This **Eighth Edition** of **Social Statistics for a Diverse Society** continues to emphasize intuition and common sense, while demonstrating that social science is a constant interplay between methods of inquiry and important social issues. Recognizing that today's students live in a world of growing diversity and richness of social differences, authors Chava Frankfort-Nachmias and Anna Leon-Guerrero use research examples that show how statistics is a tool for understanding the ways in which race, class, gender, and other categories of experience shape our social world and influence social behavior. In addition, guides for reading and interpreting the research literature help students acquire statistical literacy, while SPSS® demonstrations and a rich variety of exercises help them hone their problem-solving skills.

New to This Edition

- New real-world examples and data are drawn from a wide range of sources including news stories, government reports, and scholarly research to help students increase their understanding of both statistics and social issues.
- New *Data at Work* boxes profile individuals who use quantitative data and methods on the job.
- The authors provide two new 2014 General Social Survey SPSS® datasets.
- Chapter 2 ("Organization of Information") and Chapter 3 ("Graphic Presentation") in the previous edition are now combined into one chapter that focuses on "The Organization and Graphic Presentation of Data."
- Chapter 5, "Normal Distribution," includes new in-chapter examples based on SAT data.
- Chapter 12, "Regression and Correlation," includes extended discussion on regression, correlation, and multiple regression.
- New end-of-chapter exercises utilize data and polling based on current events.

©iStockphoto.com/sdominick

ISBN 978-1-5063-4720-2

90000

9 781506 347202

SAGE www.sagepublishing.com

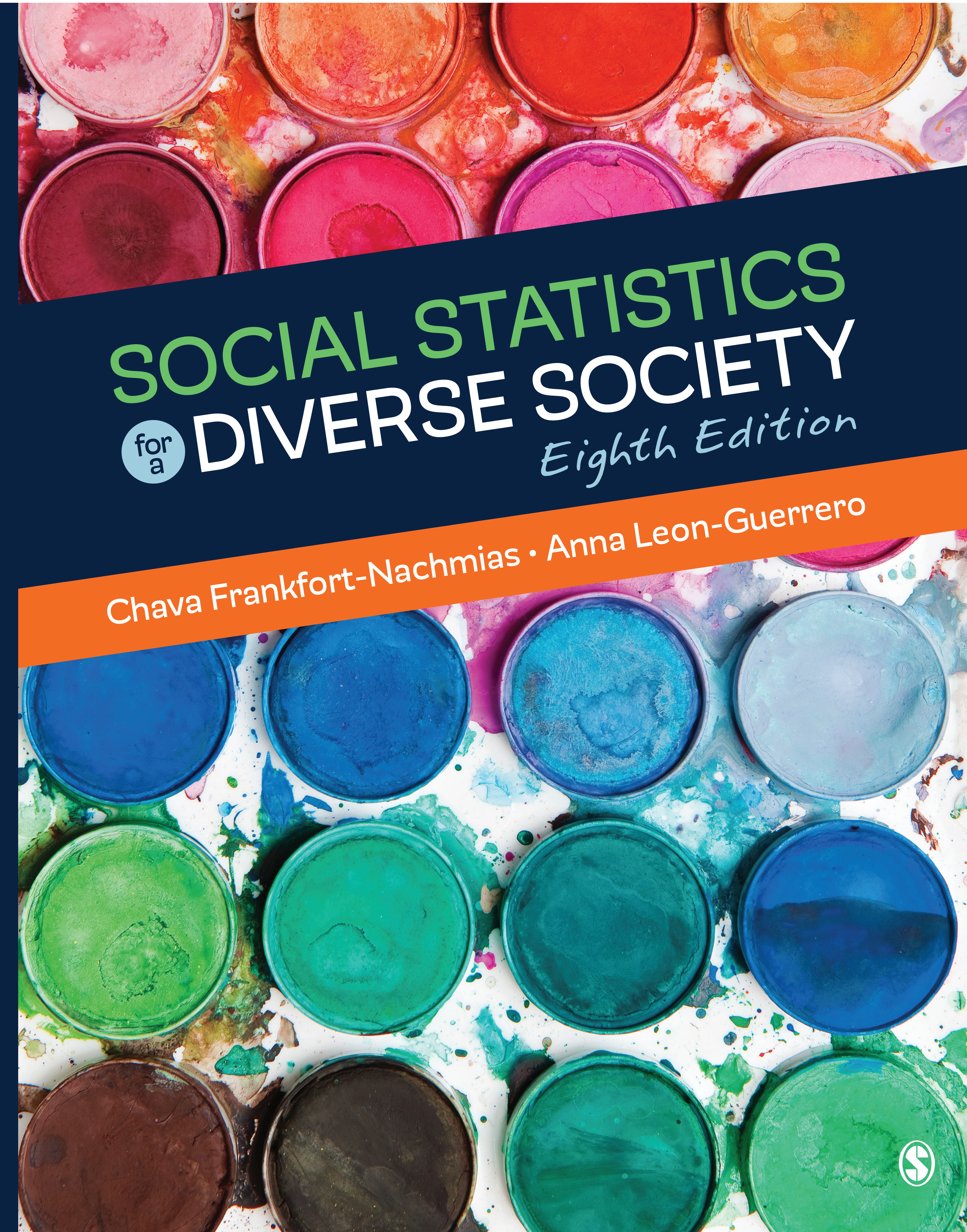
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Frankfort-Nachmias
Leon-Guerrero

SOCIAL STATISTICS for a DIVERSE SOCIETY

Eighth Edition

SAGE



SOCIAL STATISTICS for a DIVERSE SOCIETY

Eighth Edition

Chava Frankfort-Nachmias • Anna Leon-Guerrero

SOCIAL STATISTICS FOR A DIVERSE SOCIETY

Eighth Edition

SOCIAL STATISTICS FOR A DIVERSE SOCIETY

Eighth Edition

Chava Frankfort-Nachmias
University of Wisconsin

Anna Leon-Guerrero
Pacific Lutheran University

SAGE

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

SAGE

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.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Acquisitions Editor: Jeff Lasser
Development Editor: Jessica Miller
Editorial Assistant: Adeline Wilson
eLearning Editor: Gabrielle Piccininni
Production Editor: Kelly DeRosa
Copy Editor: QuADS Prepress (P) Ltd.
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Jennifer Grubba
Indexer: Sheila Bodell
Cover Designer: Candice Harman
Marketing Manager: Kara Kindstrom

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Frankfort-Nachmias, Chava, author. | Leon-Guerrero, Anna, author.

Title: Social statistics for a diverse society / Chava Frankfort-Nachmias, University of Wisconsin, Anna Leon-Guerrero, Pacific Lutheran University.

Description: Eighth edition. | Los Angeles : SAGE, [2016] | Includes bibliographical references and index.

Identifiers: LCCN 2016039109 | ISBN 978-1-5063-4720-2 (pbk. : alk. paper)

Subjects: LCSH: Social sciences—Statistical methods. | Statistics.

Classification: LCC HA29 .N25 2016 | DDC 519.5—dc23 LC record available at https://lccn.loc.gov/2016039109

This book is printed on acid-free paper.

17 18 19 20 21 10 9 8 7 6 5 4 3 2 1

BRIEF CONTENTS

DETAILED CONTENTS

CHAPTER 6 • Sampling and Sampling Distributions 151

CHAPTER 7 • Estimation 179

PREFACE

You may be reading this introduction on your first day of class. We know you have some questions and concerns about what your course will be like. Math, formulas, and calculations? Yes, those will be part of your learning experience. But there is more.

Throughout our text we highlight the relevance of statistics in our daily and professional lives. Data are used to predict public opinion, consumer spending, and even a presidential election. How Americans feel about a variety of political and social topics—race relations, gun control, immigration, the economy, health care reform, or terrorism—are measured by surveys and polls and reported daily by the news media. Your recent Amazon purchase didn't go unnoticed. The study of consumer trends, specifically focusing on young adults, helps determine commercial programming, product advertising and placement, and, ultimately, consumer spending. And as we prepare this text, just months before the 2016 Presidential election, weekly polls have begun predicting the historic election between Hillary Clinton and Donald Trump.

Statistics are not just a part of our lives in the form of news bits or information. And it isn't just numbers either. As social scientists we rely on statistics to help us understand our social world. We use statistical methods and techniques to track demographic trends, to assess social differences, and to better inform social policy. We encourage you to move beyond just being a consumer of statistics and determine how you can use statistics to gain insight into important social issues that affect you and others.

TEACHING AND LEARNING GOALS

Three teaching and learning goals continue to be the guiding principles of our book, as they were in previous editions.

Our first goal is to introduce you to social statistics and demonstrate its value. Although most of you will not use statistics in your own student research, you will be expected to read and interpret statistical information presented by others in professional and scholarly publications, in the workplace, and in the popular media. This book will help you understand the concepts behind the statistics so that you will be able to assess the circumstances in which certain statistics should and should not be used.

A special characteristic of this book is its integration of statistical techniques with substantive issues of particular relevance in the social sciences. Our second goal is to demonstrate that substance and statistical techniques are truly related in social science research. Your learning will not be limited to statistical calculations and formulas. Rather, you will become proficient in statistical techniques while learning about social differences and inequality through numerous substantive examples and real-world data applications. Because the world we live in is characterized by a growing diversity—where personal and social realities are increasingly shaped by race, class, gender, and other categories of experience—this book teaches you basic statistics while incorporating social science research related to the dynamic interplay of our social worlds.

Our third goal is to enhance your learning by using straightforward prose to explain statistical concepts and by emphasizing intuition, logic, and common sense over rote memorization and derivation of formulas.

DISTINCTIVE AND UPDATED FEATURES OF OUR BOOK

Our learning goals are accomplished through a variety of specific and distinctive features throughout this book.

A Close Link Between the Practice of Statistics, Important Social Issues, and Real-World Examples. This book is distinct for its integration of statistical techniques with pressing social issues of particular concern to society and social science. We emphasize how the conduct of social science is the constant interplay between social concerns and methods of inquiry. In addition, the examples throughout the book—mostly taken from news stories, government reports, public opinion polls, scholarly research, and the National Opinion Research Center's General Social Survey—are formulated to emphasize to students like you that we live in a world in which statistical arguments are common. Statistical concepts and procedures are illustrated with real data and research, providing a clear sense of how questions about important social issues can be studied with various statistical techniques.

A Focus on Diversity: The United States and International. A strong emphasis on race, class, and gender as central substantive concepts is mindful of a trend in the social sciences toward integrating issues of diversity in the curriculum. This focus on the richness of social differences within our society and our global neighbors is manifested in the application of statistical tools to examine how race, class, gender, and other categories of experience shape our social world and explain social behavior.

Chapter Reorganization and Content. Each revision presents many opportunities to polish and expand the content of our text. In this edition, we have made a number of changes in response to feedback from reviewers and fellow instructors. We merged frequency distributions and graphic presentation into one chapter. We expanded the discussion of probability in Chapters 6 and 7. We refined the discussion on the interpretation and application of descriptive statistics (variance and standard deviation) and inferential tests (t, Z, F ratio, and regression and correlation). End-of-chapter exercises have been organized into calculation and interpretation problems.

Reading the Research Literature, Statistics in Practice, A Closer Look, and Data at Work. In your student career and in the workplace, you may be expected to read and interpret statistical information presented by others in professional and scholarly publications. These statistical analyses are a good deal more complex than most class and textbook presentations. To guide you in reading and interpreting research reports written by social scientists, most of our chapters include a Reading the Research Literature and a Statistics in Practice feature, presenting excerpts of published research reports or specific SPSS calculations using the statistical concepts under discussion. Being statistically literate involves more than just completing a calculation; it also includes learning how to apply and interpret statistical information and being able to say what it means. We include A Closer Look discussion in each chapter, advising students about the common errors and limitations in quantitative data collection and analysis. A new chapter feature for this eighth edition is Data at Work, profiling men and women who use data in their work settings and professions.

SPSS and GSS 2014. IBM® SPSS® Statistics* is used throughout this book, although the use of computers is not required to learn from the text. Real data are used to motivate and make concrete the coverage of statistical topics. As a companion to the eighth edition's SPSS demonstrations and exercises, we provide two GSS 2014 data sets on the study site at **http://edge.sagepub.com/frankfort8e**. SPSS exercises at the end of each chapter rely on variables from both data modules. There is ample opportunity for instructors to develop their own exercises using these data.

Tools to Promote Effective Study. Each chapter concludes with a list of Main Points and Key Terms discussed in that chapter. Boxed definitions of the Key Terms also appear in the body of the chapter, as do Learning Checks keyed to the most important points. Key Terms are also clearly defined and explained in the Glossary, another special feature in our book. Answers to all the Odd-Numbered Exercises and Learning Checks in the text are included at the end of the book, as well as on the study site at **http://edge.sagepub.com/frankfort8e**. Complete step-by-step solutions are provided in the instructor's manual, available on the study site.

A NOTE ABOUT ROUNDING

Throughout this text and in ancillary materials, we followed these rounding rules: If the number you are rounding is followed by 5, 6, 7, 8, or 9, round the number up. If the number you are rounding is followed by 0, 1, 2, 3, or 4, do not change the number. For rounding long decimals, look only at the number in the place you are rounding to and the number that follows it.

SAGE edge™

edge.sagepub.com/frankfort8e

SAGE edge offers a robust online environment featuring an impressive array of tools and resources for review, study, and further exploration, keeping both instructors and students on the cutting edge of teaching and learning. **SAGE edge** content is open access and available on demand. Learning and teaching has never been easier!

SAGE edge for students provides a personalized approach to help students accomplish their coursework goals in an easy-to-use learning environment.

- Mobile-friendly **eFlashcards** strengthen understanding of key terms and concepts.
- Mobile-friendly practice **quizzes** allow for independent assessment by students of their mastery of course material.
- A customized online **action plan** includes tips and feedback on progress through the course and materials, which allows students to individualize their learning experience.
- **Web exercises** and meaningful web links facilitate student use of Internet resources, further exploration of topics, and responses to critical thinking questions.

*SPSS is a registered trademark of International Business Machines Corporation.

- EXCLUSIVE! Access to full-text **SAGE journal articles** that have been carefully selected to support and expand on the concepts presented in each chapter.
- Access to **GSS 2014 data sets.**

SAGE edge for instructors supports teaching by making it easy to integrate quality content and create a rich learning environment for students.

- **Test banks** provide a diverse range of pre-written options as well as the opportunity to edit any question and/or insert personalized questions to effectively assess students' progress and understanding.
- **Sample syllabus** provides a suggested model for instructors to use when creating the syllabi for their courses.
- Editable, chapter-specific **PowerPoint® slides** offer complete flexibility for creating a multimedia presentation for the course.
- EXCLUSIVE! Access to full-text **SAGE journal articles** have been carefully selected to support and expand on the concepts presented in each chapter to encourage students to think critically.
- **Multimedia content** includes web resources and web exercises that appeal to students with different learning styles.
- **Lecture notes** summarize key concepts by chapter to ease preparation for lectures and class discussions.
- Lively and stimulating ideas for **class activities** that can be used in class to reinforce active learning.
- Chapter-specific **discussion questions** help launch classroom interaction by prompting students to engage with the material and by reinforcing important content.
- A **course cartridge** provides easy LMS (Learning Management System) integration.

ACKNOWLEDGMENTS

We are both grateful to Jerry Westby, Series Editor for SAGE Publications, for his commitment to our book and for his invaluable assistance through the production process.

Many manuscript reviewers recruited by SAGE provided invaluable feedback. For their thoughtful comments to the eighth edition, we thank

Andrew S. Fullerton, Oklahoma State University

David A. Gay, University of Central Florida

Dr. Lindsey Peterson, Mississippi State University

Heather Macpherson Parrott, Long Island University-Post

Christopher Donoghue, Montclair State University

S. Michael Gaddis, The Pennsylvania State University

Jann W. MacInnes, University of Florida

Laura Sullivan, Brandeis University

Warren Waren, Texas A&M University

Joe Weinberg, University of Southern Mississippi

For their comments to the seventh edition, we thank

Walter F. Carroll, Bridgewater State University

Andrew S. Fullerton, Oklahoma State University

David A. Gay, University of Central Florida

Judith G. Gonyea, Boston University

Megan Henly, University of New Hampshire

Patricia A. Jaramillo, The University of Texas at San Antonio

Brett Lehman, Louisiana State University

James W. Love, California State University, Fullerton

Kay Kei-Ho Pih, California State University, Northridge

For their comments to the sixth edition, we thank

Diane Balduzy, Massachusetts College of Liberal Arts

Ellen Berg, California State University–Sacramento

Robert Carini, University of Louisville

Melissa Evans-Andris, University of Louisville

Meredith Greif, Georgia State University

Kristen Kenneavy, Ramapo College

Dave Rausch, West Texas A&M University

Billy Wagner, California State University–Channel Islands

Kevin Yoder, University of North Texas

For their comments to the fifth edition, we thank

Anna A. Amirkhanyan, The American University

Robert Carini, University of Louisville

Patricia Case, University of Toledo

Stanley DeViney, University of Maryland Eastern Shore

David Gay, University of Central Florida

Dusten R. Hollist, University of Montana

Ross Koppel, University of Pennsylvania

Benny Marcus, Temple University

Matt G. Mutchler, California State University Dominguez Hills

Mahasin C. Owens-Sabir, Jackson State University

Dave Rausch, West Texas A&M University

Kevin Yoder, University of North Texas

We are grateful to Jessica Miller and Kelly DeRosa for guiding the book through the production process. We would also like to acknowledge Laura Kirkhuff, Krishna Pradeep Joghee, and the rest of the SAGE staff for their assistance on this edition.

We extend our deepest appreciation to Michael Clark for his fine editing and data work. Among his many contributions, Michael would relate our revision goals to his student experience, reminding us of how students can learn and successfully master this material.

Chava Frankfort-Nachmias would like to thank and acknowledge her friends and colleagues for their unending support; she also would like to thank her students:

I am grateful to my students at the University of Wisconsin–Milwaukee, who taught me that even the most complex statistical ideas can be simplified. The ideas presented in this book are the products of many years of classroom testing. I thank my students for their patience and contributions.

Finally, I thank my partner, Marlene Stern, for her love and support.

Anna Leon-Guerrero would like to thank her Pacific Lutheran University students for inspiring her to be a better teacher. My love and thanks to my husband, Brian Sullivan.

Chava Frankfort-Nachmias

University of Wisconsin–Milwaukee

Anna Leon-Guerrero

Pacific Lutheran University

ABOUT THE AUTHORS

Chava Frankfort-Nachmias is an Emeritus Professor of Sociology at the University of Wisconsin–Milwaukee. She is the coauthor of *Research Methods in the Social Sciences* (with David Nachmias), coeditor of *Sappho in the Holy Land* (with Erella Shadmi), and numerous publications on ethnicity and development, urban revitalization, science and gender, and women in Israel. She was the recipient of the University of Wisconsin System teaching improvement grant on integrating race, ethnicity, and gender into the social statistics and research methods curriculum. She is also the coauthor (with Anna Leon-Guerrero) of *Essentials of Social Statistics.*

Anna Leon-Guerrero is Professor of Sociology at Pacific Lutheran University in Washington. She received her Ph.D. in sociology from the University of California–Los Angeles. A recipient of the university's Faculty Excellence Award and the K.T. Tang Award for Excellence in Research, she teaches courses in statistics, social theory, and social problems. She is also the author of *Social Problems: Community, Policy, and Social Action*.

1

THE WHAT AND THE WHY OF STATISTICS

Chapter Learning Objectives

1. Describe the five stages of the research process
2. Define independent and dependent variables
3. Distinguish between the three levels of measurement
4. Apply descriptive and inferential statistical procedures

Are you taking statistics because it is required in your major—not because you find it interesting? If so, you may be feeling intimidated because you associate statistics with numbers, formulas, and abstract notations that seem inaccessible and complicated. Perhaps you feel intimidated not only because you're uncomfortable with math but also because you suspect that numbers and math don't leave room for human judgment or have any relevance to your own personal experience. In fact, you may even question the relevance of statistics to understanding people, social behavior, or society.

In this book, we will show you that statistics can be a lot more interesting and easy to understand than you may have been led to believe. In fact, as we draw on your previous knowledge and experience and relate statistics to interesting and important social issues, you'll begin to see that statistics is not just a course you have to take but a useful tool as well.

There are two reasons why learning statistics may be of value to you. First, you are constantly exposed to statistics every day of your life. Marketing surveys, voting polls, and social research findings appear daily in the news media. By learning statistics, you will become a sharper consumer of statistical material. Second, as a major in the social sciences, you may be expected to read and interpret statistical information related to your occupation or work. Even if conducting research is not a part of your work, you may still be expected to understand and learn from other people's research or to be able to write reports based on statistical analyses.

Just what is statistics anyway? You may associate the word with numbers that indicate birthrates, conviction rates, per capita income, marriage and divorce rates, and so on. But the word **statistics** also refers to a set of procedures used by social scientists to organize, summarize, and communicate numerical information. Only information represented by numbers can be the subject of statistical analysis. Such information is called **data**; researchers use statistical procedures to analyze data to answer research questions and test theories. It is the latter usage—answering research questions and testing theories—that this textbook explores.

Statistics A set of procedures used by social scientists to organize, summarize, and communicate numerical information.

Data Information represented by numbers, which can be the subject of statistical analysis.

THE RESEARCH PROCESS

To give you a better idea of the role of statistics in social research, let's start by looking at the **research process**. We can think of the research process as a set of activities in which social scientists engage so that they can answer questions, examine ideas, or test theories.

Research process A set of activities in which social scientists engage to answer questions, examine ideas, or test theories.

As illustrated in Figure 1.1, the research process consists of five stages:

1. Asking the research question
2. Formulating the hypotheses
3. Collecting data
4. Analyzing data
5. Evaluating the hypotheses

Each stage affects the theory and is affected by it as well. Statistics is most closely tied to the data analysis stage of the research process. As we will see in later chapters, statistical analysis of the data helps researchers test the validity and accuracy of their hypotheses.

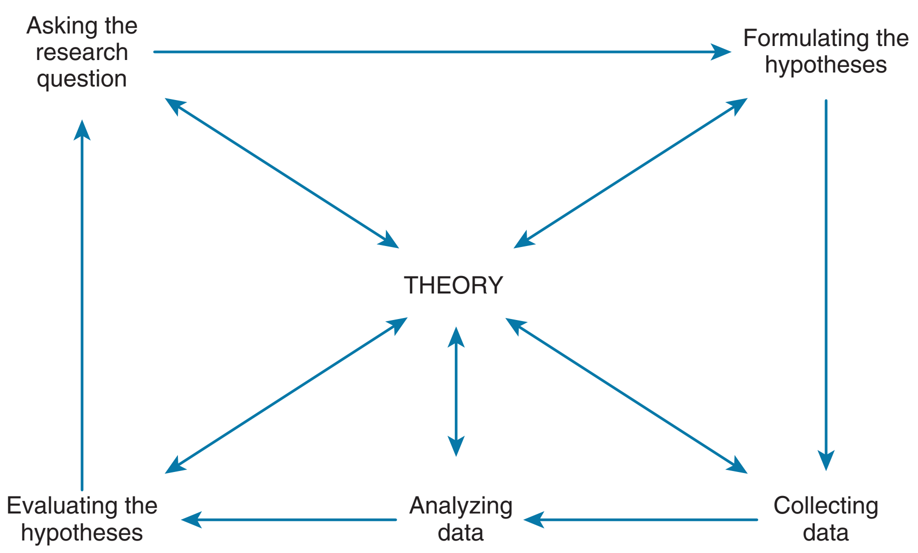
ASKING RESEARCH QUESTIONS

The starting point for most research is asking a research question. Consider the following research questions taken from a number of social science journals:

How will the Affordable Care Act influence the quality of health care?

Has support for same-sex marriage increased during the past decade?

Figure 1.1 The Research Process



Does race or ethnicity predict voting behavior?

What factors affect the economic mobility of female workers?

These are all questions that can be answered by conducting **empirical research**—research based on information that can be verified by using our direct experience. To answer research questions, we cannot rely on reasoning, speculation, moral judgment, or subjective preference. For example, the questions "Is racial equality good for society?" and "Is an urban lifestyle better than a rural lifestyle?" cannot be answered empirically because the terms good and better are concerned with values, beliefs, or subjective preference and, therefore, cannot be independently verified. One way to study these questions is by defining good and better in terms that can be verified empirically. For example, we can define good in terms of economic growth and better in terms of psychological well-being. These questions could then be answered by conducting empirical research.

Empirical research Research based on evidence that can be verified by using our direct experience.

You may wonder how to come up with a research question. The first step is to pick a question that interests you. If you are not sure, look around! Ideas for research problems are all around you, from media sources to personal experience or your own intuition. Talk to other people, write down your own observations and ideas, or learn what other social scientists have written about.

Take, for instance, the relationship between gender and work. As a college student about to enter the labor force, you may wonder about the similarities and differences between women's and men's work experiences and about job opportunities when you graduate. Here are some facts and observations based on research reports: In 2015, women who were employed full time earned about $726 (in current dollars) per week on average; men who were employed full time earned $895 (in current dollars) per week on average.[1] Women's and men's work are also very different. Women continue to be the minority in many of the higher ranking and higher salaried positions in professional and managerial occupations. For example, in 2014, women made up 25.3% of architects, 16.5% of civil engineers, 12.4% of police and sheriff's patrol officers, and 2.4% of electricians. In comparison, among all those employed as preschool and kindergarten teachers, 98% were women. Among all receptionists and information clerks in 2014, 91% were women.[2] Another noteworthy development in the history of labor in the United States took place in January 2010: Women outnumbered men for the first time in the labor force by holding 50.3% of all nonfarm payroll jobs.[3] These observations may prompt us to ask research questions such as the following: How much change has there been in women's work over time? Are women paid, on average, less than men for the same type of work?

LEARNING CHECK 1.1

Identify one or two social science questions amenable to empirical research. You can almost bet that you will be required to do a research project sometime in your college career.

THE ROLE OF THEORY

You may have noticed that each preceding research question was expressed in terms of a relationship. This relationship may be between two or more attributes of individuals or groups, such as gender and income or gender segregation in the workplace and income disparity. The relationship between attributes or characteristics of individuals and groups lies at the heart of social scientific inquiry.

Most of us use the term theory quite casually to explain events and experiences in our daily life. You may have a theory about why your roommate has been so nice to you lately or why you didn't do so well on your last exam. In a somewhat similar manner, social scientists attempt to explain the nature of social reality. Whereas our theories about events in our lives are commonsense explanations based on educated guesses and personal experience, to the social scientist, a theory is a more precise explanation that is frequently tested by conducting research.

Theory A set of assumptions and propositions used to explain, predict, and understand social phenomena.

A **theory** is a set of assumptions and propositions used by social scientists to explain, predict, and understand the phenomena they study.[4] The theory attempts to establish a link between what we observe (the data) and our conceptual understanding of why certain phenomena are related to each other in a particular way.

For instance, suppose we wanted to understand the reasons for the income disparity between men and women; we may wonder whether the types of jobs men and women have and the organizations in which they work have something to do with their wages. One explanation for gender wage inequality is gender segregation in the workplace—the fact that American men and women are concentrated in different kinds of jobs and occupations. What is the significance of gender segregation in the workplace? In our society, people's occupations and jobs are closely associated with their level of prestige, authority, and income. The jobs in which women and men are segregated are not only different but also unequal. Although the proportion of women in the labor force has markedly increased, women are still concentrated in occupations with low pay, low prestige, and few opportunities for promotion. Thus, gender segregation in the workplace is associated with unequal earnings, authority, and status. In particular, women's segregation into different jobs and occupations from those of men is the most immediate cause of the pay gap. Women receive lower pay than men do even when they have the same level of education, skill, and experience as men in comparable occupations.

FORMULATING THE HYPOTHESES

So far, we have come up with a number of research questions about the income disparity between men and women in the workplace. We have also discussed a possible explanation—a theory—that helps us make sense of gender inequality in wages. Is that enough? Where do we go from here?

Hypothesis A statement predicting the relationship between two or more observable attributes.

Our next step is to test some of the ideas suggested by the gender segregation theory. But this theory, even if it sounds reasonable and logical to us, is too general and does not contain enough specific information to be tested. Instead, theories suggest specific concrete predictions or **hypotheses** about the way that observable attributes of people or groups are interrelated in real life. Hypotheses are tentative because they can be verified only after they have been tested empirically.[5] For example, one hypothesis we can derive from the gender segregation theory is that wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Not all hypotheses are derived directly from theories. We can generate hypotheses in many ways—from theories, directly from observations, or from intuition. Probably, the greatest source of hypotheses is the professional or scholarly literature. A critical review of the scholarly literature will familiarize you with the current state of knowledge and with hypotheses that others have studied.

Table 1.1 Variables and Value Categories

Variable	Categories
Social class	Lower Working Middle Upper
Gender	Male Female
Education	Less than high school High school Some college College graduate

Let's restate our hypothesis:

> Wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Note that this hypothesis is a statement of a relationship between two characteristics that vary: wages and gender composition of occupations. Such characteristics are called variables. A **variable** is a property of people or objects that takes on two or more values. For example, people can be classified into a number of social class categories, such as upper class, middle class, or working class. Family income is a variable; it can take on values from zero to hundreds of thousands of dollars or more. Similarly, gender composition is a variable. The percentage of females (or males) in an occupation can vary from 0 to 100. Wages is a variable, with values from zero to thousands of dollars or more. See Table 1.1 for examples of some variables and their possible values.

Variable A property of people or objects that takes on two or more values.

Social scientists must also select a **unit of analysis**; that is, they must select the object of their research. We often focus on individual characteristics or behavior, but we could also examine groups of people such as families, formal organizations like elementary schools or corporations, or social artifacts such as children's books or advertisements. For example, we may be interested in the relationship between an individual's educational degree and annual income. In this case, the unit of analysis is the individual. On the other hand, in a study of how corporation profits are associated with employee benefits, corporations are the unit of analysis. If we examine how often women are featured in prescription drug advertisements, the advertisements are the unit of analysis. Figure 1.2 illustrates different units of analysis frequently employed by social scientists.

Unit of analysis The object of research, such as individuals, groups, organizations, or social artifacts.

LEARNING CHECK 1.2

Remember that research question you came up with? Formulate a testable hypothesis based on your research question. Remember that your variables must take on two or more values and you must determine the unit of analysis. What is your unit of analysis?

Figure 1.2 Examples of Units of Analysis

Independent and Dependent Variables: Causality

Hypotheses are usually stated in terms of a relationship between an independent and a dependent variable. The distinction between an independent and a dependent variable is important in the language of research. Social theories often intend to provide an explanation for social patterns or causal relations between variables. For example, according to the gender segregation theory, gender segregation in the workplace is the primary explanation (although certainly not the only one) of the male-female earning gap. Why should jobs where the majority of workers are women pay less than jobs that employ mostly men? One explanation is that

> societies undervalue the work women do, regardless of what those tasks are, because women do them. . . . For example, our culture tends to devalue caring or nurturant work at least partly because it is done by women. This tendency accounts for child care workers' low rank in the pay hierarchy.[6]

In the language of research, the variable the researcher wants to explain (the "effect") is called the **dependent variable**. The variable that is expected to "cause" or account for the dependent variable is called the **independent variable**. Therefore, in our example, *gender composition of occupations* is the independent variable, and *wages* is the dependent variable.

Dependent variable The variable to be explained (the effect).

Independent variable The variable expected to account for (the cause of) the dependent variable.

Cause-and-effect relationships between variables are not easy to infer in the social sciences. To establish that two variables are causally related, your analysis must meet three conditions: (1) The cause has to precede the effect in time, (2) there has to be an empirical relationship between the cause and the effect, and (3) this relationship cannot be explained by other factors.

Let's consider the decades-old debate about controlling crime through the use of prevention versus punishment. Some people argue that special counseling for youths at the first sign of trouble and strict controls on access to firearms would help reduce crime. Others argue that overhauling federal and state sentencing laws to stop early prison releases is the solution. In the early 1990s, Washington and California adopted "three strikes and you're out" legislation, imposing life prison terms on three-time felony offenders. Such laws are also referred to as habitual or persistent offender laws. Twenty-six other states and the federal government adopted similar measures, all advocating a "get tough" policy on crime; the most recent legislation was in 2012 in the state of Massachusetts. In 2012, California voters supported a revision to the original law, imposing a life sentence only when the new felony conviction is serious or violent. Let's suppose that years after the measure was introduced, the crime rate declined in some of these states (in fact, advocates of the measure have identified declining crime rates as evidence of its success). Does the observation that the incidence of crime declined mean that the new measure caused this reduction? Not necessarily! Perhaps the rate of crime had been going down for other reasons, such as improvement in the economy, and the new measure had nothing to do with it. To demonstrate a cause-and-effect relationship, we would need to show three things: (1) The reduction of crime actually occurred after the enactment of this measure, (2) the enactment of the "three strikes and you're out" measure was empirically associated with a decrease in crime, and (3) the relationship between the reduction in crime and the "three strikes and you're out" policy is not due to the influence of another variable (e.g., the improvement of overall economic conditions).

Independent and Dependent Variables: Guidelines

Because it is difficult to infer cause-and-effect relationships in the social sciences, be cautious about using the terms cause and effect when examining relationships between variables. However, using the terms independent variable and dependent variable is still appropriate even when this relationship is not articulated in terms of direct cause and effect. Here are a few guidelines that may help you identify the independent and dependent variables:

1. The dependent variable is always the property that you are trying to explain; it is always the object of the research.
2. The independent variable usually occurs earlier in time than the dependent variable.
3. The independent variable is often seen as influencing, directly or indirectly, the dependent variable.

The purpose of the research should help determine which is the independent variable and which is the dependent variable. In the real world, variables are neither dependent nor independent; they can be switched around depending on the research problem. A variable defined as independent in one research investigation may be a dependent variable in another.[7] For instance, *educational attainment* may be an independent variable in a study attempting to explain how education influences political attitudes. However, in an investigation of whether a person's level of education is influenced by the social status of his or her family of origin, *educational attainment* is the dependent variable. Some variables, such as race, age, and ethnicity, because they are primordial characteristics that cannot be explained by social scientists, are never considered dependent variables in a social science analysis.

LEARNING CHECK 1.3

Identify the independent and dependent variables in the following hypotheses:

- *Older Americans are more likely to support stricter immigration laws than younger Americans.*
- *People who attend church regularly are more likely to oppose abortion than people who do not attend church regularly.*
- *Elderly women are more likely to live alone than elderly men.*
- *Individuals with postgraduate education are likely to have fewer children than those with less education.*

What are the independent and dependent variables in your hypothesis?

COLLECTING DATA

Once we have decided on the research question, the hypothesis, and the variables to be included in the study, we proceed to the next stage in the research cycle. This step includes measuring our variables and collecting the data. As researchers, we must decide how to measure the variables of interest to us, how to select the cases for our research, and what kind of data collection techniques we will be using. A wide variety of data collection techniques are available to us, from direct observations to survey research, experiments, or secondary sources. Similarly, we can construct numerous measuring instruments. These instruments can be as simple as a single question included in a questionnaire or as complex as a composite measure constructed through the combination of two or more questionnaire items. The choice of a particular data collection method or instrument to measure our variables depends on the study objective. For instance, suppose we decide to study how one's social class is related to attitudes about women in the labor force. Since attitudes about working women are not directly observable, we need to collect data by asking a group of people questions about their attitudes and opinions. A suitable method of data collection for this project would be a survey that uses some kind of questionnaire or interview guide to elicit verbal reports from respondents. The questionnaire could include numerous questions designed to measure attitudes toward working women, social class, and other variables relevant to the study.

How would we go about collecting data to test the hypothesis relating the gender composition of occupations to wages? We want to gather information on the proportion of men and women in different occupations and the average earnings for these occupations. This kind of information is routinely collected and disseminated by the U.S. Department of Labor, the Bureau of Labor Statistics, and the U.S. Census Bureau. We could use these data to test our hypothesis.

Levels of Measurement

The statistical analysis of data involves many mathematical operations, from simple counting to addition and multiplication. However, not every operation can be used with every variable. The type of statistical operation we employ depends on how our variables are measured. For example, for the variable *gender*, we can use the number 1 to represent females and the number 2 to represent males. Similarly, 1 can also be used as a numerical code for the category "one child" in the variable *number of children*. Clearly, in the first example, the number is an arbitrary symbol that does not correspond to the property "female," whereas in the second example the number 1 has a distinct numerical meaning that does correspond to the property "one child." The correspondence between the properties we measure and the numbers representing these properties determines the type of statistical operations we can use. The degree of correspondence also leads to different ways of measuring—that is, to distinct levels of measurement. In this section, we will discuss three levels of measurement: (1) nominal, (2) ordinal, and (3) interval-ratio.

NOMINAL LEVEL OF MEASUREMENT At the **nominal** level of measurement, numbers or other symbols are assigned a set of categories for the purpose of naming, labeling, or classifying the observations. *Gender* is an example of a nominal-level variable (Table 1.2). Using the numbers 1 and 2, for instance, we can classify our observations into the categories "females" and "males," with 1 representing females and 2 representing males. We could use any of a variety of symbols to represent the different categories of a nominal variable; however, when numbers are used to represent the different categories, we do not imply anything about the magnitude or quantitative difference between the categories. Nominal categories cannot be rank-ordered. Because the different categories (e.g., males vs. females) vary in the quality inherent in each but not in quantity, nominal variables are often called qualitative. Other examples of nominal-level variables are political party, religion, and race.

Nominal measurement
Numbers or other symbols are assigned to a set of categories for the purpose of naming, labeling, or classifying the observations. Nominal categories cannot be rank-ordered.

Nominal variables should include categories that are both exhaustive and mutually exclusive. Exhaustiveness means that there should be enough categories composing the variables to classify every observation. For example, the common classification of the variable marital status into the categories "married," "single," and "widowed" violates the requirement of exhaustiveness. As defined, it does not allow us to classify same-sex couples or heterosexual couples who are not legally married. We can make every variable exhaustive by adding the category "other"

Table 1.2 Nominal Variables and Value Categories

Variable	Categories
Gender	Male Female
Religion	Protestant Christian Jewish Muslim
Marital status	Married Single Widowed Other

to the list of categories. However, this practice is not recommended if it leads to the exclusion of categories that have theoretical significance or a substantial number of observations.

Mutual exclusiveness means that there is only one category suitable for each observation. For example, we need to define religion in such a way that no one would be classified into more than one category. For instance, the categories Protestant and Methodist are not mutually exclusive because Methodists are also considered Protestant and, therefore, could be classified into both categories.

LEARNING CHECK 1.4

Review the definitions of exhaustive and mutually exclusive. Now look at Table 1.2. What other categories could be added to each variable to be exhaustive and mutually exclusive?

Ordinal measurement Numbers are assigned to rank-ordered categories ranging from low to high.

ORDINAL LEVEL OF MEASUREMENT Whenever we assign numbers to rank-ordered categories ranging from low to high, we have an **ordinal** level of measurement. *Social class* is an example of an ordinal variable. We might classify individuals with respect to their social class status as "upper class," "middle class," or "working class." We can say that a person in the category "upper class" has a higher class position than a person in a "middle-class" category (or that a "middle-class" position is higher than a "working-class" position), but we do not know the magnitude of the differences between the categories—that is, we don't know how much higher "upper class" is compared with the "middle class."

Many attitudes that we measure in the social sciences are ordinal-level variables. Take, for instance, the following statement used to measure attitudes toward working women: "Women should return to their traditional role in society." Respondents are asked to identify the number representing their degree of agreement or disagreement with this statement. One form in which a number might be made to correspond with the answers can be seen in Table 1.3. Although the differences between these numbers represent higher or lower degrees of agreement with the statement, the distance between any two of those numbers does not have a precise numerical meaning.

Like nominal variables, ordinal variables should include categories that are mutually exhaustive and exclusive.

Table 1.3 Ordinal Ranking Scale

Rank	Value
1	Strongly agree
2	Agree
3	Neither agree nor disagree
4	Disagree
5	Strongly disagree

INTERVAL-RATIO LEVEL OF MEASUREMENT If the categories (or values) of a variable can be rank-ordered and if the measurements for all the cases are expressed in the same units, and equally spaced, then an **interval-ratio** level of measurement has been achieved. Examples of variables measured at the interval-ratio level are *age*, *income*, and *SAT scores*. With all these variables, we can compare values not only in terms of which is larger or smaller but also in terms of how much larger or smaller one is compared with another. In some discussions of levels of measurement, you will see a distinction made between interval-ratio variables that have a natural zero point (where zero means the absence of the property) and those variables that have zero as an arbitrary point. For example, weight and length have a natural zero point, whereas temperature has an arbitrary zero point. Variables with a natural zero point are also called *ratio variables*. In statistical practice, however, ratio variables are subjected to operations that treat them as interval and ignore their ratio properties. Therefore, we make no distinction between these two types in this text.

Interval-ratio measurement Measurements for all cases are expressed in the same units and equally spaced. Interval-ratio values can be rank-ordered.

CUMULATIVE PROPERTY OF LEVELS OF MEASUREMENT Variables that can be measured at the interval-ratio level of measurement can also be measured at the ordinal and nominal levels. As a rule, properties that can be measured at a higher level (interval-ratio is the highest) can also be measured at lower levels, but not vice versa. Let's take, for example, *gender composition of occupations*, the independent variable in our research example. Table 1.4 shows the percentage of women in five major occupational groups.

The variable *gender composition* (measured as the percentage of women in the occupational group) is an interval-ratio variable and, therefore, has the properties of nominal, ordinal, and interval-ratio measures. For example, we can say that the management group differs from the natural resources group (a nominal comparison), that service occupations have more women than the other occupational categories (an ordinal comparison), and that service occupations have 35 percentage points more women (56.7 - 21.7) than production occupations (an interval-ratio comparison).

The types of comparisons possible at each level of measurement are summarized in Table 1.5 and Figure 1.3. Note that differences can be established at each of the three levels, but only at the interval-ratio level can we establish the magnitude of the difference.

LEVELS OF MEASUREMENT OF DICHOTOMOUS VARIABLES A variable that has only two values is called a **dichotomous variable**. Several key social factors, such as gender, employment status, and marital status, are dichotomies—that is, you are male or female, employed or unemployed, married or not married. Such variables may seem to be measured at the nominal level: You fit in either one category or the other. No category is naturally higher or lower than the other, so they can't be ordered.

Dichotomous variable A variable that has only two values.

Table 1.4 Gender Composition of Five Major Occupational Groups, 2014

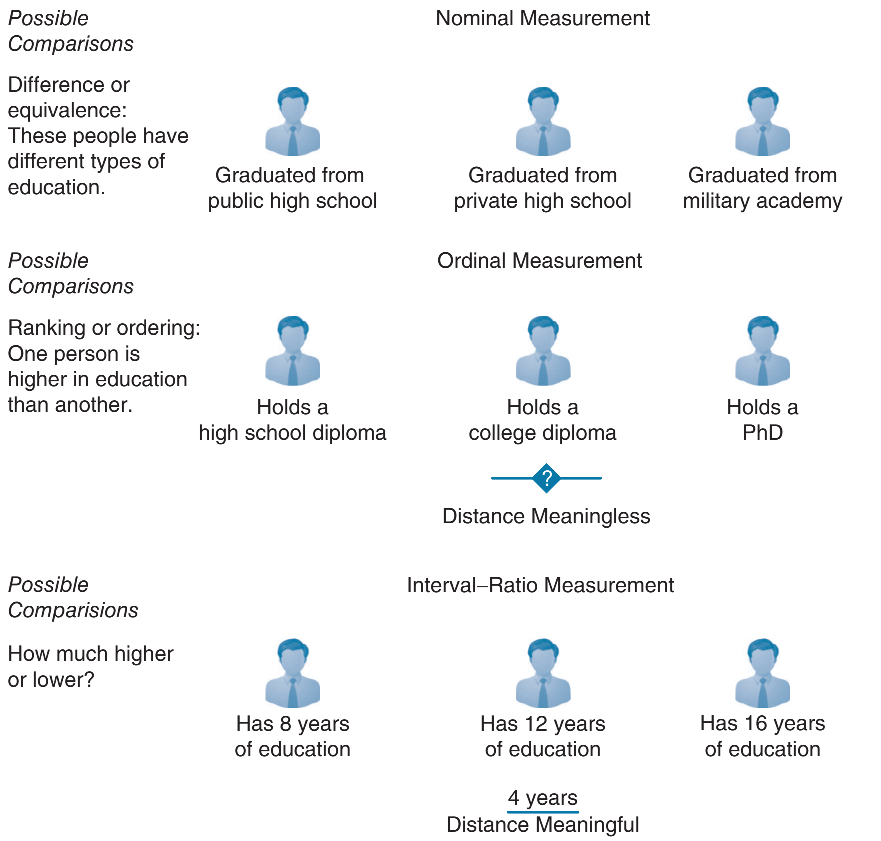
Occupational Group	Women in Occupation (%)
Management, professional, and related occupations	51.6
Service occupations	56.7
Production, transportation, and materials occupations	21.7
Sales and office occupations	61.8
Natural resources, construction, and maintenance occupations	4.4

Source: U.S. Department of Labor, 2015, *Labor Force Statistics from the Current Population Survey 2014*, Table 11

Table 1.5 Levels of Measurement and Possible Comparisons

Level	Different or Equivalent	Higher or Lower	How Much Higher
Nominal	Yes	No	No
Ordinal	Yes	Yes	No
Interval-ratio	Yes	Yes	Yes

Figure 1.3 Levels of Measurement and Possible Comparisons: Education Measured on Nominal, Ordinal, and Interval-Ratio Levels



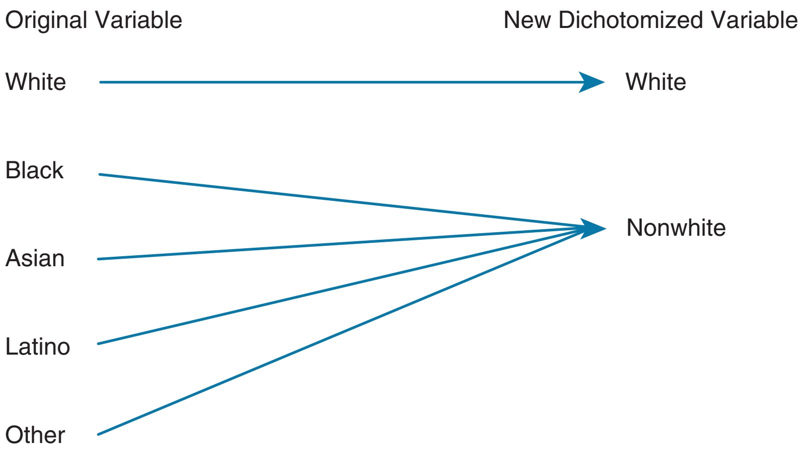
LEARNING CHECK 1.5

Make sure you understand these levels of measurement. As the course progresses, your instructor is likely to ask you what statistical procedure you would use to describe or analyze a set of data. To make the proper choice, you must know the level of measurement of the data.

However, because there are only two possible values for a dichotomy, we can measure it at the ordinal or the interval-ratio level. For example, we can think of "femaleness" as the ordering principle for gender, so that "female" is higher and "male" is lower. Using "maleness" as the ordering principle, "female" is lower and "male" is higher. In either case, with only two classes, there is no way to get them out of order; therefore, gender could be considered at the ordinal level.

Dichotomous variables can also be considered to be interval-ratio level. Why is this? In measuring interval-ratio data, the size of the interval between the categories is meaningful: The distance between 4 and 7, for example, is the same as the distance between 11 and 14. But with a dichotomy, there is only one interval. Therefore, there is really no other distance to which we can compare it. Mathematically, this gives the dichotomy more power than other nominal-level variables (as you will notice later in the text).

For this reason, researchers often dichotomize some of their variables, turning a multicategory nominal variable into a dichotomy. For example, you may see race dichotomized into "white" and "nonwhite." Though we would lose the ability to examine each unique racial category and we may collapse categories that are not similar, it may be the most logical statistical step to take. When you dichotomize a variable, be sure that the two categories capture a distinction that is important to your research question (e.g., a comparison of the number of white vs. nonwhite U.S. senators).



Discrete and Continuous Variables

The statistical operations we can perform are also determined by whether the variables are continuous or discrete. Discrete variables have a minimum-sized unit of measurement, which cannot be subdivided. The number of children per family is an example of a discrete variable because the minimum unit is one child. A family may have two or three children, but not 2.5 children. The variable *wages* in our research example is a discrete variable because currency has a minimum unit (1 cent), which cannot be subdivided. One can have $101.21 or $101.22 but not $101.21843. Wages cannot differ by less than 1 cent—the minimum-sized unit.

Unlike discrete variables, continuous variables do not have a minimum-sized unit of measurement; their range of values can be subdivided into increasingly smaller fractional values. *Length* is an example of a continuous variable because there is no minimum unit of length. A particular object may be 12 in. long, it may be 12.5 in. long, or it may be 12.532011 in. long. Although we cannot always measure all possible length values with absolute accuracy, it is possible for objects to exist at an infinite number of lengths.[8] In principle, we can speak of a tenth of an inch, a ten thousandth of an inch, or a ten trillionth of an inch. The variable *gender composition of occupations* is a continuous variable because it is measured in proportions or percentages (e.g., the percentage of women civil engineers), which can be subdivided into smaller and smaller fractions.

This attribute of variables—whether they are continuous or discrete—affects subsequent research operations, particularly measurement procedures, data analysis, and methods of inference and generalization. However, keep in mind that, in practice, some discrete variables can be treated as if they were continuous, and vice versa.

LEARNING CHECK 1.6

Name three continuous and three discrete variables. Determine whether each of the variables in your hypothesis is continuous or discrete.

A CLOSER LOOK 1.1

A CAUTIONARY NOTE: MEASUREMENT ERROR

Social scientists attempt to ensure that the research process is as error free as possible, beginning with how we construct our measurements. We pay attention to two characteristics of measurement: (1) reliability and (2) validity.

Reliability means that the measurement yields consistent results each time it is used. For example, asking a sample of individuals "Do you approve or disapprove of President Barack Obama's job performance?" is more reliable than asking "What do you think of President Donald Trump's job performance?" While responses to the second question are meaningful, the answers might be vague and could be subject to different interpretation. Researchers look for the consistency of measurement over time, in relationship with other related measures, or in measurements or observations made by two or more researchers. Reliability is a prerequisite for validity: We cannot measure a phenomenon if the measure we are using gives us inconsistent results.

Validity refers to the extent to which measures indicate what they are intended to measure. While standardized IQ tests are reliable, it is still debated whether such tests measure intelligence or one's test-taking ability. A measure may not be valid due to individual error (individuals may want to provide socially desirable responses) or method error (questions may be unclear or poorly written).

Specific techniques and practices for determining and improving measurement reliability and validity are the subject of research methods courses.

ANALYZING DATA AND EVALUATING THE HYPOTHESES

Following the data collection stage, researchers analyze their data and evaluate the hypotheses of the study. The data consist of codes and numbers used to represent their observations. In our example, each occupational group would be represented by two scores: (1) the percentage of women and (2) the average wage. If we had collected information on 100 occupations, we would end up with 200 scores, 2 per occupational group. However, the typical research project includes more variables; therefore, the amount of data the researcher confronts is considerably larger. We now must find a systematic way to organize these data, analyze them, and use some set of procedures to decide what they mean. These last steps make up the statistical analysis stage, which is the main topic of this textbook. It is also at this point in the research cycle that statistical procedures will help us evaluate our research hypothesis and assess the theory from which the hypothesis was derived.

Descriptive and Inferential Statistics

Statistical procedures can be divided into two major categories: (1) descriptive statistics and (2) inferential statistics. Before we can discuss the difference between these two types of statistics, we need to understand the terms population and sample. A **population** is the total set of individuals, objects, groups, or events in which the researcher is interested. For example, if we were interested in looking at voting behavior in the last presidential election, we would probably define our population as all citizens who voted in the election. If we wanted to understand the employment patterns of Latinas in our state, we would include in our population all Latinas in our state who are in the labor force.

Population The total set of individuals, objects, groups, or events in which the researcher is interested.

Although we are usually interested in a population, quite often, because of limited time and resources, it is impossible to study the entire population. Imagine interviewing all the citizens of the United States who voted in the last election or even all the Latinas who are in the labor force in our state. Not only would that be very expensive and time-consuming, but we would also probably have a very hard time locating everyone! Fortunately, we can learn a lot about a population if we carefully select a subset from that population. A subset of cases selected from a population is called a **sample**. The process of identifying and selecting this subset is referred to as **sampling**. Researchers usually collect their data from a sample and then generalize their observations to the population. The ultimate goal of sampling is to have a subset that closely resembles the characteristics of the population. Because the sample is intended to represent the population that we are interested in, social scientists take sampling seriously. We'll explore different sampling methods in Chapter 6.

Sample A subset of cases selected from a population.

Sampling The process of identifying and selecting the subset of the population for study.

Descriptive statistics includes procedures that help us organize and describe data collected from either a sample or a population. Occasionally data are collected on an entire population, as in a census. **Inferential statistics**, on the other hand, make predictions or inferences about a population based on observations and analyses of a sample. For instance, the General Social Survey (GSS), from which numerous examples presented in this book are drawn, is conducted every other year by the National Opinion Research Center (NORC) on a representative sample of several thousands of respondents (e.g., a total of 3,842 cases were included in the GSS 2014). The survey, which includes several hundred questions (the data collection interview takes approximately 90 minutes), is designed to provide social science researchers with a readily accessible database of socially relevant attitudes, behaviors, and attributes of a cross section of the U.S. adult population. NORC has verified that the composition of the GSS samples closely resembles census data. But because the data are based on a sample rather than on the entire population, the average of the sample does not equal the average of the population as a whole.

Descriptive statistics Procedures that help us organize and describe data collected from either a sample or a population.

Inferential statistics The logic and procedures concerned with making predictions or inferences about a population from observations and analyses of a sample.

Evaluating the Hypotheses

At the completion of these descriptive and inferential procedures, we can move to the next stage of the research process: the assessment and evaluation of our hypotheses and theories in light of the analyzed data. At this next stage, new questions might be raised about unexpected trends in the data and about other variables that may have to be considered in addition to our original variables. For example, we may have found that the relationship between gender composition of occupations and earnings can be observed with respect to some groups of occupations but not others. Similarly, the relationship between these variables may apply for some racial/ethnic groups but not for others.

These findings provide evidence to help us decide how our data relate to the theoretical framework that guided our research. We may decide to revise our theory and hypothesis to take account of these later findings. Recent studies are modifying what we know about gender segregation in the workplace. These studies suggest that race as well as gender shapes the occupational structure in the United States and helps explain disparities in income. This reformulation of the theory calls for a modified hypothesis and new research, which starts the circular process of research all over again.

Statistics provide an important link between theory and research. As our example on gender segregation demonstrates, the application of statistical techniques is an indispensable part of the research process. The results of statistical analyses help us evaluate our hypotheses and theories, discover unanticipated patterns and trends, and provide the impetus for shaping and reformulating our theories. Nevertheless, the importance of statistics should not diminish the significance of the preceding phases of the research process. Nor does the use of statistics lessen the importance of our own judgment in the entire process. Statistical analysis is a relatively small part of the research process, and even the most rigorous statistical procedures cannot speak for themselves. If our research questions are poorly conceived or our data are flawed due to errors in our design and measurement procedures, our results will be useless.

EXAMINING A DIVERSE SOCIETY

The increasing diversity of American society is relevant to social science. By the middle of this century, if current trends continue unchanged, the United States will no longer be comprised predominantly of European immigrants and their descendants. Due mostly to renewed immigration and higher birthrates, in time, nearly half the U.S. population will be of African, Asian, Latino, or Native American ancestry.

Less partial and distorted explanations of social relations tend to result when researchers, research participants, and the research process itself reflect that diversity. A consciousness of social differences shapes the research questions we ask, how we observe and interpret our findings, and the conclusions we draw. Though diversity has been traditionally defined by race, class, and gender, other social characteristics such as sexual identity, physical ability, religion, and age have been identified as important dimensions of diversity. Statistical procedures and quantitative methodologies can be used to describe our diverse society, and we will begin to look at some applications in the next chapter. For now, we will preview some of these statistical procedures.

In Chapter 2, we will learn how to organize information using descriptive statistics and graphic techniques. These statistical tools can also be employed to learn about the characteristics and experiences of groups in our society that have not been as visible as other groups. For example, in a series of special reports published by the U.S. Census Bureau over the past few years, these descriptive statistical techniques have been used to describe the characteristics and experiences of ethnic minorities and those who are foreign born. Using data published by the U.S. Census Bureau, we discuss various graphic devices that can be used to explore the differences and similarities among the many social groups coexisting within the American society. These devices are also used to emphasize the changing age composition of the U.S. population.

Whereas the similarities and commonalities in social experiences can be depicted using measures of central tendency (Chapter 3), the differences and diversity within social groups can be described using statistical measures of variation (Chapter 4). In Chapters 3 and 4, we examine a variety of social demographic variables including the ethnic composition of the 50 U.S. states.

We will learn about inferential statistics and bivariate analyses in Chapters 5 through 12. First, we review the bases of inferential statistics—the normal distribution, sampling and probability, and estimation—in Chapters 5 to 7. In Chapters 8 to 12, we examine the ways in which class, sex, or ethnicity influence various social behaviors and attitudes. Inferential statistics, such as the *t* test,

A CLOSER LOOK 1.2

A TALE OF SIMPLE ARITHMETIC: HOW CULTURE MAY INFLUENCE HOW WE COUNT

A second-grade schoolteacher posed this problem to the class: "There are four blackbirds sitting in a tree. You take a slingshot and shoot one of them. How many are left?"

"Three," answered the seven-year-old European with certainty. "One subtracted from four leaves three."

"Zero," answered the seven-year-old African with equal certainty. "If you shoot one bird, the others will fly away."[9]

chi-square, and the *F* statistic, help us determine the error involved in using our samples to answer questions about the population from which they are drawn. In addition, we review several methods of bivariate analysis, which are especially suited for examining the association between different social behaviors and attitudes and variables such as race, class, ethnicity, gender, and religion. We use these methods of analysis to show not only how each of these variables operates independently in shaping behavior but also how they interlock to shape our experience as individuals in society.[10]

Whichever model of social research you use—whether you follow a traditional one or integrate your analysis with qualitative data, whether you focus on social differences or any other aspect of social behavior—remember that any application of statistical procedures requires a basic understanding of the statistical concepts and techniques. This introductory text is intended to familiarize you with the range of descriptive and inferential statistics widely applied in the social sciences. Our emphasis on statistical techniques should not diminish the importance of human judgment and your awareness of the person-made quality of statistics. Only with this awareness can statistics become a useful tool for understanding diversity and social life.

LEARNING STATISTICS[11]

After years of teaching statistics, we have learned that what underlies many of the difficulties students have in learning statistics is the belief that it involves mainly memorization of meaningless formulas. There is no denying that statistics involves many strange symbols and unfamiliar terms. It is also true that you need to know some math to do statistics. But although the subject involves some mathematical computations, we will not ask you to know more than four basic operations: (1) addition, (2) subtraction, (3) multiplication, and (4) division.

The language of statistics may appear difficult because these operations (and how they are combined) are written in a code that is unfamiliar to you. These abstract notations are simply part of the language of statistics; much like learning any foreign language, you need to learn the alphabet before you can speak the language. Once you understand the vocabulary and are able to translate the symbols and codes into terms that are familiar to you, you will begin to see how statistical techniques simply provide another source of information with which you can analyze the diverse world around you.

Another strategy for increasing your statistical knowledge is to frame your new learning in a context that is relevant and interesting. Therefore, you will find that we rely on examples from recent sociological literature, pressing social issues, and current events to make real connections to your coursework and your life. A hallmark of our text is the use of real-world examples and data; there are some, but not many, cases of fictional data in this book. We emphasize intuition, logic, and common sense over rote memorization and the derivation of formulas. In each chapter,

DATA AT WORK

At the end of each chapter, the Data at Work feature will introduce you to women and men who use quantitative data and research methods in their professional lives. They represent a wide range of career fields—education, clinical psychology, international studies, public policy, publishing, politics, and research. Some may have been led to their current positions because of the explicit integration of quantitative data and research, while others are accidental data analysts—quantitative data became part of their work portfolio. Though "data" or "statistics" are not included in their job titles, these individuals are collecting, disseminating, and/or analyzing data.

We encourage you to review each profile and imagine how you could use quantitative data and methods at work.

you'll see "Learning Check" boxes where you can apply or test your new knowledge. The chapters also include "A Closer Look" boxes where we provide more detailed or background information about a particular statistical technique or interpretation. Beginning with Chapter 2, we include "Statistics in Practice" and "Reading the Research Literature" features, highlighting the interpretation of data, specific statistical calculations, or published research. We believe being statistically literate involves more than just completing a calculation; it also means learning how to apply and interpret statistical information and being able to say what it means.

What might also help develop confidence in your statistical ability is working with other students. We encourage you to collaborate with your peers as you learn this course material. We have learned that students who are intimidated by statistics do not like to admit it or talk about it. This avoidance mechanism may be an obstacle to overcoming statistics anxiety. Talking about your feelings with other students will help you realize that you are not the only one intimidated by the course. This sharing process is at the heart of the treatment of statistics anxiety—talking to others in a safe group setting will help you take risks and trust your own intuition and judgment. Ultimately, your judgment and intuition lie at the heart of your ability to translate statistical symbols and concepts into a language that makes sense and to interpret data using your newly acquired statistical tools.

MAIN POINTS

- Statistics are procedures used by social scientists to organize, summarize, and communicate information. Only information represented by numbers can be the subject of statistical analysis.
- The research process is a set of activities in which social scientists engage to answer questions, examine ideas, or test theories. It consists of the following stages: asking the research question, formulating the hypotheses, collecting data, analyzing data, and evaluating the hypotheses.
- A theory is a set of assumptions and propositions used for explanation, prediction, and understanding of social phenomena. Theories offer specific concrete predictions about the way observable attributes of people or groups would be interrelated in real life. These predictions, called hypotheses, are tentative answers to research problems.
- A variable is a property of people or objects that takes on two or more values. The variable that the researcher wants to explain (the "effect") is called the dependent variable. The variable that is expected to "cause" or account for the dependent variable is called the independent variable.

- Three conditions are required to establish causal relations: (1) The cause has to precede the effect in time; (2) there has to be an empirical relationship between the cause and the effect; and (3) this relationship cannot be explained by other factors.
- At the nominal level of measurement, numbers or other symbols are assigned to a set of categories to name, label, or classify the observations. At the ordinal level of measurement, categories can be rank-ordered from low to high (or vice versa). At the interval-ratio level of measurement, measurements for all cases are expressed in the same unit.
- A population is the total set of individuals, objects, groups, or events in which the researcher is interested. A sample is a relatively small subset selected from a population. Sampling is the process of identifying and selecting the subset.
- Descriptive statistics includes procedures that help us organize and describe data collected from either a sample or a population. Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.

KEY TERMS

DIGITAL RESOURCES

SAGE edge™

Get the edge on your studies. **edge.sagepub.com/frankfort8e**

Take a quiz to find out what you've learned.
Review key terms with eFlashcards.
Dive into real research with SAGE Journal Articles.

SPSS DEMONSTRATION [GSS14SSDS-A]

Introduction to Data Sets and Variables

We'll be using a set of computer data and exercises at the end of each chapter. All computer exercises are based on the program IBM SPSS Version 24 or IBM SPSS Statistics Base Student Edition 23.

Throughout this textbook, you'll be working with two data sets. (1) The GSS14SSDS-A and (2) GSS14SSDS-B each contain a selection of 49 variables[12] and 1,500 cases from the 2014 General Social Survey (GSS). The GSS has been conducted biennially since 1972. Conducted by the NORC at the University of Chicago, with principal funding from the National Science Foundation, the GSS is designed to provide social science researchers with a readily accessible database of socially relevant attitudes, behaviors, and attributes of a cross section of the U.S. population. Next to the U.S. Census data, the GSS is the most frequently analyzed source of social science information by educators, legislators, and media outlets.

The SPSS appendix found on this text's study site explains the basic operation and procedures for SPSS for Windows Student Version. We strongly recommend that you refer to this appendix before beginning the SPSS exercises.

When you begin using a data set, you should take the time to review your variables. What are the variables called? What do they measure? What do they mean? There are several ways to do this.

To review your data, you must first open the data file. Files are opened in SPSS by clicking on *File*, then *Open*, and then *Data*. After switching directories and drives to the appropriate location of the files (which may be on a hard disk or on a ZIP drive), you select one data file and click on *Open*. This routine is the same each time you open a data file. SPSS automatically opens each data file in the SPSS Data Editor window labeled Data View. We'll use GSS14SSDS-A.SAV for this demonstration.

One way to review the complete list of variables in a file is to click on the *Utilities* choice from the main menu, then on *Variables* in the list of submenu choices. The SPSS variable names, which are limited to eight characters or less, are listed in the scroll box (refer to Figure 1.4). When a variable name is highlighted, the descriptive label for that variable is listed, along with any missing values and, if available, the value labels for each variable category. (As you use this feature, please note that sometimes SPSS mislabels the variable's measurement level. Always confirm that the reported SPSS measurement level is correct.) SPSS allows you to display data in alphabetical order (based on the variable name) or in the order presented in the file (which may not be alphabetical).

A second way to review all variables is through the Variable View window. Notice on the bottom of your screen that there are two tabs, one for *Data View* and the other for *Variable View*. Click on *Variable View*, and you'll see all the variables listed in the order in which they appear in the Data View window (as depicted in Figure 1.5). Each column provides specific information about the variables. The columns labeled "Label" and "Values" provide the variable label (a brief label of what it's measuring) and value labels (for each variable category).

Figure 1.4 Utilities-Variables Dialog Box

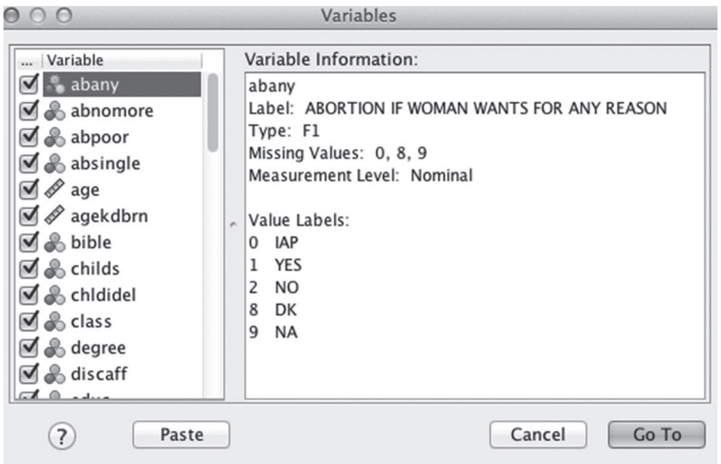


Figure 1.5 Variable View Window for GSS14SDSS-A

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	abany	Numeric	1	0	ABORTION IF W...	{0, IAP}...	0, 8, 9	8	Right	Nominal	Input
2	abnomore	Numeric	1	0	MARRIED--WA...	{0, IAP}...	0, 8, 9	8	Right	Nominal	Input
3	abpoor	Numeric	1	0	LOW INCOME-...	{0, IAP}...	0, 8, 9	8	Right	Nominal	Input
4	absingle	Numeric	1	0	NOT MARRIED	{0, IAP}...	0, 8, 9	8	Right	Nominal	Input
5	age	Numeric	2	0	AGE OF RESPO...	{89, 89 OR ...	0, 98, 99	8	Right	Scale	Input

SPSS PROBLEM [GSS14SSDS-A]

Based on the *Utilities-Variables* option, review the variables from the GSS14SSDS-A. Can you identify three nominal variables, three ordinal variables, and at least one interval-ratio variable? Based on the information in the dialog box or Variable View window, you should be able to identify the variable name, variable label, and category values.

CHAPTER EXERCISES

1. In your own words, explain the relationship of data (collecting and analyzing) to the research process. (Refer to Figure 1.1.)
2. Construct potential hypotheses or research questions to relate the variables in each of the following examples. Also, write a brief statement explaining why you believe there is a relationship between the variables as specified in your hypotheses.
    a. Political party and support of the Affordable Care Act
    b. Income and race/ethnicity
    c. The crime rate and the number of police in a city
    d. Life satisfaction and marital status
    e. Age and support for marijuana legalization
    f. Care of elderly parents and ethnicity
3. Determine the level of measurement for each of the following variables:
    a. The number of people in your statistics class
    b. The percentage of students who are first-generation college students at your school
    c. The name of each academic major offered in your college
    d. The rating of the overall quality of a textbook, on a scale from “Excellent” to “Poor”
    e. The type of transportation a person takes to school (e.g., bus, walk, car)
    f. The number of hours you study for a statistics exam
    g. The rating of the overall quality of your campus coffee shop, on a scale from “Excellent” to “Poor”

4. For each of the variables in Exercise 3 that you classified as interval-ratio, identify whether it is discrete or continuous.
5. Why do you think men and women, on average, do not earn the same amount of money? Develop your own theory to explain the difference. Use three independent variables in your theory, with annual income as your dependent variable. Construct hypotheses to link each independent variable with your dependent variable.
6. For each of the following examples, indicate whether it involves the use of descriptive or inferential statistics. Justify your answer.
   a. The number of unemployed people in the United States
   b. Determining students' opinion about the quality of food at the cafeteria based on a sample of 100 students
   c. The national incidence of breast cancer among Asian women
   d. Conducting a study to determine the rating of the quality of a new smartphone, gathered from 1,000 new buyers
   e. The average GPA of various majors (e.g., sociology, psychology, English) at your university
   f. The change in the number of immigrants coming to the United States from Southeast Asian countries between 2010 and 2015
7. Adela García-Aracil (2007)[13] identified how several factors affected the earnings of young European higher education graduates. Based on data from several EU (European Union) countries, her statistical models included the following variables: annual income (actual dollars), gender (male or female), the number of hours worked per week (actual hours), and years of education (actual years) for each graduate. She also identified each graduate by current job title (senior officials and managers, professionals, technicians, clerks, or service workers).
   a. What is García-Aracil's dependent variable?
   b. Identify two independent variables in her research. Identify the level of measurement for each.
   c. Based on her research, García-Aracil can predict the annual income for other young graduates with similar work experiences and characteristics like the graduates in her sample. Is this an application of descriptive or inferential statistics? Explain.
8. Construct measures of political participation at the nominal, ordinal, and interval-ratio levels. (*Hint:* You can use behaviors such as voting frequency or political party membership.) Discuss the advantages and disadvantages of each.
9. Variables can be measured according to more than one level of measurement. For the following variables, identify at least two levels of measurement. Is one level of measurement better than another? Explain.
   a. Individual age
   b. Annual income
   c. Religiosity
   d. Student performance
   e. Social class
   f. Number of children

2

THE ORGANIZATION AND GRAPHIC PRESENTATION OF DATA

Chapter Learning Objectives

1. Construct and analyze frequency, percentage, and cumulative distributions
2. Calculate proportions and percentages
3. Compare and contrast frequency and percentage distributions for nominal, ordinal, and interval-ratio variables
4. Construct and interpret a pie chart, bar graph, histogram, the statistical map, line graph, and time-series chart

Demographers examine the size, composition, and distribution of human populations. Changes in the birth, death, and migration rates of a population affect its composition and social characteristics.[1] In order to examine a large population, researchers often have to deal with very large amounts of data. For example, imagine the amount of data it takes to describe the immigrant or elderly population in the United States. To make sense out of these data, a researcher has to organize and summarize the data in some systematic fashion. In this chapter, we review two such methods used by social scientists: (1) the creation of frequency distributions and (2) the use of graphic presentation.

FREQUENCY DISTRIBUTIONS

The most basic way to organize data is to classify the observations into a frequency distribution. A **frequency distribution** is a table that reports the number of observations that fall into each category of the variable we are analyzing. Constructing a frequency distribution is usually the first step in the statistical analysis of data.

Frequency distribution A table reporting the number of observations that fall into each category of the variable.

Immigration has been described as "remaking America with political, economic, and cultural ramifications."[2] Globalization has fueled migration, particularly since the beginning of the 21st century. Workers migrate because of the promise of employment and higher standards of living than what is attainable in their home countries. Data reveal that many migrants seek specifically to move to the United States.[3] The U.S. Census Bureau uses the term foreign born to refer to those who are not U.S. citizens at birth. The U.S. Census estimates that nearly 13% of the U.S. population or approximately 41 million people are foreign born.[4] Immigrants are not one homogeneous group but are many diverse groups. Table 2.1 shows the frequency distribution of the world region of birth for the foreign-born population.

The frequency distribution is organized in a table, which has a number (2.1) and a descriptive title. The title indicates the kind of data

Table 2.1 Frequency Distribution for Categories of Region of Birth for Foreign-Born Population, 2012

Region of Birth	Frequency (f)
Mexico	11,489,387
South and East Asia	10,443,902
Caribbean	3,882,592
Central America	3,172,307
South America	2,731,619
Middle East	1,578,801
All other	7,439,616
Total	40,738,224

Source: Anna Brown and Eileen Patton, *Statistical Portrait of the Foreign-Born Population of the United States, 2012*, 2014.

presented: "Categories of Region of Birth for Foreign-Born Population." The table consists of two columns. The first column identifies the variable (world region of birth) and its categories. The second column, with the heading "Frequency (f)," tells the number of cases in each category as well as the total number of cases (N = 40,738,224). Note also that the source of the table is clearly identified. It tells us that the data are from a 2014 report by Anna Brown and Eileen Patton (though the information is based on 2012 U.S. Census data). The source of the data can be reported as a source note or in the title of the table.

What can you learn from the information presented in Table 2.1? The table shows that as of 2012, approximately 41 million people were classified as foreign born. Out of this group, the majority, about 11.5 million people, were from Mexico, 10.4 million were from Asia, followed by 7.4 million from the category of all other countries.

PROPORTIONS AND PERCENTAGES

Frequency distributions are helpful in presenting information in a compact form. However, when the number of cases is large, the frequencies may be difficult to grasp. To standardize these raw frequencies, we can translate them into relative frequencies—that is, proportions or percentages.

Proportion A relative frequency obtained by dividing the frequency in each category by the total number of cases.

A **proportion** is a relative frequency obtained by dividing the frequency in each category by the total number of cases. To find a proportion (p), divide the frequency (f) in each category by the total number of cases (N):

$$p = \frac{f}{N} \qquad (2.1)$$

where

f = frequency

N = total number of cases

We've calculated the proportion for the three largest groups of foreign born. First, the proportion of foreign born originally from Mexico is

$$\frac{11,489,387}{40,738,224} = .282$$

The proportion of foreign born who were originally from South and East Asia is

$$\frac{10,443,902}{40,738,224} = .256$$

The proportion of foreign born who were originally from all other countries is

$$\frac{7,439,616}{40,738,224} = .183$$

The proportion of foreign born who were originally from other reported areas (combining the Caribbean, Central and South America, and Middle East) is

$$\frac{11,365,319}{40,738,224} = .279$$

Proportions should always sum to 1.00 (allowing for some rounding errors). Thus, in our example, the sum of the six proportions is

$$0.28 + 0.26 + 0.18 + 0.28 = 1.0$$

To determine a frequency from a proportion, we simply multiply the proportion by the total N:

$$f = p(N) \tag{2.2}$$

Thus, the frequency of foreign born from South and East Asia can be calculated as

$$0.26(40,738,224) = 10,591,938$$

The obtained frequency differs somewhat from the actual frequency of 10,443,902. This difference is due to rounding off of the proportion. If we use the actual proportion instead of the rounded proportion, we obtain the correct frequency:

$$0.256366158\,(40,738,224) = 10,443,902$$

We can also express frequencies as percentages. A **percentage** is a relative frequency obtained by dividing the frequency in each category by the total number of cases and multiplying by 100. In most statistical reports, frequencies are presented as percentages rather than proportions. Percentages express the size of the frequencies as if there were a total of 100 cases.

Percentage A relative frequency obtained by dividing the frequency in each category by the total number of cases and multiplying by 100.

To calculate a percentage, multiply the proportion by 100:

$$\text{Percentage (\%)} = \frac{f}{N}(100) \tag{2.3}$$

or

$$\text{Percentage (\%)} = p(100) \tag{2.4}$$

Thus, the percentage of respondents who were originally from Mexico is

$$0.28(100) = 28\%$$

LEARNING CHECK 2.1

Calculate the proportion and percentage of males and females in your statistics class. What proportion is female?

PERCENTAGE DISTRIBUTIONS

Percentage distribution A table showing the percentage of observations falling into each category of the variable.

Percentages are usually displayed as percentage distributions. A **percentage distribution** is a table showing the percentage of observations falling into each category of the variable. For example, Table 2.2 presents the frequency distribution of categories of places of origin (Table 2.1) along with the corresponding percentage distribution. Percentage distributions (or proportions) should always show the base (N) on which they were computed. Thus, in Table 2.2, the base on which the percentages were computed is $N = 40{,}738{,}224$.

THE CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

In this section, you will learn how to construct frequency distributions. Most often, we can use statistical software to accomplish this, but it is important to go through the process to understand how frequency distributions are actually constructed.

Table 2.2 Frequency and Percentage Distributions for Categories of Region of Birth for Foreign Born, 2012

Region of Birth	Frequency (f)	Percentage (%)
Mexico	11,489,387	28
South and East Asia	10,443,902	26
Caribbean	3,882,592	9
Central America	3,172,307	8
South America	2,731,619	7
Middle East	1,578,801	4
All other	7,439,616	18
Total	40,738,224	100

Source: Anna Brown and Eileen Patton, *Statistical Portrait of the Foreign-Born Population of the United States, 2012*, 2014. Retrieved from http://www.pewhispanic.org/2014/04/29/statistical-portrait-of-the-foreign-born-population-in-the-united-states-2012/

For nominal and ordinal variables, constructing a frequency distribution is quite simple. To do so, count and report the number of cases that fall into each category of the variable along with the total number of cases (*N*). For the purpose of illustration, let's take a small random sample of 40 cases from a General Social Survey (GSS) sample and record their scores on the following variables: gender, a nominal-level variable; degree, an ordinal measurement of education; and age and number of children, both interval-ratio variables. The use of "male" and "female" in parts of this book is in keeping with the GSS categories for the variable "sex" (respondent's sex).

The interviewer recorded the gender of each respondent at the beginning of the interview. To measure degree, researchers asked each individual to indicate the highest degree completed: less than high school, high school, some college, bachelor's degree, and graduate degree. The first category represented the lowest level of education. Researchers calculated respondents' age based on the respondent's birth year. The number of children was determined by the question, "How many children have you ever had?" The answers given by our subsample of 40 respondents are displayed in Table 2.3. Note that each row in the table represents a respondent, whereas each column represents a variable. This format is conventional in the social sciences.

You can see that it is going to be difficult to make sense of these data just by eyeballing Table 2.3. How many of these 40 respondents are males? How many said that they had a graduate degree? How many were older than 50 years of age? To answer these questions, we construct a frequency distribution for each variable.

Frequency Distributions for Nominal Variables

Let's begin with the nominal variable, gender. First, we tally the number of males, then the number of females (the column of tallies has been included in Table 2.4 for the purpose of illustration). The tally results are then used to construct the frequency distribution presented in Table 2.4. The table has a title describing its content ("Frequency Distribution of the Variable Gender: GSS Subsample"). Its categories (male and female) and their associated frequencies are clearly listed; in addition, the total number of cases (*N*) is also reported. The Percentage column is the percentage distribution for this variable. To convert the Frequency column to percentages, simply divide each frequency by the total number of cases and multiply by 100. Percentage distributions are routinely added to almost any frequency table and are especially important if comparisons with other groups are to be considered. Immediately, we can see that it is easier to read the information. There are 25 females and 15 males in this sample. Based on this frequency distribution, we can also conclude that the majority of sample respondents are female.

LEARNING CHECK 2.2

Construct a frequency and percentage distribution for males and females in your statistics class.

Frequency Distributions for Ordinal Variables

To construct a frequency distribution for ordinal-level variables, follow the same procedures outlined for nominal-level variables. Table 2.5 presents the frequency distribution for the variable degree. The table shows that 60.0%, a majority, indicated that their highest degree was a high school degree.

Table 2.3 A GSS Subsample of 40 Respondents

Gender of Respondent	Degree	Number of Children	Age
M	Bachelor	1	43
F	High school	2	71
F	High school	0	71
M	High school	0	37
M	High school	0	28
F	High school	6	34
F	High school	4	69
F	Graduate	0	51
F	Bachelor	0	76
M	Graduate	2	48
M	Graduate	0	49
M	Less than high school	3	62
F	Less than high school	8	71
F	High school	1	32
F	High school	1	59
F	High school	1	71
M	High school	0	34
M	Bachelor	0	39
F	Bachelor	2	50
M	High school	3	82
F	High school	1	45
M	High school	0	22
M	High school	2	40
F	High school	2	46
M	High school	0	29
F	High school	1	75
F	High school	0	23
M	Bachelor	2	35
M	Bachelor	3	44
F	High school	3	47
M	High school	1	84
F	Graduate	1	45
F	Less than high school	3	24
F	Graduate	0	47
F	Less than high school	5	67
F	High school	1	21
F	High school	0	24
F	High school	3	49
F	High school	3	45
F	Graduate	3	37

Note: M, male; F, female.

Table 2.4 Frequency Distribution of the Variable Gender, GSS Subsample

Gender	Tallies	Frequency (f)	Percentage (%)
Male	𝍸 𝍸 𝍸	15	37.5
Female	𝍸 𝍸 𝍸 𝍸 𝍸	25	62.5
Total (*N*)		40	100.0

Table 2.5 Frequency Distribution of the Variable Degree, GSS Subsample

Degree	Tallies	Frequency (f)	Percentage (%)
Less than high school	||||	4	10.0
High school	𝍸 𝍸 𝍸 𝍸 ||||	24	60.0
Bachelor	𝍸 |	6	15.0
Graduate	𝍸 |	6	15.0
Total (*N*)		40	100.0

The major difference between frequency distributions for nominal and ordinal variables is the order in which the categories are listed. The categories for nominal-level variables do not have to be listed in any particular order. For example, we could list females first and males second without changing the nature of the distribution. Because the categories or values of ordinal variables are rank-ordered, however, they must be listed in a way that reflects their rank—from the lowest to the highest or from the highest to the lowest. Thus, the data on degree in Table 2.5 are presented in declining order from "less than high school" (the lowest educational category) to "graduate" (the highest educational category).

Frequency Distributions for Interval-Ratio Variables

We hope that you agree by now that constructing frequency distributions for nominal- and ordinal-level variables is rather straightforward. Simply list the categories and count the number of observations that fall into each category. Building a frequency distribution for interval-ratio variables with relatively few values is also easy. For example, when constructing a frequency distribution for number of children, simply list the number of children and report the corresponding frequency, as shown in Table 2.6.

Very often interval-ratio variables have a wide range of values, which makes simple frequency distributions very difficult to read. For example, take a look at the frequency distribution for the variable *age* in Table 2.7. The distribution contains age values ranging from 21 to 84 years. For a more concise picture, the large number of different scores could be reduced into a smaller number of groups, each containing a range of scores. Table 2.8 displays such a grouped

Table 2.6 Frequency Distribution of Variable Number of Children, GSS Subsample

Number of Children	Frequency (f)	Percentage (%)
0	13	32.5
1	9	22.5
2	6	15.0
3	8	20.0
4	1	2.5
5	1	2.5
6	1	2.5
7+	1	2.5
Total (*N*)	40	100.0

Table 2.7 Frequency Distribution of the Variable Age, GSS Subsample

Age of Respondent	Frequency (f)	Age of Respondent	Frequency (f)
21	1	59	1
22	1	62	1
23	1	67	1
24	2	69	1
28	1	71	4
29	1	75	1
32	1	76	1
34	2	82	1
35	1	84	1
37	2		
39	1		
40	1		
43	1		
44	1		
45	3		
46	1		
47	2		
48	1		
49	2		
50	1		
51	1		

frequency distribution of the data in Table 2.7. Each group, known as a class interval, now contains 10 possible scores instead of 1. Thus, the ages of 21, 22, 23, 24, 28, and 29 all fall into a single class interval of 20–29. The second column of Table 2.8, Frequency, tells us the number of respondents who fall into each of the intervals—for example, that seven respondents fall into the class interval of 20–29. Having grouped the scores, we can clearly see that the biggest single age group is between 40 and 49 years (12 out of 40, or 30% of sample). The percentage distribution that we have added to Table 2.8 displays the relative frequency of each interval and emphasizes this pattern as well.

LEARNING CHECK 2.3

Can you verify that Table 2.8 was constructed correctly? Use Table 2.7 to determine the frequency of cases that fall into the categories of Table 2.8.

The decision as to how many groups to use and, therefore, how wide the intervals should be is usually up to the researcher and depends on what makes sense in terms of the purpose of the research. The rule of thumb is that an interval width should be large enough to avoid too many categories but not so large that significant differences between observations are concealed. Obviously, the number of intervals depends on the width of each. For instance, if you are working with scores ranging from 10 to 60 and you establish an interval width of 10, you will have five intervals.

Table 2.8 Grouped Frequency Distribution of the Variable Age, GSS Subsample

Age Category	Frequency (f)	Percentage (%)
20–29	7	17.5
30–39	7	17.5
40–49	12	30.0
50–59	3	7.5
60–69	3	7.5
70–79	6	15.0
80–89	2	5.0
Total (*N*)	40	100.0

LEARNING CHECK 2.4

If you are having trouble distinguishing between nominal, ordinal, and interval-ratio variables, review the section on levels of measurement in Chapter 1. The distinction between these levels of measurement will be important throughout the book.

CUMULATIVE DISTRIBUTIONS

Cumulative frequency distribution A distribution showing the frequency at or below each category (class interval or score) of the variable.

Sometimes, we may be interested in locating the relative position of a given score in a distribution. For example, we may be interested in finding out how many or what percentage of our sample was younger than 40 or older than 60. Frequency distributions can be presented in a cumulative fashion to answer such questions. A **cumulative frequency distribution** shows the frequencies at or below each category of the variable.

Cumulative frequencies are appropriate only for variables that are measured at an ordinal level or higher. They are obtained by adding to the frequency in each category the frequencies of all the categories below it.

Let's look at Table 2.9. It shows the cumulative frequencies based on the frequency distribution from Table 2.8. The cumulative frequency column, denoted by *Cf*, shows the number of persons at or below each interval. For example, you can see that 14 of the 40 respondents were 39 years old or younger, and 29 respondents were 59 years old or younger.

To construct a cumulative frequency distribution, start with the frequency in the lowest class interval (or with the lowest score, if the data are ungrouped), and add to it the

Table 2.9 Grouped Frequency Distribution and Cumulative Frequency for the Variable Age, GSS Subsample

Age Category	Frequency (f)	Cumulative Frequency (Cf)
20–29	7	7
30–39	7	14
40–49	12	26
50–59	3	29
60–69	3	32
70–79	6	38
80–89	2	40
Total (*N*)	40	

A CLOSER LOOK 2.1

REAL LIMITS, STATED LIMITS, AND MIDPOINTS OF CLASS INTERVALS

The intervals presented in Table 2.8 constitute the categories of the variable age that we used to classify the survey's respondents. In Chapter 1 ("The What and the Why of Statistics"), we noted that our variables need to be both exhaustive and mutually exclusive. These principles apply to the intervals here as well. This means that each of the 40 respondents can be classified into one and only one category. In addition, we should be able to classify all the possible scores.

In our example, these requirements are met: Each observation score fits into only one interval, and there is an appropriate category to classify each individual score as recorded in Table 2.8. However, if you looked closely at Table 2.8, you may have noticed that there is actually a gap of 1 year between adjacent intervals. A gap could create a problem with scores that have fractional values. Though age is conventionally rounded down, let's suppose for a moment that respondent's age had been reported with more precision. Where would you classify a woman who was 49.25 years old? Notice that her age would actually fall between the intervals 40–49 and 50–59! To avoid this potential problem, use the real limits shown in the following table rather than the stated limits listed in Table 2.8.

Real limits extend the upper and lower limits of the intervals by .5. For instance, the real limits for the interval 40–49 are 39.5–49.5; the real limits for the interval 50–59 are 49.5–59.5; and so on. (Scores that fall exactly at the upper real limit or the lower real limit of the interval [e.g., 59.5 or 49.5] are usually rounded to the closest even number. The number 59.5 would be rounded to 60 and would thus be included in the interval 59.5–69.5.) In the following table, we include both the stated limits and real limits for the grouped frequency distribution of respondent's age. So where would you classify a respondent who was 49.25 years old? (*Answer:* In the interval 39.5–49.5.) How about 19.9? (In the interval 19.5–29.5.)

The midpoint is a single number that represents the entire interval. A midpoint is calculated by adding the lower and upper real limits of the interval and dividing by 2. The midpoint of the interval 19.5–29.5, for instance, is $(19.5 + 29.5) \div 2 = 24.5$. The midpoint for all the intervals of the table are displayed in the third column.

Even though grouped frequency distributions are very helpful in summarizing information, remember that they are only a summary and therefore involve a considerable loss of detail. Since most researchers and students have access to computers, grouped frequencies are used only when the raw data are not available. Most of the statistical procedures described in later chapters are based on the raw scores.

Respondent's Age			
Stated Limits	Real Limits	Midpoint	Frequency (f)
20–29	19.5–29.5	24.5	7
30–39	29.5–39.5	34.5	7
40–49	39.5–49.5	44.5	12
50–59	49.5–59.5	54.5	3
60–69	59.5–69.5	64.5	3
70–79	69.5–79.5	74.5	6
80–89	79.5–89.5	84.5	2
Total (*N*)			40

frequencies in the next highest class interval. Continue adding the frequencies until you reach the last class interval. The cumulative frequency in the last class interval will be equal to the total number of cases (*N*). In Table 2.9, the frequency associated with the first class interval (20–29) is 7. The cumulative frequency associated with this interval is also 7, since there are no cases below this class interval. The frequency for the second class interval is 7. The cumulative frequency for this interval is 7 + 7 = 14. To obtain the cumulative frequency of 26 for the third interval, we add its frequency (12) to the cumulative frequency associated with the second class interval (14). Continue this process until you reach the last class interval. Therefore, the cumulative frequency for the last interval is equal to 40, the total number of cases (*N*).

Cumulative percentage distribution A distribution showing the percentage at or below each category (class interval or score) of the variable.

We can also construct a cumulative percentage distribution (*C*%), which has wider applications than the cumulative frequency distribution (*Cf*). A **cumulative percentage distribution** shows the percentage at or below each category (class interval or score) of the variable. A cumulative percentage distribution is constructed using the same procedure as for a cumulative frequency distribution except that the percentages—rather than the raw frequencies—for each category are added to the total percentages for all the previous categories.

In Table 2.10, we have added the cumulative percentage distribution to the frequency and percentage distributions shown in Table 2.8. The cumulative percentage distribution shows, for example, that 35% of the sample was 39 years or younger. Like the percentage distributions described earlier, cumulative percentage distributions are especially useful when you want to compare differences between groups. For an example of how cumulative percentages are used in a comparison, we used GSS data to contrast the opinions of whites and blacks about whether they believe immigrants take jobs away from native-born Americans. Respondents were asked, "How much do you agree or disagree with the following statement? Immigrants take jobs away." The percentage distribution and the cumulative percentage distribution for whites and blacks are shown in Table 2.11. (This table is referred to as a bivariate table, reporting the overlap between two variables—[1] respondent race and [2] level of agreement to the immigration statement. We'll discuss bivariate tables in depth in Chapter 9 ["Bivariate Tables"].)

The cumulative percentage distributions suggest that a higher percentage of blacks agree to the statement that immigrants take away jobs. The two groups are separated by 5 percentage

Table 2.10 Grouped Frequency Distribution and Cumulative Percentages for the Variable Age, GSS Subsample

Age Category	Frequency (f)	Percentage (%)	Cumulative Percentage (C%)
20–29	7	17.5	17.5
30–39	7	17.5	35.0
40–49	12	30.0	65.0
50–59	3	7.5	72.5
60–69	3	7.5	80.0
70–79	6	15.0	95.0
80–89	2	5.0	100.0
Total (*N*)	40	100.0	