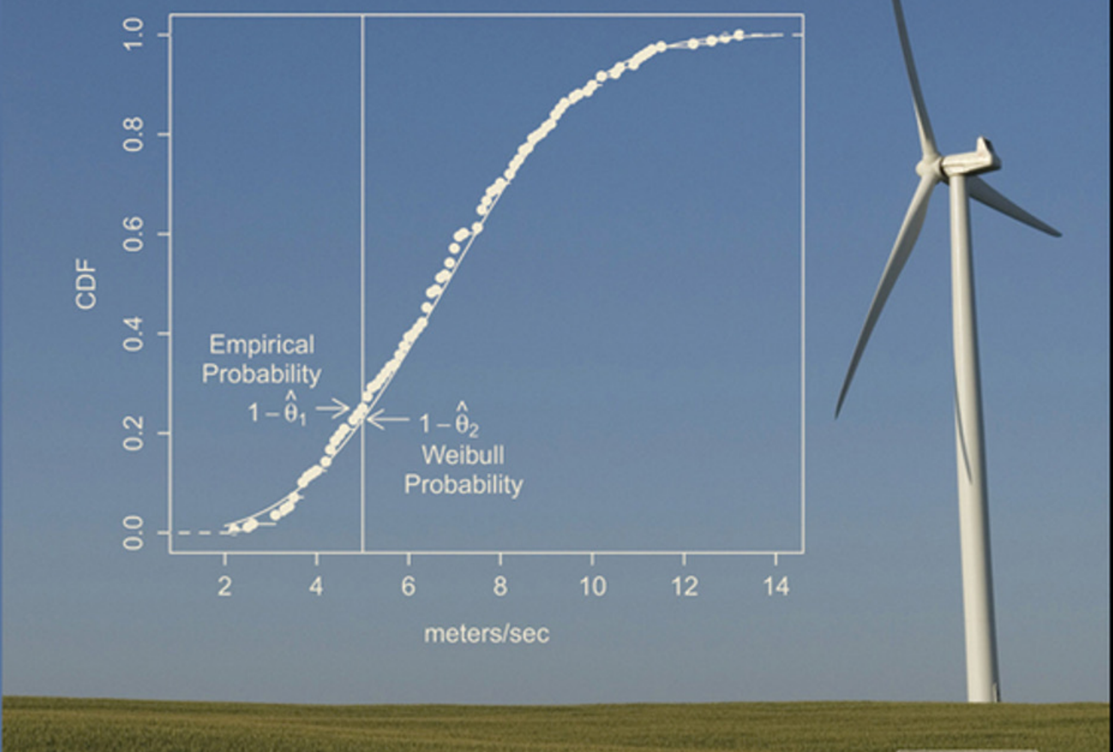
Mathematical Statistics

with Resampling and R

Laura Chihara and Tim Hesterberg



WILEY

WWW.
LINK AVAILABLE

MATHEMATICAL STATISTICS WITH RESAMPLING AND R

MATHEMATICAL STATISTICS WITH RESAMPLING AND R

LAURA CHIHARA
Carleton College

TIM HESTERBERG
Google

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

ISBN: 978-1-118-02985-5

Printed in Singapore.

10 9 8 7 6 5 4 3 2 1

The world seldom notices who teachers are;
but civilization depends on what they do.
— Lindley Stiles[*]

To
Theodore S. Chihara

To
Bev Hesterberg

[*] Stiles founded The Best Should Teach Initiative `http://thebestshouldteach.org`

CONTENTS

PREFACE

Mathematical Statistics with Resampling and R is a one-term undergraduate statistics textbook for sophomores or juniors who have taken a course in probability (at the level of, for instance, Ross (2009), Ghahramani (2004), and Scheaffer and Young (2010)) but may not have had any previous exposure to statistics.

What sets this book apart from other mathematical statistics texts is the use of modern resampling techniques—permutation tests and bootstrapping. We begin with permutation tests and bootstrap methods before introducing classical inference methods. Resampling helps students understand the meaning of sampling distributions, sampling variability, P-values, hypothesis tests, and confidence intervals. We are inspired by the textbooks of Waldrop (1995) and Chance and Rossman (2005), two innovative introductory statistics books that also take a nontraditional approach in the sequencing of topics.

We believe the time is ripe for this book. Many faculty have learned resampling and simulation-based methods in graduate school and use them in their own work and are eager to incorporate these ideas into a mathematical statistics course. Students and faculty today have access to computers that are powerful enough to perform resampling quickly.

A major topic of debate about the mathematical statistics course is how much theory to introduce. We want mathematically talented students to get excited about statistics, so we try to strike a balance between theory, computing, and applications. We feel that it is important to demonstrate some rigor in developing some of the statistical ideas presented here, but that mathematical theory should not dominate the text. And of course, if additions are made to a syllabus, then deletions must also be made. Thus, some topics such as sufficiency, Fisher information, and ANOVA have been omitted in order to make room for permutation testing, bootstrap; and other modern computing methods (though we plan to make some of these omitted topics available as supplements on the text web page https://sites.google.com/site/ChiharaHesterberg). This site will also contain R scripts for the text, and errata.

We have compiled the definitions and theorems of the important probability distributions in Appendix B. Instructors who want to prove results on distributional theory can refer to this appendix. Instructors who wish to skip the theory can continue without interrupting the flow of the statistical discussion.

Incorporating resampling and bootstrapping methods requires that students use statistical software. We use R because it is freely available (http://www.r-project.org/), powerful, flexible, and a valuable tool in future careers. One of us works at Google where there is an explosion in the use of R, with more and more nonstatisticians learning R (the statisticians already know it). We realize that the learning curve for R is high, but believe that the time invested in mastering R is worth the effort. We have written some basic materials on R that are available on the web site for this text. We recommend that instructors work through the introductory worksheet with the students on the first or second day of the term, in a computer lab if possible. We also provide R script files with code found in the text and additional examples.

Statistical computing is necessary in statistical practice and for people working with data in a wide variety of fields. There is an explosion of data—more and more data—and new computational methods are continuously being developed to handle this explosion. Statistics is an exciting field, dare we even say sexy?[1]

ACKNOWLEDGMENTS

This textbook could not have been completed without the assistance of many colleagues and students. In particular, we would like to thank Professor Katherine St. Clair of Carleton College who bravely tested an early rough draft in her Introduction to Statistical Inference class during Winter 2010. In addition, Professor Julie Legler of St. Olaf College adopted the manuscript in her Statistical Theory class during Fall 2010. Both instructors and students provided valuable feedback that improved the exposition and content of this textbook.

We would also like to thank Siyuan (Ernest) Liu and Chen (Daisy) Sun, two Carleton College students, for solving many of the exercises and writing up the solutions with LaTeX.

Finally, the staff at Wiley, including Steve Quigley, Sanchari Sil, Dean Gonzalez, and Jackie Palmieri, provided valuable assistance in preparing this book.

LAURA CHIHARA
Northfield, MN
TIM HESTERBERG
Seattle, WA

[1] Try googling "statistics sexy profession."

1

DATA AND CASE STUDIES

Statistics is the art and science of collecting and analyzing data and understanding the nature of variability. Mathematics, especially probability, governs the underlying theory, but statistics is driven by applications to real problems.

In this chapter, we introduce several data sets that we will encounter throughout the text in the examples and exercises.

1.1 CASE STUDY: FLIGHT DELAYS

If you have ever traveled by air, you probably have experienced the frustration of flight delays. The Bureau of Transportation Statistics maintains data on all aspects of air travel, including flight delays at departure and arrival (`http://www.bts.gov/xml/ontimesummarystatistics/src/index.xml`).

LaGuardia Airport (LGA) is one of three major airports that serves the New York City metropolitan area. In 2008, over 23 million passengers and over 375,000 planes flew in or out of LGA. United Airlines and American Airlines are two major airlines that schedule services at LGA. The data set `FlightDelays` contains information on all 4029 departures of these two airlines from LGA during May and June 2009 (Tables 1.1 and 1.2).

Each row of the data set is an *observation*. Each column represents a *variable*—some characteristic that is obtained for each observation. For instance, on the first observation listed, the flight was a United Airlines plane, flight number 403, destined for

Mathematical Statistics with Resampling and R, First Edition. By Laura Chihara and Tim Hesterberg.

TABLE 1.1 Partial View of `FlightDelays` Data

Flight	Carrier	FlightNo	Destination	DepartTime	Day
1	UA	403	DEN	4–8 a.m.	Fri
2	UA	405	DEN	8–noon	Fri
3	UA	409	DEN	4–8 p.m.	Fri
4	UA	511	ORD	8–noon	Fri
		⋮			

Denver, and departing on Friday between 4 a.m. and 8 a.m. This data set consists of 4029 observations and 9 variables.

Questions we might ask include the following: Are flight delay times different between the two airlines? Are flight delay times different depending on the day of the week? Are flights scheduled in the morning less likely to be delayed by more than 15 min?

1.2 CASE STUDY: BIRTH WEIGHTS OF BABIES

The birth weight of a baby is of interest to health officials since many studies have shown possible links between this weight and conditions in later life, such as obesity or diabetes. Researchers look for possible relationships between the birth weight of a baby and the age of the mother or whether or not she smoked cigarettes or drank alcohol during her pregnancy. The Centers for Disease Control and Prevention (CDC), using data provided by the U.S. Department of Health and Human Services, National Center for Health Statistics, the Division of Vital Statistics as well as the CDC, maintain a database on all babies born in a given year (`http://wonder.cdc.gov/natality-current.html`). We will investigate different samples taken from the CDC's database of births.

One data set we will investigate consists of a random sample of 1009 babies born in North Carolina during 2004 (Table 1.3). The babies in the sample had a gestation period of at least 37 weeks and were single births (i.e., not a twin or triplet).

TABLE 1.2 Variables in Data Set `FlightDelays`

Variable	Description
Carrier	UA=United Airlines, AA=American Airlines
FlightNo	Flight number
Destination	Airport code
DepartTime	Scheduled departure time in 4 h intervals
Day	Day of week
Month	September or October
Delay	Minutes flight delayed (negative indicates early departure)
Delayed30	Departure delayed more than 30 min?
FlightLength	Length of time of flight (minutes)

TABLE 1.3 Variables in Data Set `NCBirths2004`

Variable	Description
Age	Mother's age
Tobacco	Mother used tobacco?
Gender	Gender of baby
Weight	Weight at birth (grams)
Gestation	Gestation time (weeks)

In addition, we will also investigate a data set, `Girls2004`, consisting of a random sample of 40 baby girls born in Alaska and 40 baby girls born in Wyoming. These babies also had a gestation period of at least 37 weeks and were single births.

The data set `TXBirths2004` contains a random sample of 1587 babies born in Texas in 2004. In this case, the sample was not restricted to single births, nor to a gestation period of at least 37 weeks. The numeric variable `Number` indicates whether the baby was a single birth, or one of a twin, triplet, and so on. The variable `Multiple` is a factor variable indicating whether or not the baby was a multiple birth.

1.3 CASE STUDY: VERIZON REPAIR TIMES

Verizon is the primary local telephone company (incumbent local exchange carrier, ILEC) for a large area of the eastern United States. As such, it is responsible for providing repair service for the customers of other telephone companies known as competing local exchange carriers (CLECs) in this region. Verizon is subject to fines if the repair times (the time it takes to fix a problem) for CLEC customers are substantially worse than those for Verizon customers.

The data set `Verizon` contains a random sample of repair times for 1664 ILEC and 23 CLEC customers (Table 1.4). The mean repair time for ILEC customers is 8.4 hours, while that for CLEC customers is 16.5 h. Could a difference this large be easily explained by chance?

1.4 SAMPLING

In analyzing data, we need to determine whether the data represent a *population* or a *sample*. A *population* represents all the individual cases, whether they are babies,

TABLE 1.4 Variables in Data Set `Verizon`

Variable	Description
Time	Repair times (in hours)
Group	ILEC or CLEC

fish, cars, or coin flips. The data from Flight Delays Case Study in Section 1.1 are *all* the flight departures of United Airlines and American Airlines out of LaGuardia Airport in May and June 2009; thus, this data set represents the population of all such flights. On the other hand, the North Carolina data set contains only a subset of 1009 births from over 100,000 births in North Carolina in 2004. In this case, we will want to know how representative statistics computed from this sample are of the entire population of North Carolina babies born in 2004.

Populations may be finite, such as births in 2004, or infinite, such as coin flips or births next year.

Throughout this chapter, we will talk about drawing random samples from a population. We will use capital letters (e.g., X, Y, Z, and so on) to denote random variables and lowercase letters (e.g., x_1, x_2, x_3, and so on) to denote actual values or data.

There are many kinds of random samples. Strictly speaking, a "random sample" is any sample obtained using a random procedure. However, in this book we use *random sample* to mean a sample of independent and identically distributed (i.i.d.) observations from the population, if the population is infinite.

For instance, suppose you toss a fair coin 20 times and consider each head a "success." Then your sample consists of the random variables $X_1, X_2, \ldots, X_{20}$, each a Bernoulli random variable with success probability 1/2. We use the notation $X_i \sim \text{Bern}(1/2)$, $i = 1, 2, \ldots, 20$.

If the population of interest is finite $\{x_1, x_2, \ldots, x_N\}$, we can choose a random sample as follows: label N balls with the numbers $1, 2, \ldots, N$ and place them in an urn. Draw a ball at random, record its value $X_1 = x_{i_1}$, and then replace the ball. Draw another ball at random, record its value, $X_2 = x_{i_2}$, and then replace. Continue until you have a sample $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$. This is *sampling with replacement*. For instance, if $N = 5$ and $n = 2$, then there are $5 \times 5 = 25$ different samples of size 2 (where order matters). (Note: By "order matters" we do not imply that order matters in practice, rather we mean that we keep track of the order of the elements when enumerating samples. For instance, the set $\{a, b\}$ is different from $\{b, a\}$.)

However, in most real situations, for example, in conducting surveys, we do not want to have the same person polled twice. So we would sample *without replacement*, in which case, we will not have independence. For instance, if you wish to draw a sample of size $n = 2$ from a population of $N = 10$ people, then the probability of any one person being selected is $1/10$. However, after having chosen that first person, the probability of any one of the remaining people being chosen is now $1/9$.

In cases where populations are very large compared to the sample size, calculations under sampling without replacement are reasonably approximated by calculations under sampling with replacement.

Example 1.1 Consider a population of 1000 people, 350 of whom are smokers and the rest are nonsmokers. If you select 10 people at random but with replacement, then the probability that 4 are smokers is $\binom{10}{4}(350/1000)^4(650/1000)^6 \approx 0.2377$. If you select without replacement, then the probability is $\binom{350}{4}\binom{650}{6}/\binom{1000}{10} \approx 0.2388$. □

1.5 PARAMETERS AND STATISTICS

When discussing numeric information, we will want to distinguish between populations and samples.

Definition 1.1 A *parameter* is a (numerical) characteristic of a population or of a probability distribution.

A *statistic* is a (numerical) characteristic of data. ||

Any function of a parameter is also a parameter; any function of a statistic is also a statistic. When the statistic is computed from a random sample, it is itself random, and hence is a random variable.

Example 1.2 μ and σ are parameters of the normal distribution with pdf $f(x) = (1/\sqrt{2\pi}\sigma)e^{-(x-\mu)^2/(2\sigma^2)}$.

The variance σ^2 and *signal-to-noise ratio* μ/σ are also parameters. □

Example 1.3 If $X_1, X_2, \ldots, X_n$ are a random sample, then the mean $\bar{X} = 1/n \sum_{i=1}^{n} X_i$ is a statistic. □

Example 1.4 Consider the population of all babies born in the United States in 2004. Let μ denote the average weight of all these babies. Then μ is a parameter. The average weight of a sample of 2000 babies born in that year is a statistic. □

Example 1.5 If we consider the population of all adults in the United States today, the proportion p who approve of the president's job performance is a parameter. The fraction $\hat{p}$ who approve in any given sample is a statistic. □

Example 1.6 The average weight of 1009 babies in the North Carolina Case Study in Section 1.2 is 3448.26 g. This average is a statistic. □

Example 1.7 If we survey 1000 adults and find that 60% intend to vote in the next presidential election, then $\hat{p} = 0.60$ is a statistic: it estimates the parameter p, the proportion of all adults who intend to vote in the next election. □

1.6 CASE STUDY: GENERAL SOCIAL SURVEY

The General Social Survey (GSS) is a major survey that has tracked American demographics, characteristics, and views on social and cultural issues since the 1970s. It is conducted by the National Opinion Research Center (NORC) at the University of Chicago. Trained interviewers meet face-to-face with the adults chosen for the survey and question them for about 90 minutes in their homes.

TABLE 1.5 Variables in Data Set GSS2002

Variable	Description
Region	Interview location
Gender	Gender of respondent
Race	Race of respondent: White, Black, other
Marital	Marital status
Education	Highest level of education
Happy	General happiness
Income	Respondent's income
PolParty	Political party
Politics	Political views
Marijuana	Legalize marijuana?
DeathPenalty	Death penalty for murder?
OwnGun	Have gun at home?
GunLaw	Require permit to buy a gun?
SpendMilitary	Amount government spends on military
SpendEduc	Amount government spends on education
SpendEnv	Amount government spends on the environment
SpendSci	Amount government spends on science
Pres00	Whom did you vote for in the 2000 presidential election?
Postlife	Believe in life after death?

The GSS Case Study includes the responses of 2765 participants selected in 2002 to about a dozen questions, listed in Table 1.5. For example, one of the questions (`SpendEduc`) asked whether the respondent believed that the amount of money being spent on the nation's education system was too little, too much, or the right amount.

We will analyze the GSS data to investigate questions such as the following: Is there a relationship between the gender of an individual and whom they voted for in the 2000 presidential election, are people who live in certain regions happier, are there educational differences in support for the death penalty? These data are archived at the Computer-Assisted Survey Methods Program at the University of California (`http://sda.berkeley.edu/`).

1.7 SAMPLE SURVEYS

"When do you plan to vote for in the next presidential election?" "Would you purchase our product again in the future?" "Do you smoke cigarettes? If yes, how old were you when you first started?" Questions such as these are typical of sample surveys. Researchers want to know something about a population of individuals, whether they are registered voters, online shoppers, or American teenagers, but to poll every individual in the population—that is, to take a *census*—is impractical and costly.

Thus, researchers will settle for a sample from the target population. But if, say, 60% of those in your sample of 1000 adults intend to vote for candidate Smith in the next election, how close is this to the actual percentage who will vote for Smith? How can we be sure that this sample is truly representative of the population of all voters? We will learn techniques for *statistical inference*, drawing a conclusion about a population based on information about a sample.

When conducting a survey, researchers will start with a *sampling frame*—a list from which the researchers will choose their sample. For example, to survey all students at a college, the campus directory listing could be a sampling frame. For pre-election surveys, many polling organizations use a sampling frame of registered voters. Note that the choice of sampling frame could introduce the problem of *undercoverage*: omitting people from the target population in the survey. For instance, young people were missed in many election surveys during the 2008 Obama–McCain presidential race because they had not yet registered to vote.

Once the researchers have a sampling frame, they will then draw a random sample from this frame. Researchers will use some type of *probability (scientific) sampling scheme*, that is, a scheme that gives everybody in the population a positive chance of being selected. For example, to obtain a sample of size 10 from a population of 100 individuals, write each person's name on a slip of paper, put the slips of paper into a basket, and then draw out 10 slips of paper. Nowadays, statistical software is used to draw out random samples from a sampling frame.

Another basic survey design uses *stratified sampling*: the population is divided into nonoverlapping strata and then random samples are drawn from each stratum. The idea is to group individuals who are similar in some characteristic into homogeneous groups, thus reducing variability. For instance, in a survey of university students, a researcher might divide the students by class: first year, sophomores, juniors, seniors, and graduate students. A market analyst for an electronics store might choose to stratify customers based on income levels.

In *cluster sampling*, the population is divided into nonoverlapping clusters and then a random sample of clusters is drawn. Every person in a chosen cluster is then interviewed for the survey. An airport wanting to conduct a customer satisfaction survey might use a sampling frame of all flights scheduled to depart from the airport on a certain day. A random sample of flights (clusters) is chosen and then all passengers on these flights are surveyed. A modification of this design might involve sampling in stages: for instance, the analysts might first choose a random sample of flights, and then from each flight choose a random sample of passengers.

The General Social Survey uses a more complex sampling scheme in which the sampling frame is a list of counties and county equivalents (standard metropolitan statistical areas) in the United States. These counties are stratified by region, age, and race. Once a sample of counties is obtained, a sample of block groups and enumeration districts is selected, stratifying these by race and income. The next stage is to randomly select blocks and then interview a specific number of men and women who live within these blocks.

Indeed, all major polling organizations such as Gallup or Roper as well as the GSS use a multistage sampling design. In this book, we use the GSS data or polling results

TABLE 1.6 Variables in Data Set `Beerwings`

Variable	Description
Gender	Male or female
Beer	Ounces of beer consumed
Hotwings	Number of hot wings eaten

for examples as if the survey design used simple random sampling. Calculations for more complex sampling scheme are beyond the scope of this book and we refer the interested reader to Lohr (1991) for details.

1.8 CASE STUDY: BEER AND HOT WINGS

Carleton student Nicki Catchpole conducted a study of hot wings and beer consumption at the Williams Bar in the Uptown area of Minneapolis (Catchpole (2004)). She asked patrons at the bar to record their consumption of hot wings and beer over the course of several hours. She wanted to know if people who ate more hot wings would then drink more beer. In addition, she investigated whether or not gender had an impact on hot wings or beer consumption.

The data for this study are in `Beerwings` (Table 1.6). There are 30 observations and 3 variables.

1.9 CASE STUDY: BLACK SPRUCE SEEDLINGS

Black spruce (*Picea mariana*) is a species of a slow-growing coniferous tree found across the northern part of North America. It is commonly found on wet organic soils. In a study conducted in the 1990s, a biologist interested in factors affecting the growth of the black spruce planted its seedlings on sites located in boreal peatlands in northern Manitoba, Canada (Camill et al. (2010)).

The data set `Spruce` contains a part of the data from the study (Table 1.7). Seventy-two black spruce seedlings were planted in four plots under varying conditions (fertilizer–no fertilizer, competition–no competition) and their heights and diameters were measured over the course of 5 years.

The researcher wanted to see whether the addition of fertilizer or the removal of competition from other plants (by weeding) affected the growth of these seedlings.

1.10 STUDIES

Researchers carry out studies to understand the conditions and causes of certain outcomes: Does smoking cause lung cancer? Do teenagers who smoke marijuana

TABLE 1.7 Variables in Data Set `Spruce`

Variable	Description
Tree	Tree number
Competition	C (competition), CR (competition removed)
Fertilizer	F (fertilized), NF (not fertilized)
Height0	Height (cm) of seedling at planting
Height5	Height (cm) of seedling at year 5
Diameter0	Diameter (cm) of seedling at planting
Diameter5	Diameter (cm) of seedling at year 5
Ht.change	Change (cm) in height
Di.change	Change (cm) in diameter

tend to move on to harder drugs? Do males eat more hot wings than females? Do black spruce seedlings grow taller in fertilized plots?

The Beer and Hot Wings Case Study in Section 1.8 is an example of an *observational study*, a study in which researchers observe participants but do not influence the outcome. In this case, the student just recorded the number of hot wings eaten and beer consumed by the patrons of Williams Bar.

Example 1.8 The first Nurse's Health Study is a major observational study funded by the National Institutes of Health. Over 120,000 registered female nurses who, in 1976, were married, between the ages of 33 and 55 years, and who lived in the 11 most populous states, have been responding every 2 years to written questions about their health and lifestyle, including smoking habits, hormone use, and menopause status. Many results on women's health have come out of this study, such as finding an association between taking estrogen after menopause and lowering the risk of heart disease, and determining that for nonsmokers there is no link between taking birth control pills and developing heart disease.

Because this is an observational study, no *cause and effect* conclusions can be drawn. For instance, we cannot state that taking estrogen after menopause will *cause* a lowering of the risk for heart disease. In an observational study, there may be many unrecorded or hidden factors that impact the outcomes. Also, because the participants in this study are registered nurses, we need to be careful about making inferences about the general female population. Nurses are more educated and more aware of health issues than the average person. □

On the other hand, the Black Spruce Case Study in Section 1.9 was an *experiment*. In an experiment, researchers will manipulate the environment in some way to observe the response of the objects of interest (people, mice, ball bearings, etc.). When the objects of interest in an experiment are people, we refer to them as *subjects*; otherwise, we call them *experimental units*. In this case, the biologist randomly assigned the experimental units—the seedlings—to plots subject to four *treatments*: fertilization with competition, fertilization without competition, and no fertilization with competition, and no fertilization with no competition. He then recorded their height over a period of several years.

A key feature in this experiment was the *random assignment* of the seedlings to the treatments. The idea is to spread out the effects of unknown or uncontrollable factors that might introduce unwanted variability into the results. For instance, if the biologist had planted all the seedlings obtained from one particular nursery in the fertilized, no competition plot and subsequently recorded that these seedlings grew the least, then he would not be able to discern whether this was due to this particular treatment or due to some possible problem with seedlings from this nursery. With random assignment of treatments, the seedlings from this particular nursery would usually be spread out over the four treatments. Thus, the differences between the treatment groups should be due to the treatments (or chance).

Example 1.9 Knee osteoarthritis (OA) that results in deterioration of cartilage in the joint is a common source of pain and disability for the elderly population. In a 2008 paper, "Tai Chi is effective in treating knee osteoarthritis: A randomized controlled trial," Wang et al. (2009) at Tufts University Medical School describe an experiment they conducted to see whether practicing Tai Chi, a style of Chinese martial arts, could alleviate pain from OA. Forty patients over the age of 65 with confirmed knee OA but otherwise in good health were recruited from the Boston area. Twenty were randomly assigned to attend twice weekly 60 min sessions of Tai Chi for 12 weeks. The remaining 20 participants, the *control group*, attended twice weekly 60 min sessions of instructions on health and nutrition, as well as some stretching exercises.

At the end of the 12 weeks, those in the Tai Chi group reported a significant decrease in knee pain. Because the subjects were randomly assigned to the two treatments, the researchers can assert that the Tai Chi sessions lead to decrease in knee pain due to OA. Note that because the subjects were recruited, we need to be careful about making an inference about the general elderly population: Somebody who voluntarily signs up to be in an experiment may be inherently different from the average person. □

1.11 EXERCISES

1. For each of the following, describe the population and, if relevant, the sample. For each number presented, determine if it is a parameter or a statistic (or something else).
   (a) A survey of 2000 high school students finds that 47% watch the television show "Glee."
   (b) The 2000 U.S. Census reports that 13.9% of the U.S. population was between the ages of 15 and 24 years.
   (c) Based on the rosters of all National Basketball Association teams for the 2006–2007 season, the average height of the players was 78.93 in.
   (d) A December 2009 Gallup poll of 1025 national adults, aged 18 and older, shows that 47% would advise their member of Congress to vote for health care legislation (or lean toward doing so).

2. Researchers reported that moderate drinking of alcohol was associated with a lower risk of dementia ((Mukamal et al. (2003)). Their sample consisted of 373 people with dementia and 373 people without dementia. Participants were asked how much beer, wine, or shot of liquor they consumed. Thus, participants who consumed 1–6 drinks a week had a lower risk of dementia than those who abstained from alcohol.
   (a) Was this study an observational study or an experiment?
   (b) Can the researchers conclude that drinking alcohol causes a lower risk of dementia?
3. Researchers surveyed 959 ninth graders who attended three large U.S. urban high schools and found that those who listened to music that had references to marijuana were almost twice as likely to have used marijuana as those who did not listen to music with references to marijuana (Primack et al. (2010)).
   (a) Was this an observational study or an experiment?
   (b) Can the researchers conclude that listening to music with references to marijuana causes students to use drugs?
   (c) Can the researchers extend their results to all urban American adolescents?
4. Duke University researchers found that diets low in carbohydrates are effective in controlling blood sugar levels (Westman et al. (2008)). Eighty-four volunteers with obesity and type 2 diabetes were randomly assigned to either a diet of less than 20 g of carbohydrates/day or a low-glycemic, reduced calorie diet (500 calories/day). Ninety-five percent of those on the low-carbohydrate diet were able to reduce or eliminate their diabetes medications compared to 62% on the low-glycemic diet.
   (a) Was this study an observational study or an experiment?
   (b) Can researchers conclude that a low-carbohydrate diet causes an improvement in type 2 diabetes?
   (c) Can researchers extend their results to a more general population? Explain.
5. In a population of size N, the probability of any subset of size n being chosen is $1/\binom{N}{n}$. Show this implies that any one person in the population has a n/N probability of being chosen in a sample. Then, in particular, every person in the population has the same probability of being chosen.
6. A typical Gallup poll surveys about $n = 1000$ adults. Suppose the sampling frame contains 100 million adults (including you). Now, select a random sample of 1000 adults.
   (a) What is the probability that you will be in this sample?
   (b) Now suppose that 2000 such samples are selected, each independent of the others. What is the probability that you will *not* be in any of the samples?
   (c) How many samples must be selected for you to have a 0.5 probability of being in at least one sample?

2

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an approach to examining and describing data to gain insight, discover structure, and detect anomalies and outliers. John Tukey (1915–2000), an American mathematician and statistician who pioneered many of the techniques now used in EDA, stated in his 1977 book *Exploratory Data Analysis* (Tukey (1977)) that "Exploratory data analysis is detective work—numerical detective work—counting detective work—or graphical detective work." In this chapter, we will learn many of the basic techniques and tools for gaining insight into data.

Statistical software packages can easily do the calculations needed for the basic plots and numeric summaries of data. We will use the software package `R`. We will assume that you have gone through the introduction to `R` available at the web site `https://sites.google.com/site/ChiharaHesterberg`.

2.1 BASIC PLOTS

In Chapter 1, we described data on the lengths of flight delays of two major airlines flying from LaGuardia Airport in New York City in 2009. Some basic questions we might ask include how many of these flights were flown by United Airlines and how many by American Airlines? How many flights flown by each of these airlines were delayed more than 30 min?

A *categorical* variable is one that places the observations into groups. For instance, in the `FlightDelays` data set, `Carrier` is a categorical variable (we will also

Mathematical Statistics with Resampling and R, First Edition. By Laura Chihara and Tim Hesterberg.

TABLE 2.1 Counts for the `Carrier` Variable

	Carrier	
	American Airlines	United Airways
Number of flights	2906	1123

call this a *factor* variable) with two *levels*, UA and AA. Other data sets might have categorical variables such as `gender` (with two levels, Male or Female) or `size` (with levels Small, Medium, and Large).

A *bar chart* is used to describe the distribution of a categorical (factor) variable. Bars are drawn for each level of the factor variable and the height of the bar is the number of observations in that level. For the `FlightDelays` data, there were 2906 American Airlines flights and 1123 United Airlines flights (Table 2.1). The corresponding bar chart is shown in Figure 2.1.

We might also be interested in investigating the relationship between a carrier and whether or not a flight was delayed more than 30 min. A *contingency table* summarizes the counts in the different categories.

From Table 2.2, we can compute that 13.5% of American Airlines flights were delayed more than 30 min compared to 18.2% of United Airlines flights. Is this difference in percentages *statistically significant*? Could the difference in percentages be due to *natural variability*, or is there a systematic difference between the two airlines? We will address this question in the following chapters.

With a numeric variable, we will be interested in its distribution: What is the range of values? What values are taken on most often? Where is the *center*? What is the *spread*?

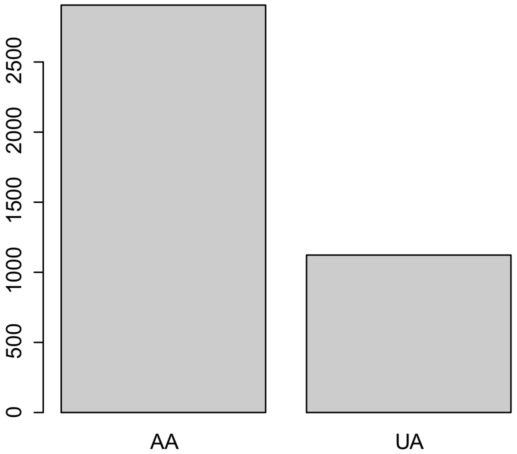


FIGURE 2.1 Bar chart of `Carrier` variable.