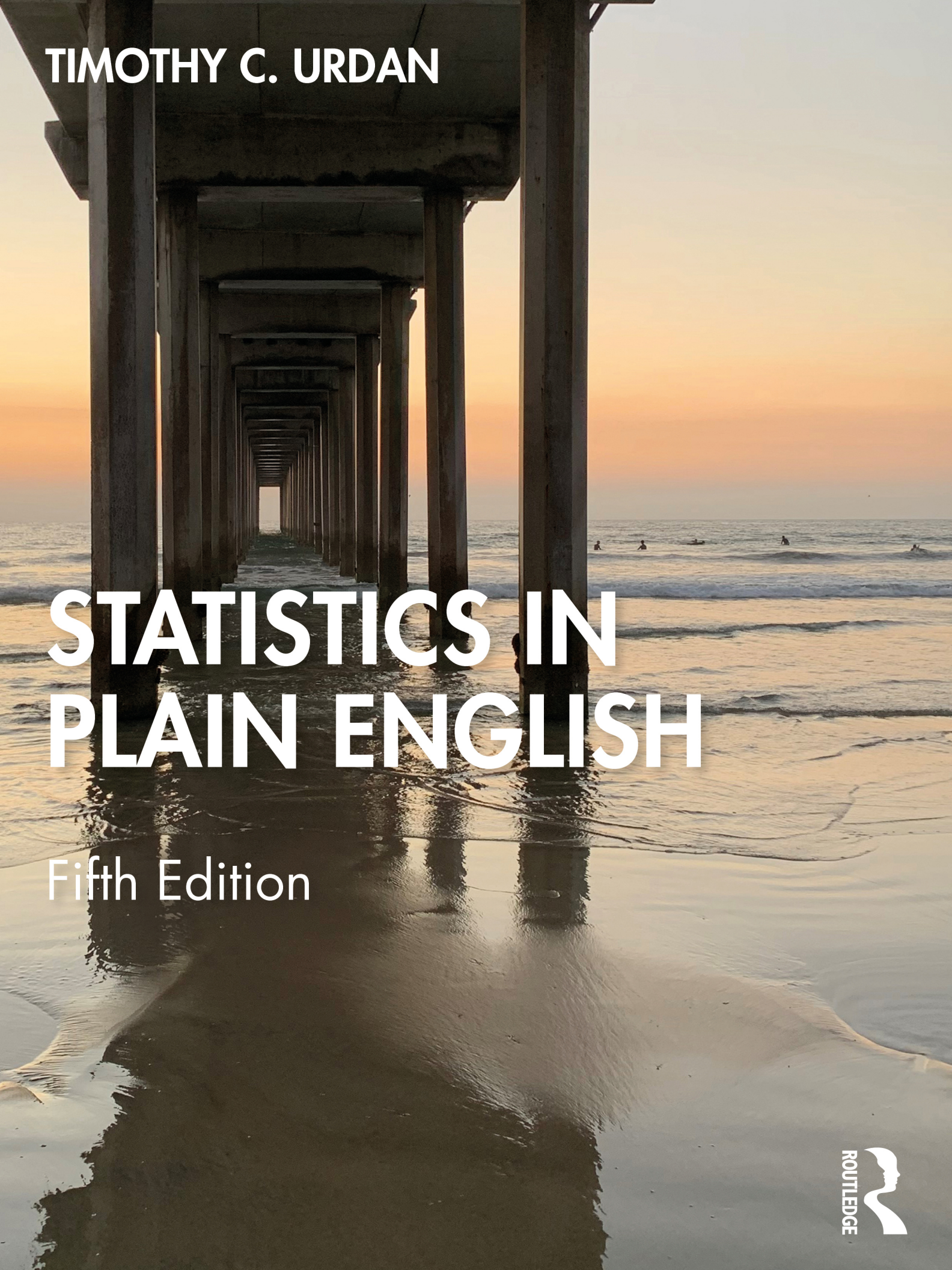
TIMOTHY C. URDAN
STATISTICS IN PLAIN ENGLISH
Fifth Edition
ROUTLEDGE

Statistics in Plain English

Statistics in Plain English is a straightforward, conversational introduction to statistics that delivers exactly what its title promises. Each chapter begins with a brief overview of a statistic (or set of statistics) that describes what the statistic does and when to use it, followed by a detailed step-by-step explanation of how the statistic works and exactly what information it provides. Chapters also include an example of the statistic (or statistics) used in real-world research, "Worked Examples," "Writing It Up" sections that demonstrate how to write about each statistic, "Wrapping Up and Looking Forward" sections, and practice work problems.

Thoroughly updated throughout, this edition features several key additions and changes. First, a new chapter on person-centered analyses, including cluster analysis and latent class analysis (LCA) has been added, providing an important alternative to the more commonly used variable-centered analyses (e.g., *t* tests, ANOVA, regression). Next, the chapter on nonparametric statistics has been enhanced with in-depth descriptions of Mann-Whitney U, Kruskal-Wallis, and Wilcoxon Signed-Rank analyses, in addition to the detailed discussion of the Chi-square statistic found in the previous edition. These nonparametric statistics are widely used when dealing with nonnormally distributed data. This edition also includes more information about the assumptions of various statistics, including a detailed explanation of the assumptions and consequences of violating the assumptions of regression, as well as more coverage of the normal distribution in statistics. Finally, the book features a multitude of real-world examples throughout to aid student understanding and provides them with a solid understanding of how several statistics techniques commonly used by researchers in the social sciences work.

Statistics in Plain English is suitable for a wide range of readers, including students taking their first statistics course, professionals who want to refresh their statistical memory, and undergraduate or graduate students who need a concise companion to a more complicated text used in their class. The text works as a standalone or as a supplement and covers a range of statistical concepts from descriptive statistics to factor analysis and person-centered analyses.

Timothy C. Urdan is a Professor at Santa Clara University. He received his Ph.D. from the Combined Program in Education and Psychology at the University of Michigan, his Master's degree in Education from Harvard University, and his B.A. in Psychology from the University of California, Berkeley. He conducts research on student motivation, classroom contexts, and teacher identity. He serves on the editorial boards of several journals and is the coeditor of the *Advances in Motivation and Achievement* book series. He is a fellow of the American Psychological Association and lives in Berkeley, California.

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

Statistics in Plain English

Fifth Edition

Timothy C. Urdan

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Cover image: © Timothy C. Urdan

Fifth edition published 2022
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Lawrence Erlbaum Association, Publishers 2001

Fourth edition published by Routledge 2016

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-0-367-34282-1 (hbk)
ISBN: 978-0-367-34283-8 (pbk)
ISBN: 978-1-003-00645-9 (ebk)
ISBN: 978-1-032-22944-7 (ebk+)

DOI: 10.4324/9781003006459

Access the Companion Website: www.routledge.com/cw/urdan

Contents

Preface

Why Use Statistics?

The first edition of this book was published in 2001. As I complete this fifth edition 20 years later, I am awed by how much the world has changed. There was no Facebook, Instagram, or Twitter back then, and the Internet was still in its early stages of what would become a world-altering technology. What the Internet, and social media platforms in particular, has ushered in is an era where everybody, regardless of their level of expertise, can voice their opinion and it can be heard. Back in 2001, I was frustrated by how often people believed things that were not supported by evidence simply because their own experience did not fit the evidence. Now that everyone has a platform to voice their viewpoint, and it is easy to find hundreds or thousands of other people who share the same view, even when there is no evidence to support it, can you imagine how frustrated I am?

Although the opinions, personal stories, and conspiracy theories that are rampant on the Internet can be compelling, as researchers we must go beyond the personal story or impassioned argument and look for broader evidence. How can we tell whether vaccines cause autism? How do we decide whether public schools are failing, whether teacher unions help or hinder valuable reform, and whether charter schools do a better job of educating students? To answer these questions, we need good data, and then we need to analyze the data using the appropriate statistics.

Many people have a general distrust of statistics, believing that crafty statisticians can "make statistics say whatever they want" or "lie with statistics." In fact, if a researcher calculates the statistics correctly, he or she cannot make them say anything other than what they say, and statistics never lie. Rather, crafty researchers can interpret what the statistics *mean* in a variety of ways, and those who do not understand statistics are forced to either accept the interpretations that statisticians and researchers offer or reject statistics completely. I believe a better option is to gain an understanding of how statistics work and then use that understanding to interpret the statistics one sees and hears for oneself. The purpose of this book is to make it a little easier to understand statistics.

Uses of Statistics

One of the potential shortfalls of anecdotal data is that they are idiosyncratic. One of our cognitive shortcomings, as people, is that we tend to believe that if something is true for us, it is just a fact. "I eat a multivitamin every day and I haven't been sick for 20 years!" "My grandmother smoked a pack a day for 50 years and she lived until she was 96!" "My parents spanked me and I turned out fine!" Although these statements may (or may not!) be true for the individuals who uttered them, that does not mean they are true for everyone, or even for most people. Statistics allow researchers to collect information, or data, from a large number of people (or other kinds of samples) and then summarize their typical experience. Do *most* people who take multivitamins live healthier lives? Do most people who smoke a pack a day live longer or shorter than people who do not smoke? Is there any association between whether one is spanked and how one "turns out," however that is defined? Statistics allow researchers to take a large batch of data and *summarize* it into a couple of numbers, such as an average. Of course, when data are summarized into a single number, a lot of information is lost, including the fact that different people have very different experiences. So, it is important to remember that, for the most part, statistics do not provide useful information about each individual's experience. Rather, researchers generally use statistics to make *general* statements about a sample

or a population. Although personal stories are often moving or interesting, it is also important to understand what the *typical* or *average* experience is. For this, we need statistics.

Statistics are also used to reach conclusions about general differences between groups. For example, suppose that in my family, there are four children, two men and two women. Suppose that the women in my family are taller than the men. This personal experience may lead me to the conclusion that women are generally taller than men. Of course, we know that, on average, men are taller than women. The reason we know this is because researchers have taken large, random samples of men and women and compared their average heights. Researchers are often interested in making such comparisons: Do cancer patients survive longer using one drug than another? Is one method of teaching children to read more effective than another? Do men and women differ in their enjoyment of a certain movie? To answer these questions, we need to collect data from samples and compare these data using statistics. The results we get from such comparisons are often more trustworthy than the simple observations people make from nonrandom samples, such as the different heights of men and women in my family.

Statistics can also be used to see if scores on two variables are related and to make predictions. For example, statistics can be used to see whether smoking cigarettes is related to the likelihood of developing lung cancer. For years, tobacco companies argued that there was no relationship between smoking and cancer. Sure, some people who smoked developed cancer. However, the tobacco companies argued that (a) many people who smoke never develop cancer, and (b) many people who smoke tend to do other things that may lead to cancer development, such as eating unhealthy foods and not exercising. With the help of statistics in a number of studies, researchers were finally able to produce a preponderance of evidence indicating that, in fact, there is a relationship between cigarette smoking and cancer. Because statistics tend to focus on overall patterns rather than individual cases, this research did not suggest that *everyone* who smokes will develop cancer. Rather, the research demonstrated that, on average, people have a greater chance of developing cancer if they smoke cigarettes than if they do not.

With a moment's thought, you can imagine a large number of interesting and important questions that statistics about relationships can help you answer. Is there a relationship between self-esteem and academic achievement? Is there a relationship between the appearance of criminal defendants and their likelihood of being convicted? Is it possible to predict the violent crime rate of a state from the amount of money the state spends on drug treatment programs? If we know the father's height, how accurately can we predict the heights of their sons? These and thousands of other questions have been examined by researchers using statistics designed to determine the relationship between variables in a population. With the rise of the Internet, data is being collected constantly. For example, most casual users of the Internet and social media provide information about their age, gender, where they live, how much money they make, how much they spend, what they like to buy, who their friends are, which web sites they visit, what they like, whether they are married or single, etc. With the help of statistics, data analysts determine what advertisements you should see when you visit a website, how to attract you to certain websites, and how to get you to encourage your friends (without your knowledge) to like or buy various products. More than ever before, statistics and data are deeply affecting many parts of your life. With this in mind, wouldn't it be nice to understand a bit more about how these statistics work?

How to Use This Book

If you are new to statistics, this book can provide an easy introduction to many of the basic, and most commonly used, statistics. Or, if you have already taken a course or two in statistics, this book may be useful as a reference book to refresh your memory about statistical concepts

you have encountered in the past. It is important to remember that this book is much less detailed than a traditional statistics textbook. This book was designed to provide a relatively short and inexpensive introduction to statistics, with a greater focus on the conceptual part of statistics than the computational, mathematical part. Most of the concepts discussed in this book are more complex than the presentation in this book would suggest, and a thorough understanding of these concepts may be acquired only with the use of a more traditional, more detailed textbook.

With that warning firmly in mind, let me describe the potential benefits of this book, and how to make the most of them. As a researcher and a teacher of statistics, I have found that statistics textbooks often contain a lot of technical information that can be intimidating to nonstatisticians. Although, as I said previously, this information is important, sometimes it is useful to have a short, simple description of a statistic, when it should be used, and how to make sense of it. This is particularly true for students taking only their first or second statistics course, those who do not consider themselves to be "mathematically inclined," and those who may have taken statistics years ago and now find themselves in need of a little refresher. My purpose in writing this book is to provide short, simple descriptions and explanations of a number of statistics that are easy to read and understand.

To help you use this book in a manner that best suits your needs, I have organized each chapter into sections. In the first section, a brief (one to two pages) description of the statistic is given, including what the statistic is used for and what information it provides. The second section of each chapter contains a slightly longer (3–12 pages) discussion of the statistic. In this section, I provide a bit more information about how the statistic works, an explanation of how the formula for calculating the statistic works, the strengths and weaknesses of the statistic, and the conditions that must exist to use the statistic. Each chapter includes an example or two in which the statistic is calculated and interpreted. The chapters conclude with illustrations of how to write up the statistical information for publication and a set of work problems.

Before reading the book, it may be helpful to note three of its features. First, some of the chapters discuss more than one statistic. For example, in Chapter 2, three measures of central tendency are described: the mean, median, and mode. Second, some of the chapters cover statistical concepts rather than specific statistical techniques. For example, in Chapter 4, the normal distribution is discussed. There is also a chapter on statistical significance, effect size, and confidence intervals. Finally, you should remember that the chapters in this book are not necessarily designed to be read in order. The book is organized such that the more basic statistics and statistical concepts are in the earlier chapters whereas the more complex concepts appear later in the book. However, it is not necessary to read one chapter before understanding the next. Rather, each chapter in the book was written to stand on its own. This was done so that you could use each chapter as needed. If, for example, you had no problem understanding *t* tests when you learned about them in your statistics class but find yourself struggling to understand one-way analysis of variance, you may want to skip the *t* test chapter (Chapter 8) and skip directly to the analysis of variance chapter (Chapter 9). If you are brand new to statistics, however, keep in mind that some statistical concepts (e.g., *t* tests, ANOVA) are easier to understand if you first learn about the mean, variance, and hypothesis testing.

New Features in This Edition

This fifth edition of *Statistics in Plain English* includes a number of features not available in the previous editions. I added a new chapter that focuses on person-centered analysis like cluster analysis and latent class analysis (see Chapter 16) to the book. In addition, the chapter on nonparametric statistics (see Chapter 14) now has much more information about

the Mann-Whitney *U* test, the Wilcoxon signed-rank test, and the Kruskal-Wallis test. I've added more information about probability to Chapter 4 and Chapter 7, a concept that can be confusing to students who are new to statistics (as well as some statistics veterans). In addition, I've added a section about assumptions of parametric statistics to Chapter 7 and more information in several other chapters about the assumptions for each statistic, and some of the consequences of violating those assumptions. Chapter 1 now begins with a longer explanation of the importance of statistics in our everyday lives, and each of the last nine chapters in the book includes an example from the real-world research that employs the statistic, or statistics, covered in each chapter. Each of the 16 chapters now includes a set of work problems with solutions provided on the website that accompanies this book.

These additions to the new edition were not at the expense of useful features that were added in previous editions of the book. These include a section called "Worked Examples" that appear in most of the chapters in the book. In this section, I work through all of the steps to calculate and interpret the statistic featured in the chapter, either by hand or using SPSS. There are also links to videos of me calculating each statistic or conducting analyses on SPSS or R and interpreting the output on the website that accompanies the book. Throughout the book, there is a greater emphasis on, and description of effect size. The support materials provided on the website have also been updated, including many new and improved videos showing how to calculate statistics, how to read and interpret the appendices, and how to understand output produced by SPSS (and in R for some of the statistics). PowerPoint and text summaries of each chapter, work problems with solutions, and a set of test questions are also provided on the website.

Acknowledgments

I would like to sincerely thank the reviewers who provided their time and expertise reading previous drafts of this book and offered very helpful feedback. I could not fit all of your suggestions into this new edition, but I incorporated many of them and the book is better as a result of your hard work. Thanks to Georgette Enriquez at Taylor & Francis for her patience, help, and guidance. As always, students and colleagues at Santa Clara University made valuable contributions to this book, so thank you to Stephanie Zhi, Alexa Desanctis, and Elwood Mills. Thanks, again, to my children for their patience as I worked on this book. Finally, thanks to you readers for using this book. We are in this statistics struggle together.

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

About the Author

Timothy C. Urdan is a Professor in the Department of Psychology at Santa Clara University. He received his Ph.D. from the Combined Program in Education and Psychology at the University of Michigan, his Master's degree in Education from Harvard University, and his B.A. in Psychology from U.C. Berkeley. He conducts research on student motivation, classroom contexts, and teacher identity. He serves on the editorial boards of several journals and is the co-editor of the *Advances in Motivation and Achievement* book series. He is a fellow of the American Psychological Association and lives in Berkeley, California.

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

Quick Guide to Statistics, Formulas, and Degrees of Freedom

Statistic	Symbol	When you use it	Formula	Degrees of freedom (*df*)
Mean	$\bar{X}$, μ	To find the average of a distribution.	$\bar{X} = \frac{\Sigma X}{n}$, $\mu = \frac{\Sigma X}{N}$	
Standard deviation (sample)	s	To use sample data to estimate the average deviation in a distribution. It is a measure of variability.	$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$	
Standard deviation (population)	σ	To find the average of deviation in a distribution. It is a measure of variability.	$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$	
Standard score for individual (*z* score)	z	To find the difference between an individual score and the mean in standard deviation units.	$z = \frac{X - \mu}{\sigma}$ or $z = \frac{X - \bar{X}}{s}$	
Standard score for mean (*z* score)	z	To find difference between a sample mean and a population mean in standard error units.	$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$	
Standard error of the mean	$s_{\bar{x}}$, $\sigma_{\bar{x}}$	To find the average difference between the population mean and sample means when samples are of a given size and randomly selected.	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ or $s_{\bar{x}} = \frac{s}{\sqrt{n}}$	
1-sample *t* test	t	To determine whether the difference between a sample mean and the population mean is statistically significant.	$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$	$n - 1$
Independent samples *t* test	t	To determine whether the difference between two independent sample means is statistically significant.	$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}1-\bar{x}2}}$	$n_1 + n_2 - 2$
Dependent (paired) samples *t* test	t	To determine whether the difference between two dependent (i.e., paired) sample means is statistically significant.	$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}}$	$N - 1$ where N is the number of pairs of scores.
One-way ANOVA	F	Two determine whether the difference between two or more independent sample means is statistically significant.	$F = \frac{MS_b}{MS_e}$	$K - 1$, $N - K$ Where K is number of groups, N is number of cases across all samples

Statistic	Symbol	When you use it	Formula	Degrees of freedom (*df*)
Cohen's *d* (effect size)	*d*	To determine the size of an effect (e.g., difference between sample means) in standard deviation units.	$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}}$	
Confidence interval for sample mean	CI	To create an interval within which one is 95% or 99% certain the population parameter (e.g., mean, difference between means, correlation) is contained.	$CI_{95} = \bar{X} \pm (t_{95})(s_{\bar{x}})$	
Correlation coefficient (Pearson)	*r*	To calculate a measure of association between two intervally scaled variables.	$r = \frac{\Sigma(z_x z_y)}{N}$	
Coefficient of determination	r^2	To determine the percentage of variance in one variable that is explained by the other variable in a correlational analysis. It is a measure of effect size.	r^2	
t test for correlation coefficient	*t*	To determine whether a sample correlation coefficient is statistically significant.	$t = (r)\sqrt{\frac{N-2}{1-r^2}}$	$N - 2$ where N is number of cases in the sample
Regression coefficient	*b*	To determine the amount of change in the *Y* variable for every change of one unit in the *X* variable in a regression analysis.	$b = r \times \frac{s_y}{s_x}$	
Regression intercept	*a*	To determine the predicted value of *Y* when *X* equals zero in a regression analysis.	$a = \bar{Y} - b\bar{X}$	
Predicted value of *Y*	$\hat{Y}$	To determine the predicted value of *Y* for a given value of *X* in a regression analysis.	$\hat{Y} = bX + a$	
Chi square	χ^2	To examine whether the frequency of scores in various categories of categorical variables are different from what would be expected.	$\chi^2 = \Sigma\left(\frac{(O-E)^2}{E}\right)$	$R - 1, C - 1$
Mann-Whitney *U*	*U*	To compare the ranks of scores in two independent samples.	$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ *or* $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$	
Wilcoxon signed-rank	*w*	To compare the medians of two paired samples.	$w_1 = \Sigma(\text{neg. different ranks})$ $w2 = \Sigma(\text{pos. different ranks})$	
Kruskal-Wallis	*H*	To compare the ranks of three or more independent groups.	$\frac{12}{n(n+1)}\left(\frac{R12}{n1} + \frac{R22}{n2} + \frac{R32}{n3}\right) - 3(n+1)$	$K - 1$ Where K is the number of groups

Chapter 1

Introduction to Social Science Research Principles and Terminology

The Importance of Statistics and Research in Our Lives

Today, more than ever, your life is being guided by statistics because in today's world, there is more information, or data, about you that is being collected and used. Every time you go onto a website on the Internet like Google, Netflix, Amazon, or a social media site, you provide information about yourself to the companies that own those sites that can be used and sold to other companies. The San Francisco Bay Area, where I live, is home to the highest concentration of billionaires in the world. Sadly, I am not one of them. However, owners of many high-tech companies—Google, Facebook, Apple, Amazon—have made tens of millions, if not billions, of dollars by collecting valuable information from users.

One of my favorite expressions, and one that I tell my own children, goes like this: "If you are not paying for the product, you are the product." What this means is that when you use a free website or an app, your personal information is collected by those sites and sold to other companies. Essentially, you are a product that companies want to buy. But what makes your information so valuable? It just so happens that the more a company knows about you, the better they are able to show you the products that you are most likely to buy. When you login to Instagram and like a video of a teenage boy riding a skateboard, you are providing the following information to the company that owns Instagram (Facebook): you are most likely a boy between the ages of 10 and 20, you may like skateboarding yourself, and you probably like shoes that skateboarders usually like to wear. If you like several other pictures and videos of skateboarders, it is even more likely that you fit these characteristics. Facebook can now sell this information about you, and everyone else who liked skateboarding photos, to skateboard makers and shoe brands that make shoes for skateboarders (e.g., Vans), and they can advertise their products directly to you.

The best way for companies to sell their products is to advertise them to people who are most likely to buy them. In the old days, before the Internet, advertisers mostly relied on media like television, radio, newspapers, and billboards for their advertising. Although companies collected some information to help them decide where and when to advertise, this information was not that accurate. For example, people who watch the news on television tend to be older, so companies that serve older people often advertise their products on news channels. You will see a lot of advertisements for drugs that older Americans use, such as medications for heart disease or diabetes, during these shows. However, advertisers do not know how many viewers of an evening newscast might be suffering from heart disease. They just know that viewers tend to be old. In contrast, someone on the Internet may search "heart disease remedies" on Google or go onto Facebook and announce their heart disease or may like their friends' posts involving heart disease. This kind of specific information allows

DOI: 10.4324/9781003006459-1

companies to target their advertising much more precisely to those who may be most interested in using their products, and this gives companies a better return on their advertising dollars.

The Age of Analytics

All the data that companies use to better target their advertising are analyzed statistically to make predictions about who is most likely to respond to the advertising and purchase their products. Using data to predict future behavior—such as purchasing a product, exercising, or watching a movie on Netflix—is known as predictive analytics. Data and statistics have long been used to predict future outcomes. For example, insurance companies have used data about their customers age, gender, driving record, and health behaviors to predict things such as their likelihood of getting into an automobile accident or dying before a certain age. These analytics are used to determine how much money individuals pay for insurance, or if they can even buy insurance at all. Today, with the tremendous increase in the amount of data that people share about themselves via the Internet, predictive analytics have become much more precise, powerful, and ubiquitous. Statistical analytics are used to recommend which videos you should watch, which products you should buy, who you should connect with on social media, when you should go to the doctor, and even for whom you should vote.

The Case of Cambridge Analytica

In 2014, a company named Cambridge Analytica used the data of 50 million Facebook users to create detailed profiles about the political orientations of these users. The data of these users included information about their age, gender, friendship network, posts on Facebook, and likes of other users' posts. These data were analyzed statistically and used to decide which political ads and posts users would see on Facebook. The presidential campaign of Donald Trump used this information to send highly targeted ads to users who were deemed, on the basis of the data collected via Facebook, to be voters who could be persuaded to vote for Trump. One of the interesting things about this case, besides how sophisticated the data analysis was, was that the 50 million Facebook users were not even aware that their data were being used for this purpose. This case provided a fascinating example of how much information almost all of us provide about ourselves through our networks of friends and our online activity, and how these data can be used to create analytical models that can then be used to influence our behavior through carefully targeted advertising. For most of us, the whole process of data collection, analytics, and behavioral manipulation occurs without our awareness or understanding of the process.

As statistics play an increasingly prominent role in our lives, it is important to understand the language of statistics and research. In this chapter, I introduce some of the basic terminologies used in research, terms that will be used repeatedly throughout this book as we examine several of the statistical methods that are used by social scientists.

Terminology Used in Statistics and Research Methods

When I was in graduate school, one of my statistics professors often repeated what passes, in statistics, for a joke: "If this is all Greek to you, well that's good." Unfortunately, most of the class was so lost we didn't even get the joke. The world of statistics and research in the social sciences, like any specialized field, has its own terminology, language, and conventions. In this chapter, I review some of the fundamental research principles and terminologies, including the distinction between samples and populations, methods of sampling, types of variables, the distinction between inferential and descriptive statistics, and a brief description about different types of research designs.

Populations, Samples, Parameters, and Statistics

A **population** is an individual or a group that represents *all* the members of a certain group or a category of interest. A sample is a subset drawn from the larger population (see Figure 1.1). For example, suppose that I wanted to know the average income of the current full-time employees at Google. There are two ways that I could find this average. First, I could get a list of every full-time employee at Google and find out the annual income of each member on this list. Because this list contains every member of the group that I am interested in, it can be considered a population. If I were to collect these data and calculate the **mean**, I would have generated a **parameter**, because a parameter is a value generated from, or applied to, a population. Another way to generate the mean income of the full-time employees at Google would be to randomly select a subset of employee names from my list and calculate the average income of this subset. The subset is known as a **sample** (in this case, it is a **random sample**), and the mean that I generate from this sample is a type of **statistic**. Statistics are values derived from sample data, whereas parameters are values that are either derived from, or applied to, population data.

It is important to keep a couple of things in mind about samples and populations. First, a population does not need to be large to count as a population. For example, if I wanted to know the average height of the students in my statistics class this term, then all of the members of the class (collectively) would comprise the population. If my class only has five students, then my population only has five cases. Second, populations (and samples) do not necessarily include people. For example, I can calculate the average age of the dogs that visited a veterinary clinic in the last year. The population in this study is made up of dogs, not people. Similarly, I may want to know the total amount of carbon monoxide produced by Ford vehicles that were assembled in the U.S. during 2005. In this example, my population is cars, but not all cars—it is limited to Ford cars, and only those actually assembled in a single country during a single calendar year.

Third, the researcher generally defines the population, either explicitly or implicitly. In the examples above, I defined my populations (of dogs and cars) explicitly. Often, however, researchers define their populations less clearly. For example, a researcher may conduct a study with the aim of examining the frequency of depression among adolescents. The researcher's sample, however, may include only a group of 15-year-olds who visited a mental health service provider in Connecticut in a given year. This presents a potential problem and leads directly into the fourth and final little thing to keep in mind about samples and populations:

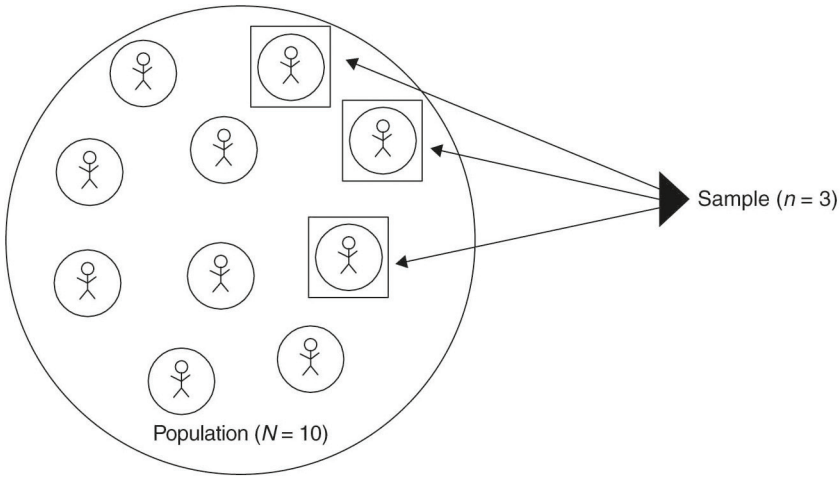

FIGURE 1.1 A population and a sample drawn from the population.

Samples are not necessarily good representations of the populations from which they were selected. In the example about the rates of depression among adolescents, notice that there are two potential populations. First, there is the population identified by the researcher and implied in the research question: all adolescents. However, notice that the population of adolescents is a very large group, including all human beings, in all countries, between the ages of, say, 13 and 20. Second, there is the much more specific population that was defined by the sample that was selected: 15-year-olds who visited a mental health service provider in Connecticut during a given year. In Figure 1.1, I offer a graphic that illustrates the concept of a sample of 3 individuals being selected from a population of 10 individuals.

Inferential and Descriptive Statistics

Why is it important to determine which of these two populations is of interest in this study? Because the consumer of this research must be able to determine how well the results from the sample **generalize** to the larger population. Clearly, depression rates among 15-year-olds who visit mental health service providers in Connecticut may be different from other adolescents. For example, adolescents who visit mental health service providers may, on average, be more depressed than those who do not seek the services of a psychologist. Similarly, adolescents in Connecticut may be more depressed, as a group, than adolescents in California, where the sun shines and Mickey Mouse keeps everyone smiling. Perhaps 15-year-olds, who have to suffer the indignities of beginning high school without yet being able to legally drive, are more depressed than their 16-year-old, driving peers. In short, there are many reasons to suspect that the adolescents who were *not* included in the study may differ in their depression rates from adolescents who were included in the study. When such differences exist, it is difficult to apply the results garnered from a sample to the larger population. In research terminology, the results may not generalize from the sample to the population, particularly if the population is not clearly defined.

So why is generalizability important? To answer this question, I need to explain the distinction between **descriptive** and **inferential** statistics. Descriptive statistics apply only to the members of a sample or population from which data have been collected. In contrast, inferential statistics refer to the use of sample data to reach some conclusions (i.e., make some inferences) about the characteristics of the larger population that the sample is supposed to represent. Although researchers are sometimes interested in simply describing the characteristics of a sample, for the most part, we are much more concerned with what our sample tells us about the population from which the sample was drawn. In the depression study, the researcher does not care so much about the depression levels of the sample *per se*. Rather, the data from the sample are used to reach some conclusions about the depression levels of adolescents *in general*. However, to make the leap from sample data to inferences about a population, one must be very clear about whether the sample accurately represents the population. If the sample accurately represents the population, then observations in the sample data should hold true for the population. But if the sample is not truly representative of the population, we cannot be confident that conclusions based on our sample data will apply to the larger population. An important first step in this process is to clearly define the population that the sample is alleged to represent.

Sampling Issues

There are several ways researchers can select samples. One of the most useful, but also the most difficult, is **random sampling**. In statistics, the term *random* has a much more specific meaning than the common usage of the term. It does not mean haphazard. In statistical jargon, *random* means that every member of a defined population has an equal chance of being selected into a sample. The major benefit of random sampling is that any differences between the sample and the population from which the sample was selected will not

be systematic. Notice that in the depression study example, the sample differed from the population in important, *systematic* (i.e., nonrandom) ways. For example, the researcher most likely systematically selected adolescents who were more likely to be depressed than the average adolescent because she selected those who had visited mental health service providers. Although randomly selected samples may differ from the larger population in important ways (especially if the sample is small), these differences are due to chance rather than to a systematic bias in the selection process.

Representative sampling is another way of selecting cases for a study. With this method, the researcher purposely selects cases so that they will match the larger population on specific characteristics. For example, if I want to conduct a study examining the average annual income of adults in San Francisco, by definition my population is "adults in San Francisco." This population includes a number of subgroups (e.g., different ethnic and racial groups, men and women, retired adults, disabled adults, parents, single adults, etc.). These different subgroups may be expected to have different incomes. To get an accurate picture of the incomes of the adult population in San Francisco, I may want to select a sample that represents the population well. Therefore, I would try to match the percentages of each group in my sample with those in my population. For example, if 15 percent of the adult population in San Francisco is retired, I would select my sample in a manner that included 15 percent retired adults. Similarly, if 55 percent of the adult population in San Francisco is male, 55 percent of my sample should be male. With random sampling, I may get a sample that looks like my population or I may not. But with representative sampling, I can ensure that my sample looks similar to my population on some important variables. This type of sampling procedure can be costly and time-consuming, but it increases my chances of being able to generalize the results from my sample to the population.

Another common method of selecting samples is called **convenience sampling**. In convenience sampling, the researcher generally selects participants on the basis of proximity, ease of access, and willingness to participate (i.e., convenience). For example, if I want to do a study on the achievement levels of eighth-grade students, I may select a sample of 200 students from the nearest middle school to my office. I might ask the parents of 300 of the eighth-grade students in the school to participate, receive permission from the parents of 220 of the students, and then collect data from the 200 students that show up at school on the day I hand out my survey. This is a convenience sample. Although this method of selecting a sample is clearly less labor-intensive than selecting a random or representative sample, that does not necessarily make it a bad way to select a sample. If my convenience sample does not differ from my population of interest *in ways that influence the outcome of the study*, then it is a perfectly acceptable method of selecting a sample.

To illustrate the importance of a sampling method in research, I offer two examples of problematic sampling methods that have led to faulty conclusions. First, a report by the American Chemical Society (2002) noted that several beach closures in southern California were caused by faulty sampling methods. To test for pollution levels in the ocean, researchers often take a single sample of water and test it. If the pollution levels are too high in the sample, the beach is declared unsafe and is closed. However, water conditions change very quickly, and a single sample may not accurately represent the overall pollution levels of water at the beach. More samples, taken at different times during the day and from different areas along the beach, would have produced results that more accurately represented the true pollution levels of the larger area of the beach, and there would have been fewer beach closures.

The second example involves the diagnosis of heart disease in women. For decades, doctors and medical researchers considered heart disease to be a problem only for men. As a result, the largest and most influential studies included only male samples (Doshi, 2015). Two consequences of this failure to include women in the samples that were researched were that doctors were less likely to order testing for heart disease for their female patients than

their male patients, and the symptoms of heart disease and cardiac failure among women, which are often different from those of men, were not understood. Many women who could have had their heart disease treated early or their symptoms of a cardiac arrest quickly diagnosed died because women were not included in the samples for research on heart disease and heart attacks. The population of people with heart disease clearly includes women, so the samples that included only men were not representative of the population.

Types of Variables and Scales of Measurement

In social science research, a number of terms are used to describe different types of variables. A **variable** is pretty much anything that can be codified and has more than a single value (e.g., income, gender, age, height, attitudes about school, score on a measure of depression, etc.). A **constant**, in contrast, has only a single score. For example, if every member of a sample is male, the "gender" category is a constant. Types of variables include **quantitative** (or **continuous**) and **qualitative** (or **categorical**). A quantitative variable is one that is scored in such a way that the numbers, or values, indicate some sort of amount. For example, height is a quantitative (or continuous) variable because higher scores on this variable indicate a greater amount of height. In contrast, qualitative variables are those for which the assigned values do not indicate more or less of a certain quality. If I conduct a study to compare the eating habits of people from Maine, New Mexico, and Wyoming, my "state" variable has three values (e.g., 1 = Maine, 2 = New Mexico, 3 = Wyoming). Notice that a value of 3 on this variable is not *more* than a value of 1 or 2—it is simply *different*. The labels represent qualitative differences in location, not quantitative differences. A commonly used qualitative variable in social science research is the **dichotomous variable**. This is a variable that has two different categories (e.g., child or adult, left-handed or right-handed).

In social science research, there are four different scales of measurement for variables: nominal, ordinal, interval, and ratio. A **nominally scaled variable** has different categories (e.g., male and female, Experimental Group 1, Experimental Group 2, Control Group, etc.). **Ordinal variables** are those whose values are placed in a meaningful order, but the distances between the values are not equal. For example, if I wanted to know the 10 richest people in America, in order from the wealthiest to the 10th richest, the wealthiest American would receive a score of 1, the next richest a score of 2, and so on through 10. Notice that while this scoring system tells me where each of the wealthiest 10 Americans stands in relation to the others (e.g., Bill Gates is 1, Michael Bloomberg is 8, etc.), it does not tell me how much *distance* there is between each score. So while I know that the wealthiest American is richer than the second wealthiest, I do not know if he has one dollar more or one billion dollars more. Variables measured with an **interval** scale have values that have order, but they also have equal distances between each unit on the scale. For example, businesses often survey their customers to gain information about how satisfied they are with the service they received. They may be asked to rate the service on a scale from 1 to 10, and this kind of rating scale is an interval scale of measurement. On such surveys, the distance between each number is presumed to be equal, such that a score of 10 would indicate twice as much satisfaction than a score of 5.[1] Variables measured using a **ratio** scale of measurement have the same properties as intervally scaled variables, but they have one additional property: Ratio scales can have a value of zero, while interval scales do not. A great deal of social science research employs measures with no zero value, such as attitude and beliefs surveys (e.g., "On a scale from 1 to 5, how much do you like orange soda?"). Examples of ratio scaled variables include temperatures (e.g., Celsius, Fahrenheit), income measured in dollars, measures of

1 There has been quite a bit of debate about whether it is accurate to treat these kinds of attitudinal scales as ratio variables. It is not clear whether people really think of the intervals between the numbers as being equal in size. Nonetheless, researchers typically treat these kinds of attitudinal measures as intervally scaled when calculating statistics that require variables using either interval or ratio scales.

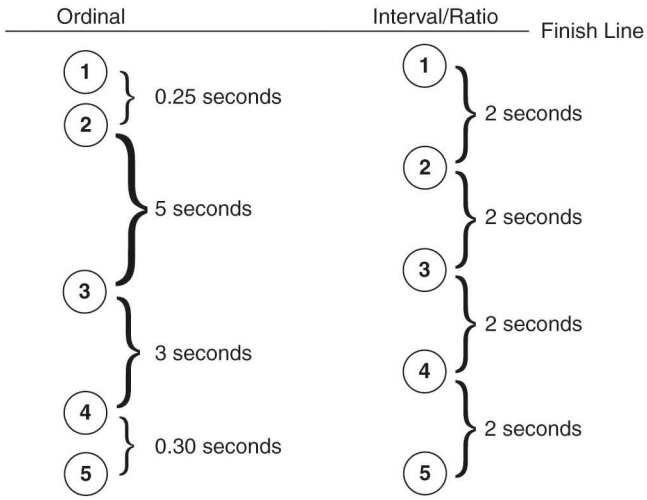

FIGURE 1.2 Difference between ordinal and interval/ratio scales of measurement.

weight and distance, and many others. Figure 1.2 illustrates a critical difference between ordinal and interval or ratio scales of measurement: Ordinal scales don't provide information about the distance between the units of measurement, but interval and ratio scales do.

One useful way to think about these different kinds of variables is in terms of how much information they provide. While nominal variables only provide labels for the different categories of the variable, ordinal variables offer a bit more information by telling us the order of the values. Variables measured using interval scales provide even more information, telling us both the order of the values and the distance between the values. Finally, variables measured with ratio scales add just a little bit more information by including the value of zero in its range of possible values. Figure 1.3 provides a graphic to help you think about the information provided by each of these four types of variables.

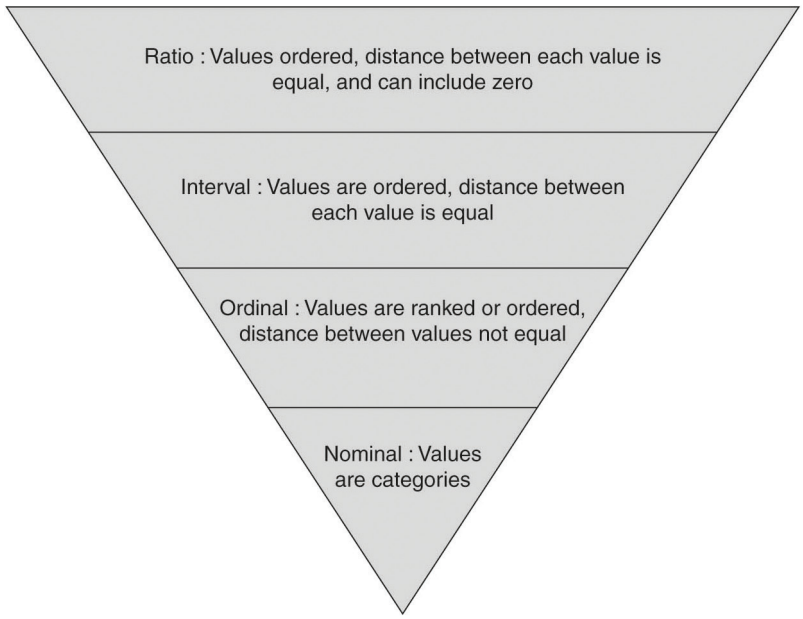

FIGURE 1.3 Hierarchical arrangement of scales of measurement.

Research Designs

There are a variety of research methods and designs employed by social scientists. Sometimes researchers use an **experimental design**. In this type of research, the experimenter divides the sample into different groups and then compares the groups on one or more variables of interest. For example, I may want to know whether my newly developed mathematics curriculum is better than the old one. I select a sample of 40 students and, using **random assignment**, teach 20 students a lesson using the old curriculum and the other 20 using the new curriculum. Then I test each group to see which group learned more mathematical concepts and find that, on average, students taught with the new curriculum had higher test scores than did students taught with the old math curriculum. This method of random assignment to groups and testing the effects of a particular treatment is known as a **randomized control trial (RCT)** experiment. It has been used for years in medical and laboratory research in the physical and social sciences, and in recent years has been used to great effect in the social sciences outside of laboratory settings. For example, Walton and Cohen (2011) used an RCT design to examine whether a brief intervention could increase first-year college students' feeling of belonging and decrease their feelings of isolation at a university. They randomly assigned about half of their participants to receive their intervention while the other half did not, then they compared the two groups and found that those who received the intervention had better psychological and academic outcomes. By assigning students to the two groups using random assignment, it is hoped that any important differences between the two groups will be distributed evenly between the two groups and that any differences in outcomes between the two groups are due to the experimental treatment that was given to one group but not the other. Of course, this may not be true.

A **quasi-experimental research design** is quite similar to an experimental design. Both of these designs typically involve manipulating a variable to see if that variable has an effect on an outcome. In addition, both research designs include some sort of random assignment. The major difference is that in a quasi-experimental design, the research usually occurs outside the lab, in a naturally occurring setting. In the earlier example used to illustrate the experimental design, I could have had the students come to my research lab and conducted my study in a controlled setting so that I could tightly control all of the conditions and make them identical between the two groups, except for the math curriculum that I used. In a quasi-experimental study, I might find two existing classrooms with 20 students each and ask the teacher in one classroom to use the old math curriculum and the teacher in another classroom to use the new curriculum. Instead of randomly assigning students to these two classrooms (which is difficult to do in a real school), I might randomly select which classroom gets the new curriculum and which one uses the old. I could take steps to try to minimize the differences between the two classrooms (e.g., conduct the study in two different classes of students that are taught by the same teacher, try to find two classrooms that are similar in terms of the gender composition of the students, etc.), but generally speaking it is more difficult to control the conditions in a quasi-experimental design than an experimental design. The benefit of a quasi-experimental design, however, is that it allows the researcher to test the effects of experimental manipulations in more natural, real-world conditions than those found in a research laboratory.

Correlational research designs are also a common method of conducting research in the social sciences. In this type of research, participants are not usually randomly assigned to groups. In addition, the researcher typically does not actually manipulate anything. Rather, the researcher simply collects data on several variables and then conducts some statistical analyses to determine how strongly different variables are related to each other. For example, I may be interested in whether employee productivity is related to how much employees sleep (at home, not on the job!). So I select a sample of 100 adult workers, measure their productivity at work, and measure how long each employee sleeps on an

average night in a given week. I may find that there is a strong relationship between sleep and productivity. Now, logically, I may want to argue that this makes sense because a more rested employee will be able to work harder and more efficiently. Although this conclusion makes sense, it is too strong a conclusion to reach based on my correlational data alone. Correlational studies can only tell us whether variables are related to each other—they cannot lead to conclusions about *causality*. After all, it is possible that being more productive at work *causes* longer sleep at home. Getting one's work done may relieve stress and perhaps even allow the worker to sleep a little longer in the morning, both of which create longer sleep duration.

Experimental research designs are good because they allow the researcher to isolate specific **independent variables** that may cause variation, or changes, in **dependent variables**. In the example above, I manipulated the independent variable of the mathematics curriculum and was able to reasonably conclude that the type of math curriculum used affected students' scores on the dependent variable, the test scores. The primary drawbacks of experimental designs are that they are often difficult to accomplish in a clean way and they often do not generalize to real-world situations. For example, in my study above, I cannot be sure whether it was the math curricula that influenced the test scores or some other factor, such as a preexisting difference in the mathematical abilities of the two groups or differences in the teaching styles that had nothing to do with the curricula, but could have influenced the test scores (e.g., the clarity or enthusiasm of the teacher). The strengths of correlational research designs are as follows: they are often easier to conduct than experimental research, allow for the relatively easy inclusion of many variables, and allow the researcher to examine many variables simultaneously. The principal drawback of correlational research is that such research does not allow for the careful controls necessary for drawing conclusions about causal associations between variables.

Making Sense of Distributions and Graphs

Statisticians spend a lot of time talking about **distributions**. A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from the smallest to largest, and then they can be presented graphically. Because distributions are so important in statistics, I want to give them some attention early on in the book and show you several examples of different types of distributions and how they are depicted in **graphs**. Note that later in this book, there are whole chapters devoted to several of the most commonly used distributions in statistics, including the **normal distribution** (Chapters 4 and 5), ***t* distributions** (Chapter 8 and parts of Chapter 7), ***F* distributions** (Chapters 9–11), and **chi-square** distributions (Chapter 14).

Let's begin with a simple example. Suppose that I am conducting a study of voters' attitudes and I select a random sample of 500 voters for my study. One piece of information I might want to know is the political affiliation of the members of my sample. I ask them if they are Republicans, Democrats, or Independents and I find that 45 percent of my sample list themselves as Democrats, 40 percent report being Republicans, and 15 percent identify themselves as Independents. Notice that political affiliation is a nominal, or categorical, variable. Because nominal variables are variables with categories that have no numerical weight, I cannot arrange my scores in this distribution from highest to lowest. The value of being a Republican is not more or less than the value of being a Democrat or an Independent—they are simply different categories. So rather than trying to arrange my data from the lowest to the highest value, I simply leave them as separate categories and report the percentage of the sample that falls into each category.

There are many different ways to graph this distribution, including a pie chart, bar graph, column graph, different sized bubbles, and so on. The key to selecting the appropriate graphic

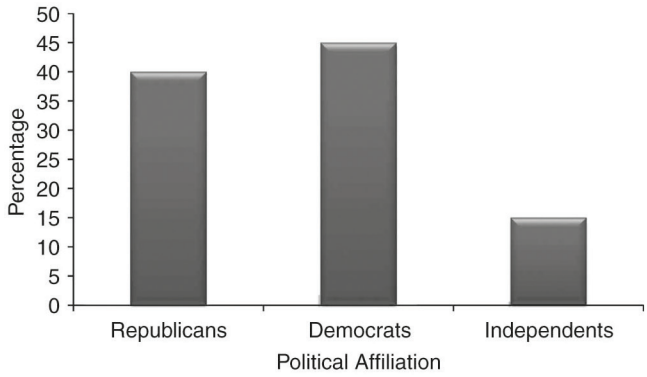

FIGURE 1.4 Column graph showing the distribution of Republicans, Democrats, and Independents.

is to keep in mind that the purpose of the graph is to make the data easy to understand. For my distribution of political affiliation, I have created two different graphs. Both are fine choices because both offer very clear and concise summaries of the distribution and are easy to understand. Figure 1.4 depicts the distribution as a column graph, and Figure 1.5 presents the data in a pie chart. Which graphic is best for these data is a matter of personal preference. As you look at Figure 1.4, notice that the *X* axis (the horizontal one) shows the party affiliations: Democrats, Republicans, and Independents. The *Y* axis (the vertical one) shows the percentage of the sample. You can see the percentages in each group and, just by quickly glancing at the columns, you can see which political affiliation has the highest percentage of this sample and get a quick sense of the differences between the party affiliations in terms of the percentage of the sample. In my opinion, the pie chart in Figure 1.5 shows the same information, but in a slightly more striking and simple manner.

Sometimes, researchers are interested in examining the distributions of more than one variable at a time. For example, suppose I wanted to know about the association between hours spent watching television and hours spent doing homework. I am particularly interested in how this association looks across different countries. So I collect data from samples of high school students in several different countries. Now I have distributions on two different variables across five different countries (the U.S., Mexico, China, Norway, and Japan). To compare these different countries, I decide to calculate the average, or **mean** (see Chapter 2), for each country on each variable. Then I graph these means using a column graph, as shown

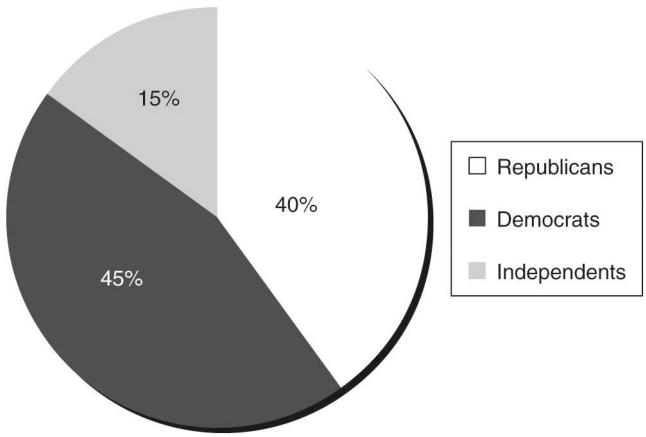

FIGURE 1.5 Pie chart showing the distribution of Republicans, Democrats, and Independents.

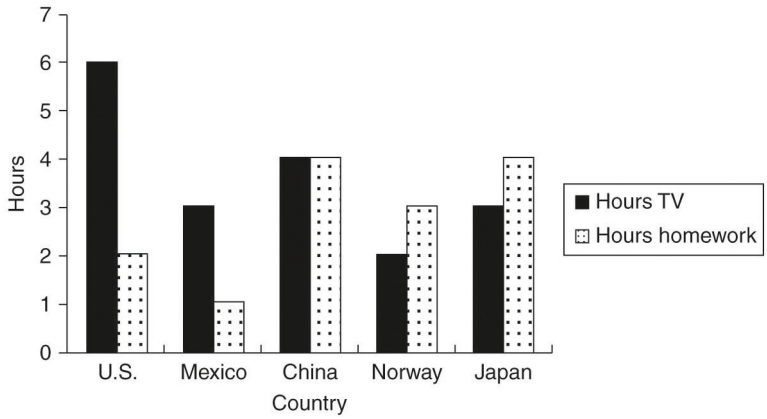

FIGURE 1.6 Average hours of television viewed and time spent on homework in five countries.

in Figure 1.6 (note that these data are fictional—I made them up). As this graph clearly shows, the disparity between the average amount of television watched and the average hours of homework completed per day is widest in the U.S. and Mexico and virtually nonexistent in China. In Norway and Japan, high school students actually spend more time on homework than they do watching TV, according to my fake data. Notice how easily this complex set of data is summarized in a single graph.

Another common method of graphing a distribution of scores is the line graph, as shown in Figure 1.7. Suppose that I select a random sample of 100 first-year college students who have just completed their first term. I ask each of them to tell me the final grades they received in each of their classes and then I calculate a grade point average (GPA) for each of them. Finally, I divide the GPAs into six groups: 1–1.4, 1.5–1.9, 2.0–2.4, 2.5–2.9, 3.0–3.4, and 3.5–4.0. When I count up the number of students in each of these GPA groups and graph these data using a line graph, I get the results presented in Figure 1.7. Notice that along the *X* axis, I have displayed the six different GPA groups. On the *Y* axis, I have the **frequency**, typically denoted by the symbol ***f***. In this graph, the *Y* axis shows how many students are in each GPA group. A quick glance at Figure 1.7 reveals that there were quite a few students (13) who really struggled in their first term in college, accumulating GPAs between 1.0 and 1.4. Only one student was in the next group category: 1.5–1.9. From there, the number of students in each GPA group generally goes up, with roughly 30 students in the 2.0–2.9 GPA categories and about 55 students in the 3.0–4.0 GPA categories. A line graph like this offers a quick way to see trends in data, either over time or across categories. In this example with GPA, we can see that the general trend is to find more students in the

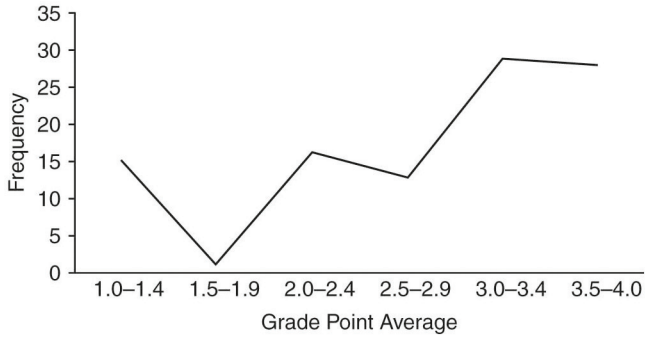

FIGURE 1.7 Line graph showing the frequency of students in different GPA groups.

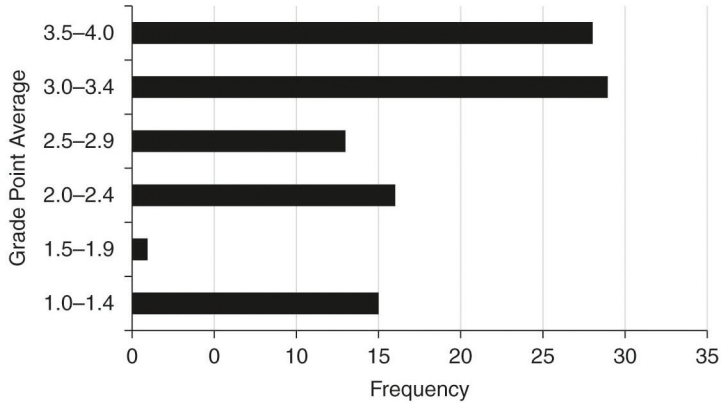

FIGURE 1.8 Bar graph showing the frequency of students in different GPA groups.

higher GPA categories, plus a fairly substantial group that is really struggling. In Figure 1.8, the same data are presented in a bar graph. Which graph gives you a clearer picture of the data?

Column graphs are another method to show trends in data. In Figure 1.9, I present a stacked column graph. This graph allows me to show several pieces of information in a single graph. For example, in this graph, I am illustrating the occurrence of two different kinds of crimes, property and violent, across the period from 1990 to 2007. On the *X* axis, I have placed the years, moving from the earliest (1990) to latest (2007) as we look from the left to the right. On the *Y* axis, I present the number of crimes committed per 100,000 people in the U.S. When presented in this way, several interesting facts jump out. First, the overall trend from 1990 to 2007 is a pretty dramatic drop in crime. From a high of nearly 6,000 crimes per 100,000 people in 1991, the crime rate dropped to well under 4,000 per 100,000 people in 2007. That is a drop of nearly 40 percent. The second noteworthy piece of information that is obvious from the graph is that violent crimes (e.g., murder, rape, assault) occur much less frequently than crimes against property (e.g., burglary, vandalism, arson) in each year of the study.

Notice that the graph presented in Figure 1.9 makes it easy to see that there has been a drop in crime *overall* from 1990 to 2007, but it is not so easy to tell whether there has been much of a drop in the violent crime rate. That is because violent crime makes up a much smaller percentage of the overall crime rate than property crimes, so the scale used in the

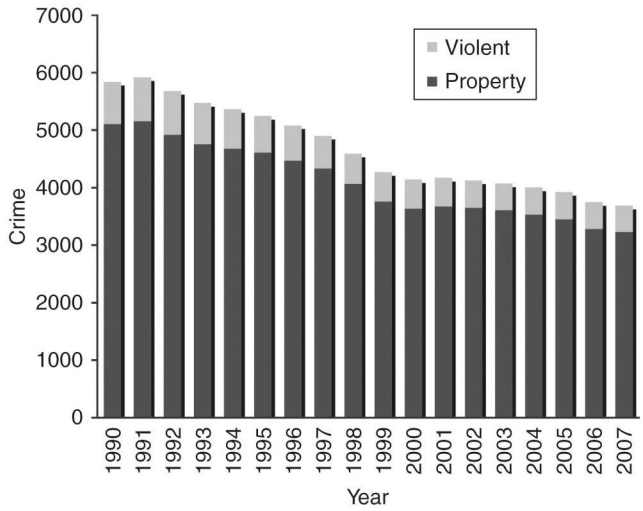

FIGURE 1.9 Stacked column graph showing crime rates from 1990 to 2007.

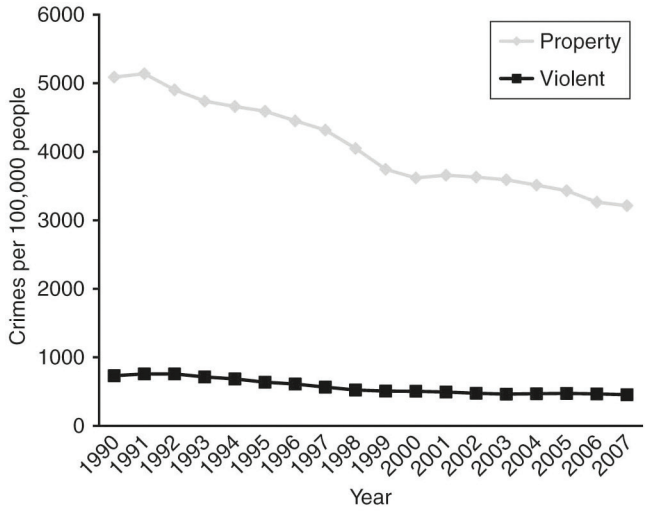

FIGURE 1.10 Line graph showing crime rates from 1990 to 2007.

Y axis is pretty large. This makes the top of the columns, the parts representing violent crimes, look quite small. To get a better idea of the trend for violent crimes over time, I created a new graph, which is presented in Figure 1.10.

In this new figure, I have presented the exact same data that were presented in Figure 1.9, this time as a line graph. The line graph separates violent crimes from property crimes completely, making it easier to see the difference in the frequency of the two types of crimes. Again, this graph clearly shows the drop in property crime over the years. However, notice that it is still difficult to tell whether there was much of a drop in violent crime over time. If you look very closely, you can see that the rate of violent crime dropped from about 800 per 100,000 people in 1990 to about 500 per 100,000 people in 2007. This is an impressive drop in the crime rate, but we have had to work too hard to see it. Remember: The purpose of a graph is to make the interesting facts in the data easy to see. If you have to work hard to see it, the graph is not that great.

The problem with Figure 1.10, just as with Figure 1.9, is that the scale on the *Y* axis is too large to clearly show the trends for violent crime rates over time. To fix this problem, we need a scale that is more appropriate for the violent crime rate data. I created one more graph that includes the data for violent crimes only, without the property crime data (Figure 1.11).

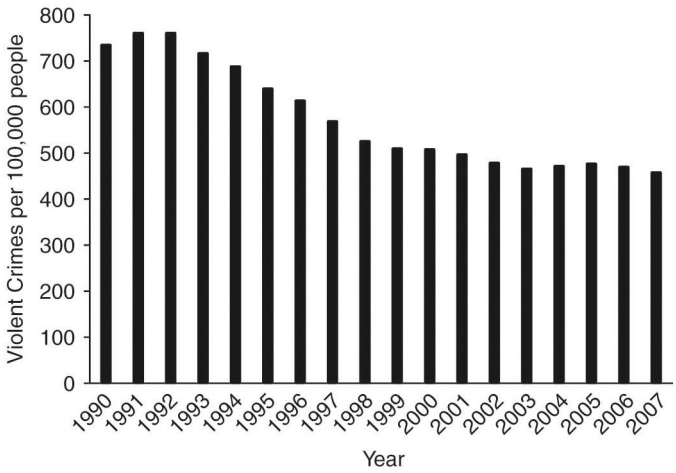

FIGURE 1.11 Column graph showing violent crime rate from 1990 to 2007.

Instead of using a scale from 0 to 6,000 or 7,000 on the *Y* axis, my new graph has a scale from 0 to 800 on the *Y* axis. In this new graph, a column graph, it is clear that the drop in violent crimes from 1990 to 2007 was also quite dramatic.

Any collection of scores on a variable, regardless of the type of variable, forms a distribution, and this distribution can be graphed. In this section of the chapter, several different types of graphs have been presented, and all of them have their strengths. The key, when creating graphs, is to select the graph that most clearly illustrates the data. When reading graphs, it is important to pay attention to the details. Try to look beyond the most striking features of the graph to the less obvious features, such as the scales used on the *X* and *Y* axes. As I discuss later (Chapter 11), graphs can be quite misleading if the details are ignored.

Wrapping Up and Looking Forward

The purpose of this chapter was to provide a quick overview of the basic principles and terminologies employed in social science research. With a foundation in the types of variables, experimental designs, and sampling methods used in social science research, it will be easier to understand the uses of the statistics described in the remaining chapters of this book. Now we are ready to talk statistics. It may still all be Greek to you, but that's not necessarily a bad thing.

For work problems with answers, please check out the book's Companion Website here: www.routledge.com/cw/urdan

References

Cadwalladar, C. & Graham-Harrison, E. (2018, March 17). How Cambridge Analytica turned Facebook "likes" into a lucrative political tool. *The Guardian*. https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm

Granville, K. (2018, March 19). Facebook and Cambridge Analytica: What you need to know as fallout widens. *The New York Times*. https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html