartments near campus	Apartment	Rent Size in square feet Number of bedrooms Number of bathrooms TV and Internet connections
All campus fast food restaurants	Restaurant	Number of employees Seating capacity Hiring/not hiring
All computers owned by students at your school	Computer	Speed of processor Size of hard disk Speed of Internet connection Screen size

Example 1 Identifying the Population and Sample

Questions concerning the effect on health of two or fewer cups of coffee a day are still largely unresolved. Current studies seek to find physiological changes that could prove harmful. An article carried the headline CAFFEINE DECREASES CEREBRAL BLOOD FLOW. It describes a study[2] that establishes a physiological side effect—a substantial decrease in cerebral blood flow for persons drinking two to three cups of coffee daily.

The cerebral blood flow was measured twice on each of 20 subjects. It was measured once after taking an oral dose of caffeine equivalent to two to three cups of coffee and then, on another day, after taking a look-alike dose but without caffeine. The order of the two tests was random and subjects were not told which dose they received. The measured decrease in cerebral blood flow was significant.

Identify the population and sample.

SOLUTION As the article implies, the conclusion should apply to you and me. The population could well be the potential decreases in cerebral blood flow for all adults living in the United States. It might even apply to all the decreases in blood flow for all caffeine users in the world, although cultural customs may vary the type of caffeine consumption from coffee breaks to tea time to kola nut chewing.

The sample consists of the decreases in blood flow for the 20 subjects who agreed to participate in the study.

Example 2 A Misleading Sample

A host of a radio music show announces that she wants to know which singer is the favorite among city residents. Listeners are asked to call in and name their favorite singer.

Identify the population and sample. Comment on how to get a sample that is more representative of the city's population.

SOLUTION The population is the collection of singer preferences of all city residents and the purported goal is to learn who is the favorite singer. Because it would be nearly impossible to question all the residents in a large city, one must necessarily settle for taking a sample.

Having residents make a local call is certainly a low-cost method of getting a sample. The sample would then consist of the singers named by each person who calls the radio station. Unfortunately, with this selection procedure, the sample is not very representative of the responses from all city residents. Those who listen to the particular radio station are already a special subgroup with similar listening tastes. Furthermore, those listeners who take the time and effort to call are usually those who feel strongest about their opinions. The resulting responses could well be much stronger in favor of a particular country western or rock singer than is the case for preference among the total population of city residents or even those who listen to the station.

If the purpose of asking the question is really to determine the favorite singer of the city's residents, we have to proceed otherwise. One procedure commonly employed is a phone survey where the phone numbers are chosen at random. For instance, one can imagine that the numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 are written on separate pieces of paper and placed in a hat. Slips are then drawn one at a time and replaced between drawings. Later, we will see that computers can mimic this selection quickly and easily. Four draws will produce a random telephone number within a three-digit exchange. Telephone numbers chosen in this manner will certainly produce a much more representative sample than the self-selected sample of persons who call the station.

Self-selected samples consisting of responses to call-in or write-in requests will, in general, not be representative of the population. They arise primarily from subjects who feel strongly about the issue in question. To their credit, many TV news and entertainment programs now state that their call-in polls are nonscientific and merely reflect the opinions of those persons who responded.

[2] A. Field et al. "Dietary caffeine consumption and withdrawal: Confounding variables in quantitative cerebral perfusion studies?" *Radiology* **227** (2003), pp. 129–135.

3.1 USING A RANDOM NUMBER TABLE TO SELECT A SAMPLE

The choice of which population units to include in a sample must be impartial and objective. When the total number of units is finite, the name or number of each population unit could be written on a separate slip of paper and the slips placed in a box. Slips could be drawn one at a time without replacement and the corresponding units selected as the sample of units. Unfortunately, this simple and intuitive procedure is cumbersome to implement. Also, it is difficult to mix the slips well enough to ensure impartiality.

Alternatively, a better method is to take 10 identical marbles, number them 0 through 9, and place them in an urn. After shuffling, select 1 marble. After replacing the marble, shuffle and draw again. Continuing in this way, we create a sequence of random digits. Each digit has an equal chance of appearing in any given position, all pairs have the same chance of appearing in any two given positions, and so on. Further, any digit or collection of digits is unrelated to any other disjoint subset of digits. For convenience of use, these digits can be placed in a table called a **random number table**.

The digits in Table 1 of Appendix B were actually generated using computer software that closely mimics the drawing of marbles. A portion of this table is shown here as Table 2.

TABLE 2 Random Digits: A Portion of Table 1, Appendix B

Row										
1	0695	7741	8254	4297	0000	5277	6563	9265	1023	5925
2	0437	5434	8503	3928	6979	9393	8936	9088	5744	4790
3	6242	2998	0205	5469	3365	7950	7256	3716	8385	0253
4	7090	4074	1257	7175	3310	0712	4748	4226	0604	3804
5	0683	6999	4828	7888	0087	9288	7855	2678	3315	6718
6	7013	4300	3768	2572	6473	2411	6285	0069	5422	6175
7	8808	2786	5369	9571	3412	2465	6419	3990	0294	0896
8	9876	3602	5812	0124	1997	6445	3176	2682	1259	1728
9	1873	1065	8976	1295	9434	3178	0602	0732	6616	7972
10	2581	3075	4622	2974	7069	5605	0420	2949	4387	7679
11	3785	6401	0504	5077	7132	4135	4646	3834	6753	1593
12	8626	4017	1544	4202	8986	1432	2810	2418	8052	2710
13	6253	0726	9483	6753	4732	2284	0421	3010	7885	8436
14	0113	4546	2212	9829	2351	1370	2707	3329	6574	7002
15	4646	6474	9983	8738	1603	8671	0489	9588	3309	5860

To obtain a random sample of units from a population of size N, we first number the units from 1 to N. Then numbers are read from the table of random digits until enough different numbers in the appropriate range are selected.

Example 3 Using the Table of Random Digits to Select Items for a Price Check

One week, the advertisement for a large grocery store contains 72 special sale items. Five items will be selected with the intention of comparing the sales price with the scan price at the checkout counter. Select the five items at random to avoid partiality.

SOLUTION The 72 sale items are first numbered from 1 to 72. Since the population size $N = 72$ has two digits, we will select random digits two at a time from Table 2. Arbitrarily, we decide to start in row 7 and columns 19 and 20. Starting with the two digits in columns 19 and 20 and reading down, we obtain

12 97 34 69 32 86 32 51

We ignore 97 and 86 because they are larger than the population size 72. We also ignore any number when it appears a second time as 32 does here. Consequently, the sale items numbered

12 34 69 32 51

are selected for the price check.

Example 4 Selecting a Sample by Random Digit Dialing

A major Internet service provider wants to learn about the proportion of people in one target area who are aware of its latest product. Suppose there is a single three-digit telephone exchange that covers the target area. Use Table 1, in Appendix B, to select six telephone numbers for a phone survey.

SOLUTION We arbitrarily decide to start at row 31 and columns 25 to 28. Proceeding upward, we obtain

7566 0766 1619 9320 1307 6435

Together with the three-digit exchange, these six numbers form the phone numbers to call in the survey. Every phone number, listed or unlisted, has the same chance of being selected. The same holds for every pair, every triplet, and so on. Commercial phones may have to be discarded and another four digits selected. If there are two exchanges in the area, separate selections could be done for each exchange.

For large sample sizes, it is better to use computer-generated random digits or even computer-dialed random phone numbers.

Data collected with a clear-cut purpose in mind are very different from **anecdotal data**. Most of us have heard people say they won money at a casino, but certainly most people cannot win most of the time as casinos are not in the business of giving away money. People tend to tell good things about themselves. In a similar vein, some drivers' lives are saved when they are thrown free of car wrecks because they were not wearing seat belts. Although such stories are told and retold, you must remember that there is really no opportunity to hear from those who would have lived if they had worn their seat belts. Anecdotal information is usually repeated because it has some striking feature that may not be representative of the mass of cases in the population. Consequently, it is not apt to provide reliable answers to questions.

Exercises

1.1 A survey of 451 men revealed that 144 men, or 31.9%, wait until Valentine's day or the day before to purchase flowers. Identify a statistical population and the sample.

1.2 A social networking site reported the number of passwords memorized by 150 visitors who agreed to participate. Identify the statistical population and the sample for the number of passwords memorized.

1.3 Twenty college students were asked for their number of close friends; persons who showed sympathy when needed and helped in hard times. The average number reported was just over 2. Identify a statistical population and the sample.

1.4 The article *What Makes a Great Tweet* concludes that only 36% of tweets are worth reading. A total of 4,220 tweets were rated in this study.[3] Identify a statistical population and the sample.

1.5 Among 40 adults 20–25 years old questioned about stress levels, 16 responded they are more stressed now than last year. Identify a statistical population and the sample.

1.6 Rap music was the favorite genre of 6 out of 56 students from a large midwestern university. Identify a statistical population and the sample.

1.7 About 42% of the members of a local hiking club own a dog. Should these people be considered as a random sample of dog ownership for persons living in the city?

1.8 Which of the following are anecdotal and which are based on a sample?

(a) Ellie told her friends that she is saving $31 a month because she changed to a prepaid cell phone.

(b) One morning, among a large number of coffee shop patrons, the orders of 47 coffee drinks were recorded and 11 of these were espresso-based drinks.

(c) Out of 200 college students at a large university, over half thought about food more than 17 times a day.

1.9 Which of the following are anecdotal and which are based on a sample?

(a) Seventy-five text messages were sent one day during a lecture by students in a large class.

(b) Erik says he gets the best bargains at online auctions by bidding on items whose termination is early in the morning.

(c) Out of 4837 college age students in California, 3,769 admitted they used a cell phone to talk or text while driving.

[3] P. Andre, M. Bernstein, and K. Luther, "What makes a great tweet," *Harvard Business Review*, May 2012, p. 36.

1.10 Twelve bicycles are available for use at the student union. Use Table 1, Appendix B, to select 4 of them for you and three of your friends to ride today.

1.11 At the last minute, 6 tickets have become available for a big football game. Use Table 1, Appendix B, to select the recipients from among 89 interested students.

4. THE PURPOSEFUL COLLECTION OF DATA

Many poor decisions are made, in both business and everyday activities, because people fail to understand and account for variability. Certainly, the purchasing habits of one person may not represent those of the population, or the reaction of one mouse to a potentially toxic chemical compound may not represent that of a large population of mice. However, despite diversity among the purchasing habits of individuals, we can obtain accurate information about the purchasing habits of the population by collecting data on a large number of persons. By the same token, much can be learned about the toxicity of a chemical if many mice are exposed.

Just making the decision to collect data is a key step to answer a question, to provide the basis for taking action, or to improve a process. Once that decision has been made, an important next step is to develop a **statement of purpose** that is both specific and unambiguous. If the subject of the study is public transportation being behind schedule, you must carefully specify what is meant by late. Is it 1 minute, 5 minutes, or more than 10 minutes behind scheduled times that should result in calling a bus or commuter train late? Words like soft or uncomfortable in a statement are even harder to quantify. One common approach, for a quality like comfort, is to ask passengers to rate the ride on public transportation on the five-point scale

1	2	3	4	5
Very uncomfortable		Neutral		Very comfortable

where the numbers 1 through 5 are attached to the scale, with 1 for very uncomfortable and so on through 5 for very comfortable.

We might conclude that the ride is comfortable if the majority of persons in the sample check either of the top two boxes.

Example 5 A Clear Statement of Purpose Concerning Water Quality

Each day, a city must sample the lake water in and around a swimming beach to determine if the water is safe for swimming. During late summer, the primary difficulty is algae growth and the safe limit has been set in terms of water clarity.

SOLUTION The problem is already well defined so the statement of purpose is straightforward.

PURPOSE: Determine whether or not the water clarity at the beach is below the safe limit.

The city will take measurements of clarity at three separated locations. In Chapter 8, we will learn how to decide if the water is safe despite the variation in the three sample values.

The overall purpose can be quite general but a specific statement of purpose is required at each step to guide the collection of data. For instance:

GENERAL PURPOSE: Design a data collection and monitoring program at a completely automated plant that handles radioactive materials.

One issue is to ensure that the production plant will shut down quickly if materials start accumulating anywhere along the production line. More specifically, the weight of materials could be measured at critical positions. A quick shutdown will be implemented if any of these exceed a safe limit. For this step, a statement of purpose could be:

> ***PURPOSE:*** Implement a fast shutdown if the weight at any critical position exceeds 1.2 kilograms.

The safe limit 1.2 kilograms should be obtained from experts; preferably it would be a consensus of expert opinion.

There still remain statistical issues of how many critical positions to choose and how often to measure the weight. These are followed with questions on how to analyze data and specify a rule for implementing a fast shutdown.

A clearly specified statement of purpose will guide the choice of what data to collect and help ensure that it will be relevant to the purpose. Without a clearly specified purpose, or terms unambiguously defined, much effort can be wasted in collecting data that will not answer the question of interest.

Exercises

1.12 What is wrong with this statement of purpose?

> ***PURPOSE:*** *Determine whether or not, over the course of the semester, the campus bus reaches your stop at the scheduled time.*

Give an improved statement of purpose.

1.13 What is wrong with this statement of purpose?

> ***PURPOSE:*** *Determine if a new style wireless mouse is comfortable.*

Give an improved statement of purpose.

1.14 Give a statement of purpose for a study to determine the favorite campus area pizza establishment.

1.15 Give a statement of purpose for determining the amount of time it takes to make hotel reservations in San Francisco using the internet.

STATISTICS IN CONTEXT

A primary health facility became aware that sometimes it was taking too long to return patients' phone calls. Patients phone in with requests for information. These requests are then turned over to doctors or nurses who collect the information and return the call. The overall objective is to understand the current procedure and then improve on it. A good first step is to find how long it takes to return calls under the current procedure. Variation in times from call to call is expected, so the purpose of the initial investigation is to benchmark the variability with the current procedure by collecting a sample of times.

> ***PURPOSE:*** Obtain a reference or benchmark for the current procedure by collecting a sample of times to return a patient's call under the current procedure.

For a sample of incoming calls collected during the week, the time received is noted along with the request. When the return call is complete, the elapsed time, in minutes, is recorded. Each of these times is represented as a dot in Figure 1. Notice that over one-third of the calls took over 120 minutes, or over two hours, to return. This is a long time to wait for information if it concerns a child with a high fever or an adult with acute symptoms. If the purpose is to determine what proportion of calls took too long to return, we need to agree on a more precise definition of "too long" in terms of number of minutes. Instead, these data clearly indicate that the process needs improvement and the next step is to proceed in that direction.

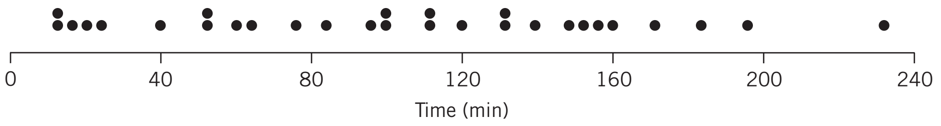


Figure 1 Time in minutes to return call.

In any context, to pursue potential improvements of a process, one needs to focus more closely on particulars. Three questions

When Where Who

should always be asked before gathering further data. More specifically, data should be sought that will answer the following questions.

When do the difficulties arise? Is it during certain hours, certain days of the week or month, or in coincidence with some other activities?

Where do the difficulties arise? Try to identify the locations of bottlenecks and unnecessary delays.

Who was performing the activity and who was supervising? The idea is not to pin blame, but to understand the roles of participants with the goal of making improvements.

It is often helpful to construct a **cause-and-effect diagram** or **fishbone diagram**. The main centerline represents the problem or the effect. A somewhat simplified fishbone chart is shown in Figure 2 for the *where* question regarding the location of delays when returning patients' phone calls. The main centerline represents the problem: Where are delays occurring? Calls come to the reception desk, but when these lines are busy, the calls go directly to nurses on the third or fourth floor. The main diagonal arms in Figure 2 represent the floors and the smaller horizontal lines more specific locations on the floor where the delay could occur. For instance, the horizontal line representing a delay in retrieving a patient's medical record connects to the second floor diagonal line. The resulting figure resembles the skeleton of a fish. Consideration of the diagram can help guide the choice of what new data to collect.

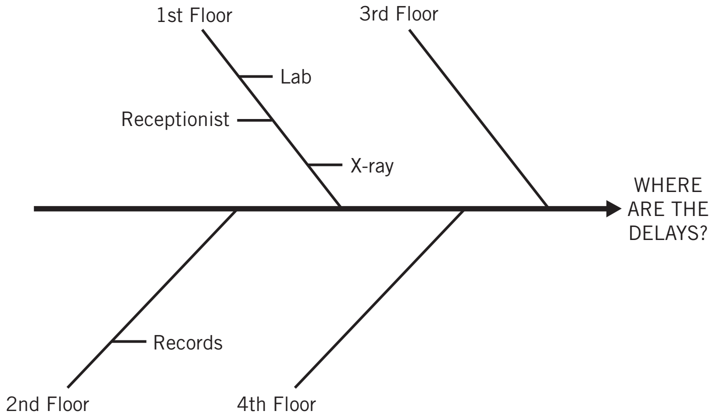


Figure 2 A cause-and-effect diagram for the location of delays.

Fortunately, the quality team conducting this study had already given preliminary consideration to the *When*, *Where*, and *Who* questions and recorded not only the time of day but also the day and person receiving the call. That is, their current data gave them a start on determining if the time to return calls depends on when or where the call is received.

Although we go no further with this application here, the quality team next developed more detailed diagrams to study the flow of paper between the time the call is received and when it is returned. They then identified bottlenecks in the flow of information that were removed and the process was improved. In later chapters, you will learn how to compare and display data from two locations or old and new processes, but the key idea emphasized here is the purposeful collection of relevant data.

Exercises

1.16 How many of the calls in the initial data set took over 125 minutes to answer? How many over 90 minutes?

1.17 According to the cause-and-effect diagram on page 13, where are the possible delays on the first floor?

5. OBJECTIVES OF STATISTICS

The subject of statistics provides the methodology to make **inferences** about the population from the collection and analysis of sample data. These methods enable one to derive plausible generalizations and then assess the extent of uncertainty underlying these generalizations. Statistical concepts are also essential during the planning stage of an investigation when decisions must be made as to the mode and extent of the sampling process.

The major objectives of statistics are:

1. To make **inferences** about a population from an analysis of information contained in sample data. This includes assessments of the extent of uncertainty involved in these inferences.
2. To **design the process and the extent of sampling** so that the observations form a basis for drawing valid inferences.

The design of the sampling process is an important step. A good design for the process of data collection permits efficient inferences to be made, often with a straightforward analysis. Unfortunately, even the most sophisticated methods of data analysis cannot, in themselves, salvage much information from data that are produced by a poorly planned experiment or survey.

The early use of statistics in the compilation and passive presentation of data has been largely superseded by the modern role of providing analytical tools with which data can be efficiently gathered, understood, and interpreted. Statistical concepts and methods make it possible to draw valid conclusions about the population on the basis of a sample. Given its extended goal, the subject of statistics has penetrated all fields of human endeavor in which the evaluation of information must be grounded in data-based evidence.

The basic statistical concepts and methods described in this book form the core in all areas of application. We present examples drawn from a wide range of applications to help develop an appreciation of various statistical methods, their potential uses, and their vulnerabilities to misuse.

USING STATISTICS WISELY

1. Compose a clear statement of purpose and use it to help decide which variables to observe.
2. Carefully define the population of interest.
3. Whenever possible, select samples using a random device or random number table.
4. Do not unquestionably accept conclusions based on self-selected samples.
5. Remember that conclusions reached in TV, magazine, or newspaper reports might not be as obvious as reported. When reading or listening to reports, you must be aware that the advocate, often a politician or advertiser, may only be presenting statistics that emphasize positive features.

KEY IDEAS

Before gathering data, on a characteristic of interest, identify a **unit** or **sampling unit**. This is usually a person or an object. The **population of units** is the complete collection of units. In statistics we concentrate on the collection of values of the characteristic, or record of a qualitative trait, evaluated for each unit in the population. We call this the **statistical population** or just the **population**.

A **sample or sample data set** from the population is the subset of measurements that are actually collected.

Statistics is a body of principles that helps to first design the process and extent of sampling and then guides the making of **inferences** about the population **(inferential statistics)**. **Descriptive statistics** help summarize the sample. Procedures for **statistical inference** allow us to make generalizations about the population from the information in the sample.

A **statement of purpose** is a key step in designing the data collection process.

REVIEW EXERCISES

1.18 A newspaper headline reads, College Students Display An Inability To Tell Fake News from Real and the article explains this conclusion was reached from the responses of 2300 students at ten colleges concerning whether or not a case of social media type news was fake or not.

(a) Specify the variable of interest.

(b) Specify the statistical population.

(c) Specify the sample.

1.19 Consider the population of all students at your college. You want to learn about total monthly entertainment expenses for a student.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.20 Consider the population of persons living in Chicago. You want to learn about the proportion of eligible voters who are registered to vote.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.21 A student is asked to estimate the mean height of all male students on campus. She decides to use the heights of members of the basketball team because they are conveniently printed in the game program.

(a) Identify the statistical population and the sample.

(b) Comment on the selection of the sample.

(c) How should a sample of males be selected?

1.22 The number of hours actively connected to social media on the previous day are recorded for 120 students at a university. Identify the population unit, statistical population, and sample.

1.23 It is often easy to put off doing an unpleasant task. At a Web site,[4] persons can take a test and receive a score that determines if they have a serious problem with procrastination. Should the scores from people who take this test online be considered a random sample from the general population? Explain your reasoning.

1.24 A magazine that features the latest electronics and computer software for homes enclosed a short questionnaire on a postcard. Readers were asked to answer questions concerning their use and ownership of various software and hardware products, and to then send the card to the publisher. A summary of the results appeared in a later issue of the magazine that used the data to make statements such as 40% of readers have purchased program X. Identify a population and sample and comment on the representativeness of the sample. Are readers who have not purchased any new products mentioned in the questionnaire as likely to respond as those who have purchased?

1.25 Each year a local weekly newspaper gives out “Best of the City” awards in categories such as restaurant, deli, pastry shop, and so on. Readers are asked to fill in their favorites on a form at the web site of the weekly paper. The establishment receiving the most votes is declared the winner in its category. Identify the population and sample and comment on the representativeness of the sample.

1.26 Which of the following are anecdotal and which are based on sample?

(a) Out of 200 students questioned, 40 admitted they lied regularly.

(b) Bobbie says the produce at Market W is the freshest in the city.

(c) Out of 50 persons interviewed at a shopping mall, 18 had made a purchase that day.

1.27 Which of the following are anecdotal and which are based on a sample?

(a) Tom says he gets the best prices on electronics at the www.bestelc.com Internet site.

(b) Out of 22 students, 6 had multiple credit cards.

(c) Among 55 people checking in at the airport, 12 were going to destinations outside of the continental United States.

1.28 What is wrong with this statement of purpose?

> ***PURPOSE:*** *Determine if it takes too long to get cash from the automated teller machine during the lunch hour.*

Give an improved statement of purpose.

1.29 Two-person sailboats are available for use at the university dock. Your group is large enough to need three of them. Use Table 1, Appendix B, to select your three boats from among the 10 that are in a line along the shore.

1.30 There are 9 sites on a large campus at which bicycles are parked as part of a shared bicycle program in the city. Use

[4] www.mindtools.com (2012) Are you a procrastinator? (online). [Accessed November 9, 2012].

Table 1, Appendix B to select 3 sites to visit and record the number of bicycles available at noon during a weekday.

1.31 Fifty band members would like to ride the band bus to an out-of-town game. However, there is room for only 44. Use Table 1, Appendix B, to select the 44 persons who will go. Determine how to make your selection by taking only a few two-digit selections.

1.32 Eight young students need mentors. Of these, there are three whom you enjoy being with while you are indifferent about the others. Two of the students will be randomly assigned to you. Label the students you like by 0, 1, and 2 and the others by 3, 4, 5, 6, and 7. Then, the process of assigning two students at random is equivalent to choosing two different digits from the table of random digits and ignoring any 8 or 9. Repeat the experiment of assigning two students 20 times by using the table of random digits. Record the pairs of digits you draw for each experiment.

(a) What is the proportion of the 20 experiments that give two students that you like?

(b) What is the proportion of the 20 experiments that give one of the students you like and one other?

(c) What is the proportion of the 20 experiments that give none of the students you like?

1.33 The United States Environmental Protection Agency[5] reports that in 2014, each American generated 4.44 pounds of solid waste.

(a) Does this mean every single American produces the same amount of garbage? What do you think this statement means?

(b) Was the number 4.44 obtained from a sample? Explain.

(c) How would you select a sample?

1.34 As a very extreme case of self-selection, imagine a five-foot-high solid wood fence surrounding a collection of Great Danes and Miniature Poodles. You want to estimate the proportion of Great Danes inside and decide to collect your sample by observing the first seven dogs to jump high enough to be seen above the fence.

(a) Explain how this is a self-selected sample that is, of course, very misleading.

(b) How is this sample selection procedure like a call-in election poll?

[5] Advancing Sustainable Management: Fact Sheet 2014, November 2016.

2

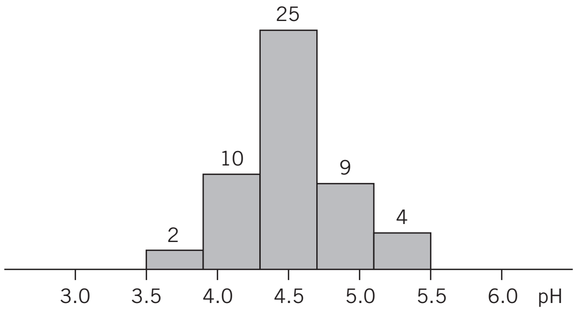
Organization and Description of Data

Bill Barley/SuperStock

Acid Rain Is Killing Our Lakes

Acid precipitation is linked to the disappearance of sport fish and other organisms from lakes. Sources of air pollution, including automobile emissions and the burning of fossil fuels, add to the natural acidity of precipitation.

The Wisconsin Department of Natural Resources initiated a precipitation monitoring program with the goal of developing appropriate air pollution controls to reduce the problem. The acidity of the first 50 rains monitored, measured on a pH scale from 1 (very acidic) to 7 (basic), are summarized by the histogram.



Histogram of acid rain data.

Notice that all the rains are more acidic than normal rain, which has a pH of 5.6. (As a comparison, apples are about pH 3 and milk is about pH 6.)

Researchers in Canada have established that lake water with a pH below 5.6 may severely affect the reproduction of game fish. More research will undoubtedly improve our understanding of the acid rain problem and lead, it is hoped, to an improved environment.

In Chapter 1, we cited several examples of situations where the collection of data by experimentation or observation is essential to acquiring new knowledge. A data set may range in complexity from a few entries to hundreds or even thousands of them. Each entry corresponds to the observation of a specified characteristic of a sampling unit. For example, a nutritionist may provide an experimental diet to 30 undernourished children and record their weight gains after two months. Here, children are the sampling units, and the data set would consist of 30 measurements of weight gains. Once the data are collected, a primary step is organizing the information and extracting a descriptive summary that highlights its salient features. In this chapter, we learn how to organize and describe a set of data by means of tables, graphs, and calculation of some numerical summary measures. □

1. MAIN TYPES OF DATA

Before introducing methods of describing data, we first distinguish between the two basic types:

1. **Qualitative** or **categorical data**
2. **Numerical, quantitative**, or **measurement** data

When the characteristic under study concerns a qualitative trait that is only classified in categories and not numerically measured, the resulting data are called categorical data. Hair color (blond, brown, red, black), employment status (employed, unemployed), and blood type (O, A, B, AB) are but some examples. If, on the other hand, the characteristic is measured on a numerical scale, the resulting data consist of a set of numbers and are called measurement data. We will use the term **numerical-valued variable** or just **variable** to refer to a characteristic that is measured on a numerical scale. The word “variable” signifies that the measurements vary over different sampling units. In this terminology, observations of a numerical-valued variable yield measurement data. A few examples of numerical-valued variables are the shoe size of an adult male, daily number of traffic fatalities in a state, intensity of an earthquake, height of a 1-year-old pine seedling, the time spent in line at an automated teller, and the number of offspring in an animal litter.

Although in all these examples the stated characteristic can be numerically measured, a close scrutiny reveals two distinct types of underlying scale of measurement. Shoe sizes are numbers such as $6, 6\frac{1}{2}, 7, 7\frac{1}{2}, \ldots$, which proceed in steps of $\frac{1}{2}$. The count of traffic fatalities can only be an integer and so is the number of offspring in an animal litter. These are examples of **discrete variables**. The name **discrete** draws from the fact that the scale is made up of distinct numbers with gaps in between. On the other hand, some variables such as height, weight, and survival time can ideally take any value in an interval. Since the measurement scale does not have gaps, such variables are called **continuous**.

We must admit that a truly continuous scale of measurement is an idealization. Measurements actually recorded in a data set are always rounded either for the sake of simplicity or because the measuring device has a limited accuracy. Still, even though weights may be recorded in the nearest pounds or time recorded in the whole hours, their actual values occur on a continuous scale so the data are referred to as continuous. Counts are inherently discrete and treated as such, provided that they take relatively few distinct values (e.g., the number of children in a family or the number of traffic violations of a driver). But when a count spans a wide range of values, it is often treated as a continuous variable. For example, the count of white blood cells, number of insects in a colony, and number of shares of stock traded per day are strictly discrete, but for practical purposes, they are viewed as continuous.

Example 1 The Various Types of Data

Consider the characteristics

(a) Height of a skyscraper.

(b) Favorite primary color.

(c) Number of visits to a dentist last year.

In each case, determine if the resulting data are categorical, discrete, or continuous.

SOLUTION

(a) Height is continuous. The numerical value is typically recorded to nearest foot, inch, or fraction of inch.

(b) Primary color is categorical. There are three choices: red, green, and blue.

(c) Number of visits is discrete. The numerical value is a count and it is usually low.

A summary description of categorical data is discussed in Section 2.1. The remainder of this chapter is devoted to a descriptive study of measurement data, both discrete and continuous. As in the case of summarization and commentary on a long, wordy document, it is difficult to prescribe concrete steps for summary descriptions that work well for all types of measurement data. However, a few important aspects that deserve special attention are outlined here to provide general guidelines for this process.

Describing a Data Set of Measurements

1. **Summarization and description of the overall pattern.**
   (a) Presentation of tables and graphs.
   (b) Noting important features of the graphed data including symmetry or departures from it.
   (c) Scanning the graphed data to detect any observations that seem to stick far out from the major mass of the data—the outliers.
2. **Computation of numerical measures.**
   (a) A typical or representative value that indicates the center of the data.
   (b) The amount of spread or variation present in the data.

2. DESCRIBING DATA BY TABLES AND GRAPHS

2.1 CATEGORICAL DATA

When a qualitative trait is observed for a sample of units, each observation is recorded as a member of one of several categories. Such data are readily organized in the form of a frequency table that shows the counts (**frequencies**) of the individual categories. Our understanding of the data is further enhanced by calculation of the proportion (also called **relative frequency**) of observations in each category.

$$\textbf{Relative frequency of a category} = \frac{\text{Frequency in the category}}{\text{Total number of observations}}$$

Example 2 Calculating Relative Frequencies to Summarize an Opinion Poll

A campus press polled a sample of 280 undergraduate students in order to study student attitude toward a proposed change in the dormitory regulations. Each student was to respond as support, oppose, or neutral in regard to the issue. The numbers were 152 support, 77 neutral, and 51 opposed. Tabulate the results and calculate the relative frequencies for the three response categories.

SOLUTION Table 1 records the frequencies in the second column, and the relative frequencies are calculated in the third column. The relative frequencies show that about 54% of the polled students supported the change, 18% opposed, and 28% were neutral.

TABLE 1 Summary Results of an Opinion Poll

Responses	Frequency	Relative Frequency
Support	152	$\frac{152}{280} = .543$
Neutral	77	$\frac{77}{280} = .275$
Oppose	51	$\frac{51}{280} = .182$
Total	280	1.000

Remark: The relative frequencies provide the most relevant information as to the pattern of the data. One should also state the sample size, which serves as an indicator of the credibility of the relative frequencies. (More on this in Chapter 8.)

Categorical data are often presented graphically as a **pie chart** in which the segments of a circle exhibit the relative frequencies of the categories. To obtain the angle for any category, we multiply the relative frequency by 360 degrees, which corresponds to the complete circle. Although laying out the angles by hand can be tedious, many software packages generate the chart with a single command. Figure 1 presents a pie chart for the data in Example 1.

When questions arise that need answering but the decision makers lack precise knowledge of the state of nature or the full ramifications of their decisions, the best procedure is often to collect more data. In the context of quality improvement, if a problem is recognized, the first step is to collect data on the magnitude and possible causes. This information is most effectively communicated through graphical presentations.

A **Pareto diagram** is a powerful graphical technique for displaying events according to their frequency. According to Pareto's empirical law, any collection of events consists of only a few that are major in that they are the ones that occur most of the time.

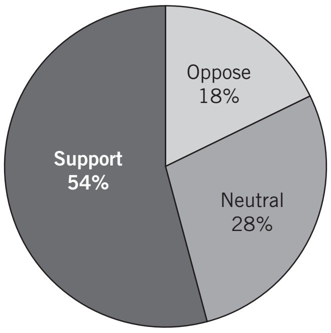


Figure 1 Pie chart of student opinion on change in dormitory regulations.

Figure 2 gives a Pareto diagram for the type of defects found in a day's production of facial tissues. The bars correspond to different causes so they do not touch.

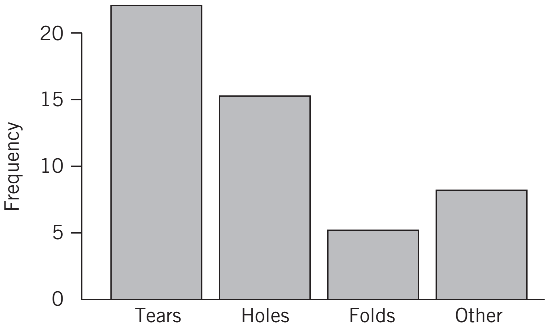


Figure 2 Pareto diagram of facial tissue defects.

The cumulative frequency is 22 for the first cause and 22 + 15 = 37 for the first and second causes combined. This illustrates Pareto's rule, with two of the causes being responsible for 37 out of 50, or 74%, of the defects.

Example 3 A Pareto Diagram Clarifies Circumstances Needing Improvement

Graduate students in a counseling course were asked to choose one of their personal habits that needed improvement. In order to reduce the effect of this habit, they were asked to first gather data on the frequency of the occurrence and the circumstances. One student collected the following frequency data on fingernail biting over a two-week period.

Frequency	Activity
58	Watching television
21	Reading newspaper
14	Talking on phone
7	Driving a car
3	Grocery shopping
12	Other

Make a Pareto diagram showing the relationship between nail biting and type of activity.

SOLUTION The cumulative frequencies are 58, 58 + 21 = 79, and so on, out of 115. The Pareto diagram is shown in Figure 3, where watching TV accounts for 50.4% of the instances.

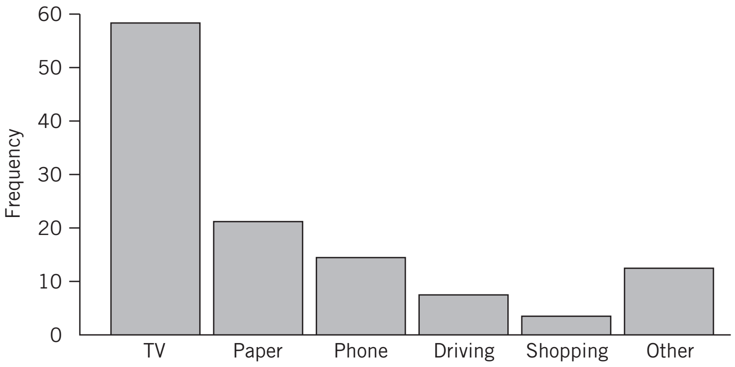


Figure 3 Pareto diagram for nail biting example.

The next step for this person would be to try and find a substitute for nail biting while watching television.

2.2 DISCRETE DATA

We next consider summary descriptions of measurement data and begin our discussion with discrete measurement scales. As explained in Section 1, a data set is identified as discrete when the underlying scale is discrete and the distinct values observed are not too numerous.

Similar to our description of categorical data, the information in a **discrete data set** can be summarized in a **frequency table**, or **frequency distribution** that includes a calculation of the **relative frequencies**. In place of the qualitative categories, we now list the distinct numerical measurements that appear in the data set and then count their frequencies.

Example 4 Creating a Frequency Distribution

Retail stores experience their heaviest returns on December 26 and December 27 each year. Most are gifts that, for some reason, did not please the recipient. The number of items returned, by a sample of 30 persons at a large discount department store, are observed and the data of Table 2 are obtained. Determine the frequency distribution.

TABLE 2 Number of Items Returned

1	4	3	2	3	4	5	1	2	1
2	5	1	4	2	1	3	2	4	1
2	3	2	3	2	1	4	3	2	5

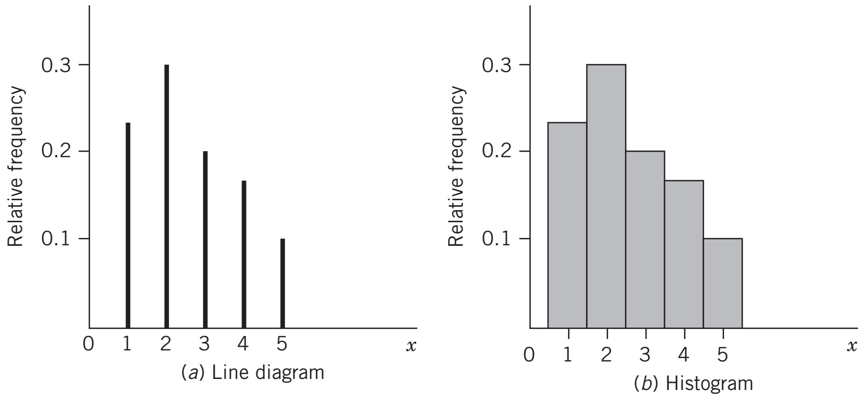
SOLUTION The **frequency distribution** of these data is presented in Table 3. The values are paired with the frequency and the calculated relative frequency.

TABLE 3 Frequency Distribution for Number (x) of Items Returned

Value x	Frequency	Relative Frequency
1	7	.233
2	9	.300
3	6	.200
4	5	.167
5	3	.100
Total	30	1.000

The frequency distribution of a discrete variable can be presented pictorially by drawing either lines or rectangles to represent the relative frequencies. First, the distinct values of the variable are located on the horizontal axis. For a **line diagram**, we draw a vertical line at each value and make the height of the line equal to the relative frequency. A **histogram** employs vertical rectangles instead of lines. These rectangles are centered at the values and their areas represent relative frequencies. Typically, the values proceed in equal steps so the rectangles are all of the same width and their heights are proportional to the relative frequencies as well as frequencies. Unlike Pareto charts, the bars of a histogram do touch each other.

Figure 4(*a*) shows the line diagram and 4(*b*) the histogram of the frequency distribution of Table 3.



(*a*) Line diagram

(*b*) Histogram

Figure 4 Graphic display of the frequency distribution of data in Table 3.

2.3 DATA ON A CONTINUOUS VARIABLE

What tabular and graphical presentations are appropriate for data sets that contain numerical measurements on a virtually continuous scale? In contrast with the discrete case, a data set of measurements on a continuous variable may contain many distinct values. Then, a table or plot of all distinct values and their frequencies will not provide a condensed or informative summary of the data.

The two main graphical methods used to display a data set of measurements are the **dot diagram** and the **histogram**. Dot diagrams are employed when there are relatively few observations (say, less than 20 or 25); histograms are used with a larger number of observations.

Dot Diagram

When the data consist of a small set of numbers, they can be graphically represented by drawing a line with a scale covering the range of values of the measurements. Individual measurements are plotted above this line as prominent dots. The resulting diagram is called a **dot diagram**.

Example 5 A Dot Diagram Reveals an Unusual Observation

The number of days the first six heart transplant patients at Stanford survived after their operations were 15, 3, 46, 623, 126, 64. Make a dot diagram.

SOLUTION These survival times extended from 3 to 623 days. Drawing a line segment from 0 to 700, we can plot the data as shown in Figure 5. This dot diagram shows a cluster of small survival times and a single, rather large value.

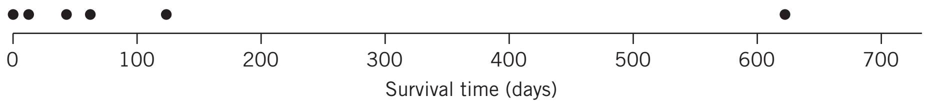


Figure 5 Dot diagram for the heart transplant data.

Paying Attention

First-grade teachers allot a portion of each day to mathematics. An educator, concerned about how students utilize this time, selected 24 students and observed them for a total of 20 minutes spread over several days. The number of minutes, out of 20, that the student was not on task was recorded (courtesy of T. Romberg). These lack-of-attention times are graphically portrayed in the dot diagram in Figure 6. The student with 13 out of 20 minutes off-task stands out enough to merit further consideration. Is this a student who finds the subject too difficult or might it be a very bright child who is bored?

Britt Erlanson/The Image Bank/Getty Images

Paying attention in class. Observations on 24 first-grade students.

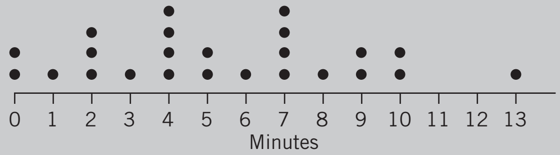


Figure 6 Time not concentrating on the mathematics assignment (out of 20 minutes).

Frequency Distribution on Intervals

When the data consist of a large number of measurements, a dot diagram may be quite tedious to construct. More seriously, overcrowding of the dots causes them to overlap and mar the clarity of the diagram. In such cases, it is convenient to condense the data by grouping the observations according to intervals and recording the frequencies of the intervals. Unlike a discrete frequency distribution, where grouping naturally takes place on points, here we use intervals of values. The main steps in this process are outlined as follows.

Constructing a Frequency Distribution for a Continuous Variable

1. Find the minimum and the maximum values in the data set.
2. Choose intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping. These are called **class intervals**, and their endpoints **class boundaries**.

3. Count the number of observations in the data that belong to each class interval. The count in each class is the **class frequency** or **cell frequency**.
4. Calculate the **relative frequency** of each class by dividing the class frequency by the total number of observations in the data:

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

The choice of the number and position of the class intervals is primarily a matter of judgment guided by the following considerations. The number of classes usually ranges from 5 to 15, depending on the number of observations in the data. Grouping the observations sacrifices information concerning how the observations are distributed within each cell. With too few cells, the loss of information is serious. On the other hand, if one chooses too many cells and the data set is relatively small, the frequencies from one cell to the next would jump up and down in a chaotic manner and no overall pattern would emerge. As an initial step, frequencies may be determined with a large number of intervals that can later be combined as desired in order to obtain a smooth pattern of the distribution.

Computers conveniently order data from smallest to largest so that the observations in any cell can easily be counted. The construction of a frequency distribution is illustrated in Example 6.

Example 6 Creating a Frequency Distribution for Hours of Sleep

Students require different amounts of sleep. A sample of 59 students at a large midwest university reported the following hours of sleep the previous night.

TABLE 4 Hours of Sleep for Fifty-Nine Students

4.5	4.7	5.0	5.0	5.3	5.5	5.5	5.7	5.7	5.7
6.0	6.0	6.0	6.0	6.3	6.3	6.3	6.5	6.5	6.5
6.7	6.7	6.7	6.7	7.0	7.0	7.0	7.0	7.3	7.3
7.3	7.3	7.5	7.5	7.5	7.5	7.7	7.7	7.7	7.7
8.0	8.0	8.0	8.0	8.3	8.3	8.3	8.5	8.5	8.5
8.5	8.7	8.7	9.0	9.0	9.0	9.3	9.3	10.0	

Construct a frequency distribution of the sleep data.

SOLUTION To construct a frequency distribution, we first notice that the minimum hours of sleep in Table 4 is 4.5 and the maximum is 10.0. We choose class intervals of length 1.2 hours as a matter of convenience.

The selection of class boundaries is a bit of fussy work. Because the data have one decimal place, we could add a second decimal to avoid the possibility of any observation falling exactly on the boundary. For example, we could end the first class interval at 5.45. Alternatively, and more neatly, we could write 4.3–5.5 and make the **endpoint convention** that the left-hand end point is included but not the right.

The first interval contains 5 observations so its frequency is 5 and its relative frequency is $\frac{5}{59} = .085$. Table 5 gives the frequency distribution. The relative frequencies add to 1, as they should (up to rounding error) for any frequency distribution. We see, for instance, that just about one-third of the students $.271 + .051 = .322$ got 7.9 hours or more of sleep.

TABLE 5 Frequency Distribution for Hours of Sleep Data (left endpoints included but right endpoints excluded)

Class Interval	Frequency	Relative Frequency
4.3–5.5	5	$\frac{5}{59} = .085$
5.5–6.7	15	$\frac{15}{59} = .254$
6.7–7.9	20	$\frac{20}{59} = .339$
7.9–9.1	16	$\frac{16}{59} = .271$
9.1–10.3	3	$\frac{3}{59} = .051$
Total	59	1.000

Remark: The rule requiring equal class intervals is inconvenient when the data are spread over a wide range but are highly concentrated in a small part of the range with relatively few numbers elsewhere. Using smaller intervals where the data are highly concentrated and larger intervals where the data are sparse helps to reduce the loss of information due to grouping.

In every application involving an **endpoint convention**, it is important that you clearly state which endpoint is included and which is excluded. This information should be presented in the title or in a footnote of any frequency distribution.

Histogram

A frequency distribution can be graphically presented as a histogram. To draw a histogram, we first mark the class intervals on the horizontal axis. On each interval, we then draw a vertical rectangle whose **area represents the relative frequency**—that is, the proportion of the observations occurring in that class interval.

To create rectangles whose area is equal to relative frequency, use the rule

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of interval}}$$

The total area of all rectangles equals 1, the sum of the relative frequencies.

The total area of a histogram is 1.

The histogram for Table 5 is shown in Figure 7. For example, the rectangle drawn on the class interval 4.3–5.5 has area $= .071 \times 1.2 = .085$, which is the relative frequency of this class. Actually, we determined the height .071 as

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of interval}} = \frac{.085}{1.2} = .071$$

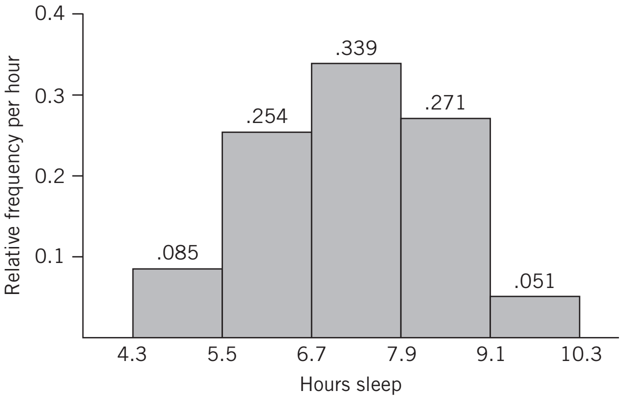


Figure 7 Histogram of the sleep data of Tables 4 and 5. Sample size = 59.

The units on the vertical axis can be viewed as relative frequencies per unit of the horizontal scale. For instance, .071 is the relative frequency per hour for the interval 4.3–5.5.

Visually, we note that the rectangle having largest area, or most frequent class interval, is 6.7–7.9. Also, proportion .085 + .254 = .339 of the students slept less than 6.7 hours.

> ***Remark:*** When all class intervals have equal widths, the heights of the rectangles are proportional to the relative frequencies that the areas represent. The formal calculation of height, as area divided by the width, is then redundant. Instead, one can mark the vertical scale according to the relative frequencies—that is, make the heights of the rectangles equal to the relative frequencies. The resulting picture also makes the areas represent the relative frequencies if we read the vertical scale as if it is in units of the class interval. This leeway when plotting the histogram is not permitted in the case of unequal class intervals.

Figure 8 shows one ingenious way of displaying two histograms for comparison. In spite of their complicated shapes, their back-to-back plot as a "pyramid" allows for easy visual comparison. Females are the clear majority in the last age groups of the male and female age distributions.

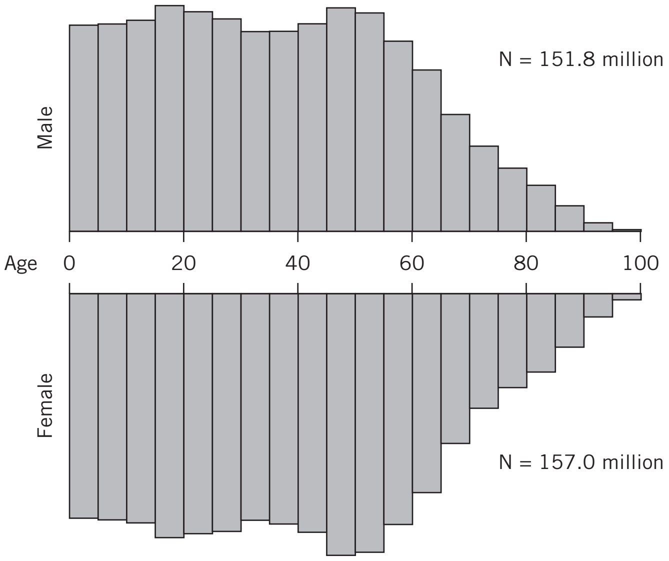


Figure 8 Population pyramid (histograms) of the male and female age distributions in the United States in 2010. (*Source*: U.S. Bureau of the Census.)

Stem-and-Leaf Display

A **stem-and-leaf display** provides a more efficient variant of the histogram for displaying data, especially when the observations are two-digit numbers. This plot is obtained by sorting the observations into rows according to their leading digit. To make a stem-and-leaf display:

1. List the digits 0 through 9 in a column and draw a vertical line. These correspond to the leading digit.
2. For each observation, record its second digit to the right of this vertical line in the row where the first digit appears.
3. Finally, arrange the second digits in each row so they are in increasing order.

The stem-and-leaf display for the data of Table 6 is shown in Table 7.

TABLE 6 Examination Scores of 50 Students

75	98	42	75	84	87	65	59	63
86	78	37	99	66	90	79	80	89
68	57	95	55	79	88	76	60	77
49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62
84	64	73	48	72				

TABLE 7 Stem-and-Leaf Display for the Examination Scores

0	
1	
2	
3	7
4	289
5	35789
6	022345689
7	01234556778899
8	00134456789
9	0023589

In the stem-and-leaf display, the column of first digits to the left of the vertical line is viewed as the stem, and the second digits as the leaves. Viewed sidewise, it looks like a histogram with a cell width equal to 10. However, it is more informative than a histogram because the actual data points are retained. In fact, every observation can be recovered exactly from this stem-and-leaf display.

A stem-and-leaf display retains all the information in the leading digits of the data. When the leaf unit = .01, 3.5 | 0 2 3 7 8 presents the data 3.50, 3.52, 3.53, 3.57, and 3.58. Leaves may also be two-digit at times. When the first leaf digit = .01, .4 | 07 13 82 90 presents the data .407, .413, .482, and .490.

Further variants of the stem-and-leaf display are described in Exercises 2.25 and 2.26. This versatile display is one of the most applicable techniques of exploratory data analysis.

When the sample size is small or moderate, no information is lost with the stem-and-leaf diagram because you can see every data point. The major disadvantage is that, when the sample size is large, diagrams with hundreds of numbers in a row cannot be constructed in a legible manner.

Exercises

2.1 Cities must find better ways to dispose of solid waste. According to the Environmental Protection Agency, in a recent year, the composition of solid municipal waste was

Paper and paperboard	28.5%
Food waste	14.9%
Yard Trimmings	13.3%
Plastics	12.9%
Metals	9.0%
Other materials	

(a) Determine the percentage of other materials in the solid waste. This category includes glass, wood, rubber, and so on.

(b) Create a Pareto chart.

(c) What percentage of the total solid waste is paper or paperboard? What percentage is from the top two categories? What percentage is from the top five categories?

2.2 Recorded here are the blood types of 40 persons who have volunteered to donate blood at a plasma center. Summarize the data in a frequency table. Include calculations of the relative frequencies.

O	O	A	B	A	O	A	A	A	O
B	O	B	O	O	A	O	O	A	A
A	A	AB	A	B	A	A	O	O	A
O	O	A	A	A	O	A	O	O	AB

2.3 A student at the University of Wisconsin surveyed 40 students in her dorm concerning their participation in extracurricular activities during the past week. The data on number of activities are

1	5	0	1	4	3	0	2	1	6	1	1	0	0
2	0	0	3	1	2	1	2	2	2	2	2	1	0
2	2	3	4	2	7	2	2	3	3	1	1		

Present these data in a frequency table and in a relative frequency bar chart.

2.4 The number of automobile accidents reported per month helps to identify intersections that require improvement. The number of crashes per month reported at an intersection near a university campus in Madison, Wisconsin, are

1	3	3	3	2	2	3	1	2	4	1	4
1	3	1	1	1	0	1	2	2	5	5	2
5	5	4	3	3	6	1	2	3	2	4	3
4	4	3	5	3	3	3	5	1	5	5	3
4	2	2	0	0	1	4	1	0	2	0	

Present these data in a frequency table and in a relative frequency bar chart.

2.5 Eighty customers at a bakery named their favorite pie. The responses are as follows:

Pie	Frequency
Apple	31
Pumpkin	28
Pecan	12
Other	9

(a) Calculate the frequency for each pie.

(b) Construct a pie chart.

2.6 Of the $207 million raised by a major university's fund drive, $117 million came from individuals and bequests, $24 million from industry and business, and $66 million from foundations and associations. Present this information in the form of a pie chart.

2.7 Data from one campus dorm on the number of burglaries are collected each week of the semester. These data are to be grouped into the classes 0–1, 2–3, 3–5, 6 or more. Both endpoints included. Explain where a difficulty might arise.

2.8 The number of goals your favorite ice hockey team scores are to be collected for each game. These game totals are to be grouped into the classes 0–1, 2–3, 4–5, 7 or more. Both endpoints are included. Explain where a difficulty might arise.

2.9 A sample of persons will each be asked to give the number of their close friends. The responses are to be grouped into the following classes: 0, 1–3, 3–5, 6 or more. Left endpoint is included. Explain where difficulties might arise.

2.10 The weights of the players on the university football team (to the nearest pound) are to be grouped into the following classes: 160–175, 175–190, 190–205, 205–220, 220–235, 235 or more. The left endpoint is included but not the right endpoint. Explain where difficulties might arise.

2.11 On flights from San Francisco to Chicago, the number of empty seats are to be grouped into the following classes: 0–4, 5–9, 10–14, 15–19, more than 19.

Is it possible to determine from this frequency distribution the exact number of flights on which there were:

(a) Fewer than 10 empty seats?

(b) More than 14 empty seats?

(c) At least 5 empty seats?

(d) Exactly 9 empty seats?

(e) Between 5 and 15 empty seats inclusively?

2.12 A major West Coast power company surveyed 50 customers who were asked to respond to the statement, "People should rely mainly on themselves to solve problems caused by power outages" with one of the following responses.

1. Definitely agree.
2. Somewhat agree.

3. Somewhat disagree.
4. Definitely disagree.

The responses are as follows:

4	2	1	3	3	2	4	2	1	1	2	2	2	2	1	3	4
1	4	4	1	3	2	4	1	4	3	3	1	1	1	2	1	1
4	4	4	4	4	1	2	2	2	4	4	4	1	3	4	2	

Construct a frequency table.

2.13 A sample of 50 departing airline passengers at the main check-in counter produced the following number of bags checked through to final destinations.

0	1	2	2	1	2	1	2	3	0	1	0
1	1	0	1	3	0	1	2	1	1	1	2
1	2	2	1	2	0	0	2	2	1	1	1
1	1	1	1	2	0	1	3	0	1	2	1
1	3										

(a) Make a relative frequency line diagram.

(b) Comment on the pattern.

(c) What proportion of passengers who check in at the main counter fail to check any bags?

2.14 A person with asthma took measurements by blowing into a peak-flow meter on seven consecutive days.

429 425 471 422 432 444 454

Display the data in a dot diagram.

2.15 Before microwave ovens are sold, the manufacturer must check to ensure that the radiation coming through the door is below a specified safe limit. The amounts of radiation leakage ($\mathrm{mW/cm^2}$) with the door closed from 25 ovens are as follows (courtesy of John Cryer):

15	9	18	10	5	12	8
5	8	10	7	2	1	
5	3	5	15	10	15	
9	8	18	1	2	11	

Display the data in a dot diagram.

2.16 A campus area merchant recorded the number of bad checks received per month, for five months

4 5 4 7 6

Display the data in a dot diagram.

2.17 The city of Madison regularly checks the water quality at swimming beaches located on area lakes. The concentration of fecal coliforms, in number of colony forming units (CFU) per 100 ml of water, was measured on fifteen days during the summer at one beach.

180	1600	90	140	50	260	400	90
380	110	10	60	20	340	80	

(a) Make a dot diagram.

(b) Comment on the pattern and any unusual features.

(c) The city closes any swimming beach if a count is over 1350. What proportion of days, among the fifteen, was this beach closed?

2.18 Tornadoes kill many people every year in the United States. The yearly number of lives lost during the 50 years 1962 through 2011 are summarized in the following table.

Number of Deaths	Frequency
24 or less	3
25–49	18
50–74	14
75–99	6
100–149	5
150–199	1
200 or more	3
Total	50

(a) Calculate the relative frequency for the intervals [0, 25), [25, 50) and so on where the right-hand endpoint is excluded. Take the last interval to be [200, 600).

(b) Plot the relative frequency histogram. (*Hint*: Since the intervals have unequal widths, make the height of each rectangle equal to the relative frequency divided by the width of the interval.)

(c) What proportion of the years had 49 or fewer deaths due to tornadoes?

(d) Comment on the shape of the distribution.

2.19 A zoologist collected wild lizards in the Southwestern United States. Thirty lizards from the genus *Phrynosoma* were placed on a treadmill and their speed measured. The recorded speed (meters/second) is the fastest time to run a half meter (courtesy of K. Bonine).

1.28	1.36	1.24	2.47	1.94	2.52	2.67	1.29
1.56	2.66	2.17	1.57	2.10	2.54	1.63	2.11
2.57	1.72	0.76	1.02	1.78	0.50	1.49	1.57
1.04	1.92	1.55	1.78	1.70	1.20		

(a) Construct a frequency distribution using the class intervals 0.45–0.90, 0.90–1.35, and so on, with the endpoint convention that the left endpoint is included and the right endpoint is excluded. Calculate the relative frequencies.

(b) Make a histogram.

2.20 In a recent year, 35 sites around the world experienced earthquakes of magnitude greater than 6.5.

6.6	6.6	6.6	6.6	6.6	6.6	6.7	6.7	6.7
6.7	6.7	6.8	6.8	6.8	6.9	6.9	6.9	6.9
7.0	7.0	7.0	7.0	7.1	7.1	7.1	7.2	7.2
7.5	7.6	7.7	7.8	7.8	7.8	7.9	7.9	

Construct a histogram using equal-length intervals starting with (6.5, 6.8], where the right-hand endpoint is included but not the left-hand endpoint.

2.21 Referring to Exercise 2.20, construct a density histogram using the intervals (6.5, 6.7], (6.7, 6.9], (6.9, 7.1], (7.1, 7.5], and (7.5, 7.9].

Using R: With the observations in **x**

```
hist(x,breaks=c(6.5,6.7,6.9,7.1,7.5,7.9),prob=T)
```

2.22 The following data represent the scores of 40 students on a college qualification test (courtesy of R. W. Johnson).

162	171	138	145	144	126	145	162	174	178
167	98	161	152	182	136	165	137	133	143
184	166	115	115	95	190	119	144	176	135
194	147	160	158	178	162	131	106	157	154

Make a stem-and-leaf display.

2.23 A federal government study of the oil reserves in Elk Hills, CA, included a study of the amount of iron present in the oil.

Amount of Iron (percent ash)				
20	18	25	26	17
14	20	14	18	15
22	15	17	25	22
12	52	27	24	41
34	20	17	20	19
20	16	20	15	34
22	29	29	34	27
13	6	24	47	32
12	17	36	35	41
36	32	46	30	51

Make a stem-and-leaf display.

2.24 The following is a stem-and-leaf display with two-digit leaves. (The leading leaf digit = 10.0.)

1					
2	46	68	93		
3	19	44	71	82	97
4	05	26	43	90	
5	04	68			
6	13				

List the corresponding measurements.

2.25 If there are too many leaves on some stems in a stem-and-leaf display, we might double the number of stems. The leaves 0–4 could hang on one stem and 5–9 on the repeated stem. For the observations

193	198	200	202	203	203	205	205	206	207	
207	208	212	213	214	217	219	220	222	226	237

we would get the **double-stem display**

19	3
19	8
20	0233
20	556778
21	234
21	79
22	02
22	6
23	
23	7

Construct a double-stem display with one-digit leaves for the data of Exercise 2.23.

2.26 If the double-stem display still has too few stems, we may wish to construct a stem-and-leaf display with a separate stem to hold leaves 0 and 1, 2 and 3, 4 and 5, 6 and 7, and a stem to hold 8 and 9. The resulting stem-and-leaf display is called a **five-stem display**. The following is a five-digit stem-and-leaf display. (Leaf unit = 1.0)

1	8
2	001
2	2233
2	444555
2	667
2	9
3	0

List the corresponding measurements.

2.27 Referring to Exercise 2.20, construct a five-stem display for the magnitude of earthquakes.

3. MEASURES OF CENTER

The graphic procedures described in Section 2 help us to visualize the pattern of a data set of measurements. To obtain a more objective summary description and a comparison of data sets, we must go one step further and obtain numerical values for the location or center of the data and the amount of variability present. Because data are normally obtained by sampling from a large population, our discussion of numerical measures is restricted to data arising in this context. Moreover, when the population is finite and completely sampled, the same arithmetic operations can be carried out to obtain numerical measures for the population.

To effectively present the ideas and associated calculations, it is convenient to represent a data set by symbols to prevent the discussion from becoming anchored to a specific set of numbers. A data set consists of a number of measurements that are symbolically represented by $x_1, x_2, \ldots, x_n$. The last subscript n denotes the number of measurements in the data, and $x_1, x_2, \ldots$ represent the first observation, the second observation, and so on. For instance, a data set consisting of the five measurements 2.1, 3.2, 4.1, 5.6, and 3.7 is represented in symbols by x_1, x_2, x_3, x_4, x_5, where $x_1 = 2.1$, $x_2 = 3.2$, $x_3 = 4.1$, $x_4 = 5.6$, and $x_5 = 3.7$.

The most important aspect of studying the distribution of a sample of measurements is locating the position of a central value about which the measurements are distributed. The two most commonly used indicators of center are the **mean** and the **median**.

The **mean**, or **average**, of a set of measurements is the sum of the measurements divided by their number. For instance, the mean of the five measurements 2.1, 3.2, 4.1, 5.6, and 3.7 is

$$\frac{2.1 + 3.2 + 4.1 + 5.6 + 3.7}{5} = \frac{18.7}{5} = 3.74$$

To state this idea in general terms, we use symbols. If a sample consists of n measurements $x_1, x_2, \ldots, x_n$, the mean of the sample is

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\text{sum of the } n \text{ measurements}}{n}$$

The notation $\overline{x}$, read x bar, represents a sample mean. To further simplify the writing of a sum, the Greek capital letter $\sum$ (sigma) is used as a statistical shorthand. With this symbol:

The sum $x_1 + x_2 + \cdots + x_n$ is denoted as $\sum_{i=1}^{n} x_i$.
Read this as "the sum of all x_i with i ranging from 1 to n."

For example, $\sum_{i=1}^{5} x_i$ represents the sum $x_1 + x_2 + x_3 + x_4 + x_5$.

Remark: When the number of terms being summed is understood from the context, we often simplify to $\sum x_i$, instead of $\sum_{i=1}^{n} x_i$. Some further operations with the $\sum$ notation are discussed in Appendix A1.

We are now ready to formally define the sample mean.

The **sample mean** of a set of n measurements $x_1, x_2, \ldots, x_n$ is the sum of these measurements divided by n. The sample mean is denoted by $\overline{x}$.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad \frac{\sum x_i}{n}$$

According to the concept of "average," the mean represents a center of a data set. If we picture the dot diagram of a data set as a thin weightless horizontal bar on which balls of equal size and weight are placed at the positions of the data points, then the mean $\overline{x}$ represents the point on which the bar will balance. The computation of the sample mean and its physical interpretation are illustrated in Example 7.

Example 7 Calculating and Interpreting the Sample Mean

The birth weights in pounds of five babies born one day in the same hospital are 9.2, 6.4, 10.5, 8.1, and 7.8. Obtain the sample mean and create a dot diagram.

SOLUTION The mean birth weight for these data is

$$\overline{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42.0}{5} = 8.4 \text{ pounds}$$

The dot diagram of the data appears in Figure 9, where the sample mean (marked by Δ) is the balancing point or center of the picture.

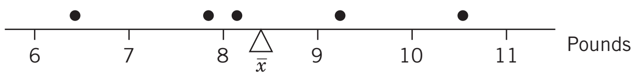


Figure 9 Dot diagram and the sample mean for the birth-weight data.

Another measure of center is the middle value.

> The **sample median** of a set of n measurements $x_1, \ldots, x_n$ is the middle value when the measurements are arranged from smallest to largest.

Roughly speaking, the median is the value that divides the data into two equal halves. In other words, 50% of the data lie below the median and 50% above it. If n is an odd number, there is a unique middle value and it is the median. If n is an even number, there are two middle values and the median is defined as their average. For instance, the ordered data 3, 5, 7, 8 have two middle values 5 and 7, so the median $= (5 + 7)/2 = 6$.

Example 8 Calculating the Sample Median

Find the median of the birth-weight data given in Example 7.

SOLUTION The measurements, ordered from smallest to largest, are

$$6.4 \quad 7.8 \quad \boxed{8.1} \quad 9.2 \quad 10.5$$

The middle value is 8.1, and the median is therefore 8.1 pounds.

Example 9 Choosing between the Mean and Median

Calculate the median of the survival times given in Example 5. Also calculate the mean and compare.

SOLUTION To find the median, first we order the data. The ordered values are

$$3 \qquad 15 \qquad 46 \qquad 64 \qquad 126 \qquad 623$$

There are two middle values, so

$$\text{Median} = \frac{46 + 64}{2} = 55 \text{ days}$$

The sample mean is

$$\bar{x} = \frac{3 + 15 + 46 + 64 + 126 + 623}{6} = \frac{877}{6} = 146.2 \text{ days}$$

Note that one large survival time greatly inflates the mean. Only 1 out of the 6 patients survived longer than $\bar{x} = 146.2$ days. Here the median of 55 days appears to be a better indicator of the center than the mean.

Example 9 demonstrates that the median is not affected by a few very small or very large observations, whereas the presence of such extremes can have a considerable effect on the mean.

For extremely asymmetrical distributions, the median is likely to be a more sensible measure of center than the mean. That is why government reports on income distribution quote the median income as a summary, rather than the mean. A relatively small number of very highly paid persons can have a great effect on the mean salary.

If the number of observations is quite large (greater than, say, 25 or 30), it is sometimes useful to extend the notion of the median and divide the **ordered data** set into quarters. Just as the point for division into halves is called the median, the points for division into quarters are called **quartiles**. The points of division into more general fractions are called **percentiles**.

The sample 100 *p*-th percentile is a value such that after the data are ordered from smallest to largest, at least 100*p*% of the observations are at or below this value and at least 100 (1 − *p*)% are at or above this value.

If we take $p = .5$, the above conceptual description of the sample $100(.5) = 50$th percentile specifies that at least half the observations are equal or smaller and at least half are equal or larger. If we take $p = .25$, the sample $100(.25) = 25$th percentile has proportion one-fourth of the observations that are the same or smaller and proportion three-fourths that are the same or larger.

We adopt the convention of taking an observed value for the sample percentile except when two adjacent values satisfy the definition, in which case their average is taken as the percentile. This coincides with the way the median is defined when the sample size is even. When all values in an interval satisfy the definition of a percentile, the particular convention used to locate a point in the interval does not appreciably alter the results in large data sets, except perhaps for the determination of extreme percentiles (those before the 5th or after the 95th percentile).

The following operating rule will simplify the calculation of the sample percentile.

Calculating the Sample 100*p*-th Percentile

1. Order the data from smallest to largest.
2. Determine the product $(\textit{sample size}) \times (\textit{proportion}) = np$.

If np is not an integer, round it up to the next integer and find the corresponding ordered value.
If np is an integer, say k, calculate the average of the kth and $(k + 1)$st ordered values.

The quartiles are simply the 25th, 50th, and 75th percentiles.

Sample Quartiles

Lower (first) quartile	Q_1	=	25th percentile
Second quartile (or median)	Q_2	=	50th percentile
Upper (third) quartile	Q_3	=	75th percentile

Example 10 Calculating Quartiles to Summarize Length of Phone Calls

An administrator wanted to study the utilization of long-distance telephone service by a department. One variable of interest is the length, in minutes, of long-distance calls made during one month. There were 38 calls that resulted in a connection. The lengths of calls, already ordered from smallest to largest, are presented in Table 8. Locate the quartiles and also determine the 90th percentile.

TABLE 8 The Lengths of Long-Distance Phone Calls in Minutes

1.6	1.7	1.8	1.8	1.9	2.1	2.5	3.0	3.0	4.4
4.5	4.5	5.9	7.1	7.4	7.5	7.7	8.6	9.3	9.5
12.7	15.3	15.5	15.9	15.9	16.1	16.5	17.3	17.5	19.0
19.4	22.5	23.5	24.0	31.7	32.8	43.5	53.3		

SOLUTION To determine the first quartile, we take $p = .25$ and calculate the product $38 \times .25 = 9.5$. Because 9.5 is not an integer, we take the next largest integer, 10. In Table 8, we see that the 10th ordered observation is 4.4 so the first quartile is $Q_1 = 4.4$ minutes.

We confirm that this observation has 10 values *at or below* it and 29 values *at or above* so that it does satisfy the conceptual definition of the first quartile.

For the median, we take $p = .5$ and calculate $38 \times .5 = 19$. Because this is an integer, we average the 19th and 20th smallest observations to obtain the median, $(9.3 + 9.5)/2 = 9.4$ minutes.

Next, to determine the third quartile, we take $p = .75$ and calculate $38 \times .75 = 28.5$. The next largest integer is 29, so the 29th ordered observation is the third quartile $Q_3 = 17.5$ minutes. More simply, we could mimic the calculation of the first quartile but now count down 10 observations starting with the largest value.

For the 90th percentile, we determine $38 \times .90 = 34.2$, which we increase to 35. The 90th percentile is 31.7 minutes. Only 10% of calls last 31.7 minutes or longer.

Exercises

2.28 Calculate the mean and median for each of the following data sets.

(a) 2 10 3 6 4

(b) 3 2 7 4

2.29 Calculate the mean and median for each of the following data sets.

(a) 3 6 2 5 4

(b) 4 3 8 5

(c) −4 0 −3 −1 2 −1 0

2.30 The height that bread rises may be one indicator of how light it will be. As a first step, before modifying her existing recipe, a student cook measured the raise height (cm) on eight occasions:

6.3 6.9 5.7 5.4 5.6 5.5 6.6 6.5

Find the mean and median of the raised heights.

2.31 With reference to the water quality in Exercise 2.17:

(a) Find the sample mean.

(b) Does the sample mean or the median give a better indication of the water quality of a "typical" day? Why?

2.32 The monthly income in dollars for seven sales persons at a car dealership are

2450 2275 2425 4700 2650 2350 2475

(a) Calculate the mean and median salary.

(b) Which of the two is preferable as a measure of center and why?

2.33 Records show that in Las Vegas, NV, the normal daily maximum temperature (°F) for each month starting in January is

56 62 68 77 87 99 105 102 95 82 66 57

Verify that the mean of these figures is 79.67. Comment on the claim that the daily maximum temperature in Las Vegas averages a pleasant 79.67.

2.34 A sample of six university students responded to the question, "How much time, in minutes, did you spend on the social network site yesterday?" One student never used the site and was dropped from the study.

100 45 60 130 30

Calculate the sample mean time for the five students that use the site.

2.35 With reference to the radiation leakage data given in Exercise 2.15:

(a) Calculate the sample mean.

(b) Which gives a better indication of the amount of radiation leakage, the sample mean or the median?

2.36 In 2017, there were eight skyscrapers in the world over 500 meters tall. The ordered heights are

508 530 541 555 594 601 632 828

(a) Calculate the sample mean height for the eight skyscrapers.

(b) Drop the largest value and recalculate the mean. Comment on the effect of dropping one very large observation.

2.37 The weights (oz) of nineteen babies born in Madison, Wisconsin, are summarized in the computer output

```
Descriptive Statistics: Weight

Variable   N    Mean   Median   StDev
Weight    19  118.05   117.00   15.47
```

Locate two measures of center tendency, or location, and interpret the values.

2.38 With reference to the extracurricular activities data in Exercise 2.3, obtain the

(a) Sample mean.

(b) Sample median.

(c) Comment on the effect of a few large observations.

2.39 With reference to the number of returns in Example 4, obtain the sample (a) mean and (b) median.

2.40 Old Faithful, the most famous geyser in Yellowstone Park, had the following durations (measured in seconds) in six consecutive eruptions:

240 248 113 268 117 253

(a) Find the sample median.

(b) Find the sample mean.

2.41 Loss of calcium is a serious problem for older women. To investigate the amount of loss, a researcher measured the initial amount of bone mineral content in the radius bone of the dominant hand of elderly women and then the amount remaining after one year. The differences, representing the loss of bone mineral content, are given in the following table (courtesy of E. Smith).

8	7	13	3	6
4	8	6	3	4
0	1	11	7	1
8	6	12	13	10
9	11	3	2	9
7	1	16	3	2
10	15	2	5	8
17	8	2	5	5

(a) Find the sample mean.

(b) Does the sample mean or the median give a better indication of the amount of mineral loss?

2.42 Physical education researchers interested in the development of the overarm throw measured the horizontal velocity of a thrown ball at the time of release. The results for first-grade children (in feet/second) (courtesy of L. Halverson and M. Roberton) are

Males

54.2	39.6	52.3	48.4	35.9	30.4	25.2	45.4	48.9
48.9	45.8	44.0	52.5	48.3	59.9	51.7	38.6	39.1
49.9	38.3							

Females

30.3	43.0	25.7	26.7	27.3	31.9	53.7	32.9	19.4
23.7	23.3	23.3	37.8	39.5	33.5	30.4	28.5	

(a) Find the sample median for males.

(b) Find the sample median for females.

(c) Find the sample median for the combined set of males and females.

2.43 On opening day one season, 10 major league baseball games were played and they lasted the following numbers of minutes.

167 211 187 176 170 158 198 218 145 232

Find the sample median.

2.44 If you were to use the data on the length of major league baseball games in Exercise 2.43 to estimate the total amount of digital memory needed to film another 10 major league baseball games, which is the more meaningful description, the sample mean or the sample median? Explain.

2.45 The following measurements of the diameters (in feet) of Indian mounds in southern Wisconsin were gathered by examining reports in the *Wisconsin Archaeologist* (courtesy of J. Williams).

22 24 24 30 22 20 28 30 24 34 36 15 37

(a) Create a dot diagram.

(b) Calculate the mean and median and then mark these on the dot diagram.

(c) Calculate the quartiles.

2.46 With reference to Exercise 2.3, calculate the quartiles.

2.47 Refer to the data of college qualification test scores given in Exercise 2.22.

(a) Find the median.

(b) Find Q_1 and Q_3.

2.48 A large mail-order firm employs numerous persons to take phone orders. Computers on which orders are entered also automatically collect data on phone activity. One variable useful for planning staffing levels is the number of calls per shift handled by each employee. From the data collected on 25 workers, calls per shift were (courtesy of Land's End)

118	118	57	92	127	109	96	68	73
69	106	91	93	94	102	105	100	104
80	50	96	82	72	108	73		

Calculate the sample mean.

2.49 With reference to Exercise 2.48, calculate the quartiles.

2.50 The speedy lizard data, from Exercise 2.19, are

1.28	1.36	1.24	2.47	1.94	2.52	2.67	1.29
1.56	2.66	2.17	1.57	2.10	2.54	1.63	2.11
2.57	1.72	0.76	1.02	1.78	0.50	1.49	1.57
1.04	1.92	1.55	1.78	1.70	1.20		

(a) Find the sample median, first quartile, and third quartile.

(b) Find the sample 90th percentile.

2.51 With reference to the water quality data in Exercise 2.17:

(a) Find the sample median, first quartile, and third quartile.

(b) Find the sample 90th percentile.

2.52 ***Some properties of the mean and median.***

1. If a fixed number c is added to all measurements in a data set, then the mean of the new measurements is

   $c \; + \; (\text{ the original mean })$.

2. If all measurements in a data set are multiplied by a fixed number d, then the mean of the new measurements is

   $d \; \times \; (\text{ the original mean })$.

(a) Verify these properties for the data set

4 8 8 7 9 6

taking $c = 4$ in property (1) and $d = 2$ in (2).

(b) The same properties also hold for the median. Verify these for the data set and the numbers c and d given in part (a).

2.53 On a day, the noon temperature measurements (in °F) reported by five weather stations in a state were

74 80 76 76 73

(a) Find the mean and median temperature in °F.

(b) The Celsius (°C) scale is related to the Fahrenheit (°F) scale by $\text{C} = \frac{5}{9}(\text{F} - 32)$. What are the mean and median temperatures in °C? (Answer without converting each temperature measurement to °C. Use the properties stated in Exercise 2.52.)

2.54 The mean and median salaries for middle management employees at two similar companies A and B in an area are as follows:

	Company A	Company B
Mean salary	\$85,000	\$81,000
Median salary	\$70,000	\$74,000

Assume that the salaries are set in accordance with job competence and the overall quality of workers is about the same in the two companies.

(a) Which company offers a better prospect to a person having superior management ability?

(b) Where can a medium-quality manager expect to earn more? Explain your answer.

2.55 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on testosterone x_4 for male alligators:

(a) Make separate dot plots for the Lake Apopka and Lake Woodruff alligators.

(b) Calculate the sample means for each group.

(c) Do the concentrations of testosterone appear to differ between the two groups? What does this suggest the contamination has done to male alligators in the Lake Apopka habitat?

2.56 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on testosterone x_4 from Lake Apopka:

(a) Make separate dot plots for the male and female alligators.

(b) Calculate the sample means for each group.

(c) Do the concentrations of testosterone appear to differ between the two groups? We would expect differences. What does your graph suggest the contamination has done to alligators in the Lake Apopka habitat?

4. MEASURES OF VARIATION

Besides locating the center of the data, any descriptive study of data must numerically measure the extent of variation around the center. Two data sets may exhibit similar positions of center but may be remarkably different with respect to variability. For example, the dots in Figure 10*b* are more scattered than those in Figure 10*a*.

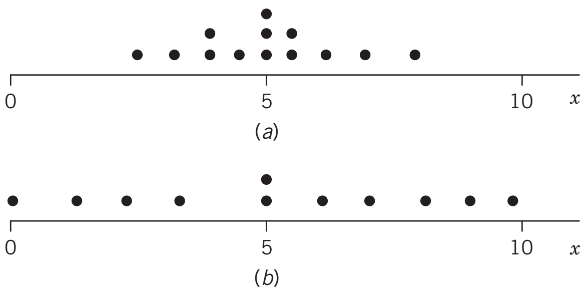


Figure 10 Dot diagrams with similar center values but different amounts of variation.

Because the sample mean $\overline{x}$ is a measure of center, the **variation** of the individual data points about this center is reflected in their deviation from the mean

$$\begin{aligned}\text{Deviation} &= \text{Observation} - (\text{Sample mean})\\ &= x - \overline{x}\end{aligned}$$

For instance, the data set 3, 5, 7, 7, 8 has mean $\overline{x} = (3 + 5 + 7 + 7 + 8)/5 = 30/5 = 6$, so the deviations are calculated by subtracting 6 from each observation. See Table 9.

TABLE 9 Calculation of Deviations

Observation x	Deviation $x - \overline{x}$
3	−3
5	−1
7	1
7	1
8	2

One might feel that the average of the deviations would provide a numerical measure of spread. However, some deviations are positive and some negative, and the total of the positive deviations exactly cancels the total of the negative ones. In the foregoing example, we see that the positive deviations add to 4 and the negative ones add to −4, so the total deviation is 0. With a little reflection on the definition of the sample mean, the reader will realize that this was not just an accident. For any data set, the total deviation is 0 (for a formal proof of this fact, see Appendix A1).

$$\sum (\text{Deviations}) = \sum (x_i - \overline{x}) = 0$$

To obtain a measure of spread, we must eliminate the signs of the deviations before averaging. One way of removing the interference of signs is to square the numbers. A measure of **spread**, called the **sample variance**, is constructed by adding the squared deviations and dividing the total by the number of observations minus one.

Sample variance of n observations:

$$\begin{aligned}s^2 &= \frac{\text{sum of squared deviations}}{n - 1}\\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}\end{aligned}$$

Example 11 Calculating Sample Variance

Calculate the sample variance of the data 3 5 7 7 8.

SOLUTION For this data set, $n = 5$. To find the variance, we first calculate the mean, then the deviations and the squared deviations. See Table 10.

TABLE 10 Calculation of Variance

	Observation	Deviation	(Deviation)^2
	x	$x - \overline{x}$	$(x - \overline{x})^2$
	3	−3	9
	5	−1	1
	7	1	1
	7	1	1
	8	2	4
Total	30	0	16
	$\sum x$	$\sum (x - \overline{x})$	$\sum (x - \overline{x})^2$

$$\overline{x} = \frac{30}{5} = 6$$

$$\text{Sample variance} \quad s^2 = \frac{16}{5 - 1} = 4$$

Remark: Although the sample variance is conceptualized as the **average squared deviation**, notice that the divisor is $n - 1$ rather than n. The divisor, $n - 1$, is called the degrees of freedom[1] associated with s^2.

Because the variance involves a sum of squares, its unit is the square of the unit in which the measurements are expressed. For example, if the data pertain to measurements of weight in pounds, the variance is expressed in $(\text{pounds})^2$. To obtain a measure of variability in the same unit as the data, we take the positive square root of the variance, called the **sample standard deviation**. The standard deviation rather than the variance serves as a basic measure of variability.

Sample Standard Deviation

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

Example 12 Calculating the Sample Standard Deviation

Calculate the standard deviation for the data of Example 11.

SOLUTION We already calculated the variance $s^2 = 4$ so the standard deviation is $s = \sqrt{4} = 2$.

To show that a larger spread of the data does indeed result in a larger numerical value of the standard deviation, we consider another data set in Example 13.

[1] The deviations add to 0 so a specification of any $n - 1$ deviations allows us to recover the one that is left out. For instance, the first four deviations in Example 11 add to −2, so to make the total 0, the last one must be +2, as it really is. In the definition of s^2, the divisor $n - 1$ represents the number of deviations that can be viewed as free quantities.

Example 13 Using Standard Deviations to Compare Variation in Two Data Sets

Calculate the standard deviation for the data 1, 4, 5, 9, 11. Plot the dot diagram of this data set and also the data set of Example 11.

SOLUTION The standard deviation is calculated in Table 11. The dot diagrams, given in Figure 11, show that the data points of Example 11 have less spread than those of Example 13. This visual comparison is confirmed by a smaller value of s for the first data set.

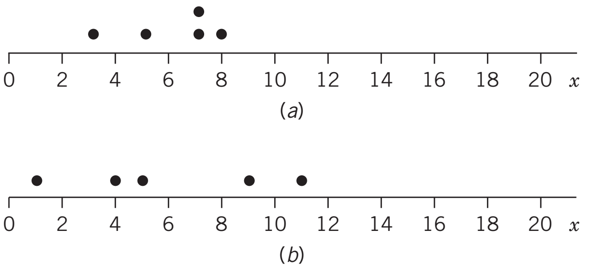


Figure 11 Dot diagrams of two data sets.

TABLE 11 Calculation of s

	x	$(x - \bar{x})$	$(x - \bar{x})^2$
	1	−5	25
	4	−2	4
	5	−1	1
	9	3	9
	11	5	25
Total	30	0	64
	$\bar{x} = 6$		

$$s^2 = \frac{64}{4} = 16$$

$$s = \sqrt{16} = 4$$

An alternative formula for the sample variance

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

does not require the calculation of the individual deviations. In hand calculation, the use of this alternative formula often reduces the arithmetic work, especially when $\bar{x}$ turns out to be a number with many decimal places. The equivalence of the two formulas is shown in Appendix A1.2.

Example 14 Calculating Sample Variance Using the Alternative Formula

In a psychological experiment a stimulating signal of fixed intensity is used on six experimental subjects. Their reaction times, recorded in seconds, are 4, 2, 3, 3, 6, 3. Calculate the standard deviation for the data by using the alternative formula.

SOLUTION These calculations can be conveniently carried out in tabular form:

	x	x^2
	4	16
	2	4
	3	9
	3	9
	6	36
	3	9
Total	21 $= \sum x$	83 $= \sum x^2$

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = \frac{83 - (21)^2/6}{5} = \frac{83 - 73.5}{5}$$

$$= \frac{9.5}{5} = 1.9$$

$$s = \sqrt{1.9} = 1.38 \text{ seconds}$$

The reader may do the calculations with the first formula and verify that the same result is obtained.

Sample z-score

The sample **z scale** (or **standard scale**), measures the position of a value relative to the sample mean in units of the standard deviation. Specifically,

$$\textbf{z-score} \text{ of a measurement } = \frac{\text{Measurement} - \bar{x}}{s}$$

Example 15 Litter Size of Lions in the Wild

Lions typically have babies in twos and threes but sometimes four or five. Many do not survive until age one. To protect the very young, the mother will take the babies away from the pride for the first 6 weeks.

The sizes of eight litters born to one pride of lions are (courtesy of Martina Trinkel)

3 5 3 3 2 3 3 1

(a) Find the sample mean, variance, and standard deviation.

(b) Calculate the *z*-score for a liter of size 2.

SOLUTION (a) We calculate

$$\bar{x} = \frac{3+5+3+3+2+3+3+1}{8} = \frac{23}{8} = 2.88 \quad \text{cubs}$$

and, to avoid the rounding error in the mean, we use the alternative formula

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

$$= \frac{3^2 + 5^2 + 3^2 + 3^2 + 2^2 + 3^2 + 3^2 + 1^2 - (23)^2/8}{8-1}$$

$$= 1.268$$

The sample standard deviation $s = \sqrt{1.268} = 1.126$ cubs.

(b) The z-score for the value 2 is $(2 - 2.88)/1.126 = -.78$ so it is .78 standard deviations below the sample mean number of cubs.

In Example 13, the data set with visibly larger variation yields a larger numerical value of s. The issue there surrounds a comparison between different data sets. In the context of a single data set, can we relate the numerical value of s to the physical closeness of the data points to the center $\bar{x}$? To this end, we view one standard deviation as a benchmark distance from the mean $\bar{x}$. For **bell-shaped distributions**, an empirical rule relates the standard deviation to the proportion of the data that lie in an interval around $\bar{x}$.

Empirical Guidelines for Symmetric Bell-Shaped Distributions

Approximately	68%	of the data lie within $\bar{x} \pm s$
	95%	of the data lie within $\bar{x} \pm 2s$
	99.7%	of the data lie within $\bar{x} \pm 3s$

Example 16 Comparing the Sleep Data with the Empirical Guidelines

Examine the 59 hours of sleep in Table 4 in the context of the empirical guideline.

SOLUTION Using a computer (see, for instance, Exercise 2.121), we obtain

$$\bar{x} = 7.18 \qquad s = 1.28 \qquad 2s = 2(1.28) = 2.56$$

The interval $7.18 - 1.28 = 5.90$ to $8.46 = 7.18 + 1.28$ contains 40 observations or $100(40/59) = 67.8\%$ of the data.

Going two standard deviations either side of $\bar{x}$ results in the interval

$$7.18 - 2.56 = 4.62 \quad \text{to} \quad 9.74 = 7.18 + 2.56$$

By actual count, all the observations except 4.5 and 10.0 fall in this interval. We find that $57/59 = .966$, or 96.6% of the observations lie within two standard deviations of $\bar{x}$.

The empirical guidelines suggest 68% and 95% respectively, so they are close.

Other Measures of Variation

Another measure of variation that is sometimes employed is

Sample range = Largest observation − Smallest observation

The range gives the length of the interval spanned by the observations.

Example 17 Calculating the Sample Range

Calculate the range for the hours of sleep data given in Example 6.

SOLUTION The data given in Table 4 contained

Smallest observation = 4.5
Largest observation = 10.0