

Introductory Business

Stat- istics

A solid red vertical band on the left side of the page, featuring a wavy, scalloped edge that separates it from the white background.

Introductory Business Statistics



Introductory Business Statistics

The world of statistics at your fingertips

SENIOR CONTRIBUTING AUTHORS

Alexander Holmes, the University of Oklahoma

Barbara Illowsky, De Anza College

Susan Dean, De Anza College

OpenStax
Rice University
6100 Main Street MS-375
Houston, Texas 77005

To learn more about OpenStax, visit <https://openstax.org>.
Individual print copies and bulk orders can be purchased through our website.

©**2018 Rice University**. Textbook content produced by OpenStax is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Under this license, any user of this textbook or the textbook contents herein must provide proper attribution as follows:

- If you redistribute this textbook in a digital format (including but not limited to PDF and HTML), then you must retain on every page the following attribution: “Download for free at <https://openstax.org/details/books/introductory-business-statistics>.”
- If you redistribute this textbook in a print format, then you must include on every physical page the following attribution: “Download for free at <https://openstax.org/details/books/introductory-business-statistics>.”
- If you redistribute part of this textbook, then you must retain in every digital format page view (including but not limited to PDF and HTML) and on every physical printed page the following attribution: “Download for free at <https://openstax.org/details/books/introductory-business-statistics>.”
- If you use this textbook as a bibliographic reference, please include <https://openstax.org/details/books/introductory-business-statistics> in your citation.

For questions regarding this licensing, please contact support@openstax.org.

Trademarks

The OpenStax name, OpenStax logo, OpenStax book covers, OpenStax CNX name, OpenStax CNX logo, OpenStax Tutor name, Openstax Tutor logo, Connexions name, Connexions logo, Rice University name, and Rice University logo are not subject to the license and may not be reproduced without the prior and express written consent of Rice University.

PRINT BOOK ISBN-10 1-947172-46-8
PRINT BOOK ISBN-13 978-1-947172-46-3
PDF VERSION ISBN-10 1-947172-47-6
PDF VERSION ISBN-13 978-1-947172-47-0
Revision Number IBS-2017-001(03/18)-LC
Original Publication Year 2017

Table of Contents

cover	1
Introductory Business Statistics	3
Introductory Business Statistics	5
Copyright page	6
Preface	9
Introduction	17
PART 1 INTRODUCTION TO STATISTICS	19
Sampling and Data	21
Descriptive Statistics	89
Appendix 1	91
Afterword	100

Preface

Welcome to *Introductory Business Statistics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP[®] courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax resources

CUSTOMIZATION

Licensing

Licensing of the source book

Introductory Business Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Making use of the open licensed source book

Because the source book is openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your

PREFACE

course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit the Instructor Resources section of your book page on OpenStax.org for more information.

Licensing of this book

This book is built from the OpenStax *Introductory Business Statistics* content, with minimal added extracts from Wikipedia to represent all types of content for an accurate design sample. You can distribute, remix, and build upon the content of this sample book, as long as you provide attribution to OpenStax and the other sources for the added extracts (quoted next to or below the relevant added content).

ERRATA

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

FORMAT

You can access this textbook for free in web view or PDF through OpenStax.org, and for a low cost in print.

About Introductory Business Statistics

Introductory Business Statistics is designed to meet the scope and sequence requirements of the one-semester statistics course for business, economics, and related majors. Core statistical concepts and skills have been augmented with

practical business examples, scenarios, and exercises. The result is a meaningful understanding of the discipline which will serve students in their business careers and real-world experiences.

COVERAGE AND SCOPE

Introductory Business Statistics began as a customized version of OpenStax *Introductory Statistics* by Barbara Illowsky and Susan Dean. Statistics faculty at The University of Oklahoma have used the business statistics adaptation for several years, and the author has continually refined it based on student success and faculty feedback.

The book is structured in a similar manner to most traditional statistics textbooks. The most significant topical changes occur in the latter chapters on regression analysis. Discrete probability density functions have been reordered to provide a logical progression from simple counting formulas to more complex continuous distributions. Many additional homework assignments have been added, as well as new, more mathematical examples.

Introductory Business Statistics places a significant emphasis on the development and practical application of formulas so that students have a deeper understanding of their interpretation and application of data. To achieve this unique approach, the author included a wealth of additional material and purposely de-emphasized the use of the scientific calculator. Specific changes regarding formula use include:

- Expanded discussions of the combinatorial formulas, factorials, and sigma notation
- Adjustments to explanations of the acceptance/rejection rule for hypothesis testing, as well as a focus on terminology regarding confidence intervals
- Deep reliance on statistical tables for the process of finding probabilities (which would not be required if probabilities relied on scientific calculators)
- Continual and emphasized links to the Central Limit Theorem throughout the book; *Introductory Business Statistics* consistently links each test statistic back to this fundamental theorem in inferential statistics

Another fundamental focus of the book is the link between statistical inference and the scientific method. Business and economics models are fundamentally

grounded in assumed relationships of cause and effect. They are developed to both test hypotheses and to predict from such models. This comes from the belief that statistics is the gatekeeper that allows some theories to remain and others to be cast aside for a new perspective of the world around us. This philosophical view is presented in detail throughout and informs the method of presenting the regression model, in particular.

The correlation and regression chapter includes confidence intervals for predictions, alternative mathematical forms to allow for testing categorical variables, and the presentation of the multiple regression model.

PEDAGOGICAL FEATURES

Examples are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics; these include examples about college life and learning, health and medicine, retail and business, and sports and entertainment.

- Practice, Homework, and Bringing It Together give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Additional resources

STUDENT AND INSTRUCTOR RESOURCES

We've compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

COMMUNITY HUBS

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a

PREFACE

platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit www.oercommons.org/hubs/OpenStax.

TECHNOLOGY PARTNERS

As allies in making high-quality learning materials accessible, our technology partners offer optional low-cost tools that are integrated with OpenStax books. To access the technology options for your text, visit your book page on OpenStax.org.

About the authors

SENIOR CONTRIBUTING AUTHORS

- Alexander Holmes, The University of Oklahoma
- Barbara Illowsky, DeAnza College
- Susan Dean, DeAnza College

CONTRIBUTING AUTHORS

Kevin Hadley, Analyst, Federal Reserve Bank of Kansas City

REVIEWERS

- Birgit Aquilonius, West Valley College
- Charles Ashbacher, Upper Iowa University - Cedar Rapids
- Abraham Biggs, Broward Community College
- Daniel Birmajer, Nazareth College
- Roberta Bloom, De Anza College
- Bryan Blount, Kentucky Wesleyan College

PREFACE

- Ernest Bonat, Portland Community College
- Sarah Boslaugh, Kennesaw State University
- David Bosworth, Hutchinson Community College
- Sheri Boyd, Rollins College
- George Bratton, University of Central Arkansas
- Franny Chan, Mt. San Antonio College
- Jing Chang, College of Saint Mary
- Laurel Chiappetta, University of Pittsburgh
- Lenore Desilets, De Anza College
- Matthew Einsohn, Prescott College
- Ann Flanigan, Kapiolani Community College
- David French, Tidewater Community College
- Mo Geraghty, De Anza College
- Larry Green, Lake Tahoe Community College
- Michael Greenwich, College of Southern Nevada
- Inna Grushko, De Anza College
- Valier Hauber, De Anza College
- Janice Hector, De Anza College
- Jim Helmreich, Marist College
- Robert Henderson, Stephen F. Austin State University
- Mel Jacobsen, Snow College
- Mary Jo Kane, De Anza College
- John Kagochi, University of Houston - Victoria
- Lynette Kenyon, Collin County Community College
- Charles Klein, De Anza College
- Alexander Kolovos
- Sheldon Lee, Viterbo University
- Sara Lenhart, Christopher Newport University
- Wendy Lightheart, Lane Community College
- Vladimir Logvenenko, De Anza College
- Jim Lucas, De Anza College
- Suman Majumdar, University of Connecticut
- Lisa Markus, De Anza College
- Miriam Masullo, SUNY Purchase
- Diane Mathios, De Anza College

PREFACE

- Robert McDevitt, Germanna Community College
- John Migliaccio, Fordham University
- Mark Mills, Central College
- Cindy Moss, Skyline College
- Nydia Nelson, St. Petersburg College
- Benjamin Ngwudike, Jackson State University
- Jonathan Oaks, Macomb Community College
- Carol Olmstead, De Anza College
- Barbara A. Osyk, The University of Akron
- Adam Pennell, Greensboro College
- Kathy Plum, De Anza College
- Lisa Rosenberg, Elon University
- Sudipta Roy, Kankakee Community College
- Javier Rueda, De Anza College
- Yvonne Sandoval, Pima Community College
- Rupinder Sekhon, De Anza College
- Travis Short, St. Petersburg College
- Frank Snow, De Anza College
- Abdulhamid Sukar, Cameron University
- Jeffery Taub, Maine Maritime Academy
- Mary Teegarden, San Diego Mesa College
- John Thomas, College of Lake County
- Philip J. Verrecchia, York College of Pennsylvania
- Dennis Walsh, Middle Tennessee State University
- Cheryl Wartman, University of Prince Edward Island
- Carol Weideman, St. Petersburg College
- Kyle S. Wells, Dixie State University
- Andrew Wiesner, Pennsylvania State University

Introduction

Introductory Business Statistics is designed to meet the scope and sequence requirements of the one-semester statistics course for business, economics, and related majors. Core statistical concepts and skills have been augmented with practical business examples, scenarios, and exercises. The result is a meaningful understanding of the discipline, which will serve students in their business careers and real-world experiences.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this book are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad".

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

INTRODUCTION

In this book, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics". You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this book, we will look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this book, you will learn how to solve probability problems using a systematic approach.

Part 1 Introduction to Statistics

In this part you'll learn the most important fundamentals of statistics, and how they are used in business. Once you've worked through this part, you'll be ready to tackle more advanced statistics, and to start applying your new skills to the world around you.




Figure 1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

1 Sampling and Data

How data are gathered and what "good" data can be distinguished from "bad"

*Alexander Holmes, The University of Oklahoma;
Barbara Illowsky, De Anza College; Susan Dean, De
Anza College; Kevin Hadley, Federal Reserve Bank of
Kansas City*

AFTER READING THIS CHAPTER YOU WILL BE ABLE TO:

- communicate key statistical concepts to non-statisticians.
- identify and articulate strategies for dealing with ethical issues that may arise.
- gain proficiency in using statistical terminology in written reports.
- calculate key statistics given relevant sample data.

OUTLINE

- 1.1 Definitions of Statistics, Probability, and Key Terms
- 1.2 Data, Sampling, and Variation in Data and Sampling
- 1.3 Levels of Measurement
- 1.4 Experimental Design and Ethics

Introduction

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad".

In this chapter we focus on answering the following key questions:

- Why is statistics important in daily life?
- What is probability?
- What is data sampling and what are the red flags and key risks when gathering data samples?
- What kinds of measurement can we use for gathering our own samples?
- How can we design experiments and what are the ethical considerations we need to take into account?

1.1 Definitions of Statistics, Probability, and Key Terms

Statistics reveal the chaotic underbelly, reveal

the tessellating uncertainties, reveal

our deepest truths hiding behind confidence levels

– *Asher Smith*

STATISTICS

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

PROBABILITY

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$

or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

KEY TERMS

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average

number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter, in this case the mean. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, or random variable, usually notated by capital letters such as X and Y , is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

EXAMPLE 1.1

Problem

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

TRY IT YOURSELF

Determine what the key terms refer to in the following study. We want to know the mean amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Reminder: This exercise requires you to identify the key terms of **population**, **sample**, **parameter**, **statistic**, **variable**, and **data**, not calculate the mean.

Shall I compare thee to a diff'rent mean?
 The control group's points art lower than thine
 From the preliminary graph I've seen.
 Do the means differ or in fact align?

First I will find the right parameter.
 Data on the sample I will collect.
 Which method to use I can't yet be sure,
 Til the assumptions are thoroughly checked.

The appropriate test statistic found,
 Thy humble knave compute the value p
 And conclude whether the diff'rence profound.
 Look at the alpha and then thou will see.

Then the confidence interval will tell
 The upper bound and the lower as well.

– *Cristina Gonzales*

EXAMPLE 1.2

Problem

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population _____ 2. Statistic _____ 3. Parameter _____ 4. Sample _____ 5. Variable _____ 6. Data _____

all students who attended the college last year

A. the cumulative GPA of one student who graduated from the college last year

B. 3.65, 2.80, 1.50, 3.90

C. a group of students who graduated from the college last year, randomly selected

D. the average cumulative GPA of students who graduated from the college last year

E. all students who graduated from the college last year

F. the average cumulative GPA of students in the study who graduated from the college last year

Solution

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

EXAMPLE 1.3

Problem

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which cars crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Table 1.1

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = whether a dummy (if it had been a real person) would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

EXAMPLE 1.4**Problem**

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = whether an individual doctor has been involved in a malpractice suit.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

1.2 Data, Sampling, and Variation in Data and Sampling

Data on the sample I will collect.

Which method to use I can't yet be sure,

Til the assumptions are thoroughly checked.

– *Cristina Gonzales*

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. **Qualitative data** are also often called categorical data. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative (categorical) data. Qualitative (categorical) data are generally described by words or letters. For instance, hair color might be

black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative (categorical) data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

EXAMPLE 1.5

Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

TRY IT YOURSELF

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

If you are not confident in your answer, recap the definitions of the key terms at the start of this section.

EXAMPLE 1.6

Data Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data.

TRY IT YOURSELF

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

EXAMPLE 1.7

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Problem

Name data sets that are quantitative discrete, quantitative continuous, and qualitative(categorical).

Solution

One Possible Solution:

The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.

- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative(categorical) data because they are categorical.

Try to identify additional data sets in this example.

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

EXAMPLE 1.10

Problem

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure 1.2. What type of data does this graph show?

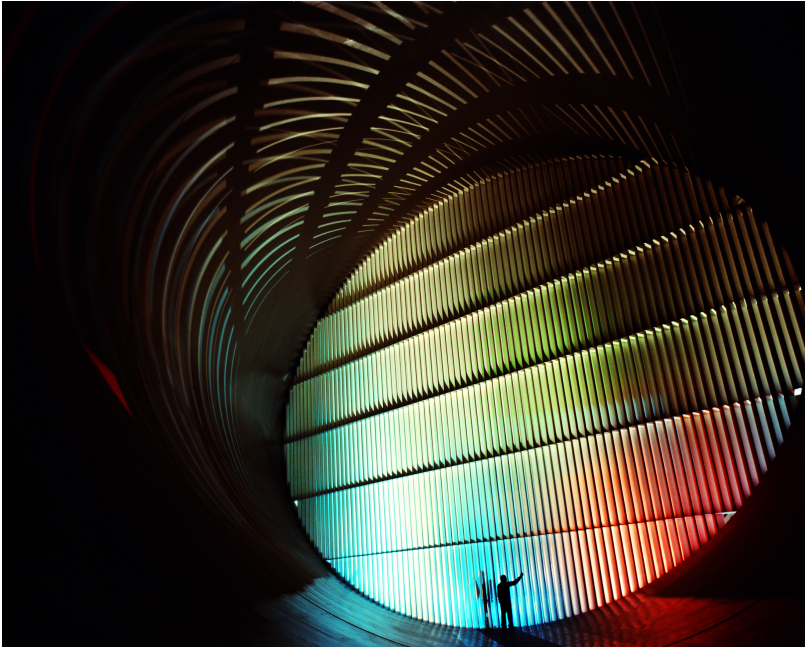


Figure 1.2

Solution

This pie chart shows the students in each year, which is **qualitative (or categorical) data**.

TRY IT YOURSELF

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

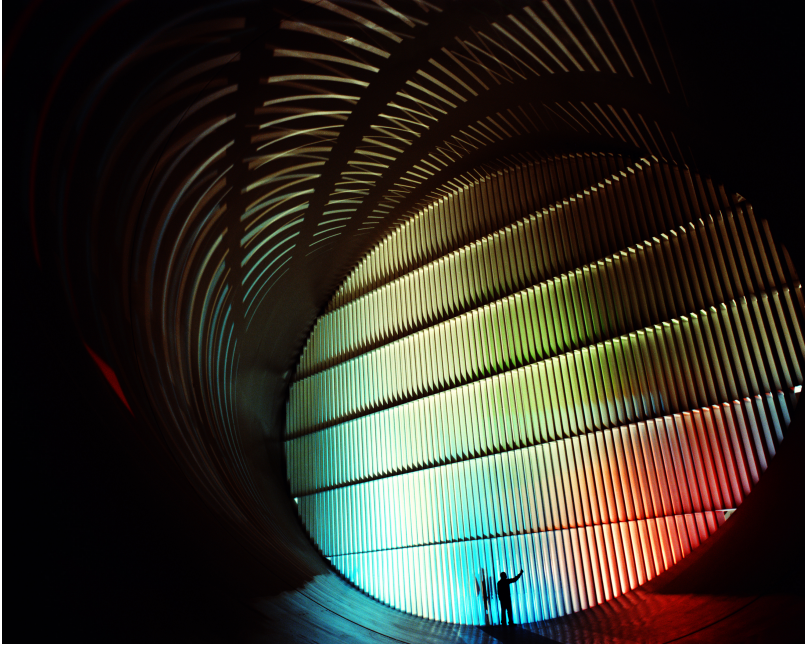


Figure 1.3

What type of data does this graph show?

QUALITATIVE DATA DISCUSSION

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College			Foothill College		
	Number	Percent		Number	Percent
Full-time	9,200	40.9%	Full-time	4,059	28.6%
Part-time	13,296	59.1%	Part-time	10,124	71.4%
Total	22,496	100%	Total	14,183	100%

Table 1.2 Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative(categorical) data are pie charts and bar graphs.

In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figure 1.4 and Figure 1.5 and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

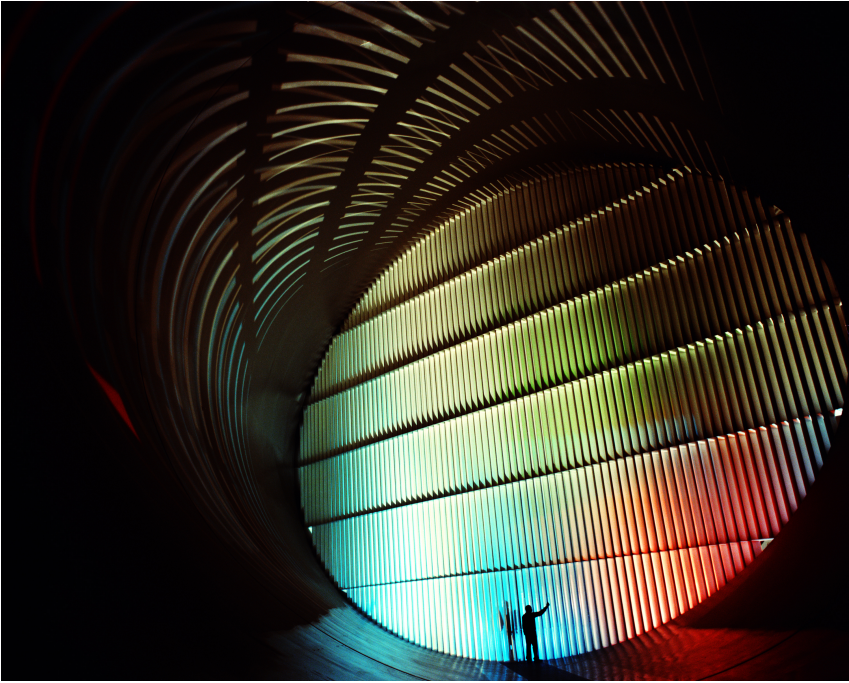


Figure 1.4

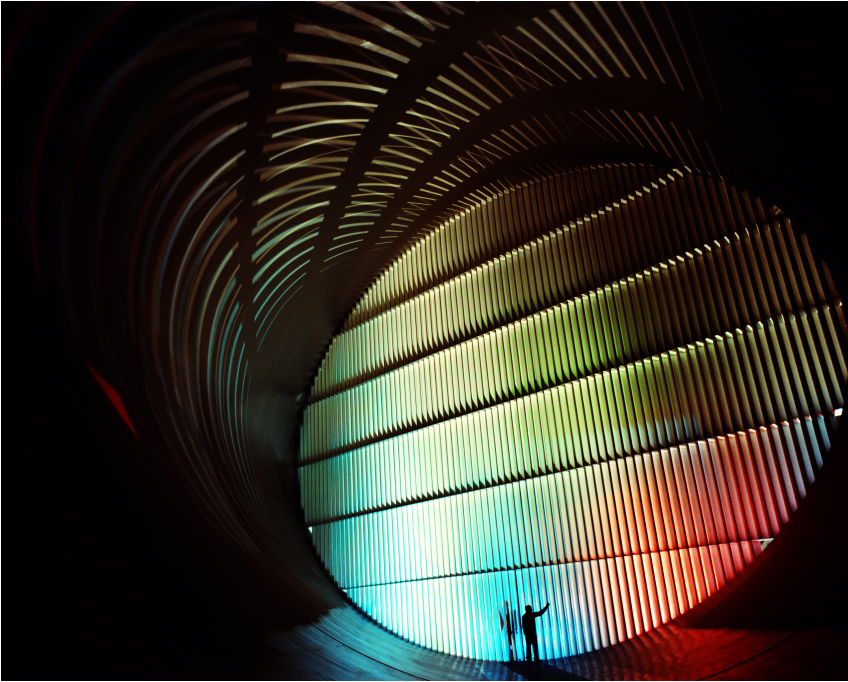


Figure 1.5

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/category	Percent
Full-time students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

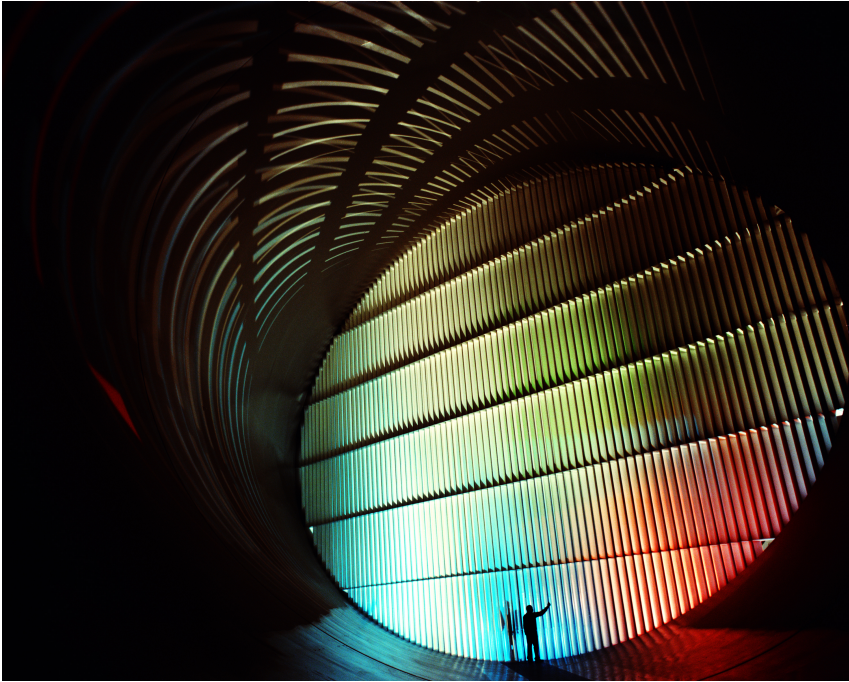
Table 1.3 De Anza College Spring 2010

Figure 1.6

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

SAMPLING AND DATA

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.4 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)



Figure 1.7

The following graph is the same as the previous graph but the “Other/Unknown” percent (9.6%) has been included. The “Other/Unknown” category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 1.8 can be difficult to understand visually. The graph in Figure 1.9 is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

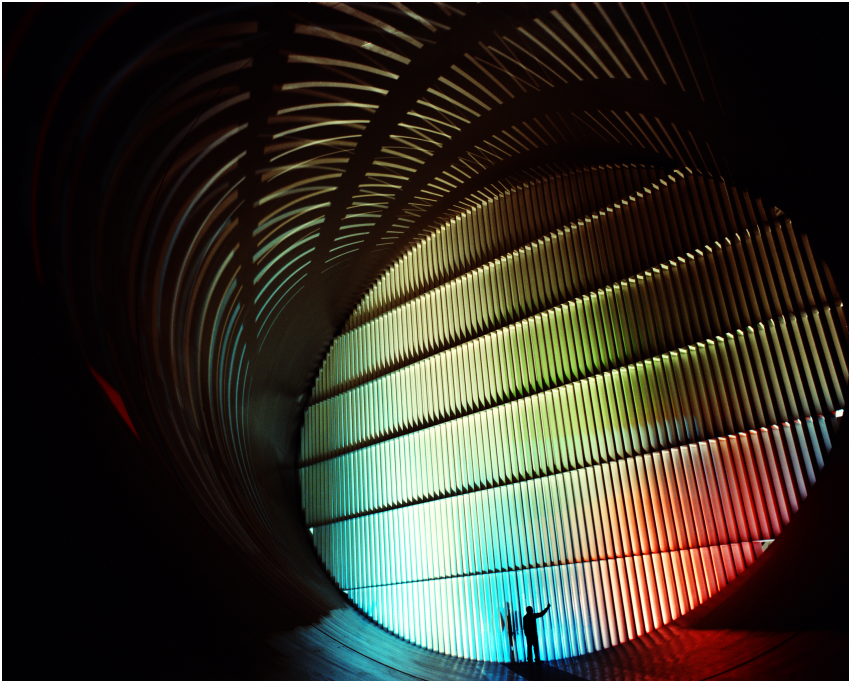


Figure 1.8

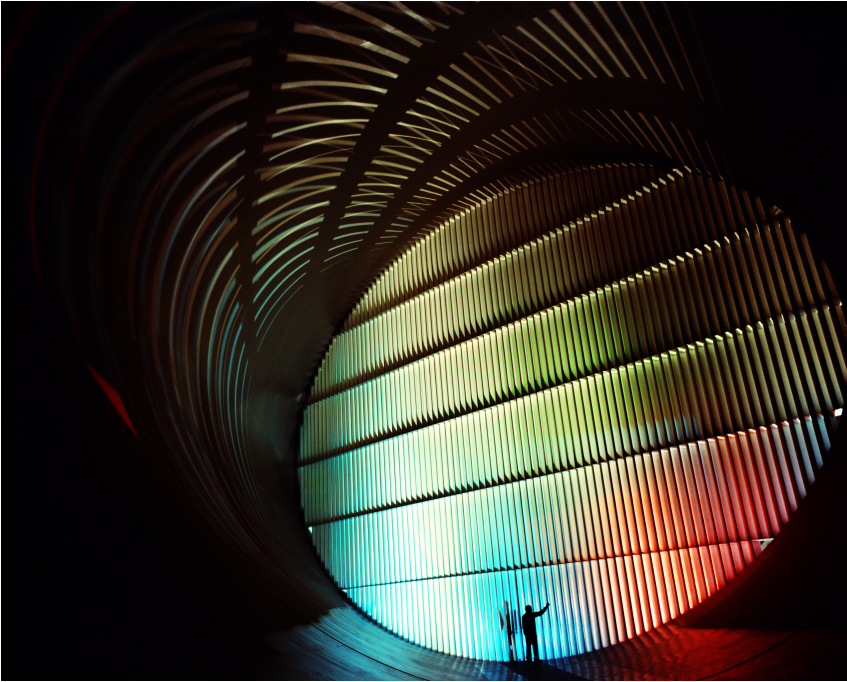


Figure 1.9

Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The chart in Figure 1.10(b) is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.10(a).

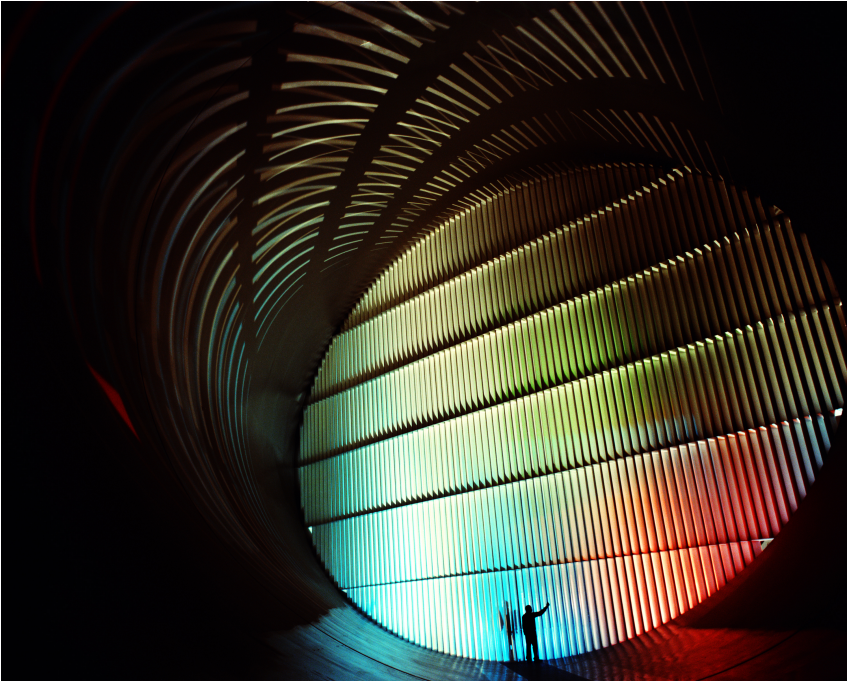


Figure 1.10

SAMPLING

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling.

Convenience sampling involves using results that are readily available. For

example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- **Problems with samples:** A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- **Self-selected samples:** Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- **Sample size issues:** Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- **Undue influence:** collecting data or asking questions in a way that influences the response
- **Non-response or refusal of subject to participate:** The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- **Causality:** A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- **Self-funded or self-interest studies:** A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- **Misleading use of data:** improperly displayed graphs, incomplete data, or lack of context
- **Confounding:** When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

EXAMPLE 1.11

Problem

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

1. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
2. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
3. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
4. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
5. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Solution

1. stratified; 2. systematic; 3. simple random; 4. cluster; 5. convenience

EXAMPLE 1.12

Problem

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution

1. stratified; 2. cluster; 3. stratified; 4. systematic; 5. simple random; 6. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

EXAMPLE 1.13

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term

calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

\$128; \$87; \$173; \$116; \$130; \$204; \$147; \$189; \$93; \$153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

\$50; \$40; \$36; \$15; \$50; \$100; \$40; \$53; \$22; \$22

It is unlikely that any student is in both samples.

Problem

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Problem

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

\$180; \$50; \$150; \$85; \$260; \$75; \$180; \$200; \$200; \$150

Problem

c. Is the sample biased?

Solution

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

TRY IT YOURSELF

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

VARIATION IN DATA

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

VARIATION IN SAMPLES

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This variability in samples cannot be stressed enough.

The following case study is a direct extract from Lookup London:

<https://lookup.london/history-scottish-widows/>

THE SCOTTISH WIDOWS FUND

In 1744 two Scottish ministers, Alexander Webster and Robert Wallace, decided to set up a life insurance fund. The idea being that it would provide for widows and children of dead clergymen.

They proposed that each minister would pay a section of his monthly salary into a savings fund, but how should they determine the amount?

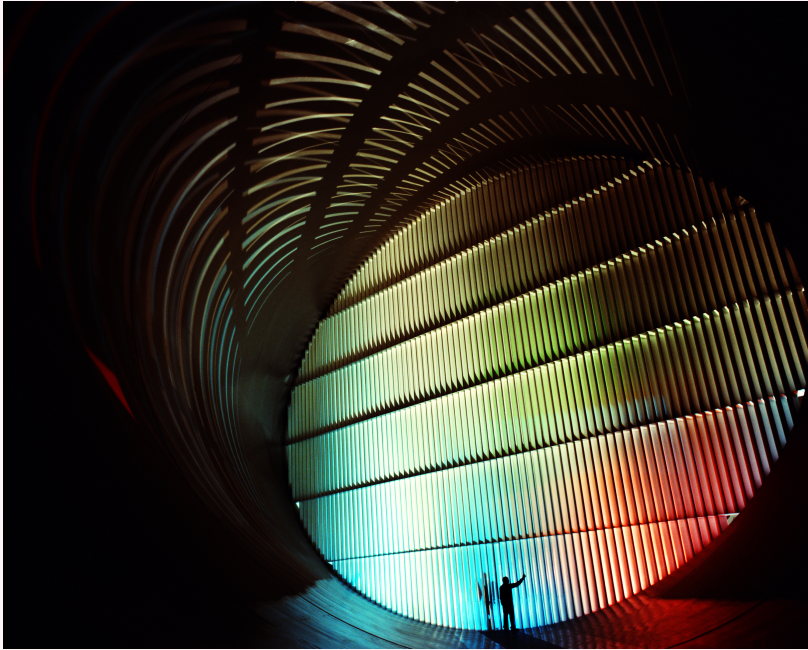
Being practical types they contacted a mathematics professor from the University of Edinburgh, Colin McClory. The three men then collected statistical data on death rates, numbers of children and years that widows outlived their husbands or remarried.

They used Jacob Bernoulli's 'Law of Large Numbers', which states that it's hard to predict (with certainty) a single event. But, it's possible to accurately predict the average outcome of many, similar events.

After their calculations were totted up. They determined that each clergyman needed to give at least £2, 12 shillings and twopence to guarantee a comfortable pension. They could also, if they wished, contribute as much as £6, 11 shillings and threepence.

They predicted that by 1765 the fund would be £58,348. In fact, the actual amount ended up being £58,347. Just £1 off the prediction!

FATHER OF STATISTICS: JACOB BERNOULLI



A Portrait of Jacob Bernoulli. Sourced from <https://mathshistory.st-andrews.ac.uk>

Jacob Bernoulli's father, Nicolaus Bernoulli (1623-1708) inherited the spice business in Basel that had been set up by his own father, first in Amsterdam and then in Basel. The family, of Belgium origin, were refugees fleeing from persecution by the Spanish rulers of the Netherlands. Philip, the King of Spain, had sent the Duke of Alba to the Netherlands in 1567 with a large army to punish those opposed to Spanish rule, to enforce adherence to Roman Catholicism, and to re-establish Philip's authority. Alba set up the Council of Troubles which was a court that condemned over 12000 people but most, like the Bernoulli family who were of the Protestant faith, fled the country.

Nicolaus Bernoulli was an important citizen of Basel, being a member of the town council and a magistrate. Jacob Bernoulli's mother also came from an important Basel family of bankers and local councillors. Jacob Bernoulli

was the brother of Johann Bernoulli and the uncle of Daniel Bernoulli. He was compelled to study philosophy and theology by his parents, which he greatly resented, and he graduated from the University of Basel with a master's degree in philosophy in 1671 and a licentiate in theology in 1676.

During the time that Jacob Bernoulli was taking his university degrees he was studying mathematics and astronomy against the wishes of his parents. It is worth remarking that this was a typical pattern for many of the Bernoulli family who made a study of mathematics despite pressure to make a career in other areas. However Jacob Bernoulli was the first to go down this road so for him it was rather different in that there was no tradition of mathematics in the family before Jacob Bernoulli. Later members of the family must have been much influenced by the tradition of studying mathematics and mathematical physics.

In 1676, after taking his theology degree, Bernoulli moved to Geneva where he worked as a tutor. He then travelled to France spending two years studying with the followers of Descartes who were led at this time by Malebranche. In 1681 Bernoulli travelled to the Netherlands where he met many mathematicians including Hudde. Continuing his studies with the leading mathematicians and scientists of Europe he went to England where, among others, he met Boyle and Hooke. At this time he was deeply interested in astronomy and produced a work giving an incorrect theory of comets. As a result of his travels, Bernoulli began a correspondence with many mathematicians which he carried on over many years.

Jacob Bernoulli returned to Switzerland and taught mechanics at the University in Basel from 1683, giving a series of important lectures on the mechanics of solids and liquids. Since his degree was in theology it would have been natural for him to turn to the Church, but although he was offered an appointment in the Church he turned it down. Bernoulli's real love was for mathematics and theoretical physics and it was in these topics that he taught and researched. During this period he studied the leading mathematical works of his time including Descartes' *Géométrie* and van Schooten's additional material in the Latin edition. Jacob Bernoulli also studied the work of Wallis and Barrow and through these he became

interested in infinitesimal geometry. Jacob began publishing in the journal *Acta Eruditorum* which was established in Leipzig in 1682.

In 1684 Jacob Bernoulli married Judith Stupanus. They were to have two children, a son who was given his grandfather's name of Nicolaus and a daughter. These children, unlike many members of the Bernoulli family, did not go on to become mathematicians or physicists.

One of the most significant events concerning the mathematical studies of Jacob Bernoulli occurred when his younger brother, Johann Bernoulli, began to work on mathematical topics. Johann was told by his father to study medicine but while he was studying that topic he asked his brother Jacob to teach him mathematics. Jacob Bernoulli was appointed professor of mathematics in Basel in 1687 and the two brothers began to study the calculus as presented by Leibniz in his 1684 paper on the differential calculus in *Nova Methodus pro Maximis et Minimis, itemque Tangentibus...* published in *Acta Eruditorum*. They also studied the publications of von Tschirnhaus. It must be understood that Leibniz's publications on the calculus were very obscure to mathematicians of that time and the Bernoullis were the first to try to understand and apply Leibniz's theories.

Although Jacob and Johann both worked on similar problems their relationship was soon to change from one of collaborators to one of rivals. Johann Bernoulli's boasts were the first cause of Jacob's attacks on him and Jacob wrote that Johann was his pupil whose only achievements were to repeat what his teacher had taught him. Of course this was a grossly unfair statement. Jacob continued to attack his brother in print in a disgraceful and unnecessary fashion, particularly after 1697. However he did not reserve public criticism for his brother. He was critical of the university authorities at Basel and again he was very public in making critical statements that, as one would expect, left him in a difficult situation at the university. Jacob probably felt that Johann was the more powerful mathematician of the two and, this hurt since Jacob's nature meant that he always had to feel that he was winning praise from all sides. Hofmann writes:

Sensitivity, irritability, a mutual passion for criticism, and an exaggerated need for recognition alienated the brothers, of whom Jacob had the slower but deeper intellect.

As suggested by this quote the brothers were equally at fault in their quarrel. Johann would have liked the chair of mathematics at Basel which Jacob held and he certainly resented having to move to Holland in 1695. This was another factor in the complete breakdown of relations in 1697.

Of course the dispute between the brothers over who could obtain the greatest recognition was a particularly stupid one in the sense that both made contributions to mathematics of the very greatest importance. Whether the rivalry spurred them on to greater things or whether they might have achieved more had they continued their initial collaboration, it is impossible to say. We shall now examine some of the major contributions made by Jacob Bernoulli at an important stage in the development of mathematics following Leibniz's work on the calculus.

Jacob Bernoulli's first important contributions were a pamphlet on the parallels of logic and algebra published in 1685, work on probability in 1685 and geometry in 1687. His geometry result gave a construction to divide any triangle into four equal parts with two perpendicular lines.

By 1689 he had published important work on infinite series and published his law of large numbers in probability theory. The interpretation of probability as relative-frequency says that if an experiment is repeated a large number of times then the relative frequency with which an event occurs equals the probability of the event. The law of large numbers is a mathematical interpretation of this result. Jacob Bernoulli published five treatises on infinite series between 1682 and 1704. The first two of these contained many results, such as fundamental result that $\sum(1/n)$ diverges, which Bernoulli believed were new but they had actually been proved by Mengoli 40 years earlier. Bernoulli could not find a closed form for $\sum(1/n^2)$ but he did show that it converged to a finite limit less than 2. Euler was the

first to find the sum of this series in 1737. Bernoulli also studied the exponential series which came out of examining compound interest.

In May 1690 in a paper published in *Acta Eruditorum*, Jacob Bernoulli showed that the problem of determining the isochrone is equivalent to solving a first-order nonlinear differential equation. The isochrone, or curve of constant descent, is the curve along which a particle will descend under gravity from any point to the bottom in exactly the same time, no matter what the starting point. It had been studied by Huygens in 1687 and Leibniz in 1689. After finding the differential equation, Bernoulli then solved it by what we now call separation of variables. Jacob Bernoulli's paper of 1690 is important for the history of calculus, since the term integral appears for the first time with its integration meaning. In 1696 Bernoulli solved the equation, now called "the Bernoulli equation", $y' = p(x)y + q(x)y^n$, and Hofmann describes this part of his work as:

... proof of Bernoulli's careful and critical work on older as well as on contemporary contributions to infinitesimal mathematics and of his perseverance and analytical ability in dealing with special pertinent problems, even those of a mechanical-dynamic nature.

Jacob Bernoulli also discovered a general method to determine evolutes of a curve as the envelope of its circles of curvature. He also investigated caustic curves and in particular he studied these associated curves of the parabola, the logarithmic spiral and epicycloids around 1692. The lemniscate of Bernoulli was first conceived by Jacob Bernoulli in 1694. In 1695 he investigated the drawbridge problem which seeks the curve required so that a weight sliding along the cable always keeps the drawbridge balanced.

The Bernoulli equation is:

$$y' = p(x)y + q(x)y^n$$

Jacob Bernoulli's most original work was *Ars Conjectandi* published in Basel in 1713, eight years after his death. The work was incomplete at the time of his death but it is still a work of the greatest significance in the

theory of probability. In the book Bernoulli reviewed work of others on probability, in particular work by van Schooten, Leibniz, and Prestet. The Bernoulli numbers appear in the book in a discussion of the exponential series. Many examples are given on how much one would expect to win playing various game of chance. There are interesting thoughts on what probability really is:

... probability as a measurable degree of certainty; necessity and chance; moral versus mathematical expectation; a priori an a posteriori probability; expectation of winning when players are divided according to dexterity; regard of all available arguments, their valuation, and their calculable evaluation; law of large numbers ...

Hofmann sums up Jacob Bernoulli's contributions as follows:

Bernoulli greatly advanced algebra, the infinitesimal calculus, the calculus of variations, mechanics, the theory of series, and the theory of probability. He was self-willed, obstinate, aggressive, vindictive, beset by feelings of inferiority, and yet firmly convinced of his own abilities. With these characteristics, he necessarily had to collide with his similarly disposed brother. He nevertheless exerted the most lasting influence on the latter.

Bernoulli was one of the most significant promoters of the formal methods of higher analysis. Astuteness and elegance are seldom found in his method of presentation and expression, but there is a maximum of integrity.

Jacob Bernoulli continued to hold the chair of mathematics at Basel until his death in 1705 when the chair was filled by his brother Johann. Jacob had always found the properties of the logarithmic spiral to be almost magical and he had requested that it be carved on his tombstone with the Latin inscription *Eadem Mutata Resurgo* meaning "Although changed, I will rise again the same".

This biography is a direct extract from https://mathshistory.st-andrews.ac.uk/Biographies/Bernoulli_Jacob/

Size of a Sample

The size of a sample (often called the number of observations, usually given the symbol n) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. Later we will find that even much smaller sample sizes will give very good results. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

1.3 Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

LEVELS OF MEASUREMENT

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative (categorical)**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60° . Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80°C is not four times as hot as 20°C (nor is 80°F four times as hot as 20°F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

FREQUENCY

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table 1.5 lists the different data values in ascending order and their frequencies.

Data value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Table 1.5 Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to Table 1.5, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Data value	Frequency	Relative frequency
2	3	2030 or 0.15
3	5	205 or 0.25
4	3	203 or 0.15
5	6	206 or 0.30
6	2	202 or 0.10
7	1	201 or 0.05

Table 1.6 Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of Table 1.6 is 2020, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table 1.7.

Heights (inches)	Frequency	Relative frequency	Cumulative relative frequency
2	3	203 or 0.15	0.15
3	5	205or 0.25	$0.15 + 0.25 = 0.40$
4	3	203 or 0.15	$0.40 + 0.15 = 0.55$
5	6	206 or 0.30	$0.55 + 0.30 = 0.85$
6	2	202 or 0.10	$0.85 + 0.10 = 0.95$
7	1	201 or 0.05	$0.95 + 0.05 = 1.00$

Table 1.7 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Table 1.8 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Heights (inches)	Frequency	Relative frequency	Cumulative relative frequency
59.95–61.95	5	$1005 = 0.05$	0.05
61.95–63.95	3	$1003 = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$10015 = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$1004 = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$10017 = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$10012 = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$1007 = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$1001 = 0.01$	$0.99 + 0.01 = 1.00$
Total = 100		Total = 1.00	

Table 1.8 Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches

- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

EXAMPLE 1.14

Problem

From Table 1.8, find the percentage of heights that are less than 65.95 inches.

Solution

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $23/100$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

TRY IT YOURSELF

Rainfall (inches)	Frequency	Relative frequency	Cumulative relative frequency
2.95–4.97	6	$506 = 0.12$	0.12
4.97–6.99	7	$507 = 0.14$	$0.12 + 0.14 = 0.26$
6.99–9.01	15	$5015 = 0.30$	$0.26 + 0.30 = 0.56$
9.01–11.03	8	$508 = 0.16$	$0.56 + 0.16 = 0.72$
11.03–13.05	9	$509 = 0.18$	$0.72 + 0.18 = 0.90$
13.05–15.07	5	$505 = 0.10$	$0.90 + 0.10 = 1.00$
Total = 50		Total = 1.00	

Table 1.9

From Table 1.9, find the percentage of rainfall that is less than 9.01 inches.

EXAMPLE 1.15**Problem**

From Table 1.8, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution

Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

TRY IT YOURSELF

From Table 1.9, find the percentage of rainfall that is between 6.99 and 13.05 inches.

EXAMPLE 1.16**Problem**

Use the heights of the 100 male semiprofessional soccer players in Table 1.8. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
2. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
3. The percentage of heights that are more than 65.95 inches is: ____.
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you count frequencies. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

EXAMPLE 1.17

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table 1.10 was produced:

Rainfall (inches)	Frequency	Relative frequency	Cumulative relative frequency
3	3	193	0.1579
4	1	191	0.2105
5	3	193	0.1579
7	2	192	0.2632
10	3	194	0.4737
12	2	192	0.7895
13	1	191	0.8421
15	1	191	0.8948
18	1	191	0.9474
20	1	191	1.0000

Table 1.10 Frequency of Commuting Distances**Problem**

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute three miles.
If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute five or seven miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution

1. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052,

SAMPLING AND DATA

0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.

3. 195

4. 197, 1912, 197

TRY IT YOURSELF

Table 1.9 represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

EXAMPLE 1.18

Table 1.11 contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total number of deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Table 1.11

Problem

Answer the following questions.

1. What is the frequency of deaths measured from 2006 through 2009?
2. What percentage of deaths occurred after 2009?
3. What is the relative frequency of deaths that occurred in 2003 or earlier?
4. What is the percentage of deaths that occurred in 2004?
5. What kind of data are the numbers of deaths?
6. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution

1. 97,118 (11.8%)
2. 41.6%
3. $67,092/823,356$ or 0.081 or 8.1 %
4. 27.8%
5. Quantitative discrete
6. Quantitative continuous

TRY IT YOURSELF

Table 1.12 contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total number of crashes	Year	Total number of crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Table 1.12

Answer the following questions.

1. What is the frequency of deaths measured from 2000 through 2004?
2. What percentage of deaths occurred after 2006?
3. What is the relative frequency of deaths that occurred in 2000 or before?
4. What is the percentage of deaths that occurred in 2011?
5. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

1.4 Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **independent variable** or **explanatory variable**. The affected variable

is called the **dependent variable** or **response variable**: stimulus, response. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Statistics is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, better scientific investigation. Whether in any given study this implies more or less mathematics is incidental.

George E. P. Box

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between

groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.

McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

EXAMPLE 1.19

Problem

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.
2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

Solution

1. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Key terms

Average also called mean or arithmetic mean; a number that describes the central tendency of the data

Blinding not telling participants which treatment a subject is receiving

Categorical Variable variables that take on values that are names or labels

Cluster Sampling a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Control Group a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Cumulative Relative Frequency The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable a random variable (RV) whose outcomes are counted

Double-blinding the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit any individual or object to be measured

Explanatory Variable the **independent variable** in an experiment; the value controlled by researchers

Frequency the number of times a value of the data occurs

Informed Consent Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board a committee tasked with oversight of research programs that involve human subjects

Lurking Variable a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Mathematical Models a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Nonsampling Error an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable variables that take on values that are indicated by numbers

Observational Study a study in which the independent variable is not manipulated by the researcher

Parameter a number that is used to represent a population characteristic and that generally cannot be determined easily

Placebo an inactive treatment that has no real effect on the explanatory variable

Population all individuals, objects, or measurements whose properties are being studied

Probability a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion the number of successes divided by the total number in the sample

Qualitative Data See Data.

Quantitative Data See Data.

Random Assignment the act of organizing experimental units into treatment groups using random methods

Random Sampling a method of selecting a sample that gives every member of the population an equal chance of being selected.

Relative Frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Representative Sample a subset of the population that has the same characteristics as the population

Response Variable the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Sample a subset of the population studied

Sampling Bias not all members of the population are equally likely to be selected

Sampling Error the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Statistic a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Statistical Models a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.

Stratified Sampling a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Survey a study in which data is collected as reported by individuals.

Systematic Sampling a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let $k = (\text{number of individuals in the population}) / (\text{number of individuals needed in the sample})$. Choose every k th individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments different values or components of the explanatory variable applied in an experiment

Variable a characteristic of interest for each person or object in a population

Chapter review

1.1 DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

1.2 DATA, SAMPLING, AND VARIATION IN DATA AND SAMPLING

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.3 LEVELS OF MEASUREMENT

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

Nominal scale level: data that cannot be ordered nor can it be used in calculations

Ordinal scale level: data that can be ordered; the differences cannot be measured

Interval scale level: data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.

Ratio scale level: data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.


When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

1.4 EXPERIMENTAL DESIGN AND ETHICS

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.” (Andrew Gelman, “Open Data and Open Methods,” *Ethics and Statistics*,

<http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf> (accessed May 1, 2013).) Ethical violations in statistics are not always easy to



spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Homework

Find a real life example from your own research or the Further reading section, and write a case study to outline each of the key terms used in this chapter, including the relevant statistics for each key term.

References

1.1 DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

The Data and Story Library,

<http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).

1.2 DATA, SAMPLING, AND VARIATION IN DATA AND SAMPLING

Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/default.asp> (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/methodology.asp> (accessed May 1, 2013).

Gallup-Healthways Well-Being Index.

<http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx> (accessed May 1, 2013).

Data from <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>

Dominic Lusinchi, “President’ Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?” *Social Science History* 36, no. 1: 23-54 (2012), <http://ssh.dukejournals.org/content/36/1/23.abstract> (accessed May 1, 2013).

“The Literary Digest Poll,” *Virtual Laboratories in Probability and Statistics* <http://www.math.uah.edu/stat/data/LiteraryDigest.html> (accessed May 1, 2013).

“Gallup Presidential Election Trial-Heat Trends, 1936–2008,” *Gallup Politics* <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4> (accessed May 1, 2013).

The Data and Story Library,

<http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html> (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011,

<http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus> (accessed May 1, 2013).

Data from San Jose Mercury News

Lookup London. History of the Scottish Widows fund. accessed 2 December 2022 <https://lookup.london/history-scottish-widows/>

St Andrews University. Jacob Bernoulli Biography. https://mathshistory.st-andrews.ac.uk/Biographies/Bernoulli_Jacob/

1.3 LEVELS OF MEASUREMENT

“State & County QuickFacts,” U.S. Census Bureau.

http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

“State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau.

<http://quickfacts.census.gov/qfd/index.html> (accessed May 1, 2013).

“Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, <http://www.nhc.noaa.gov/gifs/table5.gif> (accessed May 1, 2013).

“Levels of Measurement,”

http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, “Levels of Measurement,” about.com,

<http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm> (accessed May 1, 2013).

David Lane. “Levels of Measurement,” Connexions,

<http://cnx.org/content/m10809/latest/> (accessed May 1, 2013).

1.4 EXPERIMENTAL DESIGN AND ETHICS

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health,

<http://www.hsph.harvard.edu/nutritionsource/vitamin-e/> (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com,

<http://www.athleteinme.com/ArticleView.aspx?id=1053> (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,”

International Business Times, July 21, 2011. Also available online at

<http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443> (accessed May 1, 2013).

The Data and Story Library,

<http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html> (accessed May 1, 2013).

M.L. Jackson et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," *Accident Analysis and Prevention Journal*, Jan no. 50 (2013), <http://www.ncbi.nlm.nih.gov/pubmed/22721550> (accessed May 1, 2013).

"Earthquake Information by Year," U.S. Geological Survey.

<http://earthquake.usgs.gov/earthquakes/eqarchives/year/> (accessed May 1, 2013).

"Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. <http://www-fars.nhtsa.dot.gov/Main/index.aspx> (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

"America's Best Small Companies," <http://www.forbes.com/best-small-companies/list/> (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

"April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," [econoclass.com](http://www.econoclass.com),

<http://www.econoclass.com/misleadingstats.html> (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, *Connexions*, <http://cnx.org/content/m15555/latest/> (accessed May 1, 2013).

Further reading

“Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

“Distance Education.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

“NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. “Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income,” GALLUP® Economy, 2013. Available online at <http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx> (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. The American Freshman: National Norms Fall 2011. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics> (accessed May 15, 2013).

“Millennials: A Portrait of Generation Next,” PewResearchCenter. Available online at <http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf> (accessed May 15, 2013).

“Millennials: Confident. Connected. Open to Change.” Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at <http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-to-change/> (accessed May 15, 2013).

“Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan,” The European Union and ICON-Institute. Available online at http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills,” UNICEF Television. Video available online at <http://www.unicefusa.org/assets/video/afghan-female-literacy-centers.html> (accessed May 15, 2013).

“ATL Fact Sheet,” Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at <http://www.atl.com/about-atl/atl-factsheet/> (accessed February 6, 2019).

Center for Disease Control and Prevention. “Teen Drivers: Fact Sheet,” Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html (accessed May 15, 2013).

“Children and Childrearing,” Ministry of Health, Labour, and Welfare. Available online at <http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html> (accessed May 15, 2013).

“Eating Disorder Statistics,” South Carolina Department of Mental Health, 2006. Available online at <http://www.state.sc.us/dmh/anorexia/statistics.htm> (accessed May 15, 2013).

“Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day,” *theguardian*, 2013. Available online at <http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2> (accessed May 15, 2013).

“How Americans Use Text Messaging,” Pew Internet, 2013. Available online at <http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx> (accessed May 15, 2013).

Lenhart, Amanda. “Teens, Smartphones & Testing: Texting volume is up while the frequency of voice calling is down. About one in four teens say they own smartphones,” Pew Internet, 2012. Available online at http://www.pewinternet.org/~media/Files/Reports/2012/PIP_Teens_Smartphones_and (accessed May 15, 2013).

“One born every minute: the maternity unit where mothers are THREE to a bed,” *MailOnline*. Available online at <http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers-bed.html> (accessed May 15, 2013).

Vanderkam, Laura. “Stop Checking Your Email, Now.” *CNNMoney*, 2013. Available online at <http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/> (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).



2 Descriptive Statistics

2.1 Display Data

2.2 Measures of the Location of the Data

2.3 Measures of the Center of the Data

2.4 Sigma Notation and Calculating the Arithmetic Mean

2.5 Geometric Mean

2.6 Skewness and the Mean, Median, and Mode

2.7 Measures of the Spread of the Data

Appendix 1

Key terms used in the book

Average also called mean or arithmetic mean; a number that describes the central tendency of the data

Blinding not telling participants which treatment a subject is receiving

Categorical Variable variables that take on values that are names or labels

Cluster Sampling a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Control Group a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Cumulative Relative Frequency The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number).

Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable a random variable (RV) whose outcomes are counted

Double-blinding the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit any individual or object to be measured

Explanatory Variable the **independent variable** in an experiment; the value controlled by researchers

Frequency the number of times a value of the data occurs

Informed Consent Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board a committee tasked with oversight of research programs that involve human subjects

Lurking Variable a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Mathematical Models a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Nonsampling Error an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable variables that take on values that are indicated by numbers

Observational Study a study in which the independent variable is not manipulated by the researcher

Parameter a number that is used to represent a population characteristic and that generally cannot be determined easily

Placebo an inactive treatment that has no real effect on the explanatory variable

Population all individuals, objects, or measurements whose properties are being studied

Probability a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion the number of successes divided by the total number in the sample

Qualitative Data See Data.

Quantitative Data See Data.

Random Assignment the act of organizing experimental units into treatment groups using random methods

Random Sampling a method of selecting a sample that gives every member of the population an equal chance of being selected.

Relative Frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Representative Sample a subset of the population that has the same characteristics as the population

Response Variable the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Sample a subset of the population studied

Sampling Bias not all members of the population are equally likely to be selected

Sampling Error the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling a straightforward method for selecting a random sample; give each member of the population a number. Use a random

number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Statistic a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Statistical Models a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.

Stratified Sampling a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Survey a study in which data is collected as reported by individuals.

Systematic Sampling a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let $k = (\text{number of individuals in the population}) / (\text{number of individuals needed in the sample})$. Choose every k th individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments different values or components of the explanatory variable applied in an experiment

Variable a characteristic of interest for each person or object in a population

F Distribution

APPENDIX 1

Degrees of freedom in the numerator										
Degrees of freedom in the denominator	p	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.99
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	241.00
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.00
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6030.0
	.001	405284	500000	540379	562500	576405	585937	592873	598144	603000
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.83
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.49
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.36
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.80
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.93
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.01
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.93
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.68
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.50

APPENDIX 1

5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33

Table A1 *F* critical values

APPENDIX 1

Degrees of freedom in the numerator										
Degrees of freedom in the denominator	p	10	12	15	20	25	30	40	50	60
1	.100	60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.7
	.050	241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252
	.025	968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009
	.010	6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	631
	.001	605621	610668	615764	620908	624017	626099	628712	630285	631
2	.100	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47
	.050	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.4
	.025	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.4
	.010	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.4
	.001	999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999
3	.100	5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15
	.050	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57
	.025	14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.9
	.010	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.3
	.001	129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.
4	.100	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79
	.050	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69
	.025	8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36
	.010	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.6
	.001	48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.7

APPENDIX 1

5	.100	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
	.050	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
	.025	6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
	.010	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
	.001	26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
6	.100	2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
	.050	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
	.025	5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
	.010	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
	.001	18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
7	.100	2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
	.050	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
	.025	4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
	.010	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
	.001	14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72

Table A2 *F* critical values (continued)

Afterword

Who we are

OpenStax is part of Rice University, which is a 501(c)(3) nonprofit charitable corporation. As an educational initiative, it's our mission to improve educational access and learning for everyone. Through our partnerships with philanthropic foundations and our alliance with other educational resource companies, we're breaking down the most common barriers to learning. Because we believe that everyone should and can have access to knowledge.

Looking for more information about OpenStax? Visit our [FAQ page](#).

What we do

We publish high-quality, peer-reviewed, openly licensed college textbooks that are absolutely free online and low cost in print. **Seriously.** We've also developed a low-cost, research-based courseware that gives students the tools they need to complete their course the first time around. Check out our current library of textbooks and explore OpenStax Tutor.

Where we're going

Our textbooks are being used in 60% of college and universities in the U.S. and over 100 countries, but there's still room to grow. Alongside our efforts to expand our library, we have been taking steps toward improving student learning and advancing research in learning science. Access is just the beginning – we want to make educational tools better for a new generation of learners. Explore our [map](#) to see where we're headed.