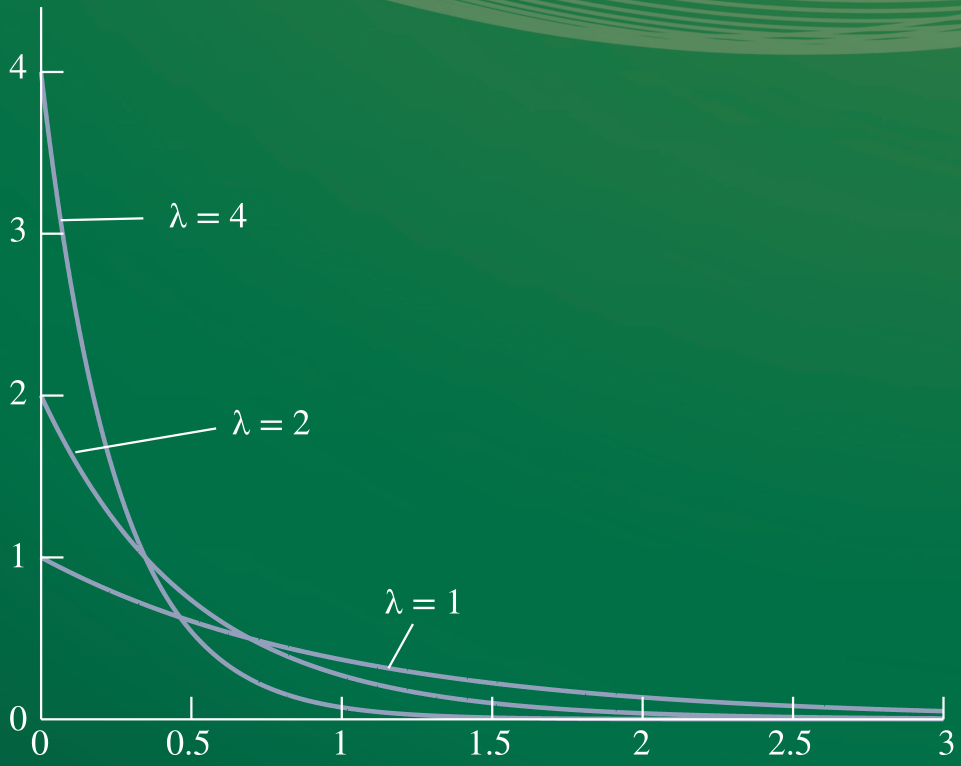
This International Student Edition is for use outside of the U.S.

Sixth Edition

STATISTICS FOR ENGINEERS AND SCIENTISTS



Mc Graw Hill

William Navidi

Statistics for Engineers and Scientists

Sixth Edition

William Navidi
Colorado School of Mines

Mc
Graw
Hill

STATISTICS FOR ENGINEERS AND SCIENTISTS

Published by McGraw Hill LLC, 1325 Avenue of the Americas, New York, NY 10019.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LCR 28 27 26 25 24 23

ISBN 978-1-266-11583-7
MHID 1-266-11583-8

Cover Image: *McGraw Hill*

mheducation.com/highered

To Catherine, Sarah, and Thomas

ABOUT THE AUTHOR

William Navidi is Professor Emeritus of Applied Mathematics and Statistics at the Colorado School of Mines. He received his B.A. degree in mathematics from New College, his M.A. in mathematics from Michigan State University, and his Ph.D. in statistics from the University of California at Berkeley. Professor Navidi has authored more than 80 research papers both in statistical theory and in a wide variety of applications including computer networks, epidemiology, molecular biology, chemical engineering, and geophysics.

BRIEF CONTENTS

501-1111

CONTENTS

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Chapter 10

PREFACE

MOTIVATION

The idea for this book grew out of discussions between the statistics faculty and the engineering faculty at the Colorado School of Mines regarding our introductory statistics course for engineers. Our engineering faculty felt that the students needed substantial coverage of propagation of error, as well as more emphasis on model-fitting skills. The statistics faculty believed that students needed to become more aware of some important practical statistical issues such as the checking of model assumptions and the use of simulation.

My view is that an introductory statistics text for students in engineering and science should offer all these topics in some depth. In addition, it should be flexible enough to allow for a variety of choices to be made regarding coverage, because there are many different ways to design a successful introductory statistics course. Finally, it should provide examples that present important ideas in realistic settings. Accordingly, the book has the following features:

- The book is flexible in its presentation of probability, allowing instructors wide latitude in choosing the depth and extent of their coverage of this topic.
- The book contains many examples that feature real, contemporary data sets, both to motivate students and to show connections to industry and scientific research.
- The book contains many examples of computer output and exercises suitable for solving with computer software.
- The book provides extensive coverage of propagation of error.
- The book presents a solid introduction to simulation methods and the bootstrap, including applications to verifying normality assumptions, computing probabilities, estimating bias, computing confidence intervals, and testing hypotheses.
- The book provides more extensive coverage of linear model diagnostic procedures than is found in most introductory texts. This includes material on examination of residual plots, transformations of variables, and principles of variable selection in multivariate models.
- The book covers the standard introductory topics, including descriptive statistics, probability, confidence intervals, hypothesis tests, linear regression, factorial experiments, and statistical quality control.

MATHEMATICAL LEVEL

Most of the book will be mathematically accessible to those whose background includes one semester of calculus. The exceptions are multivariate propagation of error, which requires partial derivatives, and joint probability distributions, which require multiple integration. These topics may be skipped on first reading, if desired.

COMPUTER USE

Over the past 50 years, the development of fast and cheap computing has revolutionized statistical practice; indeed, this is one of the main reasons that statistical methods have been penetrating ever more deeply into scientific work. Scientists and engineers today must not only be adept with computer software packages, they must also have the skill to draw conclusions from computer output and to state those conclusions in words. Accordingly, the book contains exercises and examples that involve interpreting, as well as generating, computer output, especially in the chapters on linear models and factorial experiments. Some of these exercises involve the analysis of large data sets, which are available in the Instructor resources. Many statistical software packages are available for instructors who wish to integrate their use into their courses, and this book can be used effectively with any of these packages.

The modern availability of computers and statistical software has produced an important educational benefit as well, by making simulation methods accessible to introductory students. Simulation makes the fundamental principles of statistics come alive. The material on simulation presented here is designed to reinforce some basic statistical ideas, and to introduce students to some of the uses of this powerful tool.

CONTENT

Chapter 1 covers sampling and descriptive statistics. The reason that statistical methods work is that samples, when properly drawn, are likely to resemble their populations. Therefore Chapter 1 begins by describing some ways to draw valid samples. The second part of the chapter discusses descriptive statistics.

Chapter 2 is about probability. There is a wide divergence in preferences of instructors regarding how much and how deeply to cover this subject. Accordingly, I have tried to make this chapter as flexible as possible. The major results are derived from axioms, with proofs given for most of them. This should enable instructors to take a mathematically rigorous approach. On the other hand, I have attempted to illustrate each result with an example or two, in a scientific context where possible, that is designed to present the intuition behind the result. Instructors who prefer a more informal approach may therefore focus on the examples rather than the proofs.

Chapter 3 covers propagation of error, which is sometimes called "error analysis" or, by statisticians, "the delta method." The coverage is more extensive than in most texts, but because the topic is so important to many engineers I thought it was worthwhile. The presentation is designed to enable instructors to adjust the amount of coverage to fit the needs of the course. In particular, Sections 3.2 through 3.4 can be omitted without loss of continuity.

Chapter 4 presents many of the probability distribution functions commonly used in practice. Point estimation, probability plots, and the Central Limit Theorem are also covered. The final section introduces simulation methods to assess normality assumptions, compute probabilities, and estimate bias.

Chapters 5 and 6 cover confidence intervals and hypothesis testing, respectively. The *P*-value approach to hypothesis testing is emphasized, but fixed-level testing and power calculations are also covered. The multiple testing problem is covered in some depth. Simulation methods to compute confidence intervals and to test hypotheses are introduced as well.

Chapter 7 covers correlation and simple linear regression. I have worked hard to emphasize that linear models are appropriate only when the relationship between the variables is linear. This point is all the more important since it is often overlooked in practice by engineers and scientists (not to mention statisticians). It is not hard to find in the scientific literature straight-line fits and correlation coefficient summaries for plots that show obvious curvature or for which the slope of the line is determined by a few influential points. Therefore this chapter includes a lengthy section on checking model assumptions and transforming variables.

Chapter 8 covers multiple regression. Model selection methods are given particular emphasis, because choosing the variables to include in a model is an essential step in many real-life analyses. The topic of confounding is given careful treatment as well.

Chapter 9 discusses some commonly used experimental designs and the methods by which their data are analyzed. One-way and two-way analysis of variance methods, along with randomized complete block designs and 2^p factorial designs, are covered fairly extensively.

Chapter 10 presents the topic of statistical quality control, discussing control charts, CUSUM charts, and process capability; and concluding with a brief discussion of six-sigma quality.

NEW FOR THIS EDITION

The sixth edition of this book is intended to extend the strengths of the fifth. Some of the changes are:

- In keeping with modern practice, the Student's *t* distribution, rather than the normal distribution, is used for large-sample confidence intervals and hypothesis tests where the variance is unknown.
- New exercises involving larger data sets have been added to some sections of the book. The data sets are available in the Instructor Resources in Connect.
- Examples of computer output from R have been added, to reflect the increasing popularity of this package.
- Additional material on the bootstrap has been added to Section 4.12.
- The exposition has been improved in a number of places.

RECOMMENDED COVERAGE

The book contains enough material for a year-long course. For a one-semester course, there are a number of options. In the three-hour course I taught at the Colorado School of Mines, we covered all of the first four chapters, except for joint distributions, the more theoretical aspects of point estimation, and the exponential, gamma, and Weibull distributions. We then covered the material on confidence intervals and hypothesis testing in Chapters 5 and 6, going quickly over the two-sample methods and power calculations and omitting distribution-free methods and the chi-square and F tests. We finished by covering as much of the material on correlation and simple linear regression in Chapter 7 as time would allow.

A course with a somewhat different emphasis can be fashioned by including more material on probability, spending more time on two-sample methods and power, and reducing coverage of propagation of error, simulation, or regression. Many other options are available; for example, one may choose to include material on factorial experiments in place of some of the preceding topics.

INSTRUCTOR RESOURCES

The following resources are available on Connect, McGraw Hill's online assignment and assessment platform. You can sign in at connect.mheducation.com. If you do not have an account, please contact your local McGraw Hill Learning Technology Representative. You can find their information on the Higher Ed support page at mhhe.com.

- Solutions Manual
- Data Sets
- PowerPoint Lecture Notes
- Suggested Syllabi
- Image Library

ACKNOWLEDGMENTS

I am indebted to many people for contributions at every stage of development. I received valuable suggestions from my colleagues Barbara Moskal, Gus Greivel, Ashlyn Munson, and Melissa Laeser at the Colorado School of Mines. Sarita Yadav of Aptara Corporation managed the production process smoothly and effectively. I am particularly grateful to Jack Miller of the University of Michigan, who has corrected many errors and made many valuable suggestions for improvement.

The staff at McGraw Hill has been extremely capable and supportive. In particular, I would like to express my thanks to Product Developer, Theresa Collins, Content Project Manager, Jeni McAtee, and Senior Portfolio Manager, Beth Bettcher, for their patience and guidance in the preparation of this edition.

William Navidi

ADDITIONAL RESOURCES

Proctorio
Remote Proctoring & Browser-Locking Capabilities

Mc Graw Hill connect® + proctorio

Remote proctoring and browser-locking capabilities, hosted by Proctorio within Connect, provide control of the assessment environment by enabling security options and verifying the identity of the student.

Seamlessly integrated within Connect, these services allow instructors to control the assessment experience by verifying identification, restricting browser activity, and monitoring student actions.

Instant and detailed reporting gives instructors an at-a-glance view of potential academic integrity concerns, thereby avoiding personal bias and supporting evidence-based claims.

ReadAnywhere®

Read or study when it's convenient for you with McGraw Hill's free ReadAnywhere® app. Available for iOS or Android smartphones or tablets, ReadAnywhere gives users access to McGraw Hill tools including the eBook and SmartBook® 2.0 or Adaptive Learning Assignments in Connect. Take notes, highlight, and complete assignments offline—all of your work will sync when you open the app with Wi-Fi access. Log in with your McGraw Hill Connect username and password to start learning—anytime, anywhere!

Writing Assignment

Available within Connect and Connect Master, the Writing Assignment tool delivers a learning experience to help students improve their written communication skills and conceptual understanding. As an instructor, you can assign, monitor, grade, and provide feedback on writing more efficiently and effectively.

Reflecting the Diverse World Around Us

McGraw Hill believes in unlocking the potential of every learner at every stage of life. To accomplish that, we are dedicated to creating products that reflect, and are accessible to, all the diverse, global customers we serve. Within McGraw Hill, we foster a culture of belonging, and we work with partners who share our commitment to equity, inclusion, and diversity in all forms. In McGraw Hill Higher Education, this includes, but is not limited to, the following:

- Refreshing and implementing inclusive content guidelines around topics including generalizations and stereotypes, gender, abilities/disabilities, race/ethnicity, sexual orientation, diversity of names, and age.

- Enhancing best practices in assessment creation to eliminate cultural, cognitive, and affective bias.
- Maintaining and continually updating a robust photo library of diverse images that reflect our student populations.
- Including more diverse voices in the development and review of our content.
- Strengthening art guidelines to improve accessibility by ensuring meaningful text and images are distinguishable and perceivable by users with limited color vision and moderately low vision.

Key Features

Content Overview

This book allows flexible coverage because there are many ways to design a successful introductory statistics course.

- **Flexible coverage of probability** addresses the needs of different courses. Allowing for a mathematically rigorous approach, the major results are derived from axioms, with proofs given for most of them. On the other hand, each result is illustrated with an example or two to promote intuitive understanding. Instructors who prefer a more informal approach may therefore focus on the examples rather than the proofs and skip the optional sections.

- **Extensive coverage of propagation of error,** sometimes called "error analysis" or "the delta method," is provided in a separate chapter. The coverage is more thorough than in most texts. The format is flexible so that the amount of coverage can be tailored to the needs of the course.

- **A solid introduction to simulation methods and the bootstrap is** presented in the final sections of Chapters 4, 5, and 6.

- **Extensive coverage of linear model diagnostic procedures** in Chapter 7 includes a lengthy section on checking model assumptions and transforming variables. The chapter emphasizes that linear models are appropriate only when the relationship between the variables is linear. This point is all the more important since it is often overlooked in practice by engineers and scientists (not to mention statisticians).

Real-World Data Sets

With a fresh approach to the subject, the author uses contemporary real-world data sets to motivate students and show a direct connection to industry and research.

Computer Output

The book contains exercises and examples that involve interpreting, as well as generating, computer output.

McGraw Hill connect®

Instructors

Student Success Starts with You

Tools to enhance your unique voice

Want to build your own course? No problem. Prefer to use an OLC-aligned, prebuilt course? Easy. Want to make changes throughout the semester? Sure. And you'll save time with Connect's auto-grading, too.

65%
Less Time Grading

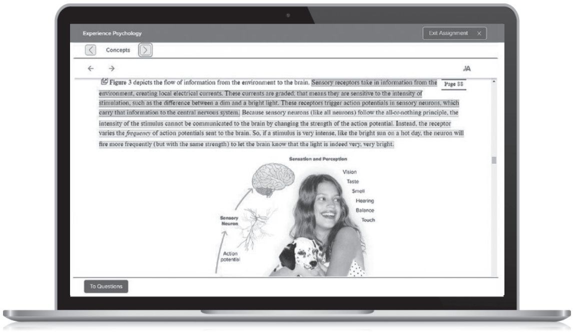


Figure 3 depicts the flow of information from the environment to the brain. Sensory receptors take in information from the environment, creating local electrical currents. These currents are graded; that means they are sensitive to the intensity of stimulation, such as the difference between a dim and a bright light. These receptors trigger action potentials in sensory neurons, which carry that information to the central nervous system. Because sensory neurons (like all neurons) follow the all-or-nothing principle, the intensity of the stimulus cannot be communicated to the brain by changing the strength of the action potential. Instead, the receptor varies the *frequency* of action potentials sent to the brain. So, if a stimulus is very intense, like the bright sun on a hot day, the neuron will fire more frequently (but with the same strength) to let the brain know that the light is indeed very, very bright.

Laptop: Getty Images; Woman/dog: George Doyle/Getty Images

A unique path for each student

In Connect, instructors can assign an adaptive reading experience with SmartBook® 2.0. Rooted in advanced learning science principles, SmartBook 2.0 delivers each student a personalized experience, focusing students on their learning gaps, ensuring that the time they spend studying is time well-spent.
mheducation.com/highered/connect/smartbook

Affordable solutions, added value

Make technology work for you with LMS integration for single sign-on access, mobile access to the digital textbook, and reports to quickly show you how each of your students is doing. And with our Inclusive Access program, you can provide all these tools at the lowest available market price to your students. Ask your McGraw Hill representative for more information.

Solutions for your challenges

A product isn't a solution. Real solutions are affordable, reliable, and come with training and ongoing support when you need it and how you want it. Visit **supportateverystep.com** for videos and resources both you and your students can use throughout the term.

SUPPORT AT every step

Students

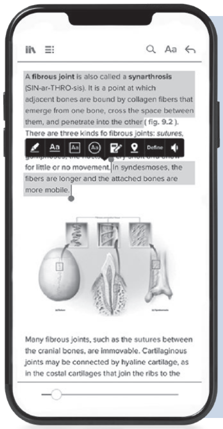
Get Learning that Fits You

Effective tools for efficient studying

Connect is designed to help you be more productive with simple, flexible, intuitive tools that maximize your study time and meet your individual learning needs. Get learning that works for you with Connect.

Study anytime, anywhere

Download the free ReadAnywhere® app and access your online eBook, SmartBook® 2.0, or Adaptive Learning Assignments when it's convenient, even if you're offline. And since the app automatically syncs with your Connect account, all of your work is available every time you open it. Find out more at **mheducation.com/readanywhere**



A **fibrous joint** is also called a **synarthrosis** (SIN-ar-THRO-sis). It is a point at which adjacent bones are bound by collagen fibers that emerge from one bone, cross the space between them, and penetrate into the other (fig. 9.2). There are three kinds fo fibrous joints: *sutures,*

for little or no movement. In syndesmoses, the fibers are longer and the attached bones are more mobile.

Many fibrous joints, such as the sutures between the cranial bones, are immovable. Cartilaginous joints may be connected by hyaline cartilage, as in the costal cartilages that join the ribs to the

iPhone: Getty Images

> ***"I really liked this app—it made it easy to study when you don't have your text-book in front of you."***
>
> - Jordan Cunningham,
> Eastern Washington University

Everything you need in one place

Your Connect course has everything you need—whether reading your digital eBook or completing assignments for class, Connect makes it easy to get your work done.

Learning for everyone

McGraw Hill works directly with Accessibility Services Departments and faculty to meet the learning needs of all students. Please contact your Accessibility Services Office and ask them to email accessibility@mheducation.com, or visit **mheducation.com/about/accessibility** for more information.

Chapter 1

Sampling and Descriptive Statistics

Introduction

The collection and analysis of data are fundamental to science and engineering. Scientists discover the principles that govern the physical world, and engineers learn how to design important new products and processes, by analyzing data collected in scientific experiments. A major difficulty with scientific data is that they are subject to random variation, or uncertainty. That is, when scientific measurements are repeated, they come out somewhat differently each time. This poses a problem: How can one draw conclusions from the results of an experiment when those results could have come out differently? To address this question, a knowledge of statistics is essential. Statistics is the field of study concerned with the collection, analysis, and interpretation of uncertain data. The methods of statistics allow scientists and engineers to design valid experiments and to draw reliable conclusions from the data they produce.

Although our emphasis in this book is on the applications of statistics to science and engineering, it is worth mentioning that the analysis and interpretation of data are playing an ever-increasing role in all aspects of modern life. For better or worse, huge amounts of data are collected about our opinions and our lifestyles, for purposes ranging from the creation of more effective marketing campaigns to the development of policies designed to improve our way of life. On almost any given day, articles are published that purport to explain social or economic trends through the analysis of data. A basic knowledge of statistics is therefore necessary not only to be an effective scientist or engineer, but also to be a well-informed member of society.

The Basic Idea

The basic idea behind all statistical methods of data analysis is to make inferences about a population by studying a relatively small sample chosen from it. As an illustration, consider a machine that makes steel rods for use in optical storage devices. The specification for the diameter of the rods is 0.45 ± 0.02 cm. During the last hour, the machine has made 1000 rods. A team of quality engineers want to know approximately how many of these rods meet the specification. They do not have time to measure all 1000 rods. So they draw a random sample of 50 rods, measure them, and find that 46 of them (92%) meet the diameter specification. Now, it is unlikely that the sample of 50 rods represents the population of 1000 perfectly. The proportion of good rods in the population is likely to differ somewhat from the sample proportion of 92%. What the engineers need to know is just how large that difference is likely to be. For example, is it plausible that the population percentage could be as high as 95%? 98%? As low as 90%? 85%?

Here are some specific questions that might need to be answered on the basis of these sample data:

1. It is necessary to compute a rough estimate of the likely size of the difference between the sample proportion and the population proportion. How large is a typical difference for this kind of sample?
2. The percentage of acceptable rods manufactured in the last hour will be recorded in a logbook. Since 92% of the sample rods were good, the percentage of acceptable rods in the population will be expressed as an interval of the form 92% ± x%, where x is a number calculated to provide reasonable certainty that the true population percentage is in the interval. How should x be calculated?
3. The engineers want to be fairly certain that the percentage of good rods is at least 90%; otherwise they will shut down the process for recalibration. How certain can they be that at least 90% of the 1000 rods are good?

Much of this book is devoted to addressing questions like these. The first of these questions requires the computation of a **standard deviation**, which we will discuss in Chapters 2 and 4. The second question requires the construction of a **confidence interval**, which we will learn about in Chapter 5. The third calls for a **hypothesis test**, which we will study in Chapter 6.

The remaining chapters in the book cover other important topics. For example, the engineers in our example may want to know how the amount of carbon in the steel rods is related to their tensile strength. Issues like this can be addressed with the methods of **correlation** and **regression**, which are covered in Chapters 7 and 8. It may also be important to determine how to adjust the manufacturing process with regard to several factors, in order to produce optimal results. This requires the design of **factorial experiments**, which are discussed in Chapter 9. Finally, the engineers will need to develop a plan for monitoring the quality of the product manufactured by the process. Chapter 10 covers the topic of **statistical quality control**, in which statistical methods are used to maintain quality in an industrial setting.

The topics listed here concern methods of drawing conclusions from data. These methods form the field of **inferential statistics**. Before we discuss these topics, we must first learn more about methods of collecting data and of summarizing clearly the basic information they contain. These are the topics of **sampling** and **descriptive statistics**, and they are covered in the rest of this chapter.

1.1 Sampling

As mentioned, statistical methods are based on the idea of analyzing a **sample** drawn from a **population**. For this idea to work, the sample must be chosen in an appropriate way. For example, let us say that we wished to study the heights of students at the Colorado School of Mines by measuring a sample of 100 students. How should we choose the 100 students to measure? Some methods are obviously bad. For example, choosing the students from the rosters of the football and basketball teams would undoubtedly result in a sample that would fail to represent the height distribution of the population of students. You might think that it would be reasonable to use some conveniently obtained sample, for example, all students living in a certain dorm or all students enrolled in engineering statistics. After all, there is no reason to think that the heights of these students would tend to differ from the heights of students in general. Samples like this are not ideal, however, because they can turn out to be misleading in ways that are not anticipated. The best sampling methods involve **random sampling**. There are many different random sampling methods, the most basic of which is **simple random sampling**.

To understand the nature of a simple random sample, think of a lottery. Imagine that 10,000 lottery tickets have been sold and that 5 winners are to be chosen. What is the fairest way to choose the winners? The fairest way is to put the 10,000 tickets in a drum, mix them thoroughly, and then reach in and one by one draw 5 tickets out. These 5 winning tickets are a simple random sample from the population of 10,000 lottery tickets. Each ticket is equally likely to be one of the 5 tickets drawn. More importantly, each collection of 5 tickets that can be formed from the 10,000 is equally likely to be the group of 5 that is drawn. It is this idea that forms the basis for the definition of a simple random sample.

Summary

- A **population** is the entire collection of objects or outcomes about which information is sought.
- A **sample** is a subset of a population, containing the objects or outcomes that are actually observed.
- A **simple random sample** of size *n* is a sample chosen by a method in which each collection of *n* population items is equally likely to make up the sample, just as in a lottery.

Since a simple random sample is analogous to a lottery, it can often be drawn by the same method now used in many lotteries: with a computer random number generator. Suppose there are *N* items in the population. One assigns to each item in the population an integer between 1 and *N*. Then one generates a list of random integers between 1 and *N* and chooses the corresponding population items to make up the simple random sample.

Example 1.1

A group of physical education professors want to study the physical fitness levels of students at their university. There are 20,000 students enrolled at the university, and they want to draw a sample of size 100 to take a physical fitness test. They obtain a list of all 20,000 students, numbered from 1 to 20,000. They use a computer random number generator to generate 100 random integers between 1 and 20,000 and then invite the 100 students corresponding to those numbers to participate in the study. Is this a simple random sample?

Solution

Yes, this is a simple random sample. Note that it is analogous to a lottery in which each student has a ticket and 100 tickets are drawn.

Example 1.2

A quality engineer wants to inspect rolls of wallpaper in order to obtain information on the rate at which flaws in the printing are occurring. She decides to draw a sample of 50 rolls of wallpaper from a day's production. Each hour for 5 hours, she takes the 10 most recently produced rolls and counts the number of flaws on each. Is this a simple random sample?

Solution

No. Not every subset of 50 rolls of wallpaper is equally likely to make up the sample. To construct a simple random sample, the engineer would need to assign a number to each roll produced during the day and then generate random numbers to determine which rolls make up the sample.

In some cases, it is difficult or impossible to draw a sample in a truly random way. In these cases, the best one can do is to sample items by some convenient method. For example, imagine that a construction engineer has just received a shipment of 1000 concrete blocks, each weighing approximately 50 pounds. The blocks have been delivered in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks. To draw a simple random sample would require removing blocks from the center and bottom of the pile, which might be quite difficult. For this reason, the engineer might construct a sample simply by taking 10 blocks off the top of the pile. A sample like this is called a **sample of convenience**.

Definition

A **sample of convenience** is a sample that is obtained in some convenient way, and not drawn by a well-defined random method.

The big problem with samples of convenience is that they may differ systematically in some way from the population. For this reason samples of convenience should not be used, except in situations where it is not feasible to draw a random sample. When it is necessary to take a sample of convenience, it is important to think carefully about all the ways in which the sample might differ systematically from the population. If it is reasonable to believe that no important systematic difference exists, then it may be acceptable to treat the sample of convenience as if it were a simple random sample. With regard to the concrete blocks, if the engineer is confident that the blocks on the top of the pile do not differ systematically in any important way from the rest, then the sample of convenience may be treated as a simple random sample. If, however, it is possible that blocks in different parts of the pile may have been made from different batches of mix or may have different curing times or temperatures, a sample of convenience could give misleading results.

Some people think that a simple random sample is guaranteed to reflect its population perfectly. This is not true. Simple random samples always differ from their populations in some ways, and occasionally may be substantially different. Two different samples from the same population will differ from each other as well. This phenomenon is known as **sampling variation**. Sampling variation is one of the reasons that scientific experiments produce somewhat different results when repeated, even when the conditions appear to be identical.

Example 1.3

A quality inspector draws a simple random sample of 40 bolts from a large shipment, measures the length of each, and finds that 34 of them, or 85%, meet a length specification. The inspector concludes that exactly 85% of the bolts in the shipment meet the specification. The inspector's supervisor concludes that the proportion of good bolts is likely to be close to, but not exactly equal to, 85%. Which conclusion is appropriate?

Solution

Because of sampling variation, simple random samples don't reflect the population perfectly. They are often fairly close, however. It is therefore appropriate to infer that the proportion of good bolts in the lot is likely to be close to the sample proportion, which is 85%. It is not likely that the population proportion is equal to 85%, however.

Example 1.4

Continuing Example 1.3, another inspector repeats the study with a different simple random sample of 40 bolts, and finds that 36 of them, or 90%, are good. The first inspector claims that something is wrong, since the previous results showed that 85%, not 90%, of bolts are good. Is this correct?

Solution

No, this is not correct. This is sampling variation at work. Two different samples from the same population will differ from each other and from the population.

Since simple random samples don't reflect their populations perfectly, why is it important that sampling be done at random? The benefit of a simple random sample is that there is no systematic mechanism tending to make the sample unrepresentative.

The differences between the sample and its population are due entirely to random variation. Since the mathematical theory of random variation is well understood, we can use mathematical models to study the relationship between simple random samples and their populations. For a sample not chosen at random, there is generally no theory available to describe the mechanisms that caused the sample to differ from its population. Therefore, nonrandom samples are often difficult to analyze reliably.

In Examples 1.1 to 1.4, the populations consisted of actual physical objects—the students at a university, the concrete blocks in a pile, the bolts in a shipment. Such populations are called **tangible populations**. Tangible populations are always finite. After an item is sampled, the population size decreases by 1. In principle, one could in some cases return the sampled item to the population, with a chance to sample it again, but this is rarely done in practice.

Engineering data are often produced by measurements made in the course of a scientific experiment, rather than by sampling from a tangible population. To take a simple example, imagine that an engineer measures the length of a rod five times, being as careful as possible to take the measurements under identical conditions. No matter how carefully the measurements are made, they will differ somewhat from one another, because of variation in the measurement process that cannot be controlled or predicted. It turns out that it is often appropriate to consider data like these to be a simple random sample from a population. The population, in these cases, consists of all the values that might possibly have been observed. Such a population is called a **conceptual population**, since it does not consist of actual objects.

> A simple random sample may consist of values obtained from a process under identical experimental conditions. In this case, the sample comes from a population that consists of all the values that might possibly have been observed. Such a population is called a **conceptual population**.

Example 1.5 involves a conceptual population.

Example 1.5

A geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a slightly different reading. Under what conditions can these readings be thought of as a simple random sample? What is the population?

Solution

If the physical characteristics of the scale remain the same for each weighing, so that the measurements are made under identical conditions, then the readings may be considered to be a simple random sample. The population is conceptual. It consists of all the readings that the scale could in principle produce.

Note that in Example 1.5, it is the physical characteristics of the measurement process that determine whether the data are a simple random sample. In general, when deciding whether a set of data may be considered to be a simple random sample, it is necessary to have some understanding of the process that generated the data. Statistical methods can sometimes help, especially when the sample is large, but knowledge of the mechanism that produced the data is more important.

Example 1.6

A new chemical process has been designed that is supposed to produce a higher yield of a certain chemical than does an old process. To study the yield of this process, we run it 50 times and record the 50 yields. Under what conditions might it be reasonable to treat this as a simple random sample? Describe some conditions under which it might not be appropriate to treat this as a simple random sample.

Solution

To answer this, we must first specify the population. The population is conceptual and consists of the set of all yields that will result from this process as many times as it will ever be run. What we have done is to sample the first 50 yields of the process. *If, and only if,* we are confident that the first 50 yields are generated under identical conditions, and that they do not differ in any systematic way from the yields of future runs, then we may treat them as a simple random sample.

Be cautious, however. There are many conditions under which the 50 yields could fail to be a simple random sample. For example, with chemical processes, it is sometimes the case that runs with higher yields tend to be followed by runs with lower yields, and vice versa. Sometimes yields tend to increase over time, as process engineers learn from experience how to run the process more efficiently. In these cases, the yields are not being generated under identical conditions and would not be a simple random sample.

Example 1.6 shows once again that a good knowledge of the nature of the process under consideration is important in deciding whether data may be considered to be a simple random sample. Statistical methods can sometimes be used to show that a given data set is *not* a simple random sample. For example, sometimes experimental conditions gradually change over time. A simple but effective method to detect this condition is to plot the observations in the order they were taken. A simple random sample should show no obvious pattern or trend.

Figure 1.1 (page 8) presents plots of three samples in the order they were taken. The plot in Figure 1.1a shows an oscillatory pattern. The plot in Figure 1.1b shows an increasing trend. Neither of these samples should be treated as a simple random sample. The plot in Figure 1.1c does not appear to show any obvious pattern or trend. It might be appropriate to treat these data as a simple random sample. However, before making that decision, it is still important to think about the process that produced the data, since there may be concerns that don't show up in the plot (see Example 1.7).

Sometimes the question as to whether a data set is a simple random sample depends on the population under consideration. This is one case in which a plot can look good, yet the data are not a simple random sample. Example 1.7 provides an illustration.

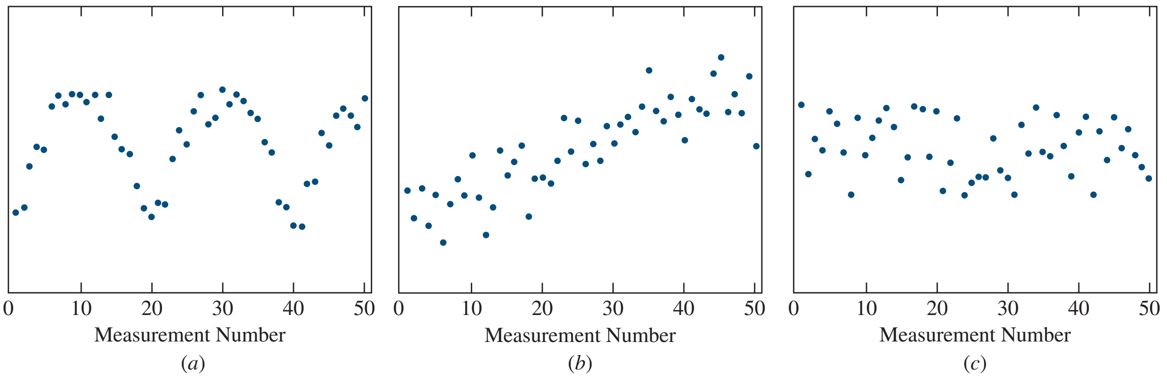


FIGURE 1.1 Three plots of observed values versus the order in which they were made. *(a)* The values show a definite pattern over time. This is not a simple random sample. *(b)* The values show a trend over time. This is not a simple random sample. *(c)* The values do not show a pattern or trend. It may be appropriate to treat these data as a simple random sample.

Example 1.7

A new chemical process is run 10 times each morning for five consecutive mornings. A plot of yields in the order they are run does not exhibit any obvious pattern or trend. If the new process is put into production, it will be run 10 hours each day, from 7 A.M. until 5 P.M. Is it reasonable to consider the 50 yields to be a simple random sample? What if the process will always be run in the morning?

Solution

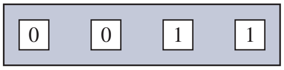
Since the intention is to run the new process in both the morning and the afternoon, the population consists of all the yields that would ever be observed, including both morning and afternoon runs. The sample is drawn only from that portion of the population that consists of morning runs, and thus it is not a simple random sample. There are many things that could go wrong if this is used as a simple random sample. For example, ambient temperatures may differ between morning and afternoon, which could affect yields.

If the process will be run only in the morning, then the population consists only of morning runs. Since the sample does not exhibit any obvious pattern or trend, it might well be appropriate to consider it to be a simple random sample.

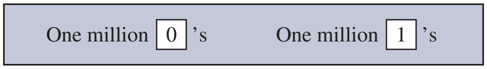
Independence

The items in a sample are said to be **independent** if knowing the values of some of them does not help to predict the values of the others. With a finite, tangible population, the items in a simple random sample are not strictly independent, because as each item is drawn, the population changes. This change can be substantial when the population is small. However, when the population is very large, this change is negligible and the items can be treated as if they were independent.

To illustrate this idea, imagine that we draw a simple random sample of 2 items from the population



For the first draw, the numbers 0 and 1 are equally likely. But the value of the second item is clearly influenced by the first; if the first is 0, the second is more likely to be 1, and vice versa. Thus the sampled items are dependent. Now assume we draw a sample of size 2 from this population:



Again on the first draw, the numbers 0 and 1 are equally likely. But unlike the previous example, these two values remain almost equally likely on the second draw as well, no matter what happens on the first draw. With the large population, the sample items are for all practical purposes independent.

It is reasonable to wonder how large a population must be in order that the items in a simple random sample may be treated as independent. A rule of thumb is that when sampling from a finite population, the items may be treated as independent so long as the sample contains 5% or less of the population.

Interestingly, it is possible to make a population behave as though it were infinitely large, by replacing each item after it is sampled. This method is called **sampling with replacement**. With this method, the population is exactly the same on every draw, and the sampled items are truly independent.

With a conceptual population, we require that the sample items be produced under identical experimental conditions. In particular, then, no sample value may influence the conditions under which the others are produced. Therefore, the items in a simple random sample from a conceptual population may be treated as independent. We may think of a conceptual population as being infinite, or equivalently, that the items are sampled with replacement.

Summary

- The items in a sample are **independent** if knowing the values of some of the items does not help to predict the values of the others.
- Items in a simple random sample may be treated as independent in many cases encountered in practice. An exception occurs when the population is finite and the sample consists of a substantial fraction (more than 5%) of the population.