THIRD EDITION

Applied Statistics I

Basic Bivariate Techniques

REBECCA M. WARNER

Applied Statistics I

Third Edition

To my students: Past, present, and future.

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Applied Statistics I

Basic Bivariate Techniques

Third Edition

Rebecca M. Warner

Professor Emerita, University of New Hampshire

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Printed in the United States of America

ISBN 978-1-5063-5280-0

Acquisitions Editor: Helen Salmon
Editorial Assistant: Megan O'Heffernan
Content Development Editor: Chelsea Neve
Production Editor: Laureen Gleason
Copy Editor: Jim Kelly
Typesetter: Hurix Digital
Proofreader: Scott Oney
Indexer: Michael Ferreira
Cover Designer: Gail Buschman
Marketing Manager: Shari Countryman

This book is printed on acid-free paper.

20 21 22 23 24 10 9 8 7 6 5 4 3 2 1

BRIEF CONTENTS

DETAILED CONTENTS

PREFACE

The set of bivariate techniques covered in this book (analyses with one predictor and one outcome) are the same as those in most introductory textbooks. This book provides an applied perspective.

What does an applied perspective involve? Textbooks often use well-behaved data (without missing values, outliers, or violations of assumptions). This book introduces, early on, the idea that real data have problems. Discussion of ways in which actual practice differs from ideal situations helps students understand statistics in the context of real-world research. Here are examples: Textbooks describe random samples from clearly defined populations, while researchers often work with convenience samples. Textbooks usually present one significance test in isolation, whereas research reports often include numerous analyses, accompanied by increased risk for Type I error. This book includes discussion of these problems.

Each chapter begins with a simple question: What kinds of questions can this analysis answer? Chapters include fully worked examples with by-hand computation for small data sets, screenshots for SPSS menu selections and output, and results sections. Technical and supplemental information, including nonparametric alternatives, is provided in appendices at the ends of most chapters.

This book devotes less space to rarely used techniques (such as frequency polygons and methods to locate medians in grouped frequency distributions) and more space to real-world decisions made during data analysis (such as outlier detection and evaluation of distribution shape). Connections are made between design decisions and results; for instance, students will see that choice of dosage levels, control over within-group variance, and sample size influence the obtained magnitude of t and F ratios (along with sampling error, of course).

Traditional use of statistical significance tests is covered. However, consistent with the New Statistics guidelines, there is greater emphasis on confidence intervals, effect sizes, and the need to document decisions made during analysis. Limitations of p values are discussed in nontechnical terms. Discussion also focuses on common researcher behaviors that affect p values (e.g., running numerous analyses and reporting only a few).

A distinction is made between "statistical significance" and practical or clinical or everyday "significance" or importance (i.e., a small p value does not necessarily indicate a strong treatment effect).

Students are encouraged to think in terms of "degree of belief" rather than yes/no decisions. To paraphrase David Hume, a wise person proportions belief to the evidence.

Notation and presentation are consistent with Volume II (*Applied Statistics II: Multivariable and Multivariate Techniques* [Warner, 2020]).

DIGITAL RESOURCES

Instructor and student support materials are available for download from **edge.sagepub.com/warner3e**. **SAGE edge** offers a robust online environment featuring an impressive array of free

tools and resources for review, study, and further explorations, enhancing use of the textbook by students and teachers.

SAGE edge for students provides a personalized approach to help you accomplish your coursework goals in an easy-to-use learning environment. Resources include the following:

- Mobile-friendly **eFlashcards** to strengthen your understanding of key terms
- **Data sets** for completing in-chapter exercises
- Links to **web resources**, including video tutorials and creative lectures, to support and enhance your learning

SAGE edge for instructors supports your teaching by providing resources that are easy to integrate into your curriculum. SAGE edge includes the following:

- Editable, chapter-specific **PowerPoint® slides** covering key information that offer you flexibility in creating multimedia presentations
- **Test banks** for each chapter with a diverse range of prewritten questions, which can be loaded into your LMS to help you assess students' progress and understanding
- **Tables and figures** pulled from the book that you can download to add to handouts and assignments
- **Answers to in-text comprehension questions**, perfect for assessing in-class work or take-home assignments

Finally, in response to feedback from instructors for R content to mirror the SPSS coverage in this book, SAGE has commissioned *An R Companion for Applied Statistics I* by Danney Rasco. This short supplement can be bundled with this main textbook.

The author welcomes communication from teachers, students, and readers; please e-mail her at rmw@unh.edu with comments, corrections, or suggestions.

ACKNOWLEDGMENTS

Writers depend on many people for intellectual preparation and moral support. My understanding of statistics was shaped by exceptional teachers, including the late Morris de Groot at Carnegie Mellon University, and my dissertation advisers at Harvard, Robert Rosenthal and David Kenny. Several people who have most strongly influenced my thinking are writers I know only through their books and journal articles. I want to thank all the authors whose work is cited in the reference list. Authors whose work has particularly influenced my understanding include Jacob and Patricia Cohen, Barbara Tabachnick, Linda Fidell, James Jaccard, Richard Harris, Geoffrey Keppel, and James Stevens.

Special thanks are due to reviewers who provided exemplary feedback on first drafts of the chapters:

For the first edition:

David J. Armor, *George Mason University*

Michael D. Biderman, *University of Tennessee at Chattanooga*

Susan Cashin, *University of Wisconsin–Milwaukee*

Ruth Childs, *University of Toronto*

Young-Hee Cho, *California State University, Long Beach*

Jennifer Dunn, *Center for Assessment*

William A. Fredrickson, *University of Missouri–Kansas City*

Robert Hanneman, *University of California, Riverside*

Andrew Hayes, *The Ohio State University*

Lawrence G. Herringer, *California State University, Chico*

Jason King, *Baylor College of Medicine*

Patrick Leung, *University of Houston*

Scott E. Maxwell, *University of Notre Dame*

W. James Potter, *University of California, Santa Barbara*

Kyle L. Saunders, *Colorado State University*

Joseph Stevens, *University of Oregon*

James A. Swartz, *University of Illinois at Chicago*

Keith Thiede, *University of Illinois at Chicago*

For the second edition:

Diane Bagwell, *University of West Florida*

Gerald R. Busheé, *George Mason University*

Evita G. Bynum, *University of Maryland Eastern Shore*

Ralph Carlson, *The University of Texas Pan American*

John J. Convey, *The Catholic University of America*

Kimberly A. Kick, *Dominican University*

Tracey D. Matthews, *Springfield College*

Hideki Morooka, *Fayetteville State University*

Daniel J. Mundfrom, *New Mexico State University*

Shanta Pandey, *Washington University*

Beverly L. Roberts, *University of Florida*

Jim Schwab, *University of Texas at Austin*

Michael T. Scoles, *University of Central Arkansas*

Carla J. Thompson, *University of West Florida*

Michael D. Toland, *University of Kentucky*

Paige L. Tompkins, *Mercer University*

For the third edition:

Linda M. Bajdo, *Wayne State University*

Timothy Ford, *University of Oklahoma*

Beverley Hale, *University of Chichester*

Dan Ispas, *Illinois State University*

Jill A. Jacobson, *Queen's University*

Seung-Lark Lim, *University of Missouri, Kansas City*

Karla Hamlen Mansour, *Cleveland State University*

Paul F. Tremblay, *University of Western Ontario*

Barry Trunk, *Capella University*

I also thank the editorial and publishing team at SAGE, including Helen Salmon, Chelsea Neve, Megan O'Heffernan, and Laureen Gleason, who provided extremely helpful advice, support, and encouragement. Copy editor Jim Kelly merits special thanks for his attention to detail.

Many people provided moral support, particularly my late parents, David and Helen Warner; and friends and colleagues at UNH, including Ellen Cohn, Ken Fuld, Jack Mayer, and Anita Remig. I hope this book is worthy of the support they have given me. Of course, I am responsible for any errors and omissions that remain.

Last but not least, I want to thank all my students, who have also been my teachers. Their questions continually prompt me to search for better explanations—and I am still learning.

Dr. Rebecca M. Warner
Professor Emerita
Department of Psychology
University of New Hampshire

ABOUT THE AUTHOR

Rebecca M. Warner is Professor Emerita at the University of New Hampshire. She has taught statistics in the UNH Department of Psychology and elsewhere for 40 years. Her courses have included Introductory and Intermediate Statistics as well as seminars in Multivariate Statistics, Structural Equation Modeling, and Time-Series Analysis. She received a UNH Liberal Arts Excellence in Teaching Award, is a Fellow of both the Association for Psychological Science and the Society of Experimental Social Psychology, and is a member of the American Psychological Association, the International Association for Statistical Education, and the Society for Personality and Social Psychology. She has consulted on statistics and data management for the World Health Organization in Geneva, Project Orbis, and other organizations; and served as a visiting faculty member at Shandong Medical University in China. Her previous book, *The Spectral Analysis of Time-Series Data*, was published in 1998. She has published articles on statistics, health psychology, and social psychology in numerous journals, including the *Journal of Personality and Social Psychology*. She has served as a reviewer for many journals, including *Psychological Bulletin*, *Psychological Methods*, *Personal Relationships*, and *Psychometrika*. She received a BA from Carnegie Mellon University in social relations in 1973 and a PhD in social psychology from Harvard in 1978. She writes historical fiction and is a hospice volunteer along with her Pet Partner certified Italian greyhound Benny, who is also the world's greatest writing buddy.

CHAPTER 1

EVALUATING NUMERICAL INFORMATION

1.1 INTRODUCTION

In everyday use, **statistics** can refer to specific pieces of numerical information, such as average income for all employed persons in the United States. In science and technical fields, the term *statistics* more often describes techniques for analyzing and interpreting numerical information. Readers should not assume that all published numerical information is correct. **Numeracy** skills are needed to understand and evaluate how numerical information is collected, analyzed, and presented.

1.2 GUIDELINES FOR NUMERACY

A report published by the American Statistical Association's Committee on Guidelines for Assessment and Instruction in Statistics Education (GAISE College Report ASA Revision Committee, 2016) described numeracy skills as follows:

> Students should become critical consumers of statistically-based results reported in popular media, recognizing whether reported results reasonably follow from the study and analysis conducted. To be a critical consumer of statistically-based results, it is necessary to understand the components that produced them: the *design* of the investigation, the *data*, its *analysis*, and its *interpretation*. Identifying the *variables* in a study, which includes consideration of the measurement units, is a necessary step to inform judgments or comparisons. Identifying the *subjects* (*cases*, *observational units*) of a study and the *population* to which the results of an analysis can be *generalized* helps the consumer to recognize whether the reported results can reasonably support the conclusions claimed for an analysis. Being able to interpret displays of data (tables, graphs, and visualizations) and statistical analyses also informs the consumer about the reasonableness of the claims being presented. (Italics added)

Italicized terms in the preceding quotation identify components of the research and data analysis process; these are discussed further in Chapter 2 and research methods courses. This chapter briefly considers other fundamental issues in the communication of numerical information: (a) sources (or communicators), (b) types of evidence, (c) questions about generalizability and causal inference, (d) quality control mechanisms, (e) ethical responsibilities, and (f) degrees of belief.

1.3 SOURCE CREDIBILITY

1.3.1 Self-Interest or Bias

Communicators can be motivated by self-interest or bias. Self-interest is often clear in mass media; messages are often intended to influence audience beliefs or behaviors (such as voting or product purchases). Science communicators can also be motivated by self-interest; for instance, some researchers receive funding from alcohol or pharmaceutical companies, and their future funding may depend on research results. Many science journals require authors to declare potential conflicts of interest.

Self-interest of information providers is not always obvious. Many webpages offer "sponsored content": paid messages from advertisers that look like news articles but in fact promote the interests of advertisers. For instance, a new diet pill might be presented as "news" when in fact the article is an advertisement. *Communicator self-interest raises concerns about credibility of messages.*

1.3.2 Bias and "Cherry-Picking"

Communicators generally cannot (or do not) present all available information. Selection of information by communicators can be influenced by **confirmation bias**, a preference for information that confirms preexisting beliefs or ideas. Biased selection of evidence can be informally called **cherry-picking**. Information and ideas that are excluded may be as important as information that is included.

As an example of cherry-picking, suppose 20 studies show no association between consuming meat and cancer risk, and 3 studies do show an association. A journalist might report only the 3 studies that showed an association or might report only the single most recent study. Whether the bias was intentional or not, the article will not provide an accurate summary of research results.

When scientists write **literature reviews** (reviews of past research), they are expected to discuss all past relevant research.[1] Literature reviews are included in the introductions to most primary source research reports; literature reviews can also be stand-alone papers or books.

1.3.3 Primary, Secondary, and Third-Party Sources

An old game called "telephone" illustrates the problem of distance from a source. People form a line; the first person whispers a message to the second person, the second person whispers it to the third, and so forth. When the final message is compared with the original message, there are changes and distortions. Transmission of information can introduce errors because of each person's biases or misunderstandings.

In science, a **primary source** is a research report written by a researcher who has firsthand knowledge of behaviors and events in a study. Primary source reports (sometimes called articles or papers) are published in **science journals**.[2] Primary source data may also appear in books written for science audiences.

A **secondary source** is a description or summary of past research, created by someone who did not experience the reported data collection or observations firsthand. In many disciplines, secondary sources are scholarly books. Some journal articles are also secondary sources because they only review past research and do not present new data about which their authors have firsthand knowledge. Literature reviews in the introductions to science journal articles are secondhand discussions of past studies. (In the sciences, *literature* refers to past published research.)

Unfortunately, primary source reports are usually long and difficult to read (particularly for readers unfamiliar with statistics and technical terms). Language in research reports is

sometimes unnecessarily obscure. Some full-length science research reports are published on the web as open-access materials; anyone can view these. However, many publishers require fees or subscriptions for access. The consequence is that many people can't easily understand most primary source information in science and sometimes cannot even gain access to it.

Much content on websites for news organizations is **third-party content**. This is content written by someone who may have examined only secondary sources or other thirdhand content, such as news reports or press releases. Often, third-party content is authored by someone who has no technical knowledge of the research field and statistical methods. Examples include articles published by news organizations. These articles usually don't provide complete or accurate information about research results.

In the past, editors of prestigious newspapers required reporters to fact-check claims carefully. Increasingly, news reports on the web are paraphrases of, or uncritical reposting of, third-party content from other news sources. Some mass media news sources specifically disclaim responsibility for accuracy. Here is an example; many other news organizations post similar disclaimers:

> CNN is a distributor (and not a publisher or creator) of content supplied by third parties and users. . . . Neither CNN nor any third-party provider of information guarantees the accuracy, completeness, or usefulness of any content. . . . (CNN, 2018)

Communicators can provide better quality information when they are closer to original sources of information, and they are likely to provide better quality information when they assume responsibility for accuracy.

In everyday life, most of us rely on thirdhand information most of the time. Because so much of what we think we know is based on thirdhand information, we should not be overly confident about things we think we know.

1.3.4 Communicator Credentials and Skills

Communicators are more believable when they have training and background related to information in the message. Researchers generally have credentials that provide evidence of this training and background, including advanced degrees such as a PhD or MD, affiliations with respected organizations such as universities, and publications in high-quality science journals. Some journalists have strong credentials in science, but many do not. People who do not have training in statistics can easily misunderstand studies that use statistical terms such as *logistic regression* and *odds ratios*.

Celebrity status is not a meaningful credential. Famous media personalities, such as Dr. Oz[3] and other self-appointed lifestyle or health experts, may base recommendations on incomplete or incorrect information.

Scientific research reports include source information (authors, university affiliations, and so forth). News reports and websites sometimes do not include source information; they provide no basis to evaluate self-interest, distance from information source, and credentials. Guidelines for evaluation of websites are provided by Kiely and Robertson (2016) and Montecino (1998).

1.3.5 Track Record for Truth-Telling

There are independent, nonpartisan organizations that evaluate communicator track records for truth-telling in journalism, for example, the Pulitzer Prize–winning site www.politifact.com. PolitiFact rates statements as true, mostly true, half true, mostly false, false, and "pants on fire" (extremely false). Other respected fact-checking sites are www.snopes.com

and www.factcheck.org. These fact-checkers do the work that information consumers usually don't have the time to do.

Information published in scientific journals can be incorrect because of fraud; fraud in science is rare, but it has occurred. A notorious example was a claim by Andrew Wakefield that vaccines cause autism (discussed by Godlee, Smith, & Marcovitch, 2011). There are severe penalties for fraud or plagiarism in science, including forced retraction of publications, withdrawal of research funds, loss of reputation, and job dismissal. Rare instances of fraud in science can be identified by a web search for the researcher name and terms such as *fraud*. *Information consumers should be skeptical of information from sources with poor records for truth-telling.*

1.4 MESSAGE CONTENT

1.4.1 Anecdotal Versus Numerical Information

Anecdote means "story," often about an individual person or situation. First-person accounts are often called **testimonials**. Audiences may find narrative stories or anecdotes more persuasive and memorable than numerical information. There are many potential problems with **anecdotes (anecdotal evidence)**. Sometimes individual situations are not reported accurately (for example, advertisements for weight loss products often include falsified before and after photos). Even when anecdotal evidence is accurate, it is difficult to know whether the experience shown is generalizable: Has this experience happened to many other people, or was this a unique situation? Diet product advertisers are required to acknowledge this and typically do so in a tiny footnote: "Individual results may vary."

In science, a detailed report of an individual person or situation is called a **case study**. The study of unique cases, such as the brain damage suffered by railway worker Phineas Gage (Kihlstrom, 2010; Twomey, 2010) can be valuable. However, generalizability concerns are still relevant.

Anecdotal evidence can dramatize genuine problems. However, anecdotal evidence can also dramatize and promote incorrect beliefs. It is obviously easy to cherry-pick anecdotes. Supporting evidence in the form of systematic numerical information can provide a more accurate overview of evidence than anecdotal reports.

1.4.2 Citation of Supporting Evidence

In science, identification of outside sources of evidence is done by **citation**. Author names and years of publication are included in the text (to identify sources of ideas and evidence), and complete information to locate each source is included in a reference list. Citation has two purposes. First, it gives credit to others for their ideas and evidence; this avoids **plagiarism**, which occurs if authors present ideas or contributions of other people as if they were the authors' own new contributions. Second, it shows how the present study builds upon an existing body of evidence.

A message is more believable when it includes or refers to specific supporting evidence. In science, the most complete and detailed supporting evidence appears in primary source research reports in science journals. Documentation of information sources is typically less detailed and systematic in journalism and mass media. (The best science journalists provide references or links to primary source research reports.)

It is possible for a writer or an advertiser to claim a spurious air of authority by citing numerous sources. However, a long list of references does not guarantee accuracy. On closer examination, readers may find that communicators have cherry-picked, misinterpreted, or

misrepresented evidence; cited sources that are not relevant to the topic; or referred only to opinion pieces that do not actually contain evidence.

To evaluate the quality of evidence, we need to know how it was collected. Collection of evidence in science is systematic; that is, there are rules and procedures that specify what researchers should do to gather evidence and limit the kinds of interpretations they are permitted to make. Rules for statistical analysis are an important part of this.

1.5 EVALUATING GENERALIZABILITY

Researchers and journalists usually want to generalize about their findings. In other words, instead of just saying: "45% of the *respondents I talked to* said they plan to vote for candidate X," they want to say something like "45% of *all registered voters* plan to vote for candidate X." **Generalizability of results** is the degree to which a researcher can claim that results obtained in a specific sample would be the same for a population of interest. Results from a sample can be generalized to an actual population of interest if the sample is representative of the population; representativeness can often be obtained using random or systematic methods to select the sample. Results from an accidental or a convenience sample may be generalizable to a hypothetical population if the sample resembles that hypothetical population. Results from a biased sample are not generalizable. In experiments, generalizability also depends on similarity of type and dosages of experimental treatment to real-world experiences with the treatment variable, setting, and other factors.

Polling organizations, such as Gallup, collect public opinion information in ways that provide a good basis for generalization. They use large samples (usually at least 1,000 individuals) and obtain these samples using combinations of random and systematic selection so that the people who responded to the survey resemble the larger population (such as all registered voters) in terms of age, income, and so forth (Gallup, n.d.).

When journalists report information from polls and demographic studies, they are (once again) in a position to cherry-pick. Because of differences in procedures and types of people contacted, various polling organizations may report different predictions about presidential candidate preference. A journalist who wants to make a case to support Candidate X may report only the poll in which Candidate X had the highest approval ratings.

In behavioral and social science, the problem of generalizability can have a different form. A researcher may want to know whether cognitive behavioral therapy (CBT) reduces depression. Typically, studies examine small to moderate numbers of cases, for instance, 35 patients who receive CBT and 35 who do not. To generalize results about effects of CBT to a large hypothetical population of "all depressed persons," ideally, we would want a random sample drawn from that population. However, participants are often convenience samples, that is, people who were easy to recruit.

It is important to know what kinds of people were (and were not) included in a study. For example, if a drug study finds evidence that a new medication is effective and safe for healthy young men, that does not necessarily mean that the drug is also effective and safe for women, elders, children, and other kinds of people not included in the study.

Be careful not to overgeneralize results, particularly when there is little information about the types and numbers of people (or cases) included. *It makes sense to generalize information from a small group to some larger population only when people in the group resemble the population of interest.* This is discussed further in Chapter 2 in sections about samples and populations.

In science communication, authors are expected to discuss limitations that must be considered before drawing any conclusions. Limitations include the number and kinds of people (or cases) included in a study. *Science writing should make limitations of evidence clear; media reporting often does not.*

1.6 MAKING CAUSAL CLAIMS

In everyday life, and in science, we often want to know about causal connections. Consider a question raised by Wootson (2017). Do diet (artificially sweetened) soft drinks cause weight gain? If you are concerned about weight gain, and if artificially sweetened soft drinks cause weight gain, then you might consider avoiding diet soft drinks to avoid weight gain. However, it is possible that the association reported in some studies did not arise because of any direct causal impact of diet soft drinks on weight. Perhaps when people drink diet soft drinks, they feel free to indulge in other high-calorie foods, and perhaps it is those other high-calorie foods, not the soft drinks in and of themselves, that cause weight gain. If that is the correct explanation, then what you need to do to avoid weight gain is to avoid consuming high-calorie foods (rather than reduce diet soda consumption).

Causal explanations are attractive because they tie events together in meaningful ways. This is useful in science as well as everyday life. Sometimes when a cause–effect relationship is known, it suggests what we can do to change outcomes.

Demonstrating that two events are causally connected can be difficult, because there are often rival possible explanations. Well-controlled experiments can rule out many rival explanations. In everyday life, people sometimes jump to conclusions about causality on the basis of insufficient evidence.

1.6.1 The "Post Hoc, Ergo Propter Hoc" Fallacy

News commentators frequently offer causal explanations for events (e.g., the stock market went down because of a blizzard the previous day). This explanation is often just an opinion of the news commentator. The stock market might have gone down for other reasons (including random variations). This is an example of a common logical fallacy called "**post hoc, ergo propter hoc**." This Latin phrase means "after this, therefore, because of this." This (incorrect) reasoning goes like this: If Event A happens, and then Event B happens, then A must have caused B. Before we can conclude that Event A caused Event B, additional conditions are required. Here is another example. If you have a cold, take a large dose of vitamin C, and then the cold goes away, you might conclude that vitamin C cured the cold. However, the cold might have gone away on its own, whether you took vitamin C or not. Post hoc, ergo propter hoc reasoning uses one instance of co-occurrence (vitamin C, end of cold) to draw a causal conclusion. That is poor-quality reasoning that often leads to mistaken beliefs in causality. To conclude that vitamin C cures colds, you would need an experiment to evaluate whether the duration of colds was less in a group that took vitamin C than in a group that did not (controlling for other factors, such as placebo effects).

1.6.2 Correlation (by Itself) Does Not Imply Causation

You may have frequently heard the warning that correlation does not imply causation. This warning should be stated more precisely. It is more accurate to say, *Existence of a statistical relationship, such as a correlation, between variables X and Y, is needed to make claims that X causes Y. However, the mere existence of a statistical relationship does not prove that X causes Y. Alternative explanations for the statistical relationship between X and Y must be ruled out before we can believe that X causes Y.*

Let's examine this idea carefully.

The word **correlation** has two meanings. First, sometimes people use the term *correlation* to refer to a specific statistic: the **Pearson product–moment correlation**, also called **Pearson's *r***. Second, the term *correlation* can be used in a broader sense; we can say that variables are correlated if they are statistically related using some statistical analysis. The statistical analysis can be something other than Pearson's *r*. For example, if we compare average height for male and female groups and find that men are taller than women, we can say that sex (X) is statistically related to height (Y) or that sex is correlated with height.

We cannot claim that an *X* variable "causes" a *Y* variable if there is no statistical relationship of any kind between *X* and *Y*. In other words, the existence of a statistical relationship between *X* and *Y* is a *necessary* condition before we can consider causal inference.

However, existence of a statistical association is not enough evidence by itself to prove causality. Sometimes variables are statistically related (correlated) just by chance, or because the *X* and *Y* variables are related to some third variable *Z*, and *Z* may be the real "cause."

Consider this example: If we measure ice cream sales (*X*) and number of homicides (*Y*) once a month for a year, there is a correlation between them. Months that have the most ice cream sales also have the largest number of homicides (Peters, 2013). Does eating ice cream cause people to commit homicide? That idea is obviously silly. A more plausible explanation is that temperature is related to both ice cream consumption and homicide. During hotter months, people may buy more ice cream; homicide rates are higher in hotter months (perhaps because people hang around outside more, or perhaps heat makes people more irritable).

Correlation (statistical association) is a **necessary but not sufficient** condition for making causal inference. Statistical association is necessary because we can't conclude that *X* causes *Y* unless *X* and *Y* go together or co-occur. Statistical association is not sufficient by itself to prove causation because, even if *X* and *Y* covary, this co-occurrence may be due to the influence of one or more other variables; one of those other variables might be the real cause of *X*, or of *Y*, or both. In this example, heat or temperature might cause (or at least predict) ice cream purchase and homicide.

The effects of rival explanatory variables can be reduced or eliminated in well-controlled experiments and reduced by statistical controls. Mere co-occurrence is not enough evidence to make a causal inference.

Sometimes the need to look for a different explanation is obvious (as in the ice cream/homicide example). It would be absurd to argue that ice cream causes homicide. However, the need to consider rival explanations also arises in situations that are not so obviously silly. In the diet soft drink/weight gain example, it is conceivable that artificial sweeteners have causal effects on appetite or metabolism that really do lead to weight gain, even though the artificial sweeteners contain zero (or negligible) calories. However, the other explanation (that drinking diet beverages leads people to indulge in other high-calorie foods) is also plausible. (It is also conceivable that both these explanations are partly correct.) Both experimental and nonexperimental studies, with humans and nonhuman animals, would be helpful in sorting out the relations among variables and whether any of the associations are causal.

1.6.3 Perfect Correlation Versus Imperfect Correlation

Perfect co-occurrence (perfect correlation or statistical association) is rare. Consider the genetic mutation for hemophilia (Table 1.1). If a male child inherits this genetic mutation, he will have hemophilia. Most other heritable diseases do not show this perfect association. (For female children, effects of the hemophilia gene are ruled out by information on the other X chromosome.)

Table 1.1 Example of Perfect Co-occurrence or Perfect Correlation (Between Gene and Disease)

	Male Child Has Hemophilia	Male Child Does Not Have Hemophilia
Hemophilia gene is present	100%	0%
Hemophilia gene is absent	0%	100%

Table 1.2 Association Between Hand Washing and Getting Sick (Imperfect Association)

	Person Does Not Get Sick	Person Gets Sick
Person washes hands regularly	75%	25%
Person does not wash hands regularly	33%	67%

If a male child does not inherit the gene for hemophilia, he will not have hemophilia. In logical terms, the mutated gene is both **necessary and sufficient** for the disease. The mutated gene is necessary for hemophilia because a person can't get hemophilia without it. The mutated gene is sufficient for hemophilia, because if a person has it, he always has hemophilia. In other words, hemophilia always occurs when the mutated gene is present and never occurs when the mutated gene is absent.

Most associations in behavioral and social sciences and medicine are *not* perfect. Consider this hypothetical example for a behavior (washing or not washing hands) and a disease outcome (getting sick).

Table 1.2 shows an imperfect association. Only 25% of regular hand washers got sick, while 67% of the those who don't regularly wash their hands got sick. While most people who washed their hands did not get sick, hand washing did not guarantee that they could avoid getting sick.

The association between lung cancer and smoking is also not perfect. The risk for getting lung cancer is much higher for smokers than for nonsmokers. However, a few nonsmokers do get lung cancer, and many smokers do not get lung cancer.

In situations where associations are not perfect, it is likely that other variables are involved. Behaviors or conditions that sometimes (but not always) precede disease are often usually called "**risk factors**" rather than causes. Smoking is a risk factor for lung cancer. Some diseases have numerous risk factors (for example, risk for heart disease is related to smoking, body weight, sex, age, high blood pressure, and other factors).

We call behaviors that reduce risk for a negative outcome "**protective factors**." For example, hand washing is a protective factor against getting sick.

1.6.4 "Individual Results Vary"

Unless there is a perfect correlation (as in the hemophilia example), statistical associations or correlations between variables do not predict exact outcomes for all individuals. Consider the results of a study by Judge and Cable (2004), informally reported in Dittman (July/ August 2014). They reported that taller persons tend to earn more money (that is, height is correlated with salary). This is not a perfect correlation. If you are short, that does not necessarily mean that you will earn very little. Mark Zuckerberg (the founder of Facebook) is reported to be 5'7", but that did not prevent him from becoming one of the wealthiest men in the world. If you think about the implications correlations might have for your own outcomes, realize that individual outcomes differ when correlations are not perfect.

1.6.5 Requirements for Evidence of Causal Inference

Training in research methods and statistics provides the skills scientists need to think carefully about the evidence needed to support causal claims. Mass media journalists often rely on secondary sources or third-party content. By the time information filters through multiple communication links, details about the nature of the evidence and concerns about

limitations that affect the ability to generalize and make causal inferences are often lost. Third-party content often does not provide accurate information about generalizability and potential causality.

1.7 QUALITY CONTROL MECHANISMS IN SCIENCE

1.7.1 Peer Review

The science research process has mechanisms for information quality control. The most important mechanism is **peer review**. Researchers submit research reports to science journals (also called academic journals) for consideration (see note 2). The editor sends papers to peer reviewers (peers are expert researchers in the same field). Reviewers provide detailed criticism of studies, including evaluation of their research methods. On the basis of reviews, editors decide whether to reject a paper as inadequate, ask authors to revise the paper to correct errors or deficiencies, or (very rarely) accept the paper with only minor corrections. Papers are rarely accepted in their initially submitted form. Rejection rates for some journals are 80% or higher.

Peer review is fallible. Reviewers can also be subject to confirmation bias (they are more likely to favor conclusions consistent with their own beliefs). Reviewers may not notice all of the problems in a research report. However, peer review weeds out much poorly conducted research and improves the quality of published papers. The community of scientists in effect systematically polices the work of all individual scientists.

1.7.2 Replication and Accumulation of Evidence

A second important mechanism for data quality control in academic research is **replication**. Replication means repeating or redoing a study. This can be an **exact replication** (keeping all methods the same) or a **conceptual replication** (changing elements of the study, such as location, measures, or type of participants, to evaluate whether the same results occur in different situations). We should not treat findings from any one study as a conclusive answer to a research question. Any single study may have unique problems or flaws. In an ideal world, before we accept a research claim, we should have a substantial body of good-quality and consistent evidence to back up that claim; this can be obtained from replications.

Peer review and replication in science are fallible. However, they provide the best ongoing quality control checks we have. In contrast to science, there are few quality control mechanisms for most mass media communication.

1.7.3 Open Science and Study Preregistration

There are recent initiatives to improve the reproducibility and quality of research results in biomedicine, psychology, and other fields (Begley & Ioannidis, 2015; Open Science Collaboration, 2015). The Open Science model includes components such as preregistration of research plans and sharing details of data and methods. For further discussion, see Cumming and Calin-Jageman (2016).

1.8 BIASES OF INFORMATION CONSUMERS

1.8.1 Confirmation Bias (Again)

Information consumers or receivers also tend to select evidence consistent with their preexisting beliefs. Media consumers need to be aware that they can systematically miss kinds of information (which may be of high or low quality) when they select news sources they

like. Ratings of many web news sources on a continuum from left/liberal to right/conservative, along with assessment of accuracy, are provided at https://mediabiasfactcheck.com/politifact/. News sources that are extremely far left or far right tend to be less accurate.

Because of confirmation bias, people can get stuck: They continue to believe "facts" that aren't true, and ideas that are wrong, because they never expose themselves to information that might prompt them to consider different possibilities. Consumers of mass media usually avoid evidence that challenges their beliefs. Philosopher of science Karl Popper argued that scientists also need to examine evidence that might falsify their beliefs. Scientists and people in general should consider evidence that challenges their beliefs.

1.8.2 Social Influence and Consensus

Should we believe something simply because many people, particularly those whom we know and respect, believe it? Not necessarily. Some incorrect beliefs are widely reported in mass media and held by millions of people. My personal favorite conspiracy theory is that alien reptiles control U.S. politics. Bump (2013) reported that more than 12 million people, or 4%, of the U.S. population said that they believed this theory in 2012–2013. To be clear, I strongly disbelieve that we are ruled by alien reptiles. (I am also not sure whether to believe Bump's report that 12 million people really believe this; surveys are not always accurate.)

Consensus among science researchers can enhance the believability of a claim. However, even in science, consensus does not always guarantee accuracy. Experts can turn out to be wrong. For example, there was a consensus among nutrition researchers that eggs are bad for health because of their cholesterol content. Some recent research suggests that this widely held belief may be incorrect[4] (Gray & Griffin, 2009), but the issue continues to be controversial.

A belief shared by millions of people is not necessarily wrong. However, consensus is neither necessary nor sufficient evidence that information is correct.

1.9 ETHICAL ISSUES IN DATA COLLECTION AND ANALYSIS

1.9.1 Ethical Guidelines for Researchers: Data Collection

Ethical issues arise when collecting data about people and nonhuman animals. For psychologists, the American Psychological Association has codes of ethics that protect the well-being of subjects (Campbell, Vasquez, Behnke, & Kinscherff, 2009). Research that involves human participants is evaluated by an **institutional review board**; research that involves nonhuman animals is evaluated by an **institutional animal care and use committee**. Ethical codes govern research in other areas such as biomedicine. Data collection cannot begin until ethics board approval of procedures has been obtained. Adherence to those rules is an ethical obligation for researchers. We should not harm the people or entities we study.

As an example of potential harm to a research participant, suppose that a study reveals that a person has a history of addiction. If that information gets into the hands of potential landlords or employers, it could have an impact on that person's search for housing and jobs. Researchers must keep such records confidential.

Researchers also have an ethical responsibility to think about the potential impact of their research (both positive and negative) on public policy and the behavior of organizations and individuals.

1.9.2 Ethical Guidelines for Statisticians: Data Analysis and Reporting

The GAISE report states, "Students should demonstrate an awareness of ethical issues associated with sound statistical practice" (GAISE College Report ASA Revision

Committee, 2016). A separate document (American Statistical Association, 2015) discusses ethical issues in detail. Here is a list of ethical practices for data analysts, paraphrased from the American Statistical Association's ethics document. You will be reminded about these issues as you continue through the book.

1. Ensure that numbers are accurate. Fully disclose data handling procedures (such as deletion of cases or replacement of missing values) that could alter conclusions.
2. Make the limitations of the type of statistical analysis clear. (As each new analysis is introduced, you will learn about its limitations.)
3. Avoid behaviors that can lead to errors (including, but not limited to, cherry-picking a few results).
4. Avoid misleading presentations (such as "lying graphs"; see Section 1.10).
5. Avoid language that obscures results.
6. Do not overgeneralize. Do not make strong claims about characteristics of a population when your sample does not resemble that population.

Real-world problems in applications of data analysis are often not clear in introductory courses; students learn to do one analysis at a time using one small set of numbers. In actual practice, data analysts often work with large sets of messy data. Data analysts need to make many choices that involve difficult judgment calls. This book points out differences between the *ideal* use of statistics in artificially simplified situations and the *actual* application of statistics to real-world data. Sometimes decisions about "best practice" are difficult.

As Harris (2001) said, "Statistics is a form of social control over the professional behavior of researchers. The ultimate justification for any statistical procedure lies in the kinds of research behavior it encourages or discourages." Science has rules and standards about good practice in collection, analysis, and presentation of evidence. These are discussed throughout this book.

Researchers should be aware that press releases from universities sometimes overhype research findings (Resnick, 2019).

This book discusses good practices in applied statistics that can potentially improve the clarity and honesty of research reports. When communicators present information in misleading, unclear, or dishonest ways, they risk loss of credibility, trust, and respect, not just for themselves but for the professions of statistics and science. When information consumers rely on incorrect information, they may make poor decisions.

1.10 LYING WITH GRAPHS AND STATISTICS

The most extreme form of lying with statistics is fabrication or falsification of data; this is rare. However, some common research practices slant information presentation in ways that can be called "lying with statistics." The classic book *How to Lie With Statistics* (Huff, 1954) presented numerous examples.

Deceptive bar graphs are among the most common ways information communicators mislead information consumers. If you will be an information producer, you need to know how to set up "honest" bar graphs. When you are an information consumer, you need to know how to examine graphs to make sure that they are not misleading. Chapter 5 provides examples of clear versus misleading graphs and guidelines for evaluation of graphs.

1.11 DEGREES OF BELIEF

People rarely have time to collect all necessary information. Even for questions in science, we often do not have enough information to be confident about conclusions. Uncertainty is more common than people realize, even in areas such as medicine. There are many questions in medicine (such as what causes autoimmune disorders) for which medical research does not have good answers (Fox, 2003).

It is useful to think about scientific knowledge in terms of **degree of belief** instead of certainty. The philosopher David Hume said that "a wise [person] . . . proportions his [or her] belief to the evidence" (Schmidt, 2004). Degree of belief should be based on the *quantity* of *consistent* and *good-quality*, *systematically collected* supporting evidence. When there is little evidence (for example, results from only one study), people should not have strong belief in a claim. As additional good-quality evidence accumulates, degree of belief can increase. People should revise degree of belief upward or downward as new (good-quality) evidence becomes available.

This rating scale illustrates the concept of degree of belief. The use of a five-point scale and the exact verbal descriptions for each numerical rating are arbitrary.

1	2	3	4	5
Probably untrue	May be untrue	Not sure; insufficient evidence	May be true	Probably true

Fairly often, the best answer to research or public policy questions is that we do not have enough high-quality evidence to be confident that we know the correct answer. We should never assume that numerical results of one single study or mass media report are conclusive.

1.12 SUMMARY

Here are some questions to keep in mind when evaluating numerical (and other) information.

1. Is there evidence of communicator bias or self-interest?
2. Is evidence cherry-picked to fit the communicator's argument?
3. Is the communicator far from the information source or not well qualified to evaluate the information?
4. Does the communicator have a good record for truth-telling?
5. What types of evidence are included. Anecdotes? Citations of specific, credible sources?
6. Have you considered your own possible biases as an information consumer? Do you accept information uncritically because it confirms when you already believe? Are you influenced by what other people believe?
7. Do data come from people (or cases) who resemble the population of interest? Are results generalizable?
8. Are causal inferences drawn when there is not enough information to prove a causal association? Remember that imperfect correlation or co-occurrence does not indicate causation.
9. Has information been subjected to quality control? (In science, this includes peer review and replication.)

10. Is the presentation of information deceptive (e.g., lying graphs)?
11. What ethical issues are at stake in the conduct and application of the research?
12. Is your degree of belief proportional to the quantity of good quality and consistent evidence? (You should never believe a claim on the basis of just one scientific study or one journalism report.)

Sometimes the best answer to questions such as "Are eggs harmful to cardiovascular health?" is that we don't have enough evidence yet to answer the question. Unfortunately, lack of evidence does not prevent some communicators from making premature claims. When claims are made on the basis of limited evidence, contradiction and confusion often arise. It is better to reserve judgment until a large quantity of good-quality evidence is available. One single media report, or one single science report, is not "proof."

Even if you do not plan to be a researcher, you can benefit from thinking like a scientist and statistician about numerical evidence you encounter in everyday life. Some decisions have high stakes. For example, you may need to decide whether to undertake a risky but potentially beneficial medical treatment. Ideally, you should have accurate information about potential outcomes. The higher the stakes, the more you need to know how to obtain trustworthy information.

The take-home message from this chapter is: *We all know a lot less than we think we do*, because most of us rely heavily on third-party content that has little or no information quality control. All of us (scientists, journalists, and information consumers) should be cautious about degree of belief. Sometimes the best answer to a question is: We don't have enough good quality evidence. Courses in statistics and research methods teach you good practice in evaluation and presentation of evidence.

COMPREHENSION QUESTIONS

1. What is cherry-picking of evidence, and why is it deceptive? (Can you think of a book or media report that seems to present cherry-picked evidence?)
2. Give examples of self-interest that might make a communicator less believable.
3. Why is distance to original source of information an important factor when you evaluate message credibility?
4. What does it mean to say that a correlation (or association) between variables is imperfect?
5. Give an example of a risk factor, and a protective factor, not discussed in the chapter.
6. Why is the existence of a correlation (existence of co-occurrence or association) between *X* and *Y* not enough evidence for us to say that *X* causes *Y*?
7. What is the post hoc, ergo propter hoc fallacy? (Give an example you have seen, different from the one in this chapter.)
8. What is confirmation bias?
9. What quality control mechanisms are used in science?
10. What is peer review? How can it improve the credibility of science reporting?
11. What is research replication? How can this improve the quality of evidence in science? How do exact replication and conceptual replication differ?
12. A researcher might say "the results of this one study prove" something. Is this justified?
13. What (approximate) degree of belief should you have on the basis of only one study?

NOTES

[1] Scientists are expected to be objective when they select information to report. However, scientists tend to focus selectively on information consistent with the most widely accepted existing theories; Kuhn and Hacking (2012) called this "selection of significant fact."

[2] Numerous predatory, for-profit online journal publishers have emerged in recent years. It has become more difficult to determine whether online publications are credible. Research reports published in predatory journals are not valued by professional colleagues and universities. Beall's List of Predatory Journals and Publishers names many publishers that are almost certainly predatory (https://beallslist.weebly.com). Additional warning signs that a publisher may be predatory:

- The journal invites you to submit your undergraduate or graduate thesis for publication (particularly if the journal title is not in your discipline or field).
- The journal offers to publish your paper without peer review.
- The journal asks you to pay for publication. (However, many legitimate publishers charge author fees to make journal articles open access on the web; therefore, a request for payment is not always an indication that a journal is predatory.)

If you are not sure whether a journal or publisher is predatory, search <journal name> or <publisher name> along with the term *predatory*. You can also ask mentors, advisers, or colleagues.

[3] About half of Dr. Oz's medical advice is not supported by medical research (Belluz, 2014). Dr. Oz was investigated in a congressional hearing and paid large settlements in lawsuits for false advertising (Cohen, 2015).

[4] This video about an imaginary time-traveling dietician makes fun of changes in dietary recommendations across the decades: https://www.youtube.com/watch?v=5Ua-WVg1SsA.

DIGITAL RESOURCES

Find **free study tools** to support your learning, including **eFlashcards**, **data sets**, and **web resources**, on the accompanying website at **edge.sagepub.com/warner3e**.

CHAPTER 2

BASIC RESEARCH CONCEPTS

2.1 INTRODUCTION

Basic understanding of research methods is needed to understand and interpret statistical results. This chapter is a brief, nontechnical introduction to selected research methods terms mentioned in the GAISE (GAISE College Report ASA Revision Committee, 2016) numeracy guidelines in Chapter 1.

The *design* of an investigation refers primarily to the distinction between designs in which investigators have a high degree of *control* over the research situation (such as *experiments*) and situations in which researchers have little or no control (*nonexperimental* studies). *Experimental methods of control* include techniques such as random assignment of participants to groups and holding variables other than the treatment variable constant. Statistical methods of control are included in some types of analysis. Other design issues are discussed in greater detail in research methods textbooks (e.g., Cozby & Bates, 2017).

Data (or **data set**) refers to information, usually in numerical form in a computer file, about multiple cases and/or multiple variables.

Analysis refers to statistical techniques.

A **variable** is a characteristic that differs or varies across subjects or cases. Examples of variables for human research participants include sex, height, heart rate, and salary.

Subjects or **cases** are the entities or observational units studied. In psychological research, cases are usually individual persons or nonhuman animals. In other disciplines, cases can be different kinds of entities; for example, in sociology, a case can be a group or an organization; in political science, a case may be a nation; in forestry, a case may be a geographic location.

The terms *sample* and *population* are often used differently in ideal textbook situations than in many real-life research situations, as discussed in Section 2.11. For now, it is sufficient to say that a sample is a subset of a population; that is, a sample consists of cases selected from a population.

A *generalization* is a statement that results obtained for people and situations included in a study are applicable to other people and situations not included in the study. Ability to generalize results from a sample to a population depends on similarity of the sample to the population of interest.

Examples of *errors in interpretation* include (a) generalizing results more widely than can be justified, (b) arguing that one variable causes another variable when there is not enough evidence to support that claim, and (c) misunderstanding the limits of research methods and statistical analyses. Other types of error are possible.

2.2 TYPES OF VARIABLES

2.2.1 Overview

It is useful to distinguish between categorical variables and quantitative variables (Jaccard & Becker, 2009). Scores for categorical variables tell us which group or category each case belongs to (e.g., whether a person is male or female). Scores for quantitative variables provide information about the amount of something (for example, height). Some psychologists make further distinctions among levels of measurement; see Appendix 2A for discussion. Two additional types of variables are discussed in this section: rating scales and ordinal (also called rank).

2.2.2 Categorical Variables

Categorical variables identify group (or category) membership for each case. They are also called **nominal variables** because numbers serve only as names or labels for groups. This is a common type of variable. Examples of categorical variables include sex (for example, with group membership coded 1 = male, 2 = female) and marital status (with values coded 1 = never married, 2 = divorced, 3 = currently married). Additional categories could be included; for example, marital status could include categories such as engaged, cohabiting, separated, and remarried. Numerical values used for categorical variables are arbitrary; we could code divorced as 3 instead of 2, and this change in group numbering will make no difference in results of analyses.

When numbers are only labels for group membership, it is not meaningful to compare these numbers in terms of "greater than" or "less than." A person whose marital status is represented by the number 2 (divorced) is not greater than or better than a person whose marital status is represented by 1 (never married). We can say only that these individuals differ in marital status. It makes no sense to apply arithmetic operations (+, –, ×, ÷) to numbers when they are used only as labels for group membership. It makes no sense to calculate statistics such as sample means for scores on categorical variables; for example, it would be nonsense to compute a mean marital status.

Often the number of different score values for a categorical variable is small. However, it is possible for categorical variables to have many different score values. For example, a categorical variable to identify choice of future career could include dozens of different possible careers.

2.2.3 Quantitative Variables

Quantitative variables indicate "how much" of some characteristic or behavior each case or person has. For example, we can measure height or blood pressure for each person. When numerical scores for these variables are compared, it makes sense to describe them in terms of "more than" and "less than." A person who is 70 inches tall is taller than a person who is 65 inches tall. It is reasonable to apply arithmetic operations to numerical values for quantitative variables; we can add, subtract, multiply, and divide scores. Thus, it is reasonable to compute a mean for variables such as height. Quantitative variables are common in behavioral and social science research.

2.2.4 Ordinal Variables

Sometimes researchers rank subjects instead of measuring amount. For example, we could tag the runners in a race as 1, 2, 3, . . . last (the order of crossing the finish line). Alternatively, we could measure running time in seconds. Variables with scores that correspond

to ranks are called **ordinal variables**. Later you will see that there are specific analyses for scores that are collected in the form of ranks or are converted to ranks to get rid of problems such as outliers. Ranks are not widely used in data collection in behavioral and social sciences; measurements of quantity are generally preferred.

2.2.5 Variable Type and Choice of Analysis

Categorical and quantitative variables require different types of descriptive statistics (Chapter 4), graphs (Chapter 5), and other statistical analyses. It is necessary to distinguish between categorical and quantitative variables to choose appropriate statistical techniques. For some variables the decision is easy. Clearly, height and age are quantitative; sex and marital status are categorical. However, there are examples of variables that can be handled as either categorical or quantitative, as noted in the next section.

2.2.6 Rating Scale Variables

A **Likert scale** is a common response format in survey and personality research. A typical Likert scale question consists of a statement (worded so that it expresses a clearly positive or negative view about an issue) followed by a choice among degree of agreement ratings, as in the following example; each person chooses the number that best represents his or her degree of agreement. Originally Likert scales included five degrees of agreement, but **multiple-point rating scales** often have different numbers of responses (such as seven).

Example: "I believe the president is doing a great job."

1	2	3	4	5
Strongly disagree	Disagree	Neutral or don't know	Agree	Strongly agree

If five-point ratings are evaluated according to the formal levels of measurement standards proposed by Stevens (1946, 1951; see Appendix 2A), they lie somewhere between the ordinal (rank) and interval levels of measurement. Rating scores provide at least rank-order information (e.g., 4 represents stronger agreement than 3). However, the differences between scores probably don't represent equal intervals; for example, the difference in degree of agreement represented by 4 versus 5 may not be the same as the difference between 3 and 4. Five-point rating scale scores fall into a gray area: probably more informative than ranks, but probably less informative than measurements that assume equal intervals. That leads to disagreement as to whether it makes sense to compute means and other statistics for variables rated on five-point scales. Authorities cited in Appendix 2B argue that is acceptable to treat rating scale variables as quantitative variables in some circumstances.

In practice, ratings on five-point scales can often be treated as either categorical or quantitative variables, whichever makes more sense in a specific research situation. Scores for the question above could be used to divide people into five groups that have different degrees of agreement (i.e., used as a categorical variable). It would also be reasonable to compute a mean for ratings.

2.2.7 Scores That Represent Counts

Consider this survey question: "How many children do you want to have in the future?" Possible responses include none, one, and so forth. This is a quantitative variable; three children are more than two children. Unlike many other quantitative variables, scores for this variable have a limited number of possible values; it is rare in the United States to encounter persons who want more than four children. In a small sample, a researcher might find that the

only responses to this question are zero, one, and two. In some analyses it may be convenient and informative to treat these scores as labels for group membership (e.g., Group 1 does not want any children, Group 2 wants only one child, and Group 3 wants two children). However, it is also reasonable to compute the mean number of children. For variables that consist of counts (e.g., number of children) and variables that represent ratings on degree of agreement or behavior frequency (as in Section 2.2.3), it sometimes makes more sense to handle them as categorical, and it sometimes makes more sense to treat them as quantitative.

2.3 INDEPENDENT AND DEPENDENT VARIABLES

The first statistical techniques you will learn are ways to describe scores for just one variable.

However, real-world research usually begins with questions about the way two or more variables are related. It often makes sense to identify one of the variables as the independent or predictor variable (X) and the other as a dependent or an outcome variable (Y). The decision about which variable to identify as independent depends on the nature of the research question about the variables.

2.4 TYPICAL RESEARCH QUESTIONS

This section describes three types of research questions about the relationship between two variables. When we distinguish between independent and dependent variables, the independent variable is often denoted X and the dependent variable Y.

2.4.1 Are *X* and *Y* Correlated?

A researcher can simply ask whether scores on two variables (X and Y) tend to co-occur or go together (without assuming any causal connection between them). There are alternative ways to word this question, such as:

- Are scores on X and Y correlated?
- Do scores for X and Y tend to co-occur?
- Are high scores on X associated with high scores on Y?
- Are X and Y associated?

I prefer this wording: Are scores on X and Y **statistically related**?

For this research question, it is not necessary to identify one variable as independent and the other variable as dependent. The term *correlated* can refer specifically to the results of a Pearson r correlation analysis. However, researchers sometimes use the word *correlated* in a much broader sense, to refer to any statistical relationship between variables, even when information about the relationship comes from some statistic other than a correlation coefficient (for example, from an independent-samples t test). We can evaluate whether X and Y are statistically related by doing whatever statistical analysis is appropriate for the types of variables (categorical vs. quantitative).

The bivariate statistics described in later chapters provide different ways to evaluate the extent to which scores on two variables are statistically related. The specific statistic that is most appropriate for a pair of X and Y variables depends on the types of variables (categorical or quantitative) and other issues; see Section 2.10. We can evaluate whether X and Y are statistically related on the basis of the outcome of any of these bivariate statistical analyses.

2.4.2 Does *X* Predict *Y*?

In this question, *X* is identified by the researcher as the predictor or independent variable; *Y* is the outcome or dependent variable. To *predict* means to anticipate or guess something that will happen in the future. A predictor should occur before the outcome (or at least not after the outcome). This is called **temporal precedence**. If *X* happens before *Y*, *X* has temporal precedence.

Consider these examples. Does height at age 10 years (*X*) predict height at age 21 years (*Y*)? Do high school grades (*X*) predict college grades (*Y*)? When temporal precedence is clear, it does not make sense to reverse these variables, that is, to ask whether height at age 21 predicts height at age 10 or whether college grades predict high school grades.

2.4.3 Does *X* Cause *Y*?

This question can be worded in several similar ways; we can replace the word *cause* with words such as *change*, *determine*, *increase*, *decrease*, or *influence*.

Here are examples of questions about cause:

- Does the death of a spouse (*X*) cause depression (*Y*)?
- Does study time (*X*) increase exam score (*Y*)?
- Does social stress (*X*) influence blood pressure (*Y*)?
- Does cigarette smoking (*X*) increase the risk for lung cancer (*Y*)? Is cigarette smoking a risk factor for lung cancer? (If a variable is called a risk factor, this usually implies that there may be other risk factors or causes.)

Note that the word order in questions can vary, for example, Is exam score (*Y*) increased by amount of study time (*X*)? In this question, study time is still the independent variable (presumed cause), and exam score is the dependent variable.

We need stronger evidence for claims that *X* causes or influences *Y* than for claims that *X* merely predicts *Y* or that *X* co-occurs with *Y*. Keep in mind that no matter what results we obtain in one study of *X* and *Y*, we should not view those results as a final answer to any of these questions.

2.5 CONDITIONS FOR CAUSAL INFERENCE

When researchers select variables to include in a study, the first consideration is:

1. There should be a plausible theory that explains why *X* and *Y* might be related (cf. Brannon, Feist, & Updegraff, 2017).

It does not make sense to choose an *X* variable and a *Y* variable at random. Variables are selected because past research or theories suggest that they may be related in meaningful ways.

Three additional conditions should be considered when interpreting research results as potential evidence for causation (Cozby & Bates, 2017).

2. We can say that *X* and *Y* are associated only if we find that *X* and *Y* are related when we do an appropriate statistical analysis. To evaluate whether *X* and *Y* are statistically related (or correlated), you will use the statistical analyses that you will learn in later chapters, such as the independent-samples *t* test and Pearson correlation.

The next condition is required for questions about prediction and causation.

3. We can say that *X* predicts *Y* only if *X* happens earlier in time than *Y* (or at least not later than *Y*) and, in addition, *X* is statistically related to *Y*.

Questions about causal relationships (does *X* cause or influence *Y*) require all these preceding types of evidence as well as this fourth additional type of evidence:

4. We can infer that *X* causes *Y* only if no other variables are plausible rival explanations for changes in *Y*. In other words, *X* must be the only possible explanation for changes in *Y*. This condition can be very difficult to satisfy, because rival explanatory variables are common in many research situations.

Rival explanatory variables (also called **confounds or confounded variables**) arise in situations where many variables (other than *X*) might cause or influence *Y*. Suppose a researcher wants to know whether social stress (*X*) causes higher blood pressure (*Y*). Many other variables, in addition to social stress, can influence blood pressure, including but not limited to cardiovascular fitness, body weight, use of caffeine, alcohol, and other drugs, smoking, and family history of high blood pressure. Smoking and use of alcohol might well be correlated with or confounded with anxiety. We can evaluate whether anxiety influences high blood pressure only if we **control** for other explanatory variables (or take them into account in statistical analysis).

In experiments, we take other rival explanatory variables into account by using **experimental controls**, such as holding the variables constant. For example, a study may include only people who do not use any drugs that may influence blood pressure. In nonexperimental studies, we use **statistical control** to try to rule out effects of rival explanatory variables. Techniques to do this are not covered in this volume; they involve more advanced forms of analysis. When a more sophisticated type of analysis is performed, the correct answer to the question "Does stress cause high blood pressure?" may be that stress is one among many variables that predict, and may possibly influence, blood pressure. Whether experimental or statistical control is used, readers need to know what variables have and have not been controlled in some way.

When scores for two potential causal or independent variables co-occur, we say that they are **confounded**. If people who experience a lot of social stress in their everyday lives tend to smoke a lot, then social stress and smoking are confounded, and it may be difficult to separate their effects. If people who report high levels of social stress have high blood pressure, the real reason for this (or at least a partial explanation for this) may be that people with high levels of stress also smoke or drink heavily.

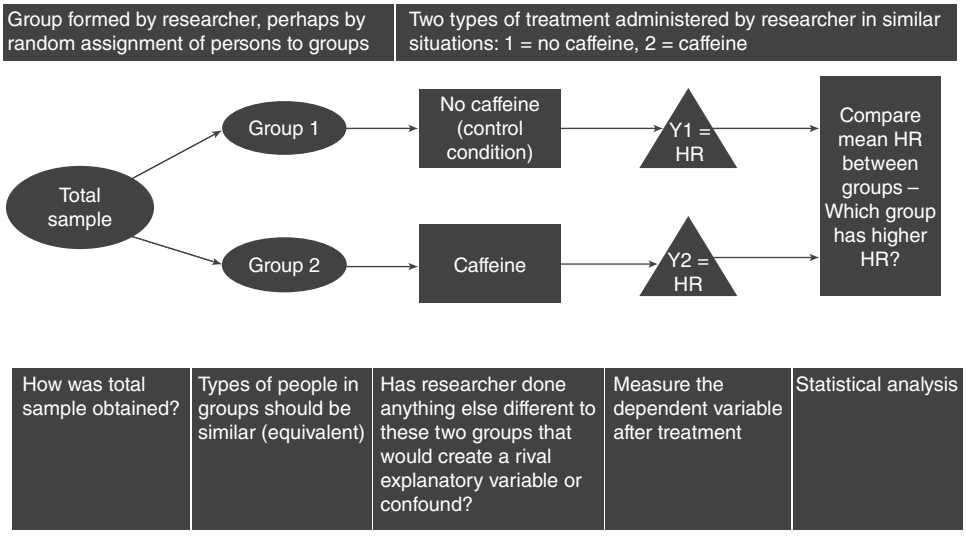
The next section describes the extent to which various **research designs** (including nonexperimental, experimental, and quasi-experimental) can provide the evidence needed to satisfy Conditions 2, 3, and 4.

2.6 EXPERIMENTAL RESEARCH DESIGN

A typical **experimental research design** includes two or more groups of cases; each group is exposed to a different type of treatment or different amount of treatment (such as a drug). Experiments require comparisons. If a researcher wants to evaluate the effects of caffeine (*X*) on heart rate (*Y*), the researcher needs to examine situations in which people do, and do not, receive caffeine (or situations in which people receive varying amounts of caffeine). In many studies, a **control group** that receives no treatment is included.

Figure 2.1 is a schematic outline of a simple experiment. Read from left to right: the researcher has a group of available participants. Participants are divided into groups using a

Figure 2.1 Schematic Outline of Simple Experimental Design



Note: HR = heart rate.

method that should ensure that similar people are included in Groups 1 and 2. Often random assignment to treatment groups is used to do this. In this example, Group 1 receives a beverage that does not contain caffeine; Group 2 receives a beverage that does contain caffeine. The outcome variable, heart rate, is measured after participants drink the beverage. Statistical analysis compares mean heart rate to see if people who consumed caffeine (Group 2) have a higher average heart rate than people who did not consume caffeine (Group 1). (A placebo control group could be added.) The independent-samples *t* test is one example of a statistic that provides information about the differences for means of *Y* across groups.

In behavioral and social sciences, experimental design typically includes several kinds of experimental control. One form of experimental control is that a researcher controls assignment of participants to groups. In many experiments, cases are assigned to groups randomly. The intended purpose of random assignment is to avoid a confound of preexisting subject characteristics with type of treatment. (Note that **random sampling of participants from a population** is not the same as random assignment of those participants to treatment groups.)

Here is an example of a potential confound of participant characteristics with type of treatment. Suppose that a researcher arbitrarily assigns people to groups. Suppose that people in Group 1 (who do not consume caffeine) have low anxiety levels; people in Group 2 (who consume caffeine) have high anxiety levels. If average heart rate is higher in Group 2, it will not be clear whether this is due to caffeine or to preexisting anxiety (or both). There is a confound between the independent variable *X* (whether caffeine is present, no or yes) and a personal characteristic (preexisting anxiety). Preexisting anxiety is a plausible rival explanatory variable; we cannot conclude that caffeine caused a higher heart rate unless we can control for, rule out, or get rid of the differences in anxiety between groups.

A common way to try to prevent confound of treatment with participant characteristics is **random assignment of participants to groups or conditions**. Random assignment means that each subject or case has an equal chance of being placed in either group. An example of a method of random assignment is tossing a coin for each person and assigning the person to the no-caffeine group for heads and to the caffeine group for tails. This should result in a mixture of high and low anxiety scores within each of the two groups, with the same average

anxiety score in Group 1 as in Group 2. This should also make the groups similar on other participant characteristics, such as age, past experience with caffeine, and body weight. When it works well, random assignment of participants to conditions prevents confounds of most participant characteristics with type of treatment.

The researcher has control over the type and amount of treatment. In this example, the researcher controls whether each participant receives caffeine and the amount of caffeine.

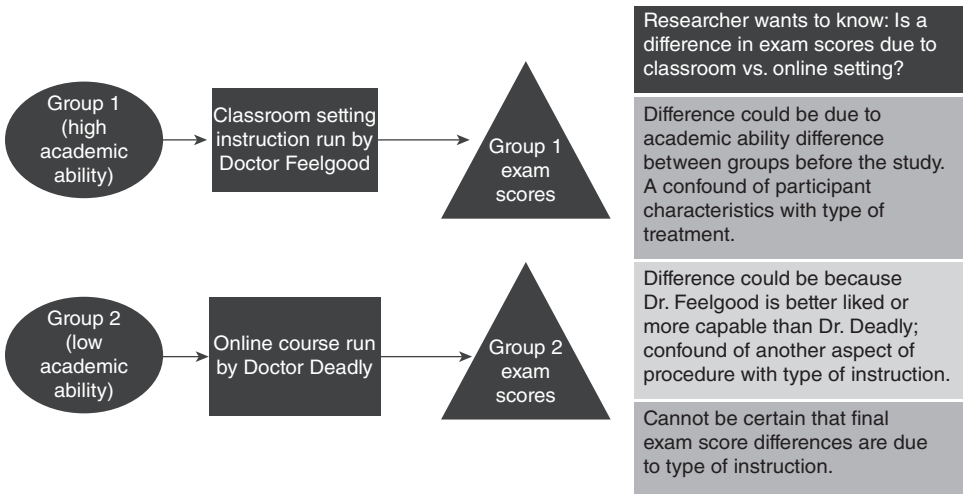
The researcher can control other variables and tries to keep them the same across participants both between groups and within groups. This is called **standardization** and **experimental control over other situational factors or extraneous variables**. Variables that are not included in the research question are extraneous (not of interest) in the present study. Many things other than the caffeine administered by the researcher could influence heart rate (for example, time of day, whether the research assistant is calm or upset, and whether participants know that they are consuming caffeine). To achieve standardization, ideally, all participants would be tested at the same time of day; the behavior of the research assistant would be made consistent, perhaps by training or even the use of a script; and neither participants nor research assistants would know which drinks contain caffeine.

Researchers need background knowledge about their variables to understand what kinds of confounds they need to anticipate and avoid. For example, if heart rate is the dependent variable, the researcher needs to know what other factors (apart from the manipulated variable, caffeine) might influence heart rate.

Sometimes experimental control does not work as well as hoped. Random assignment of participants to groups can result in "**unlucky randomization**," that is, groups that are not similar on one or more participant characteristics. In implementation of a treatment, variables may be unintentionally confounded with type of treatment. Consider the hypothetical flawed study in Figure 2.2.

Figure 2.2 illustrates two possible confounds. First, Groups 1 and 2 include different types of students (high vs. low academic ability). Second, Groups 1 and 2 had different instructors (Dr. Feelgood vs. Dr. Deadly). Any differences we find between final exam scores in these groups might be due to one or more of the following rival explanatory

Figure 2.2 Flawed Study to Compare Mean Exam Scores (Y) Between Classroom and Online Instruction (X)



Note: There are two confounds with X: participant characteristics and teacher identity.

variables: classroom versus online setting, academic ability levels of students, and behaviors of the different instructors. We cannot conclude that classroom and online instruction cause different results on final exams unless we can rule out or get rid of the effects of the two confounded, rival explanatory variables (student ability and teacher identity). In many experimental situations, there are large numbers of potential confounds. See research methods textbooks (such as Cozby & Bates, 2017) for further discussion of experimental control.

When potential confounds and extraneous variables can be ruled out by these forms of experimental control, an experiment can provide good-quality evidence that may be consistent with a researcher hypothesis about causal inference. (The results of a single study should not be considered proof of causal influence.) Nonexperimental designs lack all these types of experimental control. Quasi-experimental studies typically have some, but not all, of these forms of control.

2.7 NONEXPERIMENTAL RESEARCH DESIGN

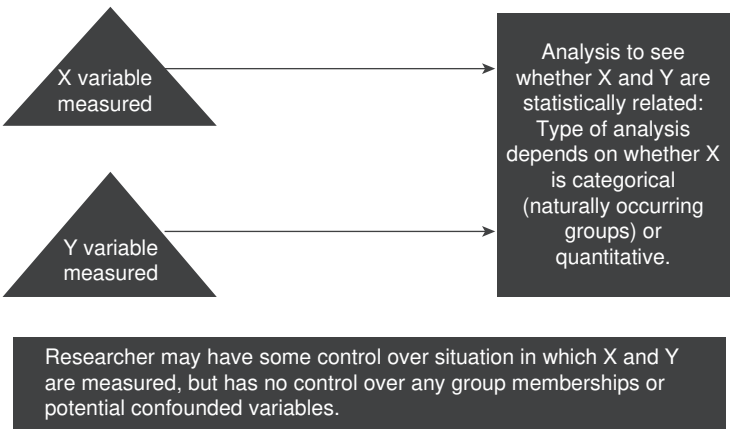
In a typical **nonexperimental research design** (also called a **correlational study**), a researcher measures two or more variables that are believed to be meaningfully related, and the researcher does not introduce a treatment or intervention.

Consider this example. Suppose that *X* is a measurement of amount of (naturally occurring) physical exercise, and *Y* is a score for depression. Both variables might be measured using self-report survey questions. A researcher may suspect that there is a causal association (getting more exercise reduces depression). See Figure 2.3.

Suppose that there is a strong correlation: People who report that they choose to exercise more tend to report lower levels of depression; people who report that they choose to exercise less tend to report higher levels of depression. That outcome cannot be interpreted as evidence that exercise causes a reduction in depression, because the data do not come from an experiment.

One requirement for causal inference is that the variable thought to be the cause should happen earlier in time than the variable thought to be the outcome. A nonexperimental study can (partly) satisfy that requirement by measuring exercise first and depression at a later point in time. Another option is to measure exercise and depression at multiple points in time.

Figure 2.3 Diagram of Nonexperimental Design With Two Variables



A more serious problem is that exercise is confounded with other variables, and those other variables might influence depression. For example, a person who experiences chronic stress may not feel like exercising, and chronic stress might cause depression. It is also possible that depression causes people to exercise less.

Advanced courses in statistics include methods for statistical control that can help separate the influences of rival explanatory variables (for example, using multiple regression). However, if all you have is a statistical relationship between amount of exercise and depression, and amount of exercise has not been manipulated, that is not sufficient evidence to conclude that lack of exercise causes depression.

It may occur to you that you could do an experiment in which you randomly assign people to high-exercise and no-exercise groups and measure later depression. That is possible, although it would be a challenge to create a good experiment for these variables.

Results from nonexperimental studies can satisfy the first two requirements in the list of conditions for causal inference. Variables *X* and *Y* can be chosen so that there is some logical or theoretical connection between them. Sometimes, but not always, there is clear temporal precedence, so that one variable can be identified as predictor and the other as outcome. Nonexperimental research can be sufficient to answer the question, Do *X* and *Y* co-occur? If a strong argument can be made for temporal precedence, data from nonexperimental studies can also be used to ask, Does *X* predict *Y*?

Researchers often identify variables in nonexperimental studies as independent and others as dependent, on the basis of theories about possible causal connections. However, distinctions between independent and dependent variables in nonexperimental studies are sometimes arbitrary (and even questionable). Consider a survey that measures self-esteem (*X*) and grades (*Y*) at the same point in time for a group of schoolchildren. If the analysis shows that higher self-esteem tends to co-occur with higher grades, and if the theory says that self-esteem causes better performance in school, a researcher may be tempted to phrase the interpretation in ways that suggest that the study proved a causal connection; the researcher might say, "High self-esteem leads to higher grades" (*leads to* is one of many synonyms for *causes*). It is plausible to theorize that grades increase self-esteem, but it is also possible that self-esteem increases grades or that both grades and self-esteem are influenced by other variables, such as intelligence. In a situation like this, I would say that neither self-esteem nor grades are clearly "the" independent variable or dependent variable. When there is no temporal precedence and no ability to rule out rival explanatory variables, it is preferable to say that *X* and *Y* are **correlated variables** (instead of calling one independent and the other dependent).

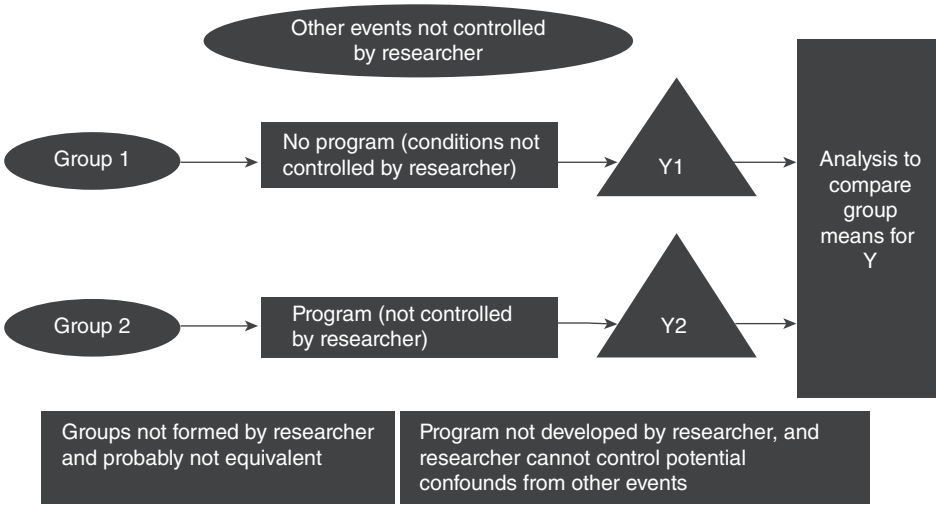
2.8 QUASI-EXPERIMENTAL RESEARCH DESIGNS

Studies that compare group outcomes but lack the full set of controls in true experiments (such as researcher control over assignment of participants to groups, researcher administration of treatments, and researcher control over other situational variables) are called quasi-experiments. **Quasi-experimental research designs** fall between experimental and nonexperimental designs in their ability to rule out rival explanatory variables. Quasi-experiments often arise when programs are evaluated in field settings. Occasionally, true experiments are run in field settings, but it is generally easier for researchers to have control over variables when they are in laboratory settings.

The simplest types of quasi-experiments involve comparison of two or more groups that receive different treatments (Figure 2.4) using preexisting groups instead of groups formed by a researcher. For example, each of two classrooms or schools may be used as a group. When preexisting groups are compared, the members of groups are likely to differ in many preexisting characteristics. This is called a **nonequivalent control group** design.

Consider potential problems in the group comparison design in Figure 2.4. Because the researcher cannot control the assignment of subjects to groups, the groups that do versus do

Figure 2.4 Quasi-Experimental Nonequivalent Control Group Design



not experience the program often include different kinds of participants (i.e., there may be a confound between participant characteristics and type of treatment).

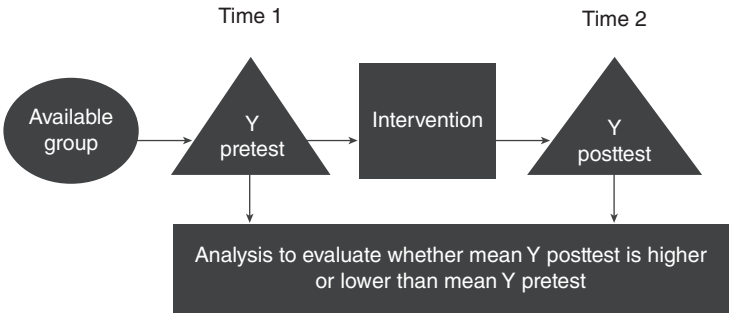
In addition, when data are collected in field settings such as schools over long periods of time, other events that might influence the outcome variable may occur. As an example, consider a hypothetical study to evaluate a drug education program (students in School 1 do not receive it; students in School 2 do receive it). The outcome measure could be self-reported intention to use drugs. To what extent does the drug education program have an impact on this? It is possible that School 1 and School 2 differ in ways that would influence drug use intention, for example, family religious backgrounds. It is possible that things happen in School 1 that did not happen in School 2 over the course of the study; for example, a popular student in School 1 dies from a drug overdose, which does not happen in School 2. These confounds would make it impossible to tell whether the drug education program causes any observed difference between groups for intention to use drugs.

A second simple quasi-experimental design compares scores for one group after the intervention with scores for the same group before the intervention (Figure 2.5). At first glance this may seem to be less problematic than the nonequivalent control group design, but this simple design is quite problematic. Many events, in addition to the intervention, may occur between Times 1 and 2, and any of these events might influence the outcome. A student may die in an alcohol-related car accident, and that event is a rival explanatory variable. If the study takes place over 3 years, there is time for maturation to occur (students are 3 years older at Time 2 than at Time 1, and changes in scores might be related to age). Shadish, Cook, and Campbell (2001) provided extensive information about the design and analysis of quasi-experimental studies.

2.9 OTHER ISSUES IN DESIGN AND ANALYSIS

Beginning students sometimes ask questions such as "Which is better, an experiment or a nonexperimental study?" It is more informative to ask, What are the potential advantages and disadvantages of experimental versus nonexperimental studies?

Figure 2.5 Within-Group or Pretest–Posttest Quasi-Experimental Design



The three types of design just reviewed (experimental, nonexperimental, and quasi-experimental) differ in the amount of control a researcher has over assignment to groups and ability to rule out rival explanatory variables. Sometimes situations in which a researcher has a substantial amount of control are in laboratory settings. Laboratory settings and experiments may be artificial or contrived situations (in other words, different from real-world situations).

Consider one highly contrived research situation in psychology: the Skinner box. A rat or pigeon is placed in a glass box. No other animals are present. Lights or tones act as signals for the performance of a specific behavior, such as lever pressing for the availability of a reward. Food, water, or other rewards drop into the box when a lever is pressed. The schedule for the availability of rewards is completely under researcher control. All other variables, for all practical purposes, are held constant: temperature, lighting conditions, the age and health of the rat, and so forth. Interactions of the human researcher with the animal may be minimal.

This situation is ideal if the goal is to make causal inferences: How does the schedule of reinforcement or reward influence the frequency of lever-pressing behavior? There are few or no rival explanatory variables. However, this situation is not ideal if we want to know about learning or food foraging in natural environments, where different factors may be important, or learning in species other than rats and pigeons.

In psychology the terms *internal validity* and *external validity* are used to describe two different aspects of research situations. A study has high **internal validity** when control of rival explanatory variables is so thorough that there are no rival explanatory variables to worry about when making a causal inference. Experiments in lab settings can potentially have high internal validity. Nonexperimental studies typically have low internal validity, because the ability to rule out rival explanatory variables is limited.

External validity refers to the similarity of the situation in the study to real-world situations we would like to be able to talk about. A study has high external validity if the situations resemble real-world situations of interest and low external validity if the situations are so artificial and contrived that they don't resemble any real-world situations of interest. Often nonexperimental research has higher external validity than experimental research, because researchers observe or ask about naturally occurring behaviors, sometimes in real-world settings.

There tends to be a trade-off between internal and external validity. Often, we have the best internal validity in experimental situations that are highly controlled and artificial, but these situations may have poor external validity. Often, we have the best external validity in uncontrolled nonexperimental studies, but these studies usually have poor internal validity.

There are things researchers can do to improve external validity in lab experiments; the goal is to make the situation as lifelike and believable as possible. There are things researchers

can do to improve internal validity in nonexperimental studies; often this involves the use of statistical control to compensate for the lack of experimental control.

We can build the strongest possible cause for a claim (for example, that crowding increases hostility) when we can show that the evidence for this claim is consistent across many different contexts: lab versus field setting, experimental versus nonexperimental design, animal and human subjects, different ways of measuring hostility, and so forth.

Another issue to consider in thinking about possible designs for a study is whether the groups in a design are between-*S*, as in Figures 2.1, 2.2, and 2.4, or within-*S* or repeated measures, as in Figure 2.5. In a typical **between-*S*** (also called independent-groups) study, each participant is assigned to just one group and contributes one score for the outcome variable. In a **within-*S*** or repeated-measures study, each case or participant receives multiple treatments or is evaluated at multiple points in time, or both. It is usually easy to tell whether a study is within-*S* or repeated measures because terms and phrases such as "each participant received all treatments," *repeated measures*, *longitudinal*, *prospective*, or *pretest–posttest* are included in descriptions of within-*S* studies.

The examples provided so far are extremely simple. However, group comparison designs can have more than two groups. In addition, research designs can include both within- and between-*S* factors (for example, pretest and posttest measures could be added to the study in Figure 2.4). Correlational or nonexperimental studies (as in Figure 2.3) usually include large numbers of variables.

You will learn statistical techniques for each of these situations one at a time. Later education in statistics shows ways to combine these simple research designs into more complex designs and analyses.

2.10 CHOICE OF STATISTICAL ANALYSIS (PREVIEW)

Chapters 9 through 17 in this book describe statistics used to assess whether two variables are related. There are four possible combinations of types of independent and dependent variables. (To select a statistical analysis, it may be necessary to identify one of your variables as an independent variable even if you do not have a causal hypothesis.) As a brief preview, here are some (not all) of the commonly used statistics for each combination of variables.

1. *X* is categorical, *Y* is categorical: χ^2 analysis of contingency table
2. *X* is categorical, *Y* is quantitative: *t* test or analysis of variance (ANOVA)
3. *X* is quantitative, *Y* is quantitative: Pearson *r*, bivariate regression

What do each of these analyses tell us?

1. A χ^2 (chi-squared) test evaluates whether membership in one type of group is statistically related to membership in another type of group. Consider sex (a group membership variable) and political party (a second group membership variable). A χ^2 analysis and examination of percentages can answer questions such as, Are women more likely to be Democrats, and are men more likely to be Republicans? The χ^2 test is more often used in nonexperimental research; however, it can be used in experiments when the outcome variable is categorical.
2. An independent-samples *t* test or analysis of variance compares mean scores on a dependent variable across two or more groups. Often this is done in an experiment in which a researcher has divided people into groups and then given a different type of treatment to each group. For example, a study might compare

mean anxiety scores between Group 1 (which received psychotherapy) and Group 2 (which did not receive psychotherapy) to see if people who received psychotherapy had lower anxiety on average. This analysis can also be used to compare means between naturally occurring groups, such as mean height between male and female groups.

3. A Pearson correlation (denoted *r*) is used to examine scores for two quantitative variables (such as *X*, height, and *Y*, salary). Pearson correlation is an appropriate analysis only when there is a linear association between *X* and *Y*, as discussed in a later chapter.

4. Chapters 9 through 17 do not cover analyses for the situation in which *X* is quantitative and *Y* is categorical. (Logistic regression can be used in this situation.)

2.11 POPULATIONS AND SAMPLES: IDEAL VERSUS ACTUAL SITUATIONS

2.11.1 Ideal Definition of Population and Sample

Statistical techniques were developed on the basis of ideal, imaginary situations. The development of statistical techniques began by specifying a **population** of interest. In an industrial quality control study, for example, the population could be all the widgets that are produced by a machine in a month. Let's assume that the variable of interest is the diameter of the widgets. If it is possible and not too expensive to measure the diameter of every single widget in the population, it makes sense to do that. However, it is often too costly or difficult to obtain information for every case in a population. Statisticians knew that it would be useful to develop techniques that can use information from a sample to make inferences (estimates) about population characteristics. A **sample** can be defined as a subset of the cases in a population, as in the following example. All members of the sample are members of the population. However, some members of the population are not included in the sample.

Population (of 7): [72, 81, 98, 67, 101, 78, 79]

Sample (subset) of size $N = 3$: [98, 72, 78]

To develop the techniques you will learn, statisticians made assumptions that simplified the problem. They assumed that *all members of the population can be identified and can potentially be included in a sample*. For the development of some statistics, they assumed that scores for the variable are normally distributed in the population. They assumed that the sample would be randomly selected from the population, in a way that gave every member of the population an equal chance of being included in the sample. Here's an example of a simple random selection method to obtain a sample that includes 50% of the population: Toss a coin for each case and include that case in the sample if the result is heads.

2.11.2 Two Real-World Research Situations Similar to the Ideal Population and Sample Situation

Industrial quality control involves a situation similar to the one imagined by statisticians. Returning to the widget example, the population of interest, all widgets produced by a machine in a month, can be identified. It is possible to select a sample of widgets randomly from this population of interest.

A second situation that is somewhat comparable with the ideal situation arises in political polling. Polling organizations such as Gallup define the population of interest in terms such as "all registered U.S. voters." It is more difficult to identify all members of this population

than in the widget example, and there are cases in this population that cannot be contacted and included in a sample. Organizations such as Gallup use complex sampling methods that include both random and systematic selection to obtain samples that should be representative of the population. A **representative sample** of a population is created if the cases in the sample have characteristics similar to those of the population. For example, if 51% of a Gallup sample are women, 10% are Hispanic, and 20% are older than 65, and the population of all registered voters includes 51% women, 10% Hispanic voters, and 20% voters older than 65, then the sample is representative of that population for those three variables. On the other hand, if the sample has 23% women, but the population has 51% women, then the sample is not representative of the population in sex distribution. This book does not deal with complex sampling issues and technical tools such as case weighting; for a comprehensive discussion, see Thompson (2012).

2.11.3 Actual Research Situations That Are Not Similar to Ideal Situations

In many behavioral and social science studies and in medicine, researchers often begin not with a well-specified population but with a **convenience sample** (sometimes called an **accidental sample**). Convenience samples consist of cases that are easy for the researcher to get. However, researchers almost always want to say something about cases not included in the study. Most textbooks don't address this question: What population can researchers talk about in this situation? Trochim (2006) suggests that researchers rely on a proximal similarity model to generalize from convenience samples. The **proximal similarity model** says that it is reasonable to generalize results from a sample to some broader **hypothetical or imaginary population** if the members of the sample have participant characteristics like those of the population of interest (i.e., if the sample is representative of the population of interest).

For example, a psychologist might run a study with a small sample of moderately depressed patients to evaluate whether cognitive behavioral therapy (CBT) (X) improves life satisfaction (Y). The psychologist can see whether patients in the study who received CBT have higher life satisfaction scores than patients in the study who did not. However, the psychologist does not want to be limited to saying, "CBT increased life satisfaction for the 30 patients in my study." The psychologist hopes to be able to say something like "CBT probably increases life satisfaction for many other depressed patients" (i.e., a broader hypothetical population of other depression patients).

How far can researchers go when making such generalizations? They should limit themselves to generalizations about populations similar to members of the study. If a CBT study finds that CBT increases life satisfaction for women ages 20 to 50 with moderate levels of depression, the researcher should not assume that CBT would have similar effects for men, older and younger persons, and persons with mild or severe depression.

When a sample is selected randomly from an actual well-specified population, cases in the sample should be like the population. In this situation we can justify generalizations beyond the sample to the population from which that sample was selected on the basis of the **sampling model** (Trochim, 2006).

In behavioral, social, and medical laboratory research situations, it is common for researchers to generalize from convenience samples to broader hypothetical populations (relying implicitly on the proximal similarity model). Research situations such as industrial quality control and political polling, where samples are obtained by random sampling from a population, can justify generalizations on the basis of the sampling model. In either case, generalizations from sample to population should be made cautiously. Even random selection of cases from a clearly defined population can sometimes yield a nonrepresentative sample.