2e

An Introduction to Statistics and Data Analysis Using Stata®

From Research Design to Final Report

Lisa Daniels • Nicholas Minot

AN INTRODUCTION TO STATISTICS AND DATA ANALYSIS USING STATA®

From Research Design to Final Report

Second Edition

To our parents, Betty, Joe, Ginny, and Steve

AN INTRODUCTION TO STATISTICS AND DATA ANALYSIS USING STATA®

From Research Design to Final Report

Second Edition

Lisa Daniels

Washington College

Nicholas Minot

International Food Policy Research Institute

Sage

FOR INFORMATION:

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

1 Oliver’s Yard

55 City Road

London, EC1Y 1SP

United Kingdom

Unit No. 323-333, Third Floor, F-Block

International Trade Tower

Nehru Place, New Delhi – 110 019

India

18 Cross Street #10-10/11/12

China Square Central

Singapore 048423

Library of Congress Control Number: 2024056491

ISBN: 978-1-0718-8370-9 (Paperback)

Printed in the United States of America

This book is printed on acid-free paper.

25 26 27 28 29 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Leah Fargotstein

Content Development Editor: Jessica Meyer

Production Editor: Neelu Sahu

Copy Editor: Jared Leighton

Typesetter: diacriTech

Cover Designer: Scott Van Atta

Marketing Manager: Victoria Velasquez

BRIEF CONTENTS

DETAILED CONTENTS

PREFACE

Does the use of ChatGPT to practice homework problems improve scores? Were mask mandates motivated by politics during the COVID-19 pandemic? Are there differences in education levels among men and women who use online dating applications? Do students from high-income families earn higher SAT scores? These are just some of the examples used in this book, *An Introduction to Statistics and Data Analysis Using Stata: From Research Design to Final Product*, second edition, to illustrate the endless number of interesting questions that can be examined with statistics.

Drawing on our 25 years of experience in teaching data analysis to undergraduate students and designing over 30 surveys in 17 countries, we have incorporated *four essential elements* in this book that we believe are fundamental to the practice of data analysis.

1). The book provides an introduction to research design and data collection, including questionnaire design, sample selection, sampling weights, and data cleaning. These topics are an important part of empirical research and provide students with the skills to conduct their own research and evaluate research carried out by others.

2). We frame data analysis within the research process—identifying gaps in the literature, examining the theory, developing research questions, designing a questionnaire or using secondary data, analyzing the data, and writing a research paper.

3). We emphasize the use of code or command files in Stata rather than the point-and-click menu features of the software. We believe that students should be taught to write programs

that document their analysis, as this allows them to reproduce their work during follow-up analyses and to facilitate collaborative work. We do, however, include brief instructions on the use of Stata menus for each command.

4). The book teaches students how to describe statistical results for technical and nontechnical audiences. Being able to explain the results to various audiences is just as important as choosing the correct statistical test and generating results.

Because the primary focus of this book is data analysis, we do not provide the same depth of treatment on research methods that may be available in other books. However, we feel that providing an integrated approach to research methods, data analysis, and interpretation of results is a worthwhile trade-off, particularly for undergraduate students who might not otherwise get exposure to research methods. We also offer resources for students who are interested in exploring any of the topics covered in this book in greater depth.

CHAPTER FEATURES

The literature on teaching statistics emphasizes the challenges students face in learning how to apply statistics to solve problems, the difficulty in understanding published results, and the inability to communicate research results. We address these problems throughout the book, as illustrated by these features:

1. *Description of the research process*

   The first chapter is devoted to the steps in the research process. These steps include choosing a general area, identifying the gaps in the literature, examining the theory, developing a research question, designing a questionnaire or using secondary data, analyzing the data, and writing the research paper. By starting with the big picture, students have a frame of reference to guide them as they then learn in detail about these steps in the chapters that follow.

2. *Summary table at the start of each chapter that includes the research question, hypothesis, statistical procedure, and Stata code*

   Each chapter related to a statistical technique (Chapters 7–12 and 14) begins with a table that identifies the research question, the research hypothesis, the statistical procedure needed to test the hypothesis, the types of variables used, the assumptions of the test, and the relevant commands in Stata. This table serves as a quick reference guide and preview of what is to come in the chapter. It also reinforces the ability to apply statistics to solve problems.

3. *Box with news article related to a statistical procedure*

   Following the summary table described previously, a portion of a newspaper article is included to illustrate the use of the

statistical technique applied to real-world data. A brief discussion of the news article follows along with the necessary statistical method to test the hypothesis and a critique of potential flaws in the research design. This is designed to help students understand published results, judge their quality, and again apply statistics to real-world problems.

4. *Tables with real-world examples from six fields of study*

   Section 2 of each chapter related to a statistical concept covers the circumstances in which that particular concept or test is appropriate. This is done by giving examples of research questions from six fields along with the null hypothesis and types of variables needed for the test. This is intended to help students identify research questions and apply statistics to solve problems. It also illustrates that the skills related to statistical techniques are applicable across multiple disciplines.

5. *Application of statistical tests using relevant data*

   We illustrate the practical application of statistical methods by employing eight data sets that captivate the interest of college students and remain relevant to their experience:

   We use data from the College Scorecard—an initiative by the United States government designed to aid students in comparing colleges based on factors such as postgraduation debt and salaries six years after graduation. This is complemented by integrating college ranking information from the U.S. News and World Report to examine variations in these metrics across different college ranks.

   We explore dating app dynamics using data from Ok Cupid, which sheds light on the characteristics of individuals

engaged in online dating.

In the context of the COVID-19 pandemic, we utilize state-level data to explore statistics concerning mask mandates, COVID-19 cases, and the political influences shaping decisions related to the pandemic.

We make use of the Admitted Student Questionnaire for 2014, which covers SAT scores, family incomes, and student perspectives on the significance of various college attributes.

The 2015–2016 Survey on Crime and Safety is used to examine the landscape of violence and discipline in US high schools.

We use a database of new and used cars for sale from Cars.com to explore the factors that affect the price of electric, hybrid, and gas-powered cars.

The issues of drug abuse and alcohol consumption are addressed using data from the National Survey on Drug Use and Health from 2015.

Finally, to illustrate examples and exercises throughout the book pertaining to trends in the attitudes and behaviors of Americans, we make use of the General Social Survey conducted in 2021.

6. *Exercises to practice techniques learned in each chapter*

   It is essential for students to practice data analysis on a regular basis in order to become proficient data analysts. This book contains more than 70 exercises that can be done in class or as homework problems. Instructors have access to the full answer key for each problem. In addition to the chapter exercises, we

also offer multiple choice quizzes for each chapter available as an instructor resource online.

7. *Instructions using Stata commands along with a brief description of menus*

   As described earlier, the use of Stata code or command files allows students to document their work, reproduce the results, and collaborate with others during the research process. Menus are also briefly illustrated for those professors who prefer to teach with the menus in each chapter.

8. *Communicating the results*

   In each chapter related to a statistical test, we include a section called “Presenting the Results,” in which we illustrate how to report the results for a nontechnical audience and for a scholarly journal with more technical language. In addition to these sections, the last chapter is devoted entirely to writing a research paper.

AUDIENCE

An Introduction to Statistics and Data Analysis Using Stata: From Research Design to Final Report is written for undergraduate students in any course that involves data analysis. Although it would be helpful to have some knowledge of statistics before using this book, the book can be used as an introduction to both statistics and Stata, a statistical software package widely used in multiple fields. The book could also be useful in an introductory graduate course or for researchers interested in learning Stata.

TEACHING RESOURCES

This text includes an array of instructor teaching materials designed to save you time and to help you keep students engaged. A list of these resources follows. To learn more, visit sagepub.com or contact your SAGE representative at sagepub.com/findmyrep.

- Access to the data sets used throughout the book
- Two sets of answer keys to the chapter exercises: A full set with all answers and output and an abbreviated set for students to check their work as they complete the exercises.
- A multiple-choice quiz for each chapter
- Suggestions for managing the homework grading load
- Sample tests and study guides
- Project instructions for designing a questionnaire and analyzing the data in groups, including
    - Data collection project instructions and timeline for a 15-week semester
    - Questionnaire design using Google Forms
    - Instructions for pretesting group questionnaires
    - Data cleaning instructions
    - Organizing PowerPoint slides for group presentations
    - Data analysis instructions
- Sample syllabus that includes a list of material covered in each class when taught by the authors.

PowerPoint slides that meet accessibility standards to accompany each chapter.

Author-created slides are available directly from the authors. To request these slides, please contact Lisa Daniels at LDaniels2@washcoll.edu. These slides provide more detailed information on each slide.

Instant polls that are built into the PowerPoint slides or for use independently outside of PowerPoint

STRUCTURE OF THE BOOK

As described previously, Part One of the book is titled “The Research Process and Data Collection.” In Chapter 1, we offer an overview of the research process by briefly describing the major steps involved at each stage. We then describe primary data collection in Chapter 2, including sampling frames, sample selection techniques, and sampling weights. In Chapter 3, we review the principles of questionnaire design along with ethical issues. In Part Two of the book, “Describing Data,” we introduce Stata in Chapter 4, discuss methods for preparing and transforming data in Chapter 5, and cover descriptive statistics in Chapter 6. Part Three, “Testing Hypotheses,” includes five chapters that cover the normal distribution followed by hypothesis testing related to a single mean, two means, analysis of variance, and the chi-square statistic. In Part Four, “Exploring Relationships,” we cover correlation, linear regression, regression diagnostics and some advanced regression topics briefly. Finally, in Part Five, a chapter is devoted to writing a research paper, including a detailed description of each section of a research paper with a special emphasis on reporting statistical results.

Instructors may choose to skip Chapters 2 and 3 if they prefer not to cover sampling techniques or questionnaire design. Similarly, the

more advanced topics in Chapters 13-15 may be excluded if time is limited or students have the option of going on to a more advanced course where they would cover those topics in depth.

WHAT'S NEW IN THE SECOND EDITION?

The second edition of *An Introduction to Statistics and Data Analysis Using Stata: From Research Design to Final Report* incorporates fresh content, updated data sets, additional data sets, enhanced explanations based on feedback from students and referees, revisions to Stata code based on software changes, and the removal of certain material. These changes are described next.

To illustrate the use of statistical concepts related to the real world, all news articles have been updated. These include new articles related the use of ChatGPT in college courses, COVID-19 practices and the politics driving these practices, updated policies related to the use of SAT scores in college admissions, characteristics of individuals using online dating apps, and factors that drive the price of gas-powered and electric vehicles.

New chapter exercises have been added.

Exercises with older data sets have been updated to use the most recent version of a data set when available.

Several new sections have been added, including

A discussion of laws, theories, and hypotheses

A more complete review of the types of surveys

A section on the treatment of missing values when generating new variables

More detail on when to use row and column percentages

A series of graphs that illustrate the *p*-value and the difference between the distribution of a single variable versus the distribution of sample means

A series of graphs that help to illustrate the standard error of the mean and the standard error of the mean difference

A table at the end of Chapters 7 through 12 that summarizes the null hypothesis, the test, the information known, and the procedure

A table at the end of Chapters 7 through 12 to illustrate the code used in each chapter and the purpose of the code

More detail about how to calculate the 95% confidence interval and why this is becoming more important compared to *p*-values

New sections in Chapters 7, 8, and 11 on testing of proportions

A new chapter 15 on advanced methods in regression analysis . The new chapter provides a brief introduction to multinominal logit, instrumental variables, analysis of time-series data, and panel data regression analysis.

Because we want to emphasize Stata code instead of menus, all screenshots of menus to produce Stata results have been removed. We continue to briefly discuss how to use the menus in words by indicating what tabs and subheadings would be needed.

When Stata updated to a more recent version, there were some minor code changes. The Stata code has been updated

throughout the book.

In Chapter 16, which covers “Writing a Research Paper,” we have incorporated a new journal article on the role of artificial intelligence in higher education as a key example to illustrate the parts of a journal article.

A new appendix has been added that describes each data set used in the book in detail.

There are several new resources for instructors as described previously.

ACKNOWLEDGMENTS

When we first started on this journey, we had no idea how many people it takes to write a textbook! We count 74 people who were involved in some aspect of writing, editing, reviewing, marketing, producing, and supporting the book. Beginning with Sage, Leah Fargotstein, our acquisitions editor, made the process painless with her patience, guidance, and continuous encouragement throughout the first and second editions of this book. We also received help from other staff at Sage and QuAds Prepress Pvt. Ltd. Elizabeth Wells and Claire Laminen exchanged endless emails with us related to permissions needed for articles in the first edition in the book, which allowed us to better understand the process for the second edition. Shelly Gupta and Tori Mirsadjadi also provided guidance in our quest for permissions early on.

Special thanks to Chelsea Neve for her role in developing the book's website and additional student resources. The Sage marketing team —Susannah Goldes, Shari Countryman, Andrew Lee, and Heather Watters—also deserve thanks for launching both the first and second editions of the book. Karen Wiley's oversight of the production process and the contributions of Izumi Sunada, Ginkhan Siam, William Ragsdale, Integra and Scott Oney in production, cover design, indexing, typesetting, and proofreading are deeply appreciated. We also extend our gratitude to copyeditors, Rajasree Ghosh and Rajeswar Krithivasan from QuADS, for their meticuolous attention to detail, enhancing the overall quality of the book.

Thanks also go to the staff and students of Washington College for their assistance with the first edition of the book, which is the foundation for the second edition. Jennifer Kaczmarczyk did the bulk of the work to get the permissions started, wading through e-mails, contracts, and phone calls to follow up. Benjamin Fizer, a Washington College student, spent more than 50 hours capturing

every dialog box, figure, and output in the first edition. He also read the entire book to help develop the glossary and changed all of the Stata code in the book to the correct format. Amanda Kramer, from the Miller Library, helped identify databases from the various fields covered in the book. We are also grateful to the students enrolled in the data analysis course who pointed out errors in the book.

We would also like to thank the administration at Washington College, which supported this project financially in a number of ways. The college funded travel to three conferences related to textbook writing and Stata, as well as two “research reassigned time” awards that allowed one of us (Lisa) to reduce her course load in two semesters along with funds to pay for a student assistant during those semesters.

Bill Rising from Stata Corporation deserves special thanks for going through the book and offering numerous suggestions to improve our Stata code and language related to statistics. Any remaining mistakes must have been introduced after Bill read the book since he did not miss anything!

We would also like to thank the people who reviewed the book over six rounds of revisions for the first edition. Their attention to detail as well as the big picture helped us improve the book in countless ways.

Eileen M. Ahlin, *Penn State Harrisburg*

Rachel Allison, *Mississippi State University*

Matthew Burbank, *University of Utah*

Hwanseok Choi, *University of Southern Mississippi*

Mengyan Dai, *Old Dominion University*

Kimberlee Everson, *Western Kentucky University*

Wendy L. Hicks, *Ashford University*

Monica L. Mispireta, *Idaho State University*

Steven P. Nawara, *Lewis University*

Holona LeAnne Ochs, *Lehigh University*

Parina Patel, *Georgetown University*

John M. Shandra, *State University of New York at Stony Brook*

Janet P. Stamatel, *University of Kentucky*

Anna Yocom, *The Ohio State University*

For the second edition, we would like to thank the following reviewers. Again, their thoughtful suggestions helped to improve the book.

Nurgul Aitalieva, *Purdue University Fort Wayne*

Matthew Burbank, *University of Utah*

Youssef Chouhoud, *Christopher Newport University*

Stacey Clifton, *Radford University*

Michael Danza, *Copper Mountain College*

Jared DeLisle, *Utah State University*

Gemma Dipoppa, *Brown University*

John Doces, *Bucknell University*

Joseph Earley, *Loyola Marymount University*

Jeffrey Glas, *University of Georgia*

Shane Gleason, *Trinity College*

Feng Hao, *University of South Florida*

Aimee Imlay, *Mississippi State University*

Kyungkook Kang, *Claremont Graduate University*

Chelsea Kelly, *The Catholic University of America*

Kyle Knight, *University of Alabama in Huntsville*

Naoru Koizumi, *George Mason University*

Veena S. Kulkarni, *Arkansas State University*

Drew Lanier, *University of Central Florida*

David Macdonald, *University of Florida*

Nathan Martin, *Arizona State University*

Jason Miller, *Eastern Kentucky University*

Michael Morgan, *Marietta College*

Joe Nedelec, *University of Cincinnati*

Adam Newmark, *Appalachian State/Virginia Tech*

Chiamaka Nwosu, *King's College London*

Hector Sandoval, *University of Florida*

David Simon, *University of Connecticut*

Sheryl Skaggs, *The University of Texas at Dallas*

Janet Stamatel, *University of Kentucky*

Jaeyun Sung, *Lyon College*

Kara Sutton, *Southern Methodist University*

James Swartz, *University of Illinois Chicago*

Joseph Taylor, *University of Colorado, Colorado Springs*

Jill Weinberg, *Tufts University*

Felix Weiss, *Aarhus University*

Cameron Wimpy, *Arkansas State University*

Jun Zhao, *Georgia State University*

Finally, we are grateful to our two children, Andrea and Alex, for patiently listening to conversations about economics and statistics throughout their lives!

PART I THE RESEARCH PROCESS AND DATA COLLECTION

1 A BRIEF OVERVIEW OF THE RESEARCH PROCESS

CHAPTER PREVIEW

Steps in the Research Process	Example from Valkenburg and Schouten, "Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem"
Step 1: Choose a research area and read the literature	• Impact of social media on self-esteem and well-being among teens
Step 2: Identify the gaps or ways to extend the literature	• Limited research on uses and consequences of social media use among adolescents • Lack of distinction between social and nonsocial Internet use
Step 3: Examine the theory	• Human beings have a desire to protect and enhance their self-esteem. • Self-esteem is strongly related to well-being.
Step 4: Develop your research questions and form hypotheses	• Does the frequency with which teens use networking sites have an impact on their self-esteem and well-being? • Does positive or negative feedback affect self-esteem?
Step 5: Develop your research method	• Online survey among adolescents between 10 and 19 years of age who have a profile on a social networking site
Step 6: Examine the data or other evidence	• Descriptive statistics of frequency of usage and types of feedback received from peers • Regression analysis to determine impact on self-esteem
Step 7: Write the research paper	• Introduction • Literature Review • Data and Methods • Results • Discussion • Conclusion

Source of example in the second column: Valkenburg, Peter, and Schouten (2006).

Steps in the Research Process	**Example from Valkenburg and Schouten, "Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem"**
Step 1: Choose a research area and read the literature	Impact of social media on self-esteem and well-being among teens

Steps in the Research Process	Example from Valkenburg and Schouten, "Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem"
Step 2: Identify the gaps or ways to extend the literature	Limited research on uses and consequences of social media use among adolescents Lack of distinction between social and nonsocial Internet use
Step 3: Examine the theory	Human beings have a desire to protect and enhance their self-esteem. Self-esteem is strongly related to well-being.
Step 4: Develop your research questions and form hypotheses	Does the frequency with which teens use networking sites have an impact on their self-esteem and well-being? Does positive or negative feedback affect self-esteem?
Step 5: Develop your research method	Online survey among adolescents between 10 and 19 years of age who have a profile on a social networking site
Step 6: Examine the data or other evidence	Descriptive statistics of frequency of usage and types of feedback received from peers Regression analysis to determine impact on self-esteem

Steps in the Research Process	Example from Valkenburg and Schouten, "Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem"
Step 7: Write the research paper	Introduction Literature Review Data and Methods Results Discussion Conclusion

Source of example in the second column: Valkenburg, Peter, and Schouten (2006).

1.1 INTRODUCTION

Does the use of ChatGPT to practice homework problems improve scores? Were mask mandates motivated by politics during the COVID-19 pandemic? Do education levels vary by gender identity among those who use online dating applications? Do students from high-income families earn higher SAT scores? These are just some of the examples used in this book to illustrate the endless number of interesting questions that can be examined with **statistics**.

Although the majority of this book is focused on statistics, it is important to understand where and how **data analysis** plays a role in the research process. We begin, therefore, by giving you a brief overview of the research process. This includes choosing a research area, identifying gaps in the literature, examining the theory, developing research questions and hypotheses, identifying your research method, analyzing data, and writing the research paper. Although this is a brief overview, some of these topics are covered in greater detail later in the book. In particular, Chapter 16 on "Writing the Research Paper," offers guidance and examples from published papers for each section of a research paper, including how to structure a literature review, examine the theory, describe your data and methods, report statistical results, discuss your results within the context of the literature and theory, and offer your final conclusions, limitations, and areas for future research.

1.2 WHAT IS RESEARCH

Research is often described as the creation of knowledge. It begins with the construction of an argument that can be supported by evidence. As described by Greenlaw (2009), scholars then create a "conversation" in scholarly journals to discuss the argument. In many cases, scholars will identify gaps in the argument and offer alternate views or evidence. In other cases, scholars may forward or extend the argument by offering new insights or examine the same argument from a different angle. Another equally valid form of research is to replicate what others have done. This can be done by conducting the same research in a different region, in a different time period, over a longer time period, or with a different set of participants. All of these may validate the original argument or disprove it.

1.3 STEPS IN THE RESEARCH PROCESS

1.3.1 Read the Literature and Identify Gaps or Ways to Extend the Literature

When starting a new research project, it is common to begin by choosing a general area, such as poverty, pollution, sports, social media, criminal justice, and so on. Before identifying a research question within the general area, you must begin reading the ***literature***. The literature can be defined as a body of articles and books, written by experts and scholars, that has been *peer reviewed*. A peer review is when two or three scholars are asked to anonymously evaluate a manuscript's suitability for publication and either reject it or accept it, typically with revisions based on their recommendations.[1] Articles in the body of literature will cite other sources and will be written for an audience of fellow scholars. Nonscholarly materials, such as newspapers, trade and professional sources, letters to the editor, and opinion-based articles are not considered part of the literature. They are sometimes used in scholarly papers, but never as a sole source of information.

Most disciplines have their own databases, with articles, book chapters, dissertations, and working papers from their field. Table 1.1 shows a list of the key databases in several fields.

TABLE 1.1 ■ Databases Of Scholarly Literature From Different Fields

Field	Database	Content	Website
Criminal Justice	ProQuest Criminal Justice Database	A comprehensive database of U.S. and international criminal justice journals	www.proquest.com/products-services/pqcriminaljustice.html
	Criminal Justice Abstracts	Titles and abstracts for articles from most significant sources in the field	www.ebsco.com/products/research-databases/criminal-justice-abstracts
Economics	Econ Lit	Over 2,000 journals, plus books, dissertations, working papers, and book reviews	www.aeaweb.org/econlit
Political Science	JSTOR	6,800 political science journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=43693417
	Academic Search Complete	340 full-text political science reference books and monographs and more than 44,000 full-text conference papers	www.ebscohost.com/academic/subjects/category/political-science
Psychology	PsycINFO	Four million bibliographic records, including more than 2 million digital object identifiers to allow for direct linking to full-text psychology articles and literature. Indexing of more than 2,500 scholarly psychology journals	www.apa.org/psycinfo
Public Health	PubMed	Access to 12 million Medline citations dating back to the 1950s	www.ncbi.nlm.nih.gov/pubmed
	PAIS	Political, social, and public policy issues	www.proquest.com
	Nexis Uni	15,000 news, business, and legal sources	www.lexisnexis.com
Sociology	Sociological Abstracts	Abstracts of sociology journal articles and citations to book reviews drawn from more than 1,800 serial publications and abstracts of books, book chapters, dissertations, and conference papers	http://proquest.libguides.com/SocAbs
	JSTOR	8,000 sociology journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=43693423
	Academic Search Complete	900 full-text sociology journals, abstracts for more than 1,500 "core" coverage journals, data from nearly 420 "priority" coverage journals and more than 2,900 "selective" coverage journals, and indexing for books/monographs, conference papers, and other nonperiodicals	www.ebscohost.com/academic/socindex

Field	Database	Content	Website
Criminal Justice	ProQuest Criminal Justice Database	A comprehensive database of U.S. and international criminal justice journals	www.proquest.com/products-services/pqcriminaljustice.html
	Criminal Justice Abstracts	Titles and abstracts for articles from most significant sources in the field	www.ebsco.com/products/research-databases/criminal-justice-abstracts
Economics	Econ Lit	Over 2,000 journals, plus books, dissertations, working papers, and book reviews	www.aeaweb.org/econlit
Political Science	JSTOR	6,800 political science journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=4369341
	Academic Search Complete	340 full-text political science reference books and monographs and more than 44,000 full-text conference papers	www.ebscohost.com/academic/subjects/category/politic science
Psychology	PsycINFO	Four million bibliographic records, including more than 2 million digital object identifiers to allow for direct linking to full-text psychology articles and literature. Indexing of more than 2,500 scholarly psychology journals	www.apa.org/psycinfo
Public Health	PubMed	Access to 12 million Medline citations dating back to the 1950s	www.ncbi.nlm.nih.gov/pubmed
	PAIS	Political, social, and public policy issues	www.proquest.com
	Nexis Uni	15,000 news, business, and legal sources	www.lexisnexis.com

Field	Database	Content	Website
Sociology	Sociological Abstracts	Abstracts of sociology journal articles and citations to book reviews drawn from more than 1,800 serial publications and abstracts of books, book chapters, dissertations, and conference papers	http://proquest.libguides.com/SocAbs
	JSTOR	8,000 sociology journals, books, and pamphlets	www.jstor.org/action/showJournals?discipline=4369342
	Academic Search Complete	900 full-text sociology journals, abstracts for more than 1,500 “core” coverage journals, data from nearly 420 “priority” coverage journals and more than 2,900 “selective” coverage journals, and indexing for books/monographs, conference papers, and other nonperiodicals	www.ebscohost.com/academic/socindex

In all of these databases, you can type in keywords from areas that interest you. You can then peruse article titles and read abstracts to get a sense of the thought-provoking questions and research in your area of interest. Once you have found some key articles that zero in on your research interests, you can review earlier articles that were referenced by the key articles (backward citation searching) and search forward in time to see what other articles have cited your key articles since they were written. For example, if an article was written in 1995, you can find every article written since 1995 that has cited the original article. This can be done through Google Scholar, PubMed, Science Direct, Scopus, and Web of Science. As you find more articles related to your specific topic, you will find that the literature will indicate what has been done in your area of interest, what questions remain, and if there are gaps or contradictions in the literature. All articles will also indicate the flaws in their own research and areas for future research. You can then identify your own research questions based on the contradictions or gaps in the literature or the need for forwarding or extending the argument. As mentioned earlier, you can also replicate what other authors have done by repeating the same study based on a different time period, a different region or country, or a different set of data.

For more information on how to identify gaps in the literature and write a literature review, refer to Chapter 16, “Writing a Research Paper,” which offers guidelines on each section of a research paper along with examples from journal articles to illustrate these concepts.

1.3.2 Examine the Theory

A *theory* can be defined as a comprehensive explanation that is supported by a large body of evidence. For example, the theory of comparative advantage used by economists suggests that countries will specialize in producing a good in which they have a lower opportunity cost and trade with each other to benefit from mutual gains. Another example is Darwin's theory of evolution, which is used to explain changes in species over time.

Theories are different from hypotheses and laws. A hypothesis is a testable prediction. For example, you could test the hypothesis that increased exposure to sunlight will lead to higher levels of vitamin D in the body. Unlike theories and hypotheses, which can be updated based on new evidence, a law describes a universal and consistent relationship between two or more variables. For example, the law of demand states that as the price of a good increases, the quantity demanded will decrease, holding all other factors constant.

Theory plays an important role in developing your research questions and hypotheses. In the article used in the chapter preview, for example, Valkenburg et al. (2006) cite the theory that humans have a desire to protect their self-esteem and that self-esteem affects well-being. From this basic theory, they develop their research question related to how social media usage affects self-esteem and thus well-being.

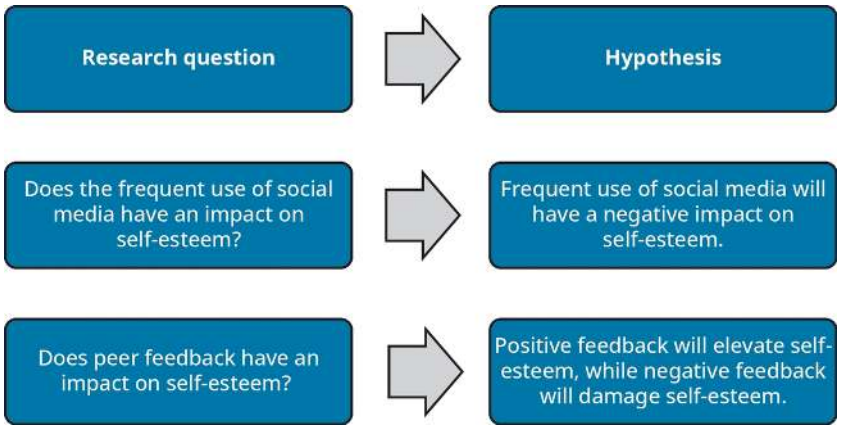
As a second example, the theory of social capital could be used to develop research questions. This theory suggests that individuals benefit from social networks that can offer emotional support, access to resources, and opportunities. Although this theory was first developed within the field of sociology, many fields use the theory of social capital including economics, public health, political science, and education. Using social capital theory as our framework, we could ask how social media usage contributes to the formation of social capital among college students and if this social capital impacts their performance. Our hypothesis could be that social capital leads to better academic performance.

Although theory offers a framework to help develop hypotheses, we also return to the theory when examining the results of our study. In other words, do your results conform to the stated theories? How do they differ? Why might they differ? These concepts are covered in more detail in Chapter 16, "Writing a Research Paper."

1.3.3 Develop your Research Questions and Hypotheses

As described in the previous sections, you begin to form your research questions as you read the literature and examine the theory. Your questions may change in the early stages of the research as you continue to find more articles on the topic or new ways that scholars have examined or answered the questions in your research area.

In the example used in the chapter preview, the authors identify two research questions that are illustrated in Figure 1.1. Each of these questions can then be restated as a *hypothesis*, or an answer to the questions. As you begin your research, you won't know the answer to your research questions, but your hypotheses indicate what you expect to find based on theory. Your research may then find evidence to support or refute your hypothesis, which is a key feature of a hypothesis. It must be testable.



Description

Figure 1.1 From Research Question To Hypothesis

Developing the research questions is often the most difficult part of the research process and requires a lot of work up front before the **questionnaire** or study design can or should begin.

In addition to identifying the research question, it is also important to begin thinking about your key variables (self-esteem, social media usage, and feedback, in this case) and how they relate to one another. In particular, self-esteem is the ***dependent variable*** because its value depends on the two independent variables: social media usage and feedback received. A dependent variable is defined in general as a variable whose variation is influenced by other variables. This is covered in more detail in later chapters.

1.3.4 Identify your Research Method

Once you have identified your research questions, your next step is to develop your research method. There are many types of research methods, such as qualitative research (narrative research, case studies, ethnographies), quantitative research (surveys and experiments with statistical analysis), and mixed methods that include both qualitative and quantitative approaches. Since this textbook focuses on quantitative analysis of *primary data* (data collected by the researcher) and *secondary data* (data collected by someone else), the remaining chapters in this book will be devoted to sampling, questionnaire design, and data analysis, with a final chapter on writing a research paper. For more complete works on the other types of research methods mentioned, see Leedy and Ormrod (2001) or Creswell and Creswell (2018).

1.3.5 Examine the Data or Other Evidence

As described above, the majority of the remainder of this book covers *data analysis*. This begins in Chapter 6 with **descriptive statistics**, such as the **mean**, **median**, and standard deviation. We then cover testing of hypotheses and exploring relationships through advanced statistical techniques. These include testing a hypothesis about a single mean, two independent means, one-way analysis of variance, chi-squared tests, **linear regression**, and an overview of some more advanced statistical methods. These will be discussed in detail in Chapters 6 through 15.[2]

1.3.6 Write the Research Paper

Once all steps of the research process are completed, you may begin to write your research paper. The typical sections in a research paper are the introduction, the literature review, the method section, the results, a discussion, and the conclusions. Each of these sections is described in Chapter 16 along with examples from published articles. We also review conventional guidelines and style guidelines for reporting statistical results.

1.4 CONCLUSION

This chapter is meant to provide a very brief overview of the research process and where data fits into this process. As mentioned in the “Research Method” section, both primary data and secondary data can be used. Primary data is often used in fields such as sociology, psychology, medicine, and marketing through surveys, experiments, interviews, and observations. Secondary data is common in fields like economics, history, and public policy, where studies employ data from government reports, historical records, or databases. Because both sources of data are important, this book offers examples of each along with chapters that focus on primary data (Chapter 2: Sampling Techniques, Chapter 3: Questionnaire Design, and Chapter 5: Preparing and Transforming Your Data). In Chapter 4, we also offer an example of entering data directly into Stata so that you can better understand the elements of a data set. In other chapters, secondary data sources such as the General Social Survey, the College Scorecard, the National Survey on Drug Use and Health, the OkCupid mobile dating app data, and COVID-19 state-level data are used to generate descriptive statistics and test hypotheses.

In conclusion, this chapter sets the stage for understanding the research process and the role of data analysis, preparing you to effectively utilize both primary and secondary sources in your own research endeavors.

EXERCISES

1. Read the article “Prevalence and Motives for Illicit Use of Prescription Stimulants in an Undergraduate Sample” by Teter, McCabe, Cranford, Boyd, and Guthrie (2005). As you read the article, answer the following questions, which are based on guidelines offered by Greenlaw (2009).
    a. What question or questions are the authors asking?
    b. Describe the theoretical approach that the authors use to develop their research question.
    c. What answers do the authors propose?
    d. In what ways does the current study improve over previous research, according to the authors of the article? In other words, what gaps do the authors identify in the current literature?
    e. What method do the authors use to answer their questions?
    f. What limitations do the authors identify in their study?
    g. What suggestions do the authors have for follow-up research that should be done?
2. Choose a general area of research that interests you. This could be sports, cancer, poverty, social media usage, gaming, and so on. Use the techniques identified in Section 1.2 to narrow your focus as you begin perusing the literature and using forward and backward searching for articles of particular interest to you. Once you have done the initial reading, you should develop a tentative research question and identify five articles that are most closely related to your question. For each of the five articles, answer the following questions:
    a. What question or questions are the authors asking?
    b. Describe the theoretical approach that the authors use to develop their research question.
    c. What is the hypothesis that the authors propose?
    d. What answers do the authors propose?
    e. In what ways does the current study improve over previous research according to the authors of the article? In other words, what gaps do the authors identify in the current literature?
    f. What method do the authors use to answer their questions?
    g. What limitations do the authors identify in their study?

h. What suggestions do the authors have for follow-up research that should be done?

KEY TERMS

data analysis

dependent variable

descriptive statistics

linear regression

literature

mean

median

questionnaire

statistics

Descriptions of Images and Figures

Back to Figure

The research questions are “Does the frequent use of social media have an impact on self-esteem?”, and “Does peer feedback have an impact on self-esteem?” The hypothesis for these are “Frequent use of social media will have a negative impact on self-esteem.”, and “Positive feedback will elevate self-esteem, while negative feedback will damage self-esteem.”

2 SAMPLING TECHNIQUES

CHAPTER PREVIEW

Terms	Definitions
Unit of observation	Type of entity being studied, such as individuals, households, or businesses
Population	The complete set of units that is the topic of a study
Sample	A subset of the population, intended to represent the population, from which data will be collected
Nonprobability sampling	Selection of units based on the discretion of researchers, which means that it is not possible to calculate the probability of selecting each unit
Probability sampling	Selection of units using random numbers, such that it is possible to calculate the probability of selecting each unit
Simple random sample	A sample in which each unit in the population has the same probability of selection
Systematic random sample	A sample in which the selected units are at constant intervals evenly spaced in a list of the units across the population
Multilevel sampling	A sample in which aggregated units (e.g., towns) are selected, followed by the selection of more disaggregated units (e.g., households)
Stratification	Division of the population into different groups, each of which may be sampled differently
Sampling weights	Weights used to calculate population averages in a way that compensates for the effect of the sampling method

Terms	Definitions
Unit of observation	Type of entity being studied, such as individuals, households, or businesses
Population	The complete set of units that is the topic of a study
Sample	A subset of the population, intended to represent the population, from which data will be collected
Nonprobability sampling	Selection of units based on the discretion of researchers, which means that it is not possible to calculate the probability of selecting each unit
Probability sampling	Selection of units using random numbers, such that it is possible to calculate the probability of selecting each unit
Simple random sample	A sample in which each unit in the population has the same probability of selection
Systematic random sample	A sample in which the selected units are at constant intervals evenly spaced in a list of the units across the population
Multilevel sampling	A sample in which aggregated units (e.g., towns) are selected, followed by the selection of more disaggregated units (e.g., households)
Stratification	Division of the population into different groups, each of which may be sampled differently

Terms	Definitions
Sampling weights	Weights used to calculate population averages in a way that compensates for the effect of the sampling method

2.1 INTRODUCTION

Primary data refer to data collected directly by the researchers. This contrasts with secondary data, which are data collected by another researcher or an organization, such as a government agency. In the social sciences, primary data are often collected through a sample survey, where the researcher interviews (or hires others to interview) a subset of the population on a topic of interest. The quality of the data depends heavily on selecting a good sample and asking the right questions. This was dramatically illustrated by the polling for the 1936 U.S. presidential elections.

As described by the National Constitutional Center in Philadelphia, *The Literary Digest* had run polls in four previous elections, successfully predicting the winner in each. In 1936, they carried out a poll of two million voters and predicted that the Republican candidate Alf Landon would beat Franklin Roosevelt, the Democratic candidate. In fact, Roosevelt won in a landslide, beating Landon in 46 of 48 states. On the other hand, George Gallup used a random sample of just 50,000 voters and correctly predicted that Roosevelt would win (see Figure 2.1).

Five biggest political polling mistakes in American history

October 2, 2012 by NCC Staff

If some political polls were truly accurate, Alf Landon would have been America's president during World War II, instead of FDR. Here's a look at an alternative universe of politics, as we examine the five biggest political poll blunders in U.S. history.

Alf Landon beats FDR in a landslide

The mother of all botched political polls was a 1936 *Literary Digest* straw poll survey that said GOP challenger Alf Landon would win in a landslide over the incumbent, Franklin Delano Roosevelt, with 57 percent of the vote.

The *Literary Digest* used national straw polls in 1920, 1924, 1928 and 1932, and it guessed the winner of each presidential election.

In 1936, a young rival pollster, George Gallup, made his own prediction before the magazine issued its poll; He said *Literary Digest* would get it all wrong, despite the *Digest's* decent track record in previous polls.

So was right? The *Literary Digest* disaster helped establish Gallup as the nation's pre-eminent pollster. The *Digest* polled about 2 million people, most of who were magazine readers, car owners or telephone customers—and had money during the Depression. It was not a representative sample.

Gallup used a random poll sample of 50,000 people.

President Roosevelt won the 1936 election easily, with 63 percent of the vote, and the *Literary Digest* was out of business the following year. If he had won, Landon could have been our wartime president.

Description

Figure 2.1 Article

NCC (National Constitution Center). 2012. "The five biggest polling mistakes in U.S. history." National Constitutional Center, Philadelphia. https://news.yahoo.com/news/five-biggest-political-polling-mistakes-u-history-132611721.html

The problem was that *The Literary Digest* relied on lists of "magazine readers, car owners, and telephone subscribers." During the Great Depression, these lists had a disproportionate number of high-income households who opposed Roosevelt and his New Deal policies. In addition, *The Literary Digest* conducted the poll by sending postcards to 10 million voters and relying on respondents to mail back

their responses. The response rate was higher among Republicans than Democrats, which also contributed to the incorrect result (Squire, 1988).

The Literary Digest was discredited by this high-profile failure and closed soon after. The success of Gallup's prediction established the national reputation of his firm, which grew to become one of the largest political polling companies. It also catalyzed the development of modern random-sample polling. The lesson for sampling methods is that it is much more important to have a representative sample than to have a large sample. In addition, this experience highlights the fact that a low response rate can distort the results of a survey. Magazine subscriber polls and online polls are not considered scientific or reliable, no matter how many people respond to them.

This chapter introduces the basic concepts of sampling, discusses some of the more common sampling methods, and explains the calculation and use of sampling weights. However, it only scratches the surface of a large and complex topic. Readers interested in a more in-depth treatment of sampling methods may wish to consult Rea and Parker (2005), Scheaffer, Mendenhall, Ott, and Gerow (2011), or Daniel (2011).

2.2 SAMPLE DESIGN

As discussed in the previous chapter, any research must begin with careful consideration of the objectives of the study. What are the research questions? What information is needed to answer those questions? What is the ***unit of observation***, defined as the type of entity about which the study will collect information? In social science research, the unit of observation is often individuals, households, businesses, or other social institutions. Table 2.1 gives four examples of units of observation, depending on the research question and information needed.

TABLE 2.1 ■ Examples of Research Questions and Surveys

	Example 1	Example 2	Example 3	Example 4
Research question	Which political candidate is favored by voters?	What is the average yield of rice farmers?	Why do students transfer from one university to another?	How do regulations affect small businesses?
Information needed	The opinions of voters regarding each candidate	The rice production and area under rice cultivation among rice farmers	The reasons that students give for wanting to transfer out	The cost of complying with a set of business regulations
Unit of observation	Voters	Rice farmers	Students	Small businesses
Population	All likely voters in the country, defined as those who voted in at least two of the past three elections	All rice farmers in a country, defined as those growing rice in the previous year	All full-time undergraduate students at the university in a year	All businesses in the state that have 10 or fewer full-time workers
Sample	1,500 likely voters	2,000 rice farmers	200 students	5,000 small businesses
Description of survey	A polling firm collects information from 1,500 likely voters about their political views	A statistical agency gathers information from 2,000 rice farmers to estimate the average yield	A university carries out a survey of 200 students to gather information on reasons for transferring	A state agency carries out a survey of 5,000 small businesses in a state

	Example 1	**Example 2**	**Example 3**	**Example 4**
Research question	Which political candidate is favored by voters?	What is the average yield of rice farmers?	Why do students transfer from one university to another?	How do regulations affect small businesses?
Information needed	The opinions of voters regarding each candidate	The rice production and area under rice cultivation among rice farmers	The reasons that students give for wanting to transfer out	The cost of complying with a set of business regulations
Unit of observation	Voters	Rice farmers	Students	Small businesses
Population	All likely voters in the country, defined as those who voted in at least two of the past three elections	All rice farmers in a country, defined as those growing rice in the previous year	All full-time undergraduate students at the university in a year	All businesses in the state that have 10 or fewer full-time workers
Sample	1,500 likely voters	2,000 rice farmers	200 students	5,000 small businesses

	Example 1	Example 2	Example 3	Example 4
Description of survey	A polling firm collects information from 1,500 likely voters about their political views	A statistical agency gathers information from 2,000 rice farmers to estimate the average yield	A university carries out a survey of 200 students to gather information on reasons for transferring	A state agency carries out a survey of 5,000 small businesses in a state

In statistics, the ***population*** is the complete set of individuals, households, businesses, or other units that is the subject of the study. Table 2.1 gives some examples of populations corresponding to the studies listed. Note that each population is defined in terms of the type of unit of observation, the geographic scope, and the period of time.

The ***sample*** is a subset of the population consisting of units from which data will be collected. *Sampling* is the process of selecting the sample in a way that ensures it will be representative of the population. One option, of course, is to collect data from every unit in the population—that is, to carry out a census. This might be feasible if the population is defined narrowly or if the budget is very large. For example, if the population is defined as all the banks in a given town, it would probably be feasible to carry out a census. Alternatively, the governments of many countries carry out a population census every 10 years. But for most purposes, it is more cost-effective to conduct a *sample survey*, defined as systematic collection of data from a limited number of units (e.g., households) to learn something about the population. Using the previous four examples, Table 2.1 provides a possible sample for each.

All surveys face a trade-off between the objectives of reducing cost and increasing accuracy. If cost were no object, then one could carry out a census (covering all units), and it would not be necessary to worry about whether the selected units were representative of the whole group. Alternatively, if accuracy were not a concern, one could just sample a handful of units in one location, which would minimize costs. In practice, most surveys are in between these two extremes. A key challenge is to ensure that the sample is selected in a way that accurately reflects the characteristics of the whole group.

2.3 SELECTING A SAMPLE

2.3.1 Probability and Nonprobability Sampling

How does the researcher select a sample for the survey? One intuitive approach is for the researcher to simply choose a set of units based on availability or subjective judgment. This is called *nonprobability sampling* because it is not possible to calculate the probability of selecting each unit. Below is a partial list of some of the various types of nonprobability sampling:

Convenience sampling involves selecting units from available but partial lists or selecting people who are passing by a location, such as a supermarket.

Purposive sampling means that the researcher uses knowledge of the field to select units to be studied.

Snowball sampling refers to picking an initial set of units, then a second round of units that are nearby or have links to the first-round selections. There may be additional rounds.

Nonprobability sampling has the advantage of being quick and inexpensive to implement. It is often used with qualitative research focused on in-depth exploration of a topic on a relatively small number of observations. Qualitative research can complement quantitative surveys in several ways. It can be

carried out before a random-sample survey to identify key issues, contributing to the design of the questionnaire. Or it can be conducted after a survey to help interpret the results or explain unexpected findings. For an in-depth discussion of qualitative research and mixed methods that combine qualitative and quantitative research, see Creswell and Creswell (2017).

The main disadvantage of nonprobability samples is that they are likely to be biased, meaning that the sampled units do not accurately reflect the characteristics of the population. (The 1936 polling by *The Literary Digest* is an example.) For this reason, it is not possible to infer characteristics of the population from the characteristics of the sample. For example, a nonprobability sample of businesses will probably include mostly large, well-known businesses—those that have more visible locations and those that advertise. Car dealers, supermarkets, and restaurants will probably be overrepresented, while shoe repair shops, cleaning services, and home-based day care providers are likely to be underrepresented or excluded.

For these reasons, almost all larger surveys carried out by researchers and professional polling companies use *probability sampling*, defined as sampling in which the selection is made randomly from a complete list of units. (Indeed, it is also known as **random sampling**.) The researcher defines the population and the selection method but does not have any discretion in deciding which individual units will be included in the sample.

If a random sample is well-designed and large enough, it will be representative of the population. In other words, the characteristics of the sample will be similar to the characteristics of the population. In the example above, the average size of businesses in the sample will be similar to the average size of businesses in the town. In technical terms, the average business size in the sample will be an *unbiased estimate* of the business size in the population. This means that if you took repeated samples using the same method, the average across samples would converge toward the population average as the number of samples increased.

Another advantage of a random sample is that we can estimate the *sampling error* of our sample-based averages—that is, the error associated with selecting a sample rather than collecting data from every unit in the population. As described in more detail in Chapter 8, the sampling error of a variable is based on (a) the size of the sample, (b) how it was selected, and (c) the variability of the variable in question. If the sample is large or the variability is low, the sample error is likely to be small. One way to describe the sampling error is the *95%* **confidence interval**, defined such that there is a 95% probability that the true average lies between the two numbers. If a political poll reveals that 45% of voters approve of a state governor with a **margin of error** of 3 percentage points, this means that the 95% confidence interval is 45% ± 3 percentage points or 42% to 48%. In other words, there is a 95% probability that this confidence interval contains the true level of approval (if you polled every voter in the state). This topic is discussed in more detail in Chapter 7.

Note that a sample does not have to represent a large percentage of the population to be precise. In national political polls, a sample of 800 to 1,200 is usually sufficient to reduce the margin of error to less than 5 percentage points, in spite of the fact that the sample is roughly 0.001% (or 1 in 100,000) of the total voting population in the United States. It is also useful to note that these calculations count only sampling error. They do not include other sources of error, such as respondents who give false answers or pollsters misidentifying who will decide to vote.

In a large majority of surveys, it is worth the additional effort to select the units randomly. The remainder of this section describes the methods used for different types of random sampling.

2.3.2 Identifying a Sampling Frame

To select a random sample, a researcher needs a *sampling frame*—that is, a list of sampling units in the population from the sample is selected. Ideally, the sampling frame would be a complete list of the units

in the population, but this is not always possible. Sometimes an available list is smaller than the target population. For example, a researcher may wish to define the population as all rice farmers in a region, but the available list may include only members of a cooperative of rice farmers, thus excluding rice farmers who are not members. It is important to either complement the list with additional sampling to capture information on nonmembers or recognize this gap in describing and interpreting the results.

Other times, an available list may include more units than the target population. For example, suppose you want to survey likely voters, but the only information available is a list of registered voters, including some who rarely vote. In this case, one option is to contact all voters, ask each respondent if they voted in two of the past three elections, and proceed with the interview only if the answer is yes. Alternatively, the researcher could collect voting patterns and opinions from all registered voters and then examine the patterns for different definitions of *likely voter* in the analysis.

In some situations, no sampling frame is available. This is particularly common when the sampling unit is a specific type of household or business. For example, if a researcher wants to conduct a survey of bicycle repair shops, fortune tellers, or beekeepers in a place where these businesses are not registered, it may not be possible to obtain a complete list to serve as a sampling frame, even at the local level. In such a situation, the researcher must create a sampling frame.

One approach is to use area sampling. The researcher obtains a set of maps of local areas, such as counties or urban neighborhoods. Using maps of each area, the researcher divides it into smaller units of similar size. One common approach is to use a grid to divide the map into equal-sized squares. Another option (relevant for urban surveys) is to use city blocks as the smaller unit. In either case, the researcher selects a sample of the smaller units and then collects information from all the sampling units within the selected unit. For example, to implement a survey of small-scale food shops, the city is divided into 80 neighborhoods using a map, and 20 neighborhoods are selected. Each selected neighborhood is divided into blocks using a street map. The survey team then visits a randomly selected set of eight blocks in each neighborhood. Within each block, every small-scale food retailer is interviewed.

In the absence of maps and a sampling frame, it may be necessary to carry out a listing exercise, in which the survey team first prepares a list of the sampling units within a given area. The sampling units are then numbered, and a random selection is made for follow-up interviews. This can be a time-consuming process, so it is useful to define the area as small as possible given the information available.

2.3.3 Determining the Sample Size

How large should a survey sample be? Not surprisingly, it depends. To explain the factors that determine the minimum sample size, it is helpful to use an example. Suppose we are designing a survey to test whether there is a gender difference in the salaries of recent graduates from a college. Would it be enough to interview 70 graduates, or do we need a sample of 700? To answer this question, we need five pieces of information:

1. How small a difference in salaries do we want to be able to measure? In our example, if we want to detect a male–female salary difference as small as 3%, the sample size will have to be relatively large. If, on the other hand, we are satisfied with only being able to detect salary differences that are 20% or more, a smaller sample will suffice.
2. How much variation is there in salaries? If all the graduates have similar salaries, then we can estimate the mean (average) salary of men and women more precisely, so a small sample would be sufficient. If, on the other hand, there is a wide variation in salaries, then we would need a larger sample to achieve the same level of precision in the estimate.
3. How small do we want to make the probability of incorrectly concluding that there *is* a difference between the salaries of men and women? The larger the sample size, the smaller the risk of making

this type of error.
4. How small do we want to make the probability of incorrectly concluding that there is *no* difference between the salaries of men and women? Again, the larger the sample, the lower the risk.
5. How was the sample selected? The sample design influences the size of sample needed to reach a given level of precision.

If we have information (or at least educated assumptions) about these five factors, we can estimate the number of graduates that need to be interviewed in the survey. We will not describe the methods here because they make use of concepts taught in later chapters. However, a brief survey of the methods can be found in Appendix 9.

2.3.4 Sample Selection Methods

This section describes four types of sampling methods: (1) simple random sampling, (2) systematic random sampling, (3) multistage (or cluster) sampling, and (4) stratified random sampling. The Stata code to implement each of these methods is shown in Appendix 7, though it requires a solid understanding of Stata. We recommend studying Chapters 4 to 7 before reading Appendix 7.

2.3.4.1 Simple Random Sampling

Once we have the sampling frame, how do we select the sample? One approach is to select a ***simple random sample***, in which the entire sample is based on a draw from the sampling frame, where each sampling unit has an equal probability of being selected. The probability of selecting each unit is n/N, where n is the number of units to be selected and N is the total number of units in the sampling frame. One disadvantage of a simple random sample is that the selected units may be "clumped" together in the sample frame, resulting in a sample that is less representative than desired. To address this problem, researchers are more likely to use a systematic random sample, as discussed next.

2.3.4.2 Systematic Random Sampling

A ***systematic random sample*** is one in which there is a fixed interval between selected units. First, a unit is randomly selected from among the first N/n units in the sampling frame. Subsequently, units are selected every N/n units. For example, a systematic random sample of 20 households from a list of 200 households starts with a randomly selected unit from the first $N/n = 10$ units. Suppose the random selection picks unit 4. After that, we select every $N/n = 10$ units, that is 14, 24, 34, and so on up to 194. The main advantage is that it spreads out the selected units evenly across the sampling frame. If the sampling frame does not follow any order, this will not make a difference. But typically, the sampling frame is sorted by some characteristic, such as location or size. In this case, a systematic random sample will ensure that the selected units are balanced in terms of that characteristic. For example, if the sampling frame is sorted by location from north to south, then a simple random sample might include a disproportionate number of units in the north. However, a systematic random sample spreads out the sample so that the number of selected units in the north and south will be proportional to the actual number of units in the north and south.

2.3.4.3 Multistage Sampling

Multistage sampling refers to a selection process in which the selection occurs in two or more steps. (This is also called cluster sampling.) For example, suppose we are carrying out a national survey. The researcher may randomly select 10 of the 50 states, 5 counties in each state, and 100 households in each county, for a total sample of 5,000 households. This represents a three-stage random sample, corresponding to the three levels of selection: states, counties, and households.

There are several possible motivations for multistage sampling:

First, it may be used to overcome limitations on the availability of a full sampling frame. Often, it is not possible to use single-stage sampling because there is no sampling frame that covers the entire population of interest. In the case above, suppose the household lists are available only from county officials. It would be very expensive and time-consuming to gather lists from every county in the country to prepare a national sampling frame for a simple random sample. In contrast, it would be much easier to randomly select a subset of counties in the first and second stages and then get the list for each selected county for third-stage selection of households.

Second, it may be used to ensure that the sample is well distributed across certain categories. In the example above, the design ensures that the sample includes 10 states and 5 counties within each state.

Third, multistage sampling may be used to reduce the cost of data collection. Even if a national sampling frame is available, visiting 5,000 randomly selected households would be much more costly than visiting households in 50 counties.

2.3.4.4 Stratified Random Sampling

Stratification refers to dividing the population into categories (or **strata**) and specifying the sample size for each one rather than allowing the distribution to be determined by chance. The strata must not overlap each other, and they must cover the entire population. For example, national household surveys are often stratified into rural and urban areas, with a separate selection of households in each area. National surveys may be stratified by region as well. Surveys of enterprises are often stratified by size, specifying the number of small, medium, and large firms that will be included.

There are three reasons to design a stratified sample. First, stratification may be used to ensure that the sample for each stratum is large enough to allow reliable estimates at the stratum level. For example, suppose a country has six administrative regions, but one of them only has 2% of the national population. In an unstratified random sample of 1,200 households, roughly 2% of the sample (24 households) would be selected from the small region. If the sample is stratified by region, the researcher can ensure that each region has 200 households, which may be enough to generate reliable results for each region. In this case, stratification would be used to oversample the small region, meaning that the percentage of households sampled in the small region is larger than its share in the overall population. The other five regions would be undersampled in this process.

Second, stratification can be used to ensure that each stratum is proportionally represented in the sample. In this sense, stratification fulfills the same function as systematic sampling where the sampling frame is organized by stratum. If the strata are internally more homogeneous than the population, stratification will improve the precision of estimates when compared with a simple random sample.

A third reason for stratification is to adapt to differences in the variability of key indicators across strata. As discussed earlier and as we will discuss in Chapter 8, the precision of survey-based estimates in measuring a variable of interest is partly determined by the variability of the variable of interest. (In the extreme, if there were no variability and all units were the same, a sample of one would be sufficient!) For example, suppose a survey is designed to estimate national income. In general, the variability of income is greater in urban areas than in rural areas. Because of this, it is useful to oversample urban households, meaning that we select a larger share of urban households than rural households. Well-designed stratification can reduce the confidence interval in survey-based estimates without increasing the overall size of the sample.

2.4 SAMPLING WEIGHTS

Sampling weights are numbers used to estimate population **parameters** (e.g., means and percentages) from sample statistics, compensating for "distortions" that may be introduced by sampling. For example, suppose 90% of the population lives in rural areas, but the sample is stratified so that it is 50% urban and 50% rural. In this case, the average income in the sample will be disproportionately affected by urban households. If urban incomes are higher, the average income for the sample will be higher than the average income of the population. In other words, the average income from the sample is biased upward because it has a disproportionately large number of urban households. Using sampling weights, however, we can calculate the weighted average, which will give greater weight to each rural household and lesser weight to each urban household, providing an unbiased estimate of the average income among the population.

2.4.1 Calculating Sampling Weights

Sampling weights are calculated as the inverse of the probability of selection. They can also be interpreted as the number of units in the population that each unit in the sample represents.

In the case of simple random sampling or one-stage systematic random sampling, the probability of selecting any one unit is n/N, where n is the size of the sample and N is the size of the population. Thus, the sampling weight (w) is calculated as the inverse:

$$w = \frac{N}{n}$$

$$w = \frac{N}{n}$$

Note that the sampling weight is the same for all units. Such a sample is considered *self-weighted* because the sample average is equal to the weighted average and represents an unbiased estimate of the population average. In this case, the main use of sampling weights is to extrapolate from sample totals to population totals. For example, suppose a survey of seniors at a university collects information on 100 out of 2,000 seniors. The weight is 2,000/100 = 20, so each senior in the sample represents 20 in the senior class. The *average* spending on books in the sample is an unbiased estimate of the average spending in the population. But if you wanted to estimate the *total* spending on books by the senior class, you would just multiply the total for the sample by 20.

In the case of a single-stage stratified sample, we carry out the calculation for each stratum. The weight for stratum i (w_i) is calculated as follows:

$$w_i = \frac{N_i}{n_i}$$

$$w_i = \frac{N_i}{n_i}$$

where N_i is the population of stratum i and n_i is the sample size for stratum i. Taking the example of urban–rural stratification, suppose there are 900,000 rural households and 100,000 urban households in the population, and the sample contains 4,000 households divided equally between urban and rural areas. The weight for rural households would be 900,000/2,000 = 450, and the weight for urban households would be 100,000/2,000 = 50. In other words, each rural household in the sample represents 450 households in the rural population, while each urban household in the sample stands for just 50 in the urban population. Calculating weighted averages would give more weight to rural households in the sample, thus compensating for the fact that they were undersampled in the survey.

For multistage sampling designs, the calculation of the sampling weights is a little more complicated, but it follows the same general rule: the sampling weight at each stage is the inverse of the probability of selection. There is a separate ratio for each stage in the sampling. Consider the example of a three-stage random sample:

In the first stage, we select 10 of the 50 states.

In the second stage, we select 5 counties in each of the 10 selected states.

In the third stage, we select 100 households in each selected county.

The sampling weight for each county (w_c) is the product of three ratios, each representing the inverse of the probability of selection in that stage of selection:

$$w_c = \frac{50}{10}\frac{C_s}{5}\frac{H_c}{100}$$

(2.3)

$$w_c = \frac{50}{10}\frac{C_s}{5}\frac{H_c}{100}$$

where 50 is the total number of states, 10 is the number of states selected, C_s is the total number of counties in state *s*, 5 is the number of counties selected in each state, H_c is the total number of households in county *c*, and 100 is the number of households selected in each county.

This equation can be adapted to other multistage sample designs, keeping in mind the fact that the number of terms should be equal to the number of stages in the sampling. A simple way to double-check the calculation of the sample weights is to sum the sample weights over the units in the sample. The total should be roughly equal to the number of units in the population.

Up to this point, we have been discussing a type of weight called inverse probability sampling weights (IPSW). The other type of weight is relative sampling weights, defined as the IPSW for each unit divided by the average IPSW. As such, the average value of relative weights is always 1.0. For estimating weighted means and percentages of the population, relative weights and IPSW give the same results. However, relative weights cannot be used to estimate population totals, while IPSW can be used for this purpose.

2.4.2 Using Sampling Weights

How are the sampling weights used? Suppose our variable of interest in a national survey is household income. We can estimate national income as a weighted sum of household income across the sample using the following equation:

$$X = \sum_{i=1}^{n} x_i w_i$$

(2.4)

$$X = \sum_{i=1}^{n} x_i w_i$$

where X is the estimate of the total for the population (e.g., national income), x_i is the value of the variable for household i (e.g., household income), and w_i is the IPSW for household i. As a reminder, $\sum$ is the summation sign, so the right side of the equation means that we should take the sum of $x_i w_i$, as i goes from 1 to n. In other words, $X = x_1 w_1 + x_2 w_2 + x_3 w_3 + \ldots + x_n w_n$.

Estimates of population means can be calculated as the weighted average:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

$$x = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

The numerator is an estimate of the sum of x across the population, as shown in Equation 2.4. The denominator is the sum of the weights across the sample, which is an estimate of the total size of the population. Thus, the overall expression is an estimate of the average value of x across the population.

If x_i is a **binary variable** taking values of 0 or 1, then this equation gives an estimate of the proportion of the population for which $x_i = 1$. In the case of categorical variables, such as region or marital status, the average has no meaning, but the variable can be broken up into a set of binary variables, one for each category. Equation 2.5 can be used to estimate the proportion of the population in each category.

However, statistical packages, such as Stata, will do these calculations for us. In Chapter 6, we show how sampling weights can be used to adjust the calculations of totals, means, and percentages in Stata.

EXERCISES

1. Suppose you have a sampling frame of 1,200 hardware stores in a state, numbered from 1 to 1,200. Describe how you would select a systematic random sample of 100 stores for a survey. Give an example of what the sample might look like, showing the store numbers of the first five selected stores.
2. Give three possible reasons why one might want to use a multistage random sample rather than a single-stage random sample.
3. Describe the general circumstances under which it would be useful to apply area sampling to select units to interview. Give an example of a situation in which area sampling would be useful.

4. You have been hired to design a survey of political opinions in 10 swing states, but you need to have a large enough sample (say, 800 respondents per state) to generate reliable results for each state. What type of sampling method do you need to use?
5. Assuming you have a list of all households in each state and can use simple random sampling in each, how would you calculate the sampling weight for each household in the survey?
6. There are 20,000 people in the country of Wakanda. Most of the population (i.e., 17,500) live in urban areas and the rest live in rural areas. If you drew a stratified sample of 250 people from urban areas and 250 people from rural areas, what would be the sampling weights for urban and rural areas?
7. Suppose that we develop a multistage sampling design and choose five states (out of 50), three counties within each state, and 300 households in each county. In the state of Pennsylvania, where there are 67 counties, we randomly select the following three counties (see Table 2.2):

TABLE 2.2 ■ Population of Three Counties in Pennsylvania

County	Population
Montgomery	819,000
Bucks	630,000
Allegheny	1,200,000

County	Population
Montgomery	819,000
Bucks	630,000
Allegheny	1,200,000

Assuming there are three people per household, what is the sampling weight for the selected households in each of these three counties?

KEY TERMS

binary variable

confidence interval

error

estimate

margin of error

parameters

population

purposive sampling

random sampling

sample

simple random sample

strata

stratification

systematic random sample

unit of observation

Descriptions of Images and Figures

Back to Figure

The text is as follows:

If some political polls were truly accurate, Alf Landon would have been America's president during World War II, instead of FDR. Here's a look at an alternative universe of politics, as we examine the five biggest political poll blunders in U.S. history.

Alf Landon beats FDR in a landslide

The mother of all botched political polls was a 1936 Literary Digest straw poll survey that said GOP challenger Alf Landon would win in a landslide over the incumbent, Franklin Delano Roosevelt, with 57 percent of the vote.

The Literary Digest used national straw polls in 1920, 1924, 1928 and 1932, and it guessed the winner of each presidential election.

In 1936, a young rival pollster, George Gallup, made his own prediction before the magazine issued its poll; He said Literary Digest would get it all wrong, despite the Digest's decent track record in previous polls.

So was right? The Literary Digest disaster helped establish Gallup as the nation's pre-eminent pollster. The Digest polled about 2 million people, most of who were magazine readers, car owners or telephone customers—and had money during the Depression. It was not a representative sample.

Gallup used a random poll sample of 50,000 people.

President Roosevelt won the 1936 election easily, with 63 percent of the vote, and the Literary Digest was out of business the following year. If he had won, Landon could have been our wartime president.

3 QUESTIONNAIRE DESIGN

CHAPTER PREVIEW

Terms	Definitions and key points
Structured questionnaire	Fixed set of questions phrased in a standardized way and given in the same order for every respondent
Semi-structured questionnaire	Some standardized questions, but some part of the questionnaire is informal and flexible. Topics may vary from one respondent to another.
Open-ended question	A question that allows the respondent to answer in any manner, with the response being recorded in the form of a narrative text
Closed-ended question	A question for which the response can be expressed as a single number or categorical response
Question order	Order questions by topic. Move from general to specific questions. Place sensitive questions last.
Question phrasing	Be specific about who is referred to. Be specific about time frame. Be clear about definitions. Avoid leading questions.
Continuous responses	Responses that can take on any numerical value Need to specify the unit of measure
Categorical responses	Responses based on predefined representative categories
Skip patterns	Guidelines that indicate which questions should be skipped when questions don't apply to all respondents based on previous answers

Terms	Definitions and key points
Structured questionnaire	Fixed set of questions phrased in a standardized way and given in the same order for every respondent
Semi-structured questionnaire	Some standardized questions, but some part of the questionnaire is informal and flexible. Topics may vary from one respondent to another.
Open-ended question	A question that allows the respondent to answer in any manner, with the response being recorded in the form of a narrative text
Closed-ended question	A question for which the response can be expressed as a single number or categorical response
Question order	Order questions by topic. Move from general to specific questions. Place sensitive questions last.

Terms	Definitions and key points
Question phrasing	Be specific about who is referred to. Be specific about time frame. Be clear about definitions. Avoid leading questions.
Continuous responses	Responses that can take on any numerical value Need to specify the unit of measure
Categorical responses	Responses based on predefined representative categories
Skip patterns	Guidelines that indicate which questions should be skipped when questions don't apply to all respondents based on previous answers

3.1 INTRODUCTION

What questions should be included in the survey? How should the questions be phrased? And how should the responses be recorded? These are some of the important issues involved in designing the questionnaire, one of the most important and time-consuming steps in implementing a survey. If a key question is omitted from the questionnaire, important information will be lost. If too many questions are included in the questionnaire, respondents may tire and stop answering, leading again to the loss of information. And if a question is poorly phrased, the results may be difficult or impossible to interpret.

This chapter provides some guidelines for the design of questionnaires, including question order, phrasing, and response codes. For additional information on questionnaire design, Grosh and Glewwe (2000) have edited a volume with detailed information on questionnaire design for developing countries. Ekinci (2015) provides a more concise review, focused on business and management research. Rea and Parker (2005) provide another valuable reference on the issues of questionnaire design.

3.2 TYPES OF QUESTIONNAIRES

Survey questionnaires can be categorized according to the type of interview, based on whether or not they are structured or not, and by the types of questions. The guidelines for designing a good questionnaire depend on what type it is.

3.2.1 Type of Interview

As shown in Table 3.1 below, interviews may be carried out online, by mail, by telephone, or with a face-to-face interview. One of the main challenges in all these approaches is getting cooperation from respondents. People are busy, and they may tune out research surveys because of the large number of commercial "surveys" designed to sell a product or solicit donations. Another challenge is to ensure that the respondents are representative of the larger population. As described later, a biased or unrepresentative sample can give highly misleading results.

TABLE 3.1 ■ Types Of Interviews

Type of interview	How responses are recorded	Comments
Online form	Respondent is sent a link and responds to questions on a web page	Questions must be carefully phrased because there is no enumerator to explain if the respondent is confused.
Mail-in questionnaire	Respondent receives a questionnaire in the mail and submits responses by mail	Many recipients will not complete the questionnaire, raising questions about the representativeness of those who do
Phone interviews	Respondent receives a phone call and answers questions over the phone	Many recipients will not answer or be unwilling to answer questions. Only suitable for short interviews.
Face-to-face interview	Respondents are contacted at home or in a public place for the interview	Home interviews are less common than they used to be. Public interviews need to be short.

Type of interview	How responses are recorded	Comments
Online form	Respondent is sent a link and responds to questions on a web page	Questions must be carefully phrased because there is no enumerator to explain if the respondent is confused.
Mail-in questionnaire	Respondent receives a questionnaire in the mail and submits responses by mail	Many recipients will not complete the questionnaire, raising questions about the representativeness of those who do
Phone interviews	Respondent receives a phone call and answers questions over the phone	Many recipients will not answer or be unwilling to answer questions. Only suitable for short interviews.
Face-to-face interview	Respondents are contacted at home or in a public place for the interview	Home interviews are less common than they used to be. Public interviews need to be short.

3.2.2 Structured and Semi-structured Questionnaires

An important distinction in questionnaire design is between structured and semi-structured questionnaires. A *structured questionnaire* has a fixed set of questions, phrased in a standardized way, and given in the same order for every respondent. Some questions may be skipped for certain types of respondents; for example, a question about the respondent's spouse would be skipped for respondents who are single. However, the same rules about skipping questions apply to all respondents. Furthermore, the questions in structured questionnaires are generally designed so that the responses are either a continuous variable or a categorical variable, rather than open-ended questions with narrative responses.

In a *semi-structured questionnaire*, some of the questions are standardized and asked in a specific order, but part of the questionnaire is more informal and flexible, with questions and topics of discussion that vary from one respondent to another. The unstructured section of the questionnaire may consist of a list of suggested questions or just topics of discussion. The questions in this section are often open-ended, and the order is flexible. This portion of the interview is more journalistic in nature, where new questions are formed in response to answers to the previous questions.

The entire interview may be informal and unstructured. However, the list of suggested questions and topics for discussion is not normally considered a questionnaire, so it is outside the scope of this chapter.

The results of unstructured interviews (and the unstructured portion of semi-structured interviews) are difficult to analyze in a systematic way for several reasons. If the questions are not standardized across respondents, then the sample varies across questions, making it difficult to summarize. Responses to open-ended questions can be summarized qualitatively, but statistical analysis requires either time-consuming classification of responses or complex computer algorithms for analysis of text. For this reason, unstructured and semi-structured surveys generally use small samples.

On the other hand, unstructured interviews provide rich information on the perceptions, beliefs, and motivations of respondents. They may uncover issues or patterns that the researcher did not anticipate at the beginning of the study. Unstructured interviews can be used to identify key issues in preparation for the design of a structured questionnaire for a large-scale formal survey. In addition, unstructured interviews can be used after a formal survey to help interpret or explain the results of the survey. Because our main interest is generating data for analysis, this chapter focuses primarily on the design of structured questionnaires. Table 3.2 summarizes some of the characteristics of each type of survey.

TABLE 3.2 ■ Characteristics Of Structured, Semi-Structured, And Unstructured Surveys

	Structured Surveys	Semi-structured Surveys	Unstructured Surveys
Types of questions	Mainly closed	Open and closed	Mainly open
Phrasing, order, and content of questions	Standardized for all respondents	Partly standardized	Varies across respondents
Sample size	Can be large	Usually small	Usually small
Type of results	Quantitative	Quantitative and qualitative	Qualitative
Strengths	Numerical results; can generate unbiased estimates of population parameters with confidence intervals	Mix of both	May reveal new issues or unexpected responses; questions adapt to earlier answers
Weaknesses	Questions and responses are fixed before the survey begins	Mix of both	Qualitative results, only practical for small sample

	Structured Surveys	**Semi-structured Surveys**	**Unstructured Surveys**
Types of questions	Mainly closed	Open and closed	Mainly open
Phrasing, order, and content of questions	Standardized for all respondents	Partly standardized	Varies across respondents
Sample size	Can be large	Usually small	Usually small
Type of results	Quantitative	Quantitative and qualitative	Qualitative

	Structured Surveys	Semi-structured Surveys	Unstructured Surveys
Strengths	Numerical results; can generate unbiased estimates of population parameters with confidence intervals	Mix of both	May reveal new issues or unexpected responses; questions adapt to earlier answers
Weaknesses	Questions and responses are fixed before the survey begins	Mix of both	Qualitative results, only practical for small sample

3.2.3 Types of Questions

Although we referred to open-ended questions above, it is useful to define the term more precisely. An ***open-ended question*** (or open question) is one that allows the respondent to answer in any manner, with the response being recorded in the form of narrative text or summary notes. Examples of open-ended questions include, “What is your view of gun control legislation?” or “Why do you think some people succeed and others do not?”

In contrast, a ***closed-ended question*** (sometimes called a closed question) is one for which the respondent either gives a number or selects from a set of predetermined responses. Closed questions can be divided into two categories depending on the type of response.

1. A closed question may generate a continuous variable, representing a measurement of a physical quantity such as weight, length, time duration, or frequency. Examples include, “How old are you?” and “How many hours a week do you watch television?”
2. A closed question may generate a categorical variable. This includes yes/no questions, such as, “Are you married?” and “Do you own a car?” It also includes multiple-choice questions, such as, “What is your education level?” and “What is your marital status?” where the respondent chooses among several options. In each case, the response is verbal, but it is coded in the database as a number.

As discussed earlier, most medium- and large-scale surveys use structured questionnaires with closed-ended questions. For this reason, we focus on this type of questionnaire for the remainder of the chapter.

3.3 GUIDELINES FOR QUESTIONNAIRE DESIGN

3.3.1 General Guidelines

Before we discuss the specifics of designing the questionnaire, it is useful to list a number of general guidelines to make the questionnaire clear and easy for the interviewer to use and for the respondent to understand.

> Whether the interview is in-person, online, or by mail, all questions should be written out in full to reduce ambiguity and ensure that the question is asked in the same way to all respondents.
>
> For in-person and phone interviews, the instructions to interviewers (also called **enumerators**) should be clearly distinguished from the questions themselves. This can be done by using a different font or putting the instructions in brackets.

For all types of questionnaires, each categorical response option should be written out in full. This helps standardize the way the questions are asked.

In paper questionnaires, it is preferable for the enumerator to record the number code of the response rather than circling or marking the response on the list. In addition, it is a good idea to use boxes to indicate where each response code should be written. This will reduce the time and increase the accuracy of data entry by making it easy to find and enter the response code in the computer.

Computer-assisted personal interview (CAPI) methods are becoming widespread. CAPI software is available to program computers, tablets, or phones to record data in the field (e.g., ODK, SurveyCTO, and Surveybe). In addition to eliminating the time and errors associated with entering data into a computer from paper questionnaires, this approach allows the researcher to incorporate quality checks into the tablet program, flagging errors and allowing the enumerator to correct them during the interview.

There are many software packages that will allow you to design an online questionnaire, and some of them are free. SurveyMonkey, for example, is free for small-scale surveys, though there is a fee if the questionnaire or the sample is large. There are many similar packages that are easy to learn and have the ability to create many types of questions, including multiple choice, rank order, slider, and tables. These packages will also allow you to print a copy of the questionnaire if you plan to do an in-person interview where an enumerator fills in the questionnaire.

The research questions of the study help determine the range of questions to be included in the survey. The questions should, of course, address the central research questions, but they should also include questions to help explain the responses to the main questions. For example, political opinion polls naturally focus on respondents' support for different candidates, but they also ask questions about the respondents' age, sex, education, and party affiliation because these characteristics often help "explain" political preferences. Similarly, a survey of college students regarding the time spent on sports could include questions about the student, including sex, age, scholarship status, high school experience with sports, and so on.

3.3.2 Question Order

The order of the questions should follow four general guidelines. First, the questions should follow an order determined by the topics, moving from one topic to the next. For example, group questions about education together before moving on to health. Whenever possible, the questionnaire should avoid returning to a topic covered earlier. This keeps the interview as close to a "natural" conversation as possible for in-person interviews. In addition, it probably reduces frustration among respondents that might be caused by going back to an earlier topic.

Second, within each topic the questionnaire should start with general questions before moving on to specific questions. The general questions will help determine which specific questions should be asked. For example, if a general question determines that the respondent does not drink alcohol, one can avoid specific questions about how much they drink. Similarly, a general question whether the respondent has children should precede questions about those children.

Third, it is better to start the interview with topics that are not sensitive, such as household composition. More sensitive topics, such as income level or use of contraception, should be asked toward the end of the interview. The respondent will probably feel more comfortable discussing sensitive topics after spending some time with the enumerator. In addition, if the sensitive topics cause the respondent to break off the interview or stop filling out the questionnaire, less information will be lost if these questions are asked toward the end of the interview.

Finally, it is preferable to have the questions most important for the intended analysis toward the beginning of the questionnaire. This way, if the interview cannot be completed, at least the essential questions will have been covered. Clearly, these guidelines may conflict with one another. This highlights the importance of testing the questionnaire one or more times before finalizing the wording, question order, and response options.

3.3.3 Phrasing the Questions

In designing the questions, it is important to make sure they are clear and unambiguous. This means avoiding research jargon or other vocabulary that might not be familiar to some of the respondents. The questions should also avoid abbreviations and acronyms unless they are universally understood. Finally, the questions need to be specific about *who* they refer to. In English, "you" can mean you (singular), referring to the respondent himself or herself, or it can mean you (plural), referring to the respondent's family. Take the following question:

"Have you taken out a loan?"

It is not clear *who* the question refers to, the respondent alone or anyone in the household. If the researcher is interested in access to credit by the household, a better phrasing would be as follows:

"Have you or anyone in your household taken out a loan?"

The questions also need to be explicit about *when*—that is, the time period referred to. In the question above, it is not clear if the respondent should include a loan received many years ago as a college student. Thus, the question above would be better phrased as follows:

"Have you or anyone in your household taken out a loan in the past 12 months?"

Note that "in the past year" is ambiguous because it could mean over the past 12 months or during the current calendar year. For this reason, "in the past 12 months" or "since this time last year" is better.

Finally, the questions should be explicit about *what* they are referring to. In the example above, what is the definition of a *loan*? Should it include $20 borrowed from a friend, or is it limited to official bank loans? If the latter, do we include loans from credit cooperatives and other nonbank financial institutions? To remove the ambiguity, the question could be rephrased as follows:

"Have you or anyone in your household taken out a loan from a bank or other financial institution in the past 12 months?"

Another important factor in phrasing questions is to avoid making any assumption about the respondent that has not been verified in a previous question. Table 3.3 gives some examples of questions that make assumptions about the respondent that may or may not be true.

TABLE 3.3 ■ Questions With Embedded Assumptions

Question	Implicit Assumption
How old is your oldest child?	The respondent has at least one child.
How much do you pay in rent?	The respondent rents his or her housing.
What is your favorite radio station?	The respondent listens to the radio.
Which state were you born in?	The respondent was born in the United States.

Question	Implicit Assumption
How old is your oldest child?	The respondent has at least one child.

Question	Implicit Assumption
How much do you pay in rent?	The respondent rents his or her housing.
What is your favorite radio station?	The respondent listens to the radio.
Which state were you born in?	The respondent was born in the United States.

One way to address this problem would be to include a response option for the excluded answer, such as, “I don’t have a child,” or “I don’t pay rent.” A better approach, however, is to add a prior question that determines whether this is a valid question. For example, see Table 3.4:

TABLE 3.4 ■ Example of a Filter Question

No.	Questions	Response Codes or Units	Response
A1	Do you have any children?	1. Yes 2. No	If no, skip to A3
A2	How many children do you have?	Number	

No.	Questions	Response Codes or Units	Response
A1	Do you have any children?	1. Yes 2. No	If no, skip to A3
A2	How many children do you have?	Number	

For in-person and phone interviews, it is important to provide clear instructions on which questions to skip based on responses to earlier questions. For online surveys and CAPI-based interviews, the skip patterns should be programmed with the software. The topic of **skip patterns** is discussed in Section 3.7.

It is also important to avoid “double-barreled” questions, meaning questions that may have two (or more) responses because the wording of the question combines multiple issues. Examples include the following:

> “Do you believe that supermarkets should sell cheaper and more nutritious food?”
>
> “Do you think the county government should spend less on salaries and more on roads?”
>
> “How often do you purchase gasoline, and how much do you spend?”

The solution is to separate the individual queries into two or more questions, so that respondents are not forced to answer two questions with one response.

Finally, the researcher should also ensure that the questions are neutral and do not “lead” the respondent to answer in a certain way. Table 3.5 provides some examples of questions that clearly express a point of view on the topic and “lead” respondents to adopt the same view. To nudge respondents one way or another, **leading questions** use terms and concepts with positive associations (e.g., “family business” and “protect people”) or ones with negative associations (e.g., “runaway spending” and “pork-barrel projects”). Some of them include reasons for supporting or opposing the statement within the question.

TABLE 3.5 ■ Examples Of Leading Questions

Topic	Leading Question Toward a "Yes" Response	Leading Question Biased Toward a "No" Response
Welfare	Do you feel the government has a moral responsibility to assist families who are in need through no fault of their own?	Do you support the use of your hard-earned tax dollars to hand out welfare checks to people?
Infrastructure spending	Do you agree that the government should be investing more in our crumbling infrastructure to promote economic growth?	Do you think the government should indulge in runaway spending on pork-barrel projects that could worsen the fiscal deficit?
City health code	Do you support the new city law that would strengthen the health code and protect people from unsanitary conditions in restaurants?	Do you support the new city regulations on restaurants that impose unnecessary costs on family businesses and threaten food service jobs?

Topic	Leading Question Toward a "Yes" Response	Leading Question Biased Toward a "No" Response
Welfare	Do you feel the government has a moral responsibility to assist families who are in need through no fault of their own?	Do you support the use of your hard-earned tax dollars to hand out welfare checks to people?
Infrastructure spending	Do you agree that the government should be investing more in our crumbling infrastructure to promote economic growth?	Do you think the government should indulge in runaway spending on pork-barrel projects that could worsen the fiscal deficit?
City health code	Do you support the new city law that would strengthen the health code and protect people from unsanitary conditions in restaurants?	Do you support the new city regulations on restaurants that impose unnecessary costs on family businesses and threaten food service jobs?

The above examples are heavily biased to demonstrate the effect, but in actual questionnaires, the bias may be less obvious. One method for testing for bias is to have someone read over the question and guess which response the researcher would give for that question. If the wording provides clues to the researcher's own views, the question should be revised.

Researchers designing questionnaires should also be aware of *social desirability bias*, which refers to the tendency of respondents to give answers that are socially acceptable rather than accurate. Questions about whether the respondent has voted may overestimate the proportion of adults who vote because people may be reluctant to admit that they did not vote. Likewise, questions about domestic violence, illegal drug use, or cruelty to animals are likely to underestimate their prevalence. Questions should be phrased to make respondents comfortable enough to admit the truth. In addition, the results should be interpreted with a recognition that the responses may overestimate socially desirable responses.

3.4 RECORDING RESPONSES

As mentioned above, closed questions can yield two types of responses: (1) a continuous variable or (2) a categorical variable. The methods for capturing information from each type of question are described below.

3.4.1 Responses in the Form of Continuous Variables

A continuous variable describes a quantity of something and requires a unit, such as kilograms or years. Examples of questions leading to a continuous variable are as follows:

How old are you (age at last birthday)?

How tall is your child, expressed in centimeters?

What is the area of your farm in hectares?

How much gasoline do you purchase each week, expressed in gallons?

How much do you spend per month on mobile phone service?

For continuous variables, it is necessary to gather information on the unit of measure (e.g., centimeters, hectares, gallons). This can be done within the question itself, when it is appropriate to assume that all responses can be given in the same unit. Table 3.6 gives an example:

TABLE 3.6 ■ Example of a Question with a Fixed Unit of Measure

No.	Questions	Response Codes or Units	Response
B1	How many times per month do you go out to the movies?	Times/month	

No.	Questions	Response Codes or Units	Response
B1	How many times per month do you go out to the movies?	Times/month	

In other cases, respondents may use different units. In this situation, it is better to convert to a standard unit in the data analysis phase than to ask enumerators or respondents to do calculations in their heads. In this case, the unit of measure is entered as a separate variable, as shown in Table 3.7.

TABLE 3.7 ■ Example of a Question with a Flexible Unit of Measure

No.	Questions	Response codes or Units	Response
B2	How frequently do you go out to the movies?	Number of times	
		1. Per week 2. Per month 3. Per year	

No.	Questions	Response codes or Units	Response
B2	How frequently do you go out to the movies?	Number of times	
		1. Per week 2. Per month 3. Per year	

In addition, for any questions that involve flows, the unit of time should be specified or the respondent should be allowed to select the time unit. For example, income, spending, driving habits, and frequency of exercise all have a time dimension, which needs to be captured in the questionnaire.

3.4.2 Responses in the Form of Categorical Variables

For categorical questions, most large-scale formal surveys use precoded response options, meaning that the possible responses to each question are specified before implementing the survey. There are three advantages of precoding:

1. During the interview, the response can be recorded quickly by checking a box or writing a number, rather than writing the full response in words.
2. After the survey, there is no need to examine all the answers and classify them into groups, a time-consuming process.
3. Finally, it avoids the situation where a response is ambiguous or covers two different response options.

The disadvantage of precoding responses is that the response options must be carefully selected to cover all likely responses. It may be useful to include an "other" option and, in some cases, allow the respondent to specify in words a response that is not listed.

In preparing the response codes for categorical questions, it is important that the response options be both mutually exclusive and exhaustive. The responses should be *mutually exclusive* in the sense of not overlapping with each other. For example, a respondent might be both divorced and a widow. It is important to include instructions on how to handle difficult cases, either in the questionnaire for online surveys or in the training for in-person interviews. For example, the enumerators could be instructed to select the first option that applies. Or the question could indicate that the respondent should "check all that apply."

In addition, the response options should be *exhaustive*, covering all—or at least the vast majority of—cases. The use of "other" ensures that the response options are exhaustive, but ideally "other" should represent only a small share of the responses—say, less than 5% of the total.

Closed questions can be used to collect information on opinions. One approach is to give a statement to the respondent and ask about the degree of agreement (Table 3.8). For example, political polls often use a 5-point **Likert scale**.

TABLE 3.8 ■ Example of a Question using the Likert Scale

No.	Questions	Response codes or Units	Response
B3	My representative in Congress stands for the interests of people like me.	1. Strongly agree 2. Agree 3. Neither agree nor disagree 4. Disagree 5. Strongly disagree 98. Do not know	

No.	Questions	Response codes or Units	Response

No.	Questions	Response codes or Units	Response
B3	My representative in Congress stands for the interests of people like me.	1. Strongly agree 2. Agree 3. Neither agree nor disagree 4. Disagree 5. Strongly disagree 6. Do not know	

Closed questions can even be used to address “why” questions if the researcher has a good idea of the most common responses. In this case, it may be useful to include “Other” as a response option. If “Other” is likely to be a common response, the researcher may wish to allow the respondent to specify what the “Other” response is. For example, see Table 3.9:

TABLE 3.9 ■ Example of a Pre-coded "Why" Question

No.	Questions	Response codes or Units	Response
B4	Why do you rent your housing rather than buying an apartment or a house?	1. I cannot afford the down payment 2. I don't think I would be able to get a mortgage 3. I don't plan to live in this area very long 4. I prefer not to make a major financial commitment 5. Other (specify)	

No.	Questions	Response codes or Units	Response
B4	Why do you rent your housing rather than buying an apartment or a house?	1. I cannot afford the down payment 2. I don’t think I would be able to get a mortgage 3. I don’t plan to live in this area very long 4. I prefer not to make a major financial commitment 5. Other (specify)	

3.5 SKIP PATTERNS

Skip patterns are guidelines in the questionnaire to tell the enumerator which questions should be skipped over based on the responses to earlier questions. It is important that these be clearly specified in the questionnaire to ensure consistency in the way questions are asked from one respondent to the next. When asking about members of the household, many questions are age-specific. For example, questions about school attendance are not appropriate for infants, while questions about occupation and marital status only make sense for adults. Rather than basing the skip patterns on vague terms such as *infant* and *adult*, the questionnaire should specify the appropriate age range for each question.

Skip pattern instructions, like other instructions to the enumerator, should be distinguished from the wording of questions. For example, instructions to the enumerator may be put in italics or in brackets to distinguish them.

The skip patterns can get complicated, particularly if there are multiple branches that the interview could take. For example, a set of questions about housing can take three paths depending on the answers to the first and third questions (See Table 3.10).

TABLE 3.10 ■ Example of Questions with Skip Patterns

No.	Questions	Response Codes or Units	Response	Skip Instructions
B5	Do you own or rent your housing?	1. Own 2. Rent		If "Own," skip to B7
B6	How much do you spend on rent each month?	$/month		Skip to B9
B7	Do you have a mortgage?	1. Yes 2. No		If "No," Skip to B9
B8	How much do you pay per month for your mortgage?	$/month		

No.	Questions	Response Codes or Units	Response	Skip Instructions
B5	Do you own or rent your housing?	1. Own 2. Rent		If “Own,” skip to B7
B6	How much do you spend on rent each month?	$/month		Skip to B9
B7	Do you have a mortgage?	1. Yes 2. No		If “No,” Skip to B9
B8	How much do you pay per month for your mortgage?	$/month		

Clearly specifying the skip pattern is important whether the responses are being recorded on a paper questionnaire or on a tablet. With paper questionnaires, the skip pattern must be included in the questionnaire so that the enumerator can clearly see which questions should be asked next, based on the previous answer. This is a common source of error in implementing paper-based questionnaires, so it is worth emphasizing the skip patterns in training the enumerators.

With tablet-based questionnaires, the skip patterns need to be incorporated into the program with a series of if–then commands so that the enumerator is automatically guided to the correct question, depending on the responses to previous questions. One of the important advantages of tablet-based questionnaires is that, by incorporating the skip patterns into the program, data collection errors are greatly reduced.

The skip patterns have implications for the analysis of the data. Questions that are skipped over in the interview will be recorded as missing values in the data. In the earlier example, the respondents are first asked whether they have any children and, if the response is “yes,” then asked how many. In this case, the variable for number of children will be missing (rather than zero) if there are no children. Calculating

the average of this variable will give the average among those respondents with children. If the researcher wants the average number of children including the zeros, the missing values will need to be replaced with zeros.

3.6 ETHICAL ISSUES

The formal review of ethics in research was prompted by a number of cases of extreme abuse of research participants, most notably the Tuskegee Syphilis Study (1932–1972). Congress passed the National Research Act of 1974, which led to the Belmont Report, outlining issues and guidelines for the use of human subjects (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Subsequent regulations require the establishment of institutional review boards (IRBs) to review and approve (or reject) research plans to protect the rights and interests of human subjects. IRBs are certified and regulated by the Office for Human Research Protections of the Department of Health and Human Services. Almost all universities, hospitals, and research institutes in the United States have created IRBs.

Biomedical research is strictly controlled to ensure that the risks associated with testing new drugs and treatments are understood by the participants and that the potential benefits outweigh the risks. Medical researchers must apply for and obtain approval for each study from their IRB. The regulations are not as tight on surveys and other forms of social science research, but approval from an IRB is required if the research involves human subjects. Even without the risks associated with new drugs or treatments, respondents are offering their time to answer questions, some of which may be on sensitive topics.

IRB approval is based on three broad criteria. First, it is necessary that respondents or participants give informed consent to participate in the study. *Informed consent* means that the respondents must give prior approval for their participation after receiving information about the study and the nature of their involvement. This typically takes the form of a paragraph explaining to the respondents about who is carrying out the survey, the goals of the survey, and any risks or benefits of participation.

Second, participants must be assured of *confidentiality*, meaning that the results of the survey will be presented in aggregate form so that the responses of individuals cannot be identified. Furthermore, if the data are shared with other researchers, any variables that allow the identification of individual respondents (e.g., names, addresses, phone numbers, or GPS [Global Positioning System] coordinates) will be removed from the data set.

Third, the IRB approval depends on an assessment that the *costs and risks to the participants are justified* by some benefit to the public at large. Although somewhat subjective, this criterion ensures that the research is worthwhile, taking into account any cost or inconveniences to the participants. There are special protections for vulnerable groups, including racial minorities, very ill people, children, and prisoners.

Additional information on IRBs is available from Qiao (2018) and Protection of Human Subjects (2009).

EXERCISES

1. Which of the following questions generate responses that are continuous variable responses and which ones generate responses that are categorical variables?
    a. How old are you?
    b. What is the highest educational degree you have completed?
    c. How many years of education do you have?
    d. What state were you born in?
    e. Are both of your parents still alive?
    f. How many times per week do you exercise?

g. What is your weight in pounds?

2. Identify the hidden assumption(s) or the flaw(s) in the following questions.
    a. How much do you earn at your job?
    b. What is the age of your oldest child?
    c. In light of his ineffectiveness, do you agree that the governor should not be reelected?
    d. How old (in years) is your car?
    e. How frequently do you go shopping?
    f. Do you think the town firefighters should be full-time workers and paid more?
3. What are skip patterns in a questionnaire, and what purpose do they serve?
4. Give an example of social desirability bias and how it might affect the accuracy of results in a survey.
5. How would you use skip patterns to design questions to gather information about the car payments of a sample of respondents?
6. What are the three main principles for research on human subjects used by IRBs to approve research? Given an example of a violation of each principle.

KEY TERMS

closed-ended question

enumerators

leading questions

Likert scale

open-ended question

skip patterns

PART II DESCRIBING DATA

4 AN INTRODUCTION TO STATA

CHAPTER PREVIEW

Stata Basics	Specifics
Opening Stata	Click on Start and search for Stata
	Double-click on any file generated with Stata
	Double-click on the Stata icon
Stata windows	Results
	Review
	Command
	Variables
	Properties
Working with existing data	Command window
	Menus
	Do-files
Entering your own data into Stata	Entering data
	Renaming variables
	Creating variable labels
	Creating value labels
Using log files and saving your work	Opening and closing a log file
	Copying output to a word processor
	Saving changes to your data
Getting help	Help command
	Search command
	Stata website
	Using the Google Search Engine
Stata command examples	**tab mditate1**
	rename var1 hand
	label define handlabel 1 "Right" 2 "Left"
	label value hand handlabel
	log using "c:\gunlaw", **text** (on a PC)
	log using "c:/gunlaw", **text** on a Mac)
	save "c:\filename.dta" (on a PC)
	save "c:/filename.dta" (on a Mac)

Stata Basics	Specifics
Opening Stata	Click on Start and search for Stata Double-click on any file generated with Stata Double-click on the Stata icon

Stata Basics	Specifics
Stata windows	Results Review Command Variables Properties
Working with existing data	Command window Menus Do-files
Entering your own data into Stata	Entering data Renaming variables Creating variable labels Creating value labels
Using log files and saving your work	Opening and closing a log file Copying output to a word processor Saving changes to your data
Getting help	Help command Search command Stata website Using the Google Search Engine

Stata Basics	Specifics
Stata command examples	**tab mditate1** **rename** var1 hand **label define** handlabel 1 “Right” 2 “Left” **label value** hand handlabel **log using** “c:\gunlaw”, **text** (on a PC) **log using** “c:/gunlaw”, **text** on a Mac) **save** “c:\filename.dta” (on a PC) **save** “c:/filename.dta” (on a Mac)

4.1 INTRODUCTION

Stata is a powerful statistical software package that is relatively easy to learn. As described in the preface, it has been growing rapidly in popularity and is used almost exclusively in some fields. It is particularly popular in the fields of biomedicine, epidemiology, economics, political science, psychology, and sociology. Learning data analysis with Stata will provide you with a distinct, marketable skill.

In this chapter, we will learn the basics of Stata and move on to more advanced skills in later chapters, where we will use Stata to examine descriptive statistics and test hypotheses. By completing the examples and exercises in this book, you will have a basic knowledge of Stata that you can build on as you develop more advanced statistical skills through further study or use.

4.2 OPENING STATA AND STATA WINDOWS

You can open Stata by clicking on Start and then searching for the Stata program. If the Stata icon is on your desktop, you can click on the Stata icon. You can also click on any file created with Stata to open the package. We will begin by double-clicking on the GSS2021.dta file and examining Stata’s five main windows.

Figure 4.1 shows the opening screen. This screen will be the same on both a PC and a Mac computer. There are a few minor differences, however, when using a PC or a Mac. For example, Ctrl + D on a PC is replaced by Command + Shift + D on a Mac. These differences are noted in the chapter when necessary.

Description

Figure 4.1 Opening Stata Screen

Like most software packages, the top row offers a set of menus followed by a second row of icons for functions that are used most frequently. In addition to these standard features, there are five windows that appear: (1) the Results Window (which is the largest window in the center without a label), (2) the History Window on the left, (3) the Command Window at the bottom, (4) the Variables Window on the upper right side, and (5) the Properties Window on the lower right side.

4.2.1 Results Window

Once you start using Stata to analyze data, all of your recent commands, output, and error messages will appear in the Results Window in the center of your screen. The slide bar or scroll bar on the right side can be used to look at earlier results that are not on the screen. However, the Results Window does not keep all of the output generated. By default, it will keep about 500 lines of the most recent output and delete any earlier output. If you want to store output in a file, you must use a **log file**, which is described in more detail later.

4.2.2 History Window

This History Window on the left lists all the recent commands. If you click on one of the commands, it will be copied to the Command Window at the bottom of the screen, where it can be executed by pressing the “Enter” key. Or you can modify the command first and then run the command. If you double-click on the command, it will be directly re-executed by Stata.

4.2.3 Command Window

This Command Window at the bottom of the screen allows you to enter commands that will be executed as soon as you press the “Enter” key. You can also use recent commands again by using the “Page Up” key (to go to the previous command that appears in the History Window) and “Page Down” key (to go to the next command). If you double-click on a variable in the Variables Window, it will appear in the Command Window.

4.2.4 Variables Window

The Variables Window on the upper-right side of your screen lists all the variables in the data set that is open. You can increase the size of this window to see the variable names and their variable labels. If you create new variables, they will be added to the list of variables. If you delete variables, they will be removed from the list. You can insert a variable into the Command Window by double-clicking on it in the Variables Window.

4.2.5 Properties Window

The Properties Window on the lower right side provides information about the variables in the open data set. If you click once on any variable in the Variables Window, the Properties Window will give you information about that variable, such as the name, label, and type of variable, along with information about the data set.

4.3 WORKING WITH EXISTING DATA

Let's begin by using the **General Social Survey** data set from 2021 (GSS2021) that we already opened above. This is a data set that explores attitudes, behaviors, and demographic information about people living in the United States. It has been collected almost every year since 1972. Because we will use the survey from just one year, 2021, it is called a cross-section data set. This means it looks at a cross-section of responses at one point in time. Every row represents the response of one individual, and all responses are from the same point in time. If instead, we followed the inflation rate, interest rates, and money supply in one country over 30 years, it would be a time series data set, since it represents data or information over time. In that case, each row represents a different year. Finally, a panel data set combines both cross-section data and time series data. For example, if you followed 100 patients after surgery for 10 years and measured their progress, you would be using cross-section data (100 patients in 1 year) and time series data (each patient's results every year over 10 years). In this case, each row represents one cross-section unit (a patient) and one time period (a year).

Let's suppose that we want to find out what proportion of the population meditates. This is called a categorical variable since it has seven categories or possible responses. Continuous variables, on the other hand, are variables that take on a specific value, such as someone's exact age or income. Types of variables and their measurement are discussed in more detail in Chapter 6.

There are three ways to obtain information on what proportion of the population meditates using Stata: (1) the Command Window, (2) menus, and (3) **do-files**.

In the Command Window at the bottom of the screen, we would type in **tab mditate1** and press "Enter." You can also type **tab mdi** and then the "Tab" key. This will fill in the rest of the variable name automatically. The information in Figure 4.2 would then appear in our Results Window. Notice that our command **tab mditate1** also appears in the History Window located to the left of the Results Window. If we double-click on the command in the History Window, the command will be executed again. If we click only once on the command in the Review Window, it will appear in the Command Window.

```
. tab mditate1
```

HOW OFTEN DO YOU MEDITATE?	Freq.	Percent	Cum.
at least once a day	336	9.40	9.40
almost every day	402	11.25	20.65
once or twice a week	461	12.90	33.55
once or twice a month	343	9.60	43.14
a few times per year	374	10.46	53.61
once a year or less	204	5.71	59.32
never	1,454	40.68	100.00
Total	3,574	100.00	

Description

Figure 4.2 Frequency Table Of Meditation Use

In the Results Window, you will notice that the command **tab mditate1** is shown above the output. When we typed the command into the Command Window, we shortened tabulate to **tab**. Stata accepts abbreviations for commands, but in some cases, such as the **table** and **tabulate** commands, **table** must be spelled out completely. Otherwise, **tab** is mistaken for **tabulate**. If you are unsure, you can use the help file to look up a command, and it will underline that portion of the command that must be typed. More information about the help file is in Section 4.6.

The second method with which to interact with Stata is by using menus. Although we will describe how to use menus to generate statistics throughout the book, we do not encourage the use of menus. Instead, we encourage students to use *do-files*, which are described later, along with their benefits. To generate the same output as in Output 4.1 using menus, we would click on the sequence listed below that would bring us to a dialog box. In this box, we would select the variable “mditate1” in the drop-down menu under “categorical variable” and then click on “OK.”

Statistics → Summaries, tables, and tests → Frequency tables → One-way table

---

Finally, the third way to interact with Stata is through the use of do-files. A do-file is a file where you type commands or code rather than using menus. By using do-files, you can save, revise, and rerun commands. This is particularly helpful if you have completed much of your analysis but then make changes to the data or if you add new observations to your data set. Instead of writing out new commands in the Command Window or clicking on menus for each analysis, you would simply highlight the commands in your do-file and run them all at once. Do-files are also important since they document changes to your data set and allow you to collaborate with others. Most data analysis should be carried out using the do-file editor.

To use a do-file to generate a **frequency table** of how often respondents meditate, we would open a do-file. The fastest way to open a new do-file is to click on the icon that shows a notepad with a pencil or by pressing Ctrl + 9 if you are using a Personal Computer (PC), or Command + 9 if you are using a Mac computer. You could also use menus by clicking on “Window → Do-File Editor → New Do-File Editor.” Once you have your do-file open, type in **tab mditate1** on the first line. Pressing on “Enter” will only take you to the second line and will not run the command. Instead, you can either put your cursor anywhere on the line and press Ctrl + D if you are using a PC and Command + Shift + D if you are using a Mac. If you have more than one line in your do-file, this will run all of the commands. If you only want to run one command, then you need to highlight at least one character on the line that you want to run before pressing Ctrl + D or Command + Shift + D. Instead of Ctrl + D or Command + Shift + D, you can also

click on the icon that shows a paper with the corner folded down and an arrow, which is the “Execute (do)” icon.

4.4 SETTING PREFERENCES IN STATA

As you continue to work with Stata, you may find that having two or three windows open at any given time (results, do-file, data editor) can be tedious if you have to find each window as you need them. If they are too large, only one window will appear on your screen and the others can only be found by hovering your cursor over the Stata icon at the bottom of your screen. Or, if you are running a command on the do-file and can’t see the results window, you will not know if the command was executed or had an error. It is standard practice, therefore, to set your window-size preferences. Once this is done, you can open Stata with your defined preferences each time that you use it. Ideally, you should have a narrow do-file open along one side of your computer screen and the opening Stata screen on the other side. Within the opening Stata screen, the Results Window should be wide enough so that statistical results in a table fit across the screen. Once you have the do-file and the opening Stata screen set to your preferred size, you should now use the menus to set the preferences by clicking on the following sequence:

Edit → Preferences → Save preference set → New preference set → Give a name to your preference set and click on “Okay”

The next time that you open Stata, click on:

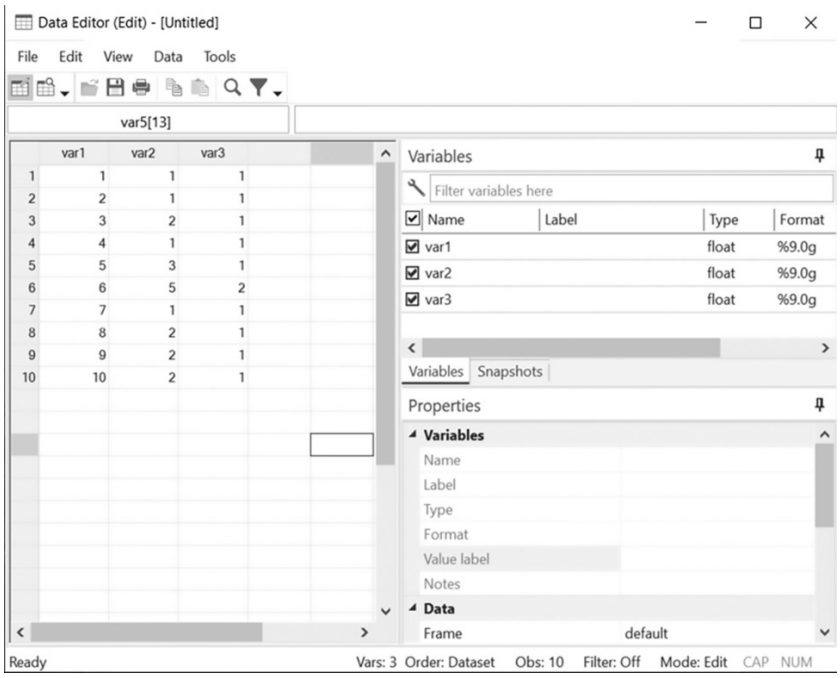
Edit → Preferences → Load preference set → Choose your preference set

---

4.5 ENTERING YOUR OWN DATA INTO STATA

To enter your own data into Stata, we would start by double-clicking on the Stata icon. We would then open the data editor by clicking on the icon that shows a spreadsheet with a pencil. We could also use the menus and click on “Data → Data Editor → Data Editor (Edit).”

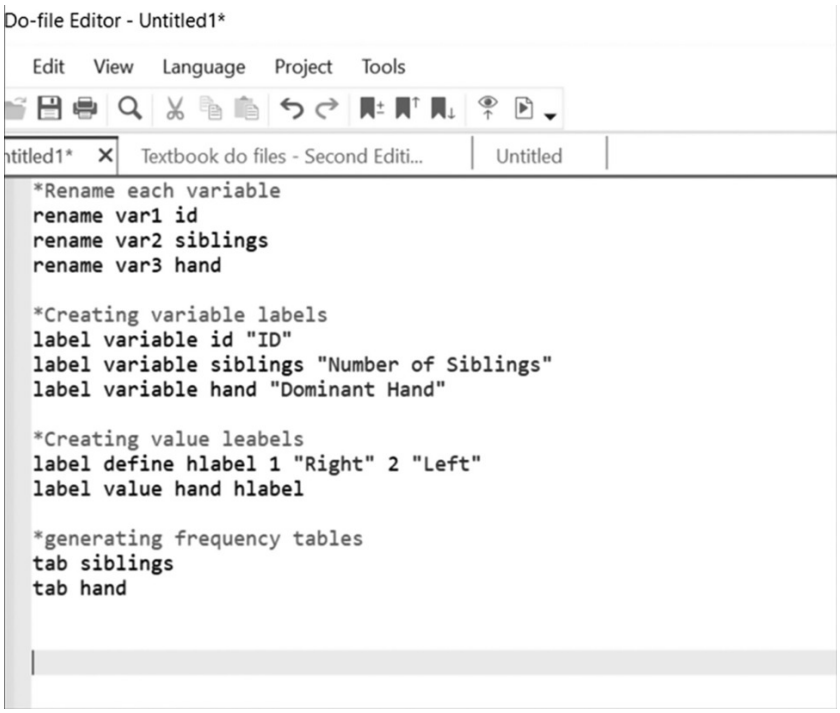
With the data editor open, we can now type data into the cells. Suppose, for example, there are 10 students in a classroom. We would first fill in the first column with numbers 1 through 10 so that each student has an identification number or ID. Alternatively, you could type in each student’s name. We could then ask each student to indicate how many siblings are in their family including themselves and record the response in the second column. Finally, we could ask each student if they are right-handed or left-handed. Right-handed would be recorded as “1” and left-handed would be recorded as “2.” The data would appear as illustrated in Figure 4.3.



Description

Figure 4.3 Entering Your Own Data For Two Variables Plus An Id

Next, we would want to give each variable a name, a variable label, and value labels for the question about dominant hand. To do this, we would open a do-file by placing the cursor on the main screen of Stata and clicking on the do-file icon. In our do-file, we would type the following commands as illustrated in Figure 4.4 and explained next.



Description

Figure 4.4 Do-File To Create Variable Names, Variable Labels, And Value Labels

First, note that we can write notes within the do-file to indicate what we are doing. If there is an asterisk at the beginning of a line, Stata will ignore the line. We can also skip lines to keep the do-file more organized.

The first four lines are used to rename the variables from “var1,” “var2,” and “var3” to “id,” “siblings,” and “hand,” respectively. In Stata, variable names are case sensitive, meaning that “siblings,” “Siblings,” and “SIBLINGS” would be considered three separate variables.

Lines 6 through 9 are used to give each variable a label, which is often shown in Stata output tables. This is useful when the variable name alone does not give enough information about the variable. In Lines 11 through 13, we are creating value labels, which indicate for the variable “hand” that each number represents right or left. Note that we first define a set of labels using **label define** hlabel 1 “Right” 2 “Left.” The word “hlabel” can be any word that we choose. We then apply these labels in Line 13. Finally, we can generate two tables in Lines 16 and 17 that use the variable labels and value labels. The output from this do-file is shown in Figure 4.5.

```
. tab siblings
```

Number of Siblings	Freq.	Percent	Cum.
1	4	40.00	40.00
2	4	40.00	80.00
3	1	10.00	90.00
5	1	10.00	100.00
Total	10	100.00	

```
. tab hand
```

Dominant Hand	Freq.	Percent	Cum.
Right	9	90.00	90.00
Left	1	10.00	100.00
Total	10	100.00	

Description

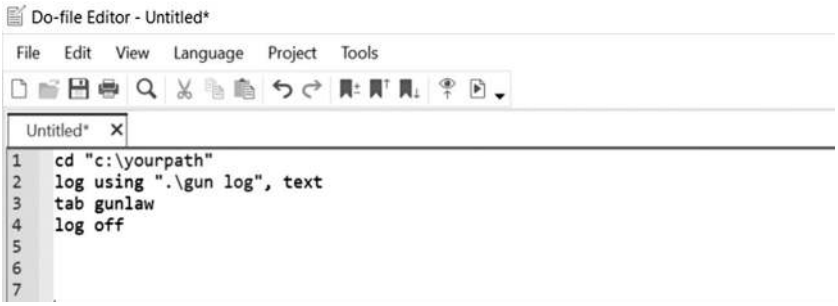
Figure 4.5 Frequency Tables After Creating Variable Names, Variable Labels, And Value Labels

4.6 USING LOG FILES AND SAVING YOUR WORK

As mentioned earlier, the Results Window does not automatically keep all of the output that you generate. It only stores about 500 lines. When it is full, it begins to delete the old results as you add new results. You can increase the amount of memory allocated to the Results Window, but even the maximum amount of memory will not be enough for a long session with Stata. To save all of the output from one session, you can use the **log** command to save our output in a *log file*.

There are several ways to start a log file. You can use the icon that shows a notebook with a spiral binding, or you can click on File and then Log from the menus. You can also use a **log** command in the Command Window. Finally, you can use **log** commands in the do-file. Because it is so important to learn to use do-files for most of your work, we will focus on the use of **log** commands within do-files.

Let's use the GSS2021.dta file to work through an example about views on gun permits in the United States. After opening the file, we could type the following commands into a do-file as shown in Figure 4.6.

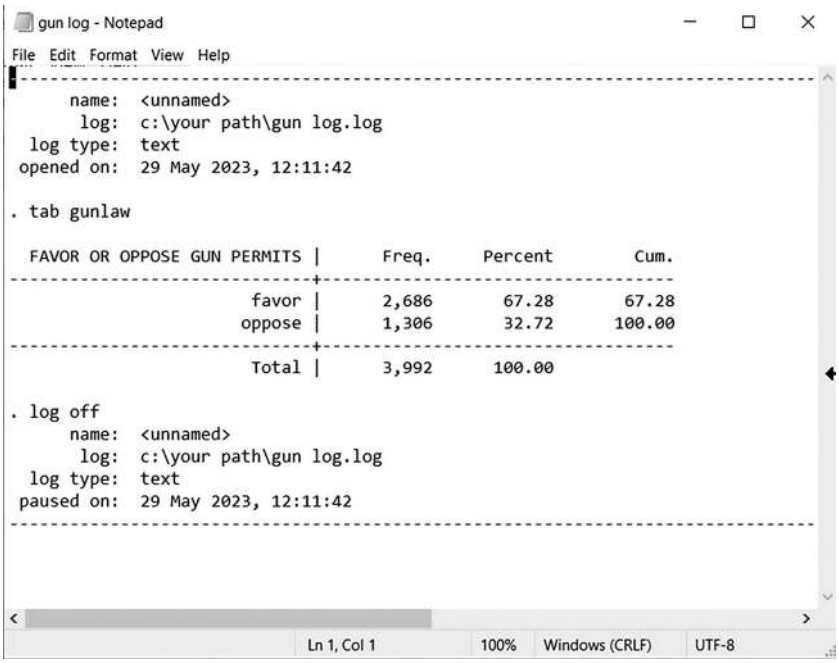


Description

Figure 4.6 Do-File To Open And Close A Log File

Line 1 tells Stata to change the directory to the location of the folder where you want to store the log file. We could indicate “your path” in Line 2 as the location. However, if we want to collaborate and allow others to use our do-file, it is better to use this technique. Each collaborator could then change the directory to their own path and run the rest of the do-file.

Line 2 tells Stata where you want to save the log file (your path) and the name of the log file. You could put it in any directory and folder. You can also give the log file any name. At the end of Line 2, we used the command **text** to let Stata know that we want the format to be a text file. If we didn’t specify this, Stata would make this a Stata Markup and Control Language (smcl) file that could only be opened in Stata. A text file, on the other hand, can be opened in Word, Notepad, or any word processor.[1] Line 3 generates a frequency table on views about favoring or opposing gun permits. Finally, the last line closes the log file. Figure 4.7 shows the full contents of the log file that is generated.



Description

Figure 4.7 Log Files In Text Format Generated By Stata

Notice that the formatting for the table shows dotted lines instead of solid lines. Although you could cut and paste this table into a report, the dotted lines are not ideal. Instead, there are several ways to copy the table into a document. Within Stata, you can highlight a table and then use menus to click on “Edit” and “Copy as picture.” If you prefer an image with a border, you could use the “Snipping Tool” that is included in all Windows operating systems. There are also several equivalent software tools for Mac, such as Grab, which is built into every operating system for a Mac computer.

In addition to the **log** commands that we illustrated above, there are several commands that you may want to use. For example, if you run the same do-file multiple times as you add to it or make changes, you must tell Stata to replace the existing log file with the newer version. Or if you want to add output to an existing file, you can do this with the **append** command. These are done as follows on a PC:

```
log using "c:\your file directory name\gunlaw", replace
log using "c:\your file directory name\gunlaw", append
```

Or as follows on a Mac:

```
log using "c:/your file directory name/gunlaw", replace
log using "c:/your file directory name/gunlaw", append
```

Also, there is a difference between **log off** and **log close**. With the command **log off** that we used previously, we can turn the log back on by running the command **log on**.[2]

If you use **log close**, you can only turn the log back on by running the command **log using**.

In addition to saving your commands in a do-file and your output in a log file, you may also want to save your data set if you made new variables or changes to any existing variables. It is good practice to always keep a copy of your original file. This allows you to start over if you make any mistakes as you modify the data. For this reason, when you save your file, you should save it under a new name that is different from the original file name. If it is the first time you are saving your new file, you would use the first command below on a PC. If you are using Mac, you would use the same commands with forward slashes. If you are saving changes to the data set again later to a file that already exists, you would need to use the second line that includes the **replace** command, again using forward slashes if using a Mac. You could also use the “save file” icon in the tool bar at the top of the screen, but it is better to get in the habit of adding all commands to your do-file so that you can document your work and rerun the do-file as you continue your analysis.

```
save "c:\your file directory name\new file name.dta"
save "c:\your file directory name\same file name as above.dta",
replace
```

If for some reason you do want to save several versions of a data file, it is convenient to put the dates in the file name. Avoid using “new” or “old” in the file names as these labels are vague and will become outdated.

4.7 GETTING HELP

Documentation for earlier versions of Stata came with a set of books that took up an entire bookshelf—about 12,000 pages! Today, all documentation is built into the software. You may also find information on the Stata website or by searching for each individual Stata command using the search engine.

4.7.1 Help Command

If you know the name of a Stata command but need more information about how to use it, you can access the **help** command in two ways. First, you can type **help** into the Command Window along with the name of the command. For example, you could type **help tabulate**. This will open a screen that shows the **tabulate** command, various options that can be used with the command, how to access it from the menus, and some examples. You can also access the help files by clicking on "Help" in the menus and then "Stata command."

4.7.2 Search Command

The **search** command can also be accessed from the Command Window or by clicking on "Help" from the menus and then "Search." If we type **search tabulate** into the Command Window, this will open a screen with a long list of resources in addition to Stata's help files. For example, it will include web resources and information from other users.

4.7.3 Stata Website

At Stata's website, www.stata.com, you can find information on Stata products, training courses, technical support, and documentation. The training courses include online courses, video tutorials, and classroom training. Finally, you can join user groups from this site where you may post questions about Stata and receive responses from other users.

4.7.4 Using a Search Engine

Full documentation for each Stata command, such as "tab," can be found by using a search engine. Simply type "Stata tab," and the search engine will display Stata's own documentation for that command.

4.8 SUMMARY OF COMMANDS USED IN THIS CHAPTER

In each chapter where we use Stata code (See Table 4.1), all of the commands used in the chapter will be summarized in this last section before the chapter exercises. In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 4.1 ■ Code Used In Chapter 4

Function	Code
Frequency table	tab mditate1
Variable names, variable labels, and value labels	rename var1 siblings label variable siblings "Number of Siblings" label define hlabel 1 "Right" 2 "Left" label value hand hlabel
Log files	log using "c:\your file directory name\gunlaw", text log using "c:\your file directory name\gunlaw", replace log using "c:\your file directory name\gunlaw", append log off (turn log back on using "log on") log close (turn log back on by running "log using"
Saving files	save "c:\your file directory name\new file name.dta" save "c:\your file directory name\same file name as above. dta", replace
Help commands	help tabulate search tabulate

Function	**Code**
Frequency table	tab mditate1
Variable names, variable labels, and value labels	rename var1 siblings label variable siblings “Number of Siblings” label define hlabel 1 “Right” 2 “Left” label value hand hlabel
Log files	log using “c:\your file directory name\gunlaw”, text log using “c:\your file directory name\gunlaw”, replace log using “c:\your file directory name\gunlaw”, append log off (turn log back on using “log on”) log close (turn log back on by running “log using”
Saving files	save “c:\your file directory name\new file name.dta” save “c:\your file directory name\same file name as above. dta”, replace
Help commands	help tabulate search tabulate

EXERCISES

1. Ten college students were asked four questions about their streaming habits (Table 4.2):

   (1) Which streaming service do you use most often to watch television shows and movies?

   (2) How many hours a week do you spend watching series or movies?

   (3) How often do you binge watch shows (watching more than three episodes of the same show in a row)? They could choose from (a) not at all, (b) sometimes—one to three times per week, and (c) frequently—more than three times per week.

   (4) Gender: How do you identify? (a) female, (b) male, (c) nonbinary, (d) other

   a. Based on their responses that are in the table below, enter the data for each of the four variables. For the three categorical variables (streaming service, frequency of binging, and genders), create a numeric code for each response. For example, for streaming service, 1 = Amazon Prime, 2 = Hulu Plus, 3 = HBO, and 4 = Netflix. (HINT: A common mistake is to type the variable names into the first row of the spreadsheet. Do not type variable names into the data editor spreadsheet. You should use Stata code to enter the names, which will appear at the top of each column in the data editor spreadsheet.)

   b. Once you have entered the data, use a do-file to rename each variable.

   c. Give each variable a variable label.

   d. Give each numeric code a value label.

   e. Save your data file (you will use this again in a later chapter).

   f. What percentage of the sample identify as female?

   g. Which streaming service is used most frequently?

   h. What percentage of students binge-watch shows frequently?

TABLE 4.2 ■ Student Responses to Survey About TV and Movie Viewing Habits

Student	TV Source	Hours per Week	Binge Frequency	Sex at Birth
1	Hulu Plus	14	Not at all	Male
2	Amazon Prime	18	Sometimes	Female
3	Hulu Plus	20	Frequently	Female
4	Netflix	5	Frequently	Nonbinary
5	Netflix	12	Frequently	Male
6	HBO	10	Not at all	Female
7	HBO	8	Frequently	Female
8	HBO	7	Sometimes	Other
9	Amazon Prime	24	Frequently	Male
10	Hulu Plus	30	Sometimes	Female

Student	TV Source	Hours per Week	Binge Frequency	Sex at Birth
1	Hulu Plus	14	Not at all	Male
2	Amazon Prime	18	Sometimes	Female
3	Hulu Plus	20	Frequently	Female
4	Netflix	5	Frequently	Nonbinary
5	Netflix	12	Frequently	Male
6	HBO	10	Not at all	Female
7	HBO	8	Frequently	Female
8	HBO	7	Sometimes	Other
9	Amazon Prime	24	Frequently	Male
10	Hulu Plus	30	Sometimes	Female

To determine the political views of people in the United States and how often they attend religious services, use the GSS2021 data set to complete the following exercises. These exercises will also allow you to practice using the **log** commands.

a. Open a do-file and then use it for all of your commands for this exercise.

b. Open a log file and name it “gss log.”

c. Open the GSS2021 data set.

d. Generate a frequency table of the variable “polviews.”

e. Stop your log file by using “**log off**.”

f. Turn your log file back on.

g. Generate a frequency table of the variable “attend.”

h. Stop your log file by using “**log close**.”

i. Submit your log file as your answer to all parts of Question 2.

The Admitted Student Questionnaire (ASQ) is administered by colleges each year to its incoming first-year students. The 2014 ASQ data set contains the responses from all students who answered the questionnaire that year— more than 5,000. Use this data set to explore how many colleges students applied to in the 2014–2015 academic year and to practice copying the output for use in other documents.

a. Generate a table that shows the percentage of students that applied to one college, the percentage that applied to two colleges, and so on for the 2014–2015 academic year. In other words, generate a frequency table for the variable Q65.

b. Copy and paste that table into a Word document by highlighting the table, right clicking on the table, and then selecting “Copy as picture.”

c. Change the font size of the output to fit it all on the results screen in Stata. To do this, highlight the table, right click, and choose font. Then set the font to a smaller size (8 or 9) so that the whole table fits on the screen.

d. If you are using a Windows operating system, copy the table you resized by using Snipping Tool. To do this, open the Snipping Tool software, click on “New,” place the cursor

in the upper-left corner of the table, and then drag the cursor to the lower-right corner while holding down the left button on the mouse. Then click on Ctrl + C to copy the table and then Ctrl + V to paste it into a Word document.

e. If you are using a Mac operating system, select the content you want to copy by highlighting it and then press Command + C simultaneously or choose Edit > Copy. To paste the material, position the cursor where you want to paste it and press Command + V simultaneously.

In the GSS2021 data set, one variable is “satjob.” Using this variable as an example, explain the difference between a variable name, a variable label, and a value label.

KEY TERMS

closed-ended question

enumerators

Likert scale

open-ended question

skip patterns

Descriptions of Images and Figures

Back to Figure

The top of the screen includes the tabs: file, edit, data, graphics, statistics, user, window, and help. The left of the screen shows the history and the right of the screen shows the variables including name, label, type, format, value label, and notes.

The center pane shows the following.

STATA 18.0 SE-Standard Edition

Statistics and Data Science

StataCorp

4905 Lakeway Drive

College Station, Texas 77845 USA

800-STATA-PC https://www.stata.com

979-696-4600 stata@stata.com

State license: Single-user, expiring 13 Oct 2023

Serial number: 401809204028

Licensed to: Lisa Daniels

Washington College

Notes:

1. Unicode is supported; see help unicode_advice.

2. Maximum number of variables is set to 5,000 but can be increased;

see help set_maxvar.

. use "C:\Users\WAC\OneDrive - Washington College\Documents\Textbook - Second ed

> ition\Data and Do files 2e\GSS2021.dta"

Back to Figure

How often do you meditate?	Freq.	Percent	Cum.
at least once a day	336	9.40	9.40
almost every day	402	11.25	20.65
once or twice a week	461	12.90	33.55
once or twice a month	343	9.60	43.14
a few times per year	374	10.46	53.61
once a year or less	204	5.71	59.32
never	1,454	40.68	100.00
Total	3,574	100.00	

.tab mditate1

Back to Figure

The data entered in the cells are as follows:

var1	var2	var3
1	1	1
2	1	1
3	2	1
4	1	1
5	3	1
6	5	2
7	1	1
8	2	1
9	2	1
10	2	1

On the right side, under the variables section, four checkboxes labeled "name," "var1," "var2," and "var3" are checked. The type and format corresponding to "var1," "var2," and "var3" are set to "float" and "%9.0g," respectively. Below that, from the list of variable properties (name, label, type, format, value label, notes), the "value label" option is selected.

Back to Figure

The commands shown in the Do-file are as follows:

*Rename each variable

rename var1 id

rename var2 siblings

rename var3 hand

*Creating variable labels

label variable id "ID"

label variable siblings "Number of Siblings" label variable hand "Dominant Hand"

*Creating value leabels

label define hlabel 1 "Right" 2 "Left"

label value hand hlabel

*generating frequency tables

tab siblings

tab hand

Back to Figure

.tab siblings

Number of siblings	Freq.	Percent	Cum.
1	4	40.00	40.00
2	4	40.00	80.00
3	1	10.00	90.00
5	1	10.00	100.00
Total	10	100.00	

.tab hand

Dominant Hand	Freq.	Percent	Cum.
Right	9	90.00	90.00
Left	1	10.00	100.00
Total	10	100.00	

Back to Figure

The command reads as follows:

cd "c:\yourpath"

log using ".\gun log", text

tab gunlaw

log off

Back to Figure

The content reads as follows:

name: <unnamed>

log: c:\your path\gun log.log

log type: text

opened on: 29 May 2023, 12:11:42

. tab gunlaw

FAVOR OR OPPOSE GUN PERMITS	Freq.	Percent	Cum.
Favor	2,686	67.28	67.28
Oppose	1,306	32.72	100.00
Total	3,992	100.00	

. log off

name: <unnamed>

log: c:\your path\gun log.log

log type: text

paused on: 29 May 2023, 12:11:42

5 PREPARING AND TRANSFORMING YOUR DATA

CHAPTER PREVIEW

Data Preparation Basics	Examples
Checking for outliers	Codebook
	Frequency tables
	Descriptive statistics
	Histograms
Creating new variables	Generate
	Using operators
	Recode
	Egen
Missing values	Missing
	Replace

Data Preparation Basics	Examples
Checking for outliers	Codebook Frequency tables Descriptive statistics Histograms
Creating new variables	Generate Using operators Recode Egen
Missing values	Missing Replace

5.1 INTRODUCTION

Whether you are using primary data that you collected yourself or secondary data, you will want to spend some time **"cleaning" your data**. This involves checking all variables for missing data, errors, or ***outliers***. An outlier is an observation that lies extremely far from the mean or other values in a variable. For example, if someone records that he or she is 125 years old or that he or she earns $25 billion, you will want to investigate these numbers and possibly make some changes to the data. Or, if someone

records that he or she watches 24 hours of television a day on average, you know there must be a mistake.

In addition to checking for missing data, errors, or outliers, you will also want to make new variables using the existing data. This could involve adding several variables together or transforming a variable, such as age, into categories or age ranges. You may also want to know how many people responded to several different variables combined.

All of these procedures are covered in this chapter, along with documenting your work through the use of do-files. Do-files are particularly important when cleaning a data set since you will need to keep track of all changes that you make. In addition, before you begin cleaning your data, you should always make an original copy of the file that will not be changed. This allows you to start over if you make any mistakes as you modify the data.

5.2 CHECKING FOR OUTLIERS

As described in the introduction, outliers occur when a value is simply far outside of the range of other observations. Because these extreme values will affect most of the statistics that we will learn about later in this book, we need to identify outliers and then decide what to do with them.

One way to examine your variables and look for outliers is to use the command **codebook**. By typing this into the Command Window or in a do-file, Stata will generate information about every variable in your data set. Figures 5.1 and 5.2 give examples of two types of variables from the GSS2021 data set—a continuous variable and a categorical variable, which are discussed in more detail in Chapter 6. In Figure 5.1, you can easily see the age range of your respondents and the number of missing values. The **missing.: 0/4,032** tells you that there are zero missing values with no explanation out of 4,032 observations. The **missing.*: 333/4,032** tells you that there are 333 answers that are missing with a code for why they are missing. If you then examine the variable with the command or **tab age, missing**, you can see that there are 107 responses listed as “no answer” and 226 listed as “iap,” or inapplicable, leading to a total of 333 respondents who have no recorded answer.

Description

Figure 5.1 Codebook Output For Continuous Variable

In Figure 5.2, respondents are asked if they have confidence in the United States Supreme Court. In this case, you can see that there are three possible answers. In addition, there are four codes for missing values: don’t know (.d), inapplicable (IAP), no answer (n) and skipped on web (.s).

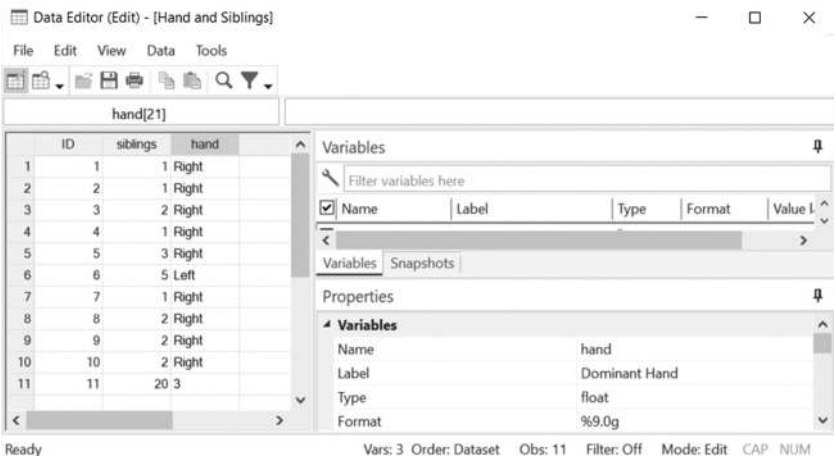
```
. codebook conjudge

conjudge                                                    CONFID. IN UNITED STATES SUPREME COURT

                  Type: Numeric (byte)
                 Label: INSTCONF

                 Range: [1,3]                         Units: 1
         Unique values: 3                         Missing .: 0/4,032
     Unique mv codes: 4                           Missing .*: 1,370/4,032

            Tabulation: Freq.   Numeric  Label
                          689         1  a great deal
                        1,437         2  only some
                          536         3  hardly any
                            3        .d  don't know
                        1,360        .i  iap
                            1        .n  no answer
                            6        .s  skipped on web
```

Description

Figure 5.2 Codebook Output For a Categorical Variable

In addition to the codebook, it is also useful to generate frequency tables for categorical variables with a limited number of responses and descriptive statistics such as the mean and standard deviation for continuous variables. Histograms are also useful to identify patterns in your data. All of these methods are discussed further in Chapter 6.

Returning to the data set that we created in Chapter 4, where the respondents indicated the number of siblings they have and their dominant hand, let's suppose that we have one additional observation whereby the 11th person accidentally entered 20 for the number of siblings and "3" instead of "1" or "2" for right-handed and left-handed, respectively. Figure 5.3 shows the data editor screen for this data set. Because there are only 11 observations, we could easily identify these errors by simply looking at the raw data. With thousands of observations, however, we would need to examine the data using the codebook, frequency tables, descriptive statistics, and histograms.



Description

Figure 5.3 Data Editor Screen Showing Errors In Data

Once we have identified these errors, we would need to make changes to the data set. If we don't know what the respondent intended to write for the number of siblings, then we would have to delete the value. If we know that the respondent meant to type "2," we could legitimately change this in our do-file. The two commands would appear as follows:

```
replace siblings =. i f id==11 (to change the value to missing)
replace siblings = 2 i f id==11 (to change the value to a 2)
```

In the first command, Stata changes the value in observation 11 to a “.” Indicating that the value is missing and to a “2” in the second command. Although we could do this directly in the data editor screen, it is better to record all changes in a do-file, as mentioned earlier. In addition to documenting all changes, if we download a data set multiple times as new observations are added to an online questionnaire, for example, it will have the same errors each time. By writing a do-file, we can correct the errors each time by simply running the do-file.

Similarly, to change the number 3 in the 11th row to a 2 for the variable hand, we would write,

```
replace hand = 2 if id==11
```

Although researchers are tempted to simply remove all outliers, there are some basic rules regarding when it is acceptable to drop an outlier and how it should be documented. If the outlier is a data entry error, such as watching 24 hours of TV a day, then you can remove this since you know it is a mistake. If, however, there is a legitimate outlier, such as $1 billion in a data set that asks for annual income, the researcher could remove the outlier and include a footnote to indicate that this one observation was removed.[1]

5.3 CREATING NEW VARIABLES

When you begin working with a data set, you will often want to create new variables. This can be done in a number of ways. In this section, we will cover the following methods:

generate

recode

egen

5.3.1 Generate

The **generate** command is used to create a new variable. The Stata code for **generate** is as follows:

```
gen newvar = expression [if expression]
```

The command **gen** must be in lowercase, like all Stata commands, while “newvar” represents the new variable being created. As discussed in the previous chapter, Stata variable names are case sensitive, meaning that uppercase letters are considered different from lowercase letters. After defining the variable, you must use the same capitalization whenever you refer to it.

The first “expression” in the code above is mathematical such as **gen BMI = (weight/height^2)*703** which is the formula for Body Mass Index (BMI). If weight or height were missing for any given observation, then the newly generated variable that uses these values would also be missing. If you wanted to only calculate the weight of adults, you could add an “if expression” such as **gen BMIAdult = weight/height^2)*703 if age > 17**. This new variable would skip respondents who are under 18. If, however, age is missing, age is treated as the value of infinity and Stata will generate a BMI for that respondent assuming they are an adult. This must be corrected as follows: **gen BMIAdult2 = weight/height^2)*703 if age > 17 & age !=.** The “!=” means “not equal to” and lets Stata know to skip any respondent that did not specify an age.

If the expression is an inequality such as **gen adult = age > 17**, then the new variable will take the value of 0 if the expression is false and 1 if it is true. Just as in the case of the missing age above, Stata will count the missing age as infinity and assign it a value of 1. Similarly, if you use the code **gen adult2 = 1 if age > 17**, Stata will count a missing age as infinity and assign a value of 1 to that case but list missing values for anyone under 18. You can correct both of these as we did above by telling Stata to skip missing values for age as follows: **gen adult3 = age > 17 & age !=**.

The do-file and corresponding data editor results that illustrate these concepts are shown in Figure 5.4. Table 5.1 shows further examples of how to use the generate command.

Description

Figure 5.4 Do File and Data Editor Illustrating the Impact of Missing Values When Generating New Variables Based on Existing Variables

TABLE 5.1 ■ Examples of the Generate Command

Command	Operation
gen pctoffers=totoffers/applications*100	Creates a new variable that shows the percentage of job offers someone receives as a proportion of their total applications submitted.
gen salaryintern=beginningsalary if internship==1 & beginningsalary !=.	Creates a new variable that shows the beginning salary for a first job after college if the student had an internship. If they did not, the value will be missing.
gen highprice = (price > 1000) & price !=.	Creates a variable equal to 1 to indicate that the price is greater than 1,000 and 0 to indicate that it is 1,000 or lower.

Source: Adapted from Minot (2012).

Command	Operation
gen pctoffers=totoffers/applications*100	Creates a new variable that shows the percentage of job offers someone receives as a proportion of their total applications submitted.
gen salaryintern=beginningsalary if internship==1 & beginningsalary !=.	Creates a new variable that shows the beginning salary for a first job after college if the student had an internship. If they did not, the value will be missing.
gen highprice = (price > 1000) & price !=.	Creates a variable equal to 1 to indicate that the price is greater than 1,000 and 0 to indicate that it is 1,000 or lower.

Source: Adapted from Minot (2012).

5.3.2 Using Operators

Operators are symbols used in equations, as shown in Table 5.1. Most of the operators are obvious (e.g., + and −), but some are not. Table 5.2 lists the most commonly used operators. In Stata, you cannot use words such as “or,” “and,” “eq,” or “gt.” Instead, you must use operator symbols.

TABLE 5.2 ■ Key Operators for Writing Equations in Stata

Operator	Meaning	Example
+	Addition	gen income = agincome + nonagincome
–	Subtraction	gen netrevenue = revenue - cost
*	Multiplication	gen value = price * quantity
/	Division	gen exppc = expenditure/hhsize
^	Power	gen agesquared = age^2
>	Greater than	gen aboveavg = 1 if income > avgincome
>=	Greater than or equal to	gen adult = 1 if age >= 18
<	Less than	gen belowavg = 1 if income < avgincome
<=	Less than or equal to	gen child = 1 if age <=10
=	Assignment operator	gen expend = foodexp + nonfoodexp
= =	Equal	gen femhead = 1 if sexhead= =2
!=	Not equal	gen error = 1 if value1 != value2
|	Or	gen age=. if age= =999 | age=9999
&	And	gen sexhead = 1 if sex= =1 & relation= =1

Source: Adapted from Minot (2012).

Operator	Meaning	Example
+	Addition	gen income = agincome + nonagincome
–	Subtraction	gen netrevenue = revenue - cost
*	Multiplication	gen value = price * quantity
/	Division	gen exppc = expenditure/hhsize
^	Power	gen agesquared = age^2
>	Greater than	gen aboveavg = 1 if income > avgincome
>=	Greater than or equal to	gen adult = 1 if age >= 18
<	Less than	gen belowavg = 1 if income < avgincome
<=	Less than or equal to	gen child = 1 if age <=10
=	Assignment operator	gen expend = foodexp + nonfoodexp
= =	Equal	gen femhead = 1 if sexhead= =2
!=	Not equal	gen error = 1 if value1 != value2
|	Or	gen age=. if age= =999 | age=9999
&	And	gen sexhead = 1 if sex= =1 & relation= =1

Source: Adapted from Minot (2012).

The most difficult rule to remember is when to use = (single equal symbol) and when to use = = (double equal symbol).

Use a single equal symbol (=) when defining a variable.

Use a double equal symbol (= =) when you are testing an equality, such as in an “if” statement and when creating a dummy variable, which are discussed in later chapters.

5.3.3 Recode

The **recode** command redefines the values of a variable according to rules that you specify. The command is as follows:

```
recode varlist (oldvalues = newvalue) (oldvalues = newvalue) … [if exp] [in range]
```

Table 5.3 lists some examples of the **recode** command.

TABLE 5.3 ■ Examples of the Recode Command

Command	Operation
recode x (1=2)	Within the *x* variable, all 1s become 2
recode x y z (1=2) (3=4)	In variables *x*, *y*, and *z*, changes 1 to 2 and 3 to 4
recode x (1=2) (2=1)	In the variable *x*, exchanges the values 1 and 2
recode x (1=2) (*=3)	In the variable *x*, changes 1 to 2 and all other values to 3
recode x 1/5=2	In the variable *x*, changes 1 through 5 to 2
recode x y (1 3 4 5 = 6)	In variables *x* and *y*, changes 1, 3, 4, and 5 to 6
recode x (.=9)	In the variable *x*, changes missing to 9
recode x (9=.)	In the variable *x*, changes 9 to missing

Source: Adapted from Minot [2012].

Command	Operation
recode x (1=2)	Within the *x* variable, all 1s become 2
recode x y z (1=2) (3=4)	In variables *x, y*, and *z*, changes 1 to 2 and 3 to 4
recode x (1=2) (2=1)	In the variable *x*, exchanges the values 1 and 2
recode x (1=2) (*=3)	In the variable *x*, changes 1 to 2 and all other values to 3
recode x 1/5=2	In the variable *x*, changes 1 through 5 to 2
recode x y (1 3 4 5 = 6)	In variables *x* and *y*, changes 1, 3, 4, and 5 to 6
recode x (.=9)	In the variable *x*, changes missing to 9
recode x (9=.)	In the variable *x*, changes 9 to missing

Source: Adapted from Minot [2012].

Notice that you can use some special symbols in the **recode** command:

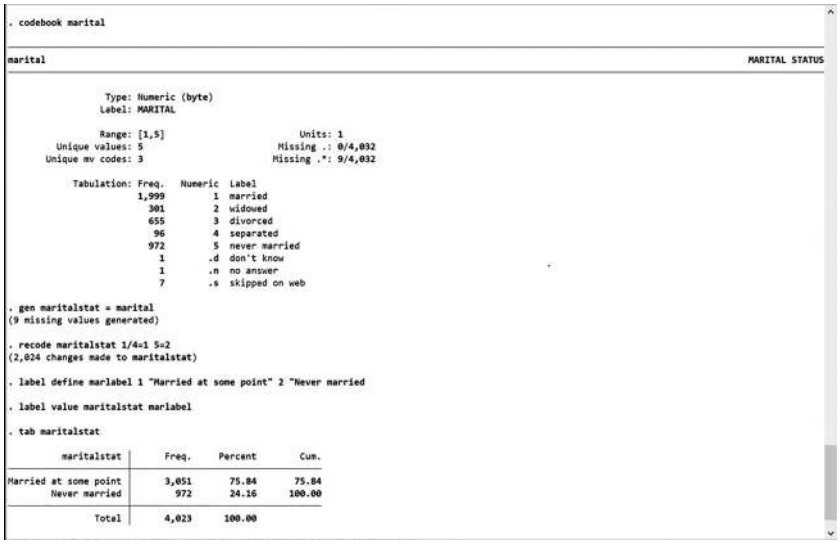
***** means all other values

x/y means all values from *x* to *y*

x y means values *x* and *y*

Figure 5.5 shows an example of creating a new variable using the **recode** command. Suppose that we are interested in knowing the happiness level of someone who is currently married and living with his or her partner versus someone who is not currently married nor living with his or her partner. Using the GSS2016 data sets, we would first check the marital status variable, mar1, using a **codebook mar1**. Notice that there are five categories: (1) married, (2) widowed, (3) divorced, (4) separated, and (5) never married. Although we would assume that the labels are given values in the order from 1 to 5, it is always

important to check this to be sure before you recode. We would now generate a new variable, “maritalstat,” that is identical to the original variable mar1. Then, we begin the recoding process indicating that values 2 through 5 will all be equal to 2. Finally, we give new value labels to the variable showing that “1” is married and “2” is not currently married or living with a spouse.[2]



```
. codebook marital

marital                                                         MARITAL STATUS

                  Type: Numeric (byte)
                 Label: MARITAL

                 Range: [1,5]                         Units: 1
         Unique values: 5                         Missing .: 0/4,032
     Unique mv codes: 3                           Missing .*: 9/4,032

            Tabulation: Freq.   Numeric  Label
                        1,999        1  married
                          301        2  widowed
                          655        3  divorced
                           96        4  separated
                          972        5  never married
                            1       .d  don't know
                            1       .n  no answer
                            7       .s  skipped on web

. gen maritalstat = marital
(9 missing values generated)

. recode maritalstat 1/4=1 5=2
(2,024 changes made to maritalstat)

. label define marlabel 1 "Married at some point" 2 "Never married

. label value maritalstat marlabel

. tab maritalstat

       maritalstat |      Freq.     Percent        Cum.
-------------------+-----------------------------------
Married at some point |      3,051       75.84       75.84
      Never married |        972       24.16      100.00
-------------------+-----------------------------------
              Total |      4,023      100.00
```

Description

Figure 5.5 Example of the Recode Command

5.3.4 Egen

The **egen** command is an extended version of the **generate** command. It is used to create a new variable by aggregating the existing data. The command is as follows:

```
egen newvar = fcn(argument) [if exp] [in range], by(var)]
```

where

newvar is the new variable to be created

fcn is one of numerous functions such as

count() max() min()

mean() median() rank()

sd() sum() rowtotal()

(See help **egen** for the full list.)

argument is normally just a variable or a variable list

var in the **by**() subcommand must be a *categorical* variable

Table 5.4 gives a few examples of the **egen** command using the mean, median, and sum functions.

TABLE 5.4 ■ Examples of the Egen Command

Command	Operation
egen avgincome = mean(income)	Creates a variable of average income over the entire sample.
by region: egen regincome = median(income)	Creates a variable of median income by region.
by household: egen hhincome = sum(income)	Creates a variable of total income for each household.

Source: Adapted from Minot (2012).

Command	Operation
egen avgincome = mean(income)	Creates a variable of average income over the entire sample.
by region: egen regincome = median(income)	Creates a variable of median income by region.
by household: egen hhincome = sum(income)	Creates a variable of total income for each household.

Source: Adapted from Minot (2012).

Figure 5.6 shows another example of the **egen** command. Using the 2014 Admitted Student Questionnaire data set (2014 ASQ Data), we may want to know how much emphasis students place on academics versus social life when choosing a college. Questions QA1, QA2, and QA3 ask students whether the quality of the faculty, majors of interest, and academic reputation are *very important* (= 1), *somewhat important* (= 2), or *not important* (= 3). Questions QA11, QA12, and QA14 ask about the importance of extracurricular opportunities, off-campus activities, and quality of social life using the same scale of *very, somewhat*, and *not important*. As shown in Figure 5.6, we first generate two new variables using the **egen** command—academic and social. For the academic variable, **egen** counts the number of times variables QA1, QA2, and QA3 are given a value of “1” or *very important*.

```
. egen academic = anycount (QA1 QA2 QA3), values (1)

. egen social = anycount (QA11 QA12 QA14), values (1)

. tab academic
```

QA1 QA2 QA3 == 1	Freq.	Percent	Cum.
0	95	1.63	1.63
1	548	9.43	11.06
2	1,893	32.56	43.62
3	3,278	56.38	100.00
Total	5,814	100.00	

```
. tab social
```

QA11 QA12 QA14 == 1	Freq.	Percent	Cum.
0	1,320	22.70	22.70
1	1,680	28.90	51.60
2	1,639	28.19	79.79
3	1,175	20.21	100.00
Total	5,814	100.00	

Description

Figure 5.6 Example of the Egen Command

A frequency table of this variable shows that 3,278 students, or 56%, ranked all three questions related to academics as *very important*. Using the same method for the "social" variable, the frequency table for "social" shows that only 1,175 students, or 20%, ranked all three variables related to social activities as *very important*. Based on this sample, students do place more emphasis on academic reputation than on social life.

5.4 MISSING VALUES IN STATA

Missing values are represented as a "." or as ".a," ".b," ".c," … ".z" in Stata. As we saw earlier in the codebook, users can also specify multiple codes for missing such as "don't know," "not applicable," or "refused to answer." Most commands ignore missing values by default. Some commands, such as **tabulate**, have an option to display missing values if you want to see how many observations are missing. This would be done using the code **tab mar1, missing**.

In some cases, you may use missing values in a way that you did not intend. For example, the **replace** command does not ignore missing values, so you must take them into account when you replace variables using a > (greater than) function as you may inadvertently replace missing values.

When there are missing values, statistical packages may eliminate that entire case (or row) from the data set. This is called a “listwise” deletion. In other cases, “pairwise deletion” is done, which eliminates a case only when it is missing a variable required for a particular analysis. In the case of Stata, the default is pairwise deletion.

Researchers will also use *imputation*, which is the practice of replacing missing data with other values. One common method is to substitute the mean value of a variable for any observation that is missing. This is a somewhat controversial procedure and is considered inappropriate by many researchers. For a more thorough discussion of how to deal with missing data, refer to Enders (2010), Little and Rubin (2014), or Sauro (2015).

5.5 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes all of the Stata code used in the chapter (Table 5.5). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 5.5 ■ Code Used In Chapter 5

Function	Command
Looking for outliers and missing data	codebook tab age, missing
Replacing or removing data	replace siblings =. if id==11 (to change the value to missing) replace siblings = 2 if id==11 (to change the value to a 2) replace hand = 2 if id==11
Creating new variables	gen pctoffers=totoffers/applications*100 gen salaryintern=beginningsalary if internship==1 & beginingsalary !=. gen highprice = (price>1000) & price !=. gen income = agincome + nonagincome gen netrevenue = revenue – cost gen value = price * quantity gen exppc = expenditure/hhsize gen agesquared = age^2 gen aboveavg = 1 if income > avgincome gen adult = 1 if age >= 18 gen belowavg = 1 if income < avgincome gen child = 1 if age <=10 gen expend = foodexp + nonfoodexp gen femhead = 1 if sexhead==2 gen error = 1 if value1 != value2 gen age=. if age==999 | age=9999 gen sexhead = 1 if sex==1 & relation==1
Recoding existing variables	recode x (1=2) recode x y z (1=2) (3=4) recode x (1=2) (2=1) recode x (1=2) (*=3) recode x 1/5=2 recode x y (1 3 4 5 = 6) recode x (.=9) recode x (9=.) recode mar1 2/5=2, generate(maritalstat)
Working with labels	label define marlabel 1 "Married" 2 "Not currently married or living with spouse" label value maritalstata marlabel
Aggregating existing data	by region: egen avgincome = mean(income) by region: by household: egen regincome = median(income) by household: egen hhincome = sum(income)

Function	Command
Looking for outliers and missing data	codebook tab age, missing
Replacing or removing data	replace siblings =. if id==11 (to change the value to missing) replace siblings = 2 if id==11 (to change the value to a 2) replace hand = 2 if id==11
Creating new variables	gen pctoffers=totoffers/applications*100 gen salaryintern=beginningsalary if internship==1 & beginingsalary !=. gen highprice = (price>1000) & price !=. gen income = agincome + nonagincome gen netrevenue = revenue – cost gen value = price * quantity gen exppc = expenditure/hhsize gen agesquared = age^2 gen aboveavg = 1 if income > avgincome gen adult = 1 if age >= 18 gen belowavg = 1 if income < avgincome gen child = 1 if age <=10 gen expend = foodexp + nonfoodexp gen femhead = 1 if sexhead==2 gen error = 1 if value1 != value2 gen age=. if age==999 | age=9999 gen sexhead = 1 if sex==1 & relation==1

Function	Command
Recoding existing variables	recode x (1=2) recode x y z (1=2) (3=4) recode x (1=2) (2=1) recode x (1=2) (*=3) recode x 1/5=2 recode x y (1 3 4 5 = 6) recode x (.=9) recode x (9=.) recode mar1 2/5=2, generate(maritalstat)
Working with labels	label define marlabel 1 “Married” 2 “Not currently married or living with spouse” label value maritalstata marlabel
Aggregating existing data	by region: egen avgincome = mean(income) by region: by household: egen regincome = median(income) by household: egen hhincome = sum(income)

EXERCISES

1. To examine the age when someone first tried smokeless tobacco, use the “National Survey on Drug Use and Health, 2015” data set to complete the following exercises.
   a. Generate a table of the age when someone first tried smokeless tobacco (smklsstry).
   b. Generate a new variable that is identical to smklsstry and call it smklssage.
   c. Recode smklssage so that the codes 994, 997, and 998 are blank.
   d. Recode smklssage so that you combine users into the following categories: never used smokeless tobacco, <10, 10 to 12, 13 to 15, 16 to 18, 19 to 21, and >21.
   e. Generate value labels for these age groups, and apply them to smklssage.
   f. Generate a table of smklssage, and notice the label that is in the left column of the table at the top.
   g. Create a variable label “Age when first tried smokeless tobacco,” and apply this to smklssage.
   h. Generate a table of smklssage again, and notice the change in the label above the left-hand column.
2. Use the GSS2021 data set to complete the following exercises that generate a categorical variable out of a continuous variable.
   a. Generate a table of how many children a respondent has (childs).

b. Generate a new variable that is equal to 1 if the respondent has any children and 2 if the respondent has no children.
c. Create a variable label “Respondent has children,” and apply it to your new variable.
d. Create value labels so that 1 is “Yes” and 2 is “No.”
e. Generate a table of your new variable.

3. Use the GSS2021 data set to complete the following exercises related to regional income disparities in the United States.
   a. Use the **egen** command to generate a variable that is the median value of real income (realinc) of all respondents in the data set.
   b. Generate a new variable that is the difference between an individual’s real income (realinc) and the median income of all individuals (the variable you created in Part A).
   c. Generate a new variable that is equal to 1 if an individual earns above the median income and 0 if the individual earns below the median income.
   d. Define and apply value labels to the variable you created in Part C.
   e. Create a table that shows region of the United States in the rows and the variable you created in Part C in the columns. Have this table add across the rows and include no frequencies.
   f. In a couple of sentences, describe the results and their meaning.

KEY TERMS

cleaning data

histograms

imputation

observation

outliers

Descriptions of Images and Figures

Back to Figure

The command reads as follows:

. use "$datapath\G552021"

.

. *Output 5.1

. codebook age

age AGE OF RESPONDENT

Type: Numeric (byte)

Label: AGE, but 71 nonmissing values are not labeled

Range: [18,89] Units: 1

Unique values: 72 Missing .: 0/4,032

Unique my codes: 2 Missing .*: 333/4,032

Examples: 36

49

61

73

Back to Figure

The command reads as follows:

. codebook conjudge

conjudge CONFID. IN UNITED STATES SUPREME COURT

Type: Numeric (byte)

Label: INSTCONF

Range: [1, 3] Units: 1

Unique values: 3 Missing .: 0/4,032

Unique my codes: 4 Missing .": 1,370/4,032

Tabulation: Freq.	Numeric	Label
689	1	a great deal
1,437	2	only some
536	3	hardly any
3	.d	don’t know
1,360	.i	iap
1	.n	no answer
6	.s	skipped on web

Back to Figure

The data entered in the cells are as follows:

ID	siblings	hand
1	1	Right
2	1	Right
3	2	Right
4	1	Right
5	3	Right
6	5	Left
7	1	Right
8	2	Right
9	2	Right
10	2	Right
11	20	3

On the right side, the tab “Variables” is selected. The variable properties displayed are as follows:

Namehand

LabelDominant Hand

Typefloat

Format%9.0g

Back to Figure

The command in the Do-file editor reads as follows:

```
1 *Mathematical statement 2
2 *If height or weight is missing, BMI is missing.
3 gen BMI = (weight/height^2)’703
4
5 *Mathematical statement with if statement
6 *If age is missing, age counts as infinity and that respondent is included as an adult
7 gen BMIAdult = (weight/heightA2)*703 if age > 17
8
9 *Let Stata know that missing values should not be counted as follows:
10 gen BMIAdult2 = (weight/height"2)703 if age > 17 & age !=.
11
12 *Inequality
13 *counts "." as infinity and returns 0 if false and 1 if true
14 gen adult = age > 17
15
16 ‘Equality with if statement
17 *Counts "." as infinity and respondents who do not specify and age will be counted as an adult.
18 *Returns 1 if true and "." if not true
19 gen adult2 = 1 if age > 17
20
21 *Inequality that rules out missing
22 *Returns 1 for adults and "." for non-adults and respondents that do not report an age.
```

23 gen adult3 = 1 if age > 17 & age !=.

The data in the data editor is as follows:

age	weight	height	BMI	BMIAdult	BMIAdult2	adult	adult2	adult3
2	15	24	18.30729	.	.	0	.	.
4	.	70	.	.	.	0	.	.
24	145	70	20.80306	20.80306	20.80306	1	1	1
28	.	62	.	.	.	1	1	1
.	195	71	27.19401	.	.	1	1	.

Back to Figure

The command reads as follows:

. codebook marital

marital MARITAL STATUS

Type: Numeric (byte)

Label: MARITAL

Range: [1,5] Units: 1

Unique values: 5 Missing .: 0/4,032

Unique mv codes: 3 Missing .*: 9/4,032

Tabulation: Freq	Numeric	Label
1,999	1	married
301	2	widow
655	3	divorced
9	4	separated
972	5	never married
1	.d	don’t know
1	.n	no answer
7	.s	skipped on web

. gen maritalstat = marital

(9 missing values generated)

. recode maritalstat 1/4=1 5=2

(2,024 changes made to maritalstat)

. label define marlabel 1 "Married at some point" 2 "Never married

. label value maritalstat marlabel

. tab maritalstat

maritalstat.	Freq.	Percent	Cum.
Married at some point	3,051	75.84	75.84
Never married	972	24.16	100.00
Total	4,023	100.00	

Back to Figure

The command reads as follows:

. egen academic = anycount (QA1 QA2 QA3), values (1)

. egen social = anycount (QA11 QAl2 QA14), values (1)

. tab academic

QA1 QA2 QA3 ==1	Freq.	Percent	Cum.
0	95	1.63	1.63
1	548	9.43	11.06
2	1,893	32.56	43.62
3	3,278	56.38	100.00
Total	5,814	100.00	

.tab social

QA11 QA12 QA14 ==1	Freq.	Percent	Cum.
0	1,320	22.70	22.70
1	1,680	28.90	51.60
2	1,639	28.19	79.79
3	1,175	20.21	100.00
Total	5,814	100.00	

6 DESCRIPTIVE STATISTICS

CHAPTER PREVIEW

Descriptive Statistics Basics	Example
What are descriptive statistics?	A summary or description of data
	Based on a sample or a population
	Used to describe a sample or population, to answer research questions, to check violation of assumptions, and to look for outliers
Types of variables and measurement	Categorical—nominal and ordinal
	Continuous—interval and ratio
Descriptive statistics for all variables	Frequency tables
	Mode
Descriptive statistics for ordinal, interval, and ratio scales	Median
	Percentile
Descriptive statistics for interval and ratio scales	Mean
	Variance
	Standard deviation
	Coefficient of variation
Descriptive statistics for nominal scales	Cross tabulation
Graphs to describe data	Bar
	Box plot
	Histogram
	Pie

Descriptive Statistics Basics	Example
What are descriptive statistics?	A summary or description of data Based on a sample or a population Used to describe a sample or population, to answer research questions, to check violation of assumptions, and to look for outliers
Types of variables and measurement	Categorical—nominal and ordinal Continuous—interval and ratio

Descriptive Statistics Basics	Example
Descriptive statistics for all variables	Frequency tables Mode
Descriptive statistics for ordinal, interval, and ratio scales	Median Percentile
Descriptive statistics for interval and ratio scales	Mean Variance Standard deviation Coefficient of variation
Descriptive statistics for nominal scales	Cross tabulation
Graphs to describe data	Bar Box plot Histogram Pie

6.1 INTRODUCTION

Descriptive statistics are used to describe or summarize data. For example, you may want to know the average age of the respondents in a study or the range of the respondents' ages. You may also want to know the percentage of respondents by gender. In some cases, you may have access to the data for an entire population, in which case descriptive statistics are used to describe the population. A census of the population, for example, is conducted every 10 years in many countries. Generally, however, descriptive statistics are based on a sample of a larger population. If you are conducting a survey of 300 students in a college where there are 1,500 students, for example, then descriptive statistics will describe that sample. They cannot be used to make generalizations or inferences about the population without further analysis or testing. We will learn how to test hypotheses and make inferences about the population in later chapters. In addition to describing data, descriptive statistics are used to answer research questions, to check for violations of assumptions, and to look for outliers. Before we cover descriptive statistics and how to use them, however, we need to know about the different types of variables and measures since this will affect which descriptive statistics can be used.

6.2 TYPES OF VARIABLES AND MEASUREMENT

In statistics, a ***variable*** is defined as a number or characteristic that can be measured and that varies over a sample or population. For example, age, income, gender, and political affiliation are variables that can be measured. Variables can be divided into two major categories: (1) categorical and (2) continuous. Within these two categories, there are different scales of measurement as illustrated in Table 6.1. These distinctions are important since they affect the type of analysis that can be done with each variable.

TABLE 6.1 ■ Variable Types, Scales of Measurement, and Analyses

Variable Type	Categorical		Continuous	
Scale of Measurement	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Definition	A measure with two or more categories that do not have a natural order	A measure with two or more categories that can be ranked or ordered, but the distance between categories can't be measured precisely	A measure that has a numerical value, and the magnitude between intervals is the same	A measure that is the same as an interval measure, but it also has a true zero value
Example	Gender Race First language	Military rank Education level (primary, some secondary, high school, etc.) Economic status (low-, middle-, or high-income)	Temperature in Fahrenheit Date Time of day	Income (exact dollar amount) Weight Sales
Frequencies	✓	✓	✓ (when limited # of values)	✓ (when limited # of values)
Mode	✓	✓	✓	✓
Median, percentiles	✓	✓	✓	✓
Mean, variance, standard deviation	✓	✓	✓	✓
Cross-tabulation	✓	✓	✓	✓
Bar graph	✓	✓	✓	✓
Box plot	✓	✓	✓	✓
Histogram	✓	✓	✓	✓
Pie chart	✓	✓	✓	✓

Variable Type	**Categorical**		**Continuous**	
Scale of Measurement	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Definition	A measure with two or more categories that do not have a natural order	A measure with two or more categories that can be ranked or ordered, but the distance between categories can't be measured precisely	A measure that has a numerical value, and the magnitude between intervals is the same	A measure that is the same as an interval measure, but it also has a true zero value

Variable Type	**Categorical**		**Continuous**	
Example	Gender Race First language	Military rank Education level (primary, some secondary, high school, etc.) Economic status (low-, middle-, or high-income)	Temperature in Fahrenheit Date Time of day	Income (exact dollar amount) Weight Sales
Frequencies	✓	✓	✓ (when limited # of values)	✓ (when limited # of values)
Mode	✓	✓	✓	✓
Median, percentiles	✓	✓	✓	✓
Mean, variance, standard deviation	✓	✓	✓	✓
Cross-tabulation	✓	✓	✓	✓
Bar graph	✓	✓	✓	✓
Box plot	✓	✓	✓	✓
Histogram	✓	✓	✓	✓
Pie chart	✓	✓	✓	✓

A ***categorical variable*** is a variable that has a limited number of possible values that fall into categories based on some qualitative property or quantitative ranges, such as 1 to 5, 6 to 10, and so on. It can be measured on a nominal or ordinal scale. A ***nominal scale*** is a measure with two or more categories that do not have a natural order. For example, gender, political affiliation, and first language are categorical variables measured on a nominal scale. An ***ordinal scale*** is a measure with two or more categories that can be ranked or ordered, but the distance between the categories can't be measured precisely. For example, education level could be measured as completion of grade school, some high school, high school, some graduate school, and so on. Although these can be ordered from the lowest to the highest level, the difference between each level is not precise or the same.

A ***continuous variable*** is often described as a variable that can take on an infinite or large number of possible values, such as temperature, age, and weight. It can be measured on an interval or ratio scale. An ***interval scale*** is a measure that has a numerical value, and the magnitude between the intervals is the same. Temperature, date, and the time of day are examples of variables measured on an interval scale. A ***ratio scale*** is a measure that is the same as an interval measure, but it also has a true zero value or a complete absence of what is being measured. For example, when income or sales is zero, it means that there is no income or sales. For interval measures, however, you can't say that there is a zero time of day or a zero date.

Many questionnaires use a 5-point Likert-type scale with the categories of "strongly agree," "agree," "neutral," "disagree," and "strongly disagree." This is considered an ordinal scale since the responses can be ranked from the lowest to the highest. Researchers sometimes consider this an interval scale, which assumes that the distance is equal between each category. This practice is controversial since the distance between each category may not be identical, and it is subject to the interpretation of each

respondent. Some statistical guides suggest that a higher number of categories (11 or more) would be sufficient to consider it an interval scale. The reason why this is important is because you can't calculate the mean or variance of an ordinal scale, but you can with an interval scale. If you can calculate the mean and the variance, it allows you to do more in-depth statistical tests.

The types of descriptive statistics that can be calculated for each type of variable are described in the following sections, followed by a section on using graphs to describe data.

6.3 DESCRIPTIVE STATISTICS FOR ALL TYPES OF VARIABLES: FREQUENCY TABLES AND MODES

All variables, regardless of their scale of measurement, can be examined with a frequency table or the mode. Each of these is described below.

6.3.1 Frequency Tables

When using a new data set, researchers often begin by generating frequency tables to examine the distribution of each variable. Using the "College Scorecard April 23 - USNews" data set, we can generate a frequency table of the variable inst_type to see how many colleges fall into the three categories of private not-for-profit, public, and private for-profit using the code **tab inst_type** as shown in Figure 6.1. This can also be done using menus by clicking on the following sequence: Statistics → Summaries, tables, and tests → Frequency table → One-way table

```
. tab inst_type
```

Type of Institution	Freq.	Percent	Cum.
Public	587	39.56	39.56
Private nonprofit	866	58.36	97.91
Private for-profit	31	2.09	100.00
Total	1,484	100.00	

Figure 6.1 Frequency Table of College Types

The frequency and percent columns in Figure 6.1 show the actual number and percentage of each type of college in the data set. The cumulative column adds up the percentages from the percent column, but for a nominal variable, this doesn't make any sense. For example, you can't report that 97.91 colleges are less than private nonprofit colleges. If, on the other hand, we generated a frequency table for a continuous variable, such as the number of students enrolled in colleges, we could report that 50% of colleges have less 2,302 students.

If we wanted to generate frequency tables for multiple variables, we could use the Stata command **tab1** and then list the variables following the command. For example, we could type **tab1 region admcon7, sort** to show the percentage of colleges in each region in the country followed by a table that shows the percentage of colleges that require standardized testing as illustrated in Figure 6.2. The **sort** command

will sort the frequencies in the table from largest to smallest. Using **tab1** eliminates the need to type tab multiple times on separate lines. This is not possible using menus.

```
. tab1 region admcon7, sort

-> tabulation of region
```

REGION	Freq.	Percent	Cum.
Southeast (AL, AR, FL, GA KY, LA, MS, N	404	27.22	27.22
Mid East (DE, DC, MD, NJ NY, PA)	278	18.73	45.96
Great Lakes (IL, IN, MI, OH, WI)	235	15.84	61.79
Plaines (IA, KS, MN, MO, NE, ND, SD)	160	10.78	72.57
New England (CT, ME, MA NH RI, VT)	129	8.69	81.27
Far West (AK, CA, HI, NV, OR, WA)	128	8.63	89.89
Southwest (AZ, NM, OK, TX)	78	5.26	95.15
Rocky Mountains (CO, ID, MT, UT, WY)	34	2.29	97.44
Outlying Areas (AS, FM, GU, MH, MP, PR,	33	2.22	99.66
U.S. Service Schools	5	0.34	100.00
Total	1,484	100.00	

```
-> tabulation of admcon7
```

Admissions Test Score Policy	Freq.	Percent	Cum.
Considered but not required	772	61.32	61.32
Neither required nor recommended	177	14.06	75.38
Recommended	170	13.50	88.88
Required	140	11.12	100.00
Total	1,259	100.00	

Figure 6.2 Multiple Frequency Tables Using The Tab1 Command

Notice that in the table of admissions test score policies, the total number of colleges reporting their policy is 1,259 compared with 1,484 in the first table that lists the region of each college. If we wanted to include the number of missing values in the table, we could add "**, missing**" at the end of the command that generates a table. Generally, when a variable has missing data, it is important to identify patterns or reasons for the missing data. It could be due to respondents who refuse to answer a question or a skip pattern in a survey where only certain individuals are asked questions. For example, a survey may ask some questions only to individuals who work full time. There are many reasons why there might be missing observations and several ways to deal with missing data. A more advanced book on statistical analysis would cover these methods, and there are entire books written just about this problem. For a brief guide, Sauro (2015) suggests "7 Ways to Handle Missing Data." For a complete overview of missing data, refer to Enders (2010) and Little and Rubin (2014).

In the previous illustration, we used a categorical variable with only three categories. If we generated a frequency table for a continuous variable, such as the number of students enrolled in colleges using the variable "ugds" from the College Scorecard April 2023 - USNews data set, there would be close to 1,480 lines in our frequency table, or one for each unique value. You may want to do this if you are looking for outliers, but in general, you would never include a frequency table with a large number of possible values in a report. In this example, a table with 1,480 lines would run over several pages and provide limited useful information. A summary of the variable (mean, median, and standard deviation) or a box plot (described at the end of this chapter) would be more appropriate.

Finally, it is important to think about labeling tables with appropriate titles and sources. The title should indicate the statistic, the variable, and possibly the unit. A source should be listed at the bottom of the

table if the table is taken from another article or if you want to cite the source of the data. One rule of thumb is that a table should be self-explanatory without any accompanying text.

6.3.2 Mode

The *mode* is the most common value in a variable. It can be used for both categorical and continuous variables. In some cases, there may not be a mode if all values appear once or the same number of times. In other cases, you could have a bimodal or multimodal distribution whereby there is more than one mode. Although the mode is sometimes called a "measure of central tendency," the mode could be at the low or high end of the distribution of a variable.

The easiest way to find a mode is to look at a frequency table and see which value appears most frequently. When a frequency table is too long to easily find a mode or multiple modes, however, you can use the Stata **egen** command that was covered in Chapter 5. In this case it would be **egen mode = mode(ugds)** if we wanted to determine the mode for size among U.S. colleges in the College Scorecard April 23 - USNews data set. If there were multiple modes, Stata would indicate this after running the command.[1]

6.4 DESCRIPTIVE STATISTICS FOR VARIABLES MEASURED AS ORDINAL, INTERVAL, AND RATIO SCALES: MEDIAN AND PERCENTILES

6.4.1 Median

In addition to frequency tables and the mode described previously, variables that are measured on an ordinal, interval, or ratio scale can be examined with the median and percentile values. A *median* is found by ranking a variable from its lowest to its highest value and then identifying the observation or number that falls exactly in the middle. If there is an odd number of observations for the variable, then there will be one number that represents the middle value. If there is an even number of observations, then you would use the average of the two middle values.

As described later in the section on means, it is often useful to calculate both the mean and the median, particularly when there are outliers. A mean or average will be skewed in one direction if there is a small number of unusually high or low values. A full example of this is given in Section 6.5.

6.4.2 Percentiles

A ***percentile*** is a value below which a percentage of the data falls within a variable. For example, if your Scholastic Aptitude Test (SAT) percentile was 85, then 85% of all students who took the test earned a lower score than you. Because the median is the exact middle value of a variable, the median is the 50th percentile.

To find the median and several percentiles for the size of a college in the College Scorecard April 23 - USNews data set, we would use the code **sum ugds, detail** as shown in Figure 6.3, or we could do this with menus by clicking on the following sequence:

Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Summary

---

```
. sum ugds, detail

                    Undergraduate enrollment
-------------------------------------------------------------
      Percentiles      Smallest
 1%           155            27
 5%           419            32
10%         656.5            38       Obs               1,480
25%        1151.5            67       Sum of wgt.       1,480

50%          2302                     Mean            5503.583
                        Largest       Std. dev.       8560.997
75%          6027         63963
90%       14318.5         64210       Variance        7.33e+07
95%         22690         72229       Skewness        4.204682
99%         37743        119248       Kurtosis        33.74106
```

Figure 6.3 Percentiles And Median

The first two columns in Figure 6.3 show the percentile and the value for each percentile. For example, the 90th percentile is 14,318.5 indicating that 90% of all colleges have fewer than 14,318.5 students. The median, or 50th percentile, is 2,302. The third column shows the four smallest numbers in the variable and the four largest. Finally, the last column shows the number of observations and the mean, along with other descriptive statistics that we will cover later.

Because there may be outliers in the data set, it is often common to consider the interquartile range, which is the range between the 25th and the 75th percentile. In Figure 6.3, for example, we see that the four smallest values seem unrealistically small, and we may want to examine them to be sure that there wasn't a data entry error. Similarly, the largest college reported 119,248 students! The interquartile range tells us that 50 percent of all colleges fall within 1,151 and 6,027 students, which helps eliminate the extreme outliers when examining school size.

6.5 DESCRIPTIVE STATISTICS FOR CONTINUOUS VARIABLES: MEAN, VARIANCE, STANDARD DEVIATION, AND COEFFICIENT OF VARIATION

In addition to frequency tables, modes, medians, and percentiles, continuous variables (those measured on the interval and ratio scale) can be examined using the mean, variance, standard deviation, and coefficient of variation. The methods to calculate these statistics are described below.

6.5.1 Mean

The calculation of the mean or average is expressed mathematically in Equation 6.1.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(6.1)

$$X = \frac{\sum_{i=1}^{n} x_i}{n}$$

where *x* is the value of each individual observation of the variable and *n* is the number of observations or values of the variable.

For those of you not familiar with the summation sign, ∑, it is a symbol that indicates that you should add the values indicated by what lies to the right of the symbol. The symbols "*i* = 1" and "*n*" in the numerator indicate that you begin with the first observation of variable *x* and add each successive value together until you reach the last value or the *n*th unit. Finally, you divide this by *n*, or the total number of observations. For example, if there are five values of *x*, Equation 6.2 shows the long form of the equation:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

(6.2)

$$X = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

In this example, X represents a sample mean and is called a statistic. If we had information about each observation in a population, we would calculate the population mean and use the Greek letter mu μ to represent it. In this case, μ is a parameter since it describes the entire population.

Although an average is a commonly used statistic, it is often useful to calculate both the mean and the median, which was discussed in the previous section. The median is particularly useful when there are outliers or extreme values in a variable. Income is a classic example since a billionaire in the data set will skew the average to a higher value than is typical for a household. By calculating a median or middle value instead, it offers a much better picture of income for typical households in the middle of the income range.

To show the difference in the mean and median, suppose that we have a sample with five respondents and we ask their age. We receive the responses of 20, 25, 35, 40, and 95. Table 6.2 shows the results for the two measures. Although four of the five respondents are 40 years or younger, the mean indicates that the average age of the person in the sample is 43. The median, 35, on the other hand, gives a better idea of someone in the middle of the group.

TABLE 6.2 ■ Calculation of Central Tendency

Measure	Calculation	Example	Result
Mean	Sum of all observations divided by the number of observations	$\frac{20+25+35+40+95}{5}$	43
Median	Observation that falls in the middle when ranked from low to high	20, 25, 35, 40, 95	35

Measure	Calculation	Example	Result
Mean	Sum of all observations divided by the number of observations	$\frac{20+25+35+40+95}{5}$	43
Median	Observation that falls in the middle when ranked from low to high	20, 25, 35, 40, 95	35

As shown in Figure 6.3 earlier, the mean can be generated through Stata using the **summarize** command. Also, if you want to display less information than in the **summarize variable**, **detail** command, you can use the **summarize** command without the **detail** option. This is illustrated in Figure 6.4.

```
. sum ugds
```

Variable	Obs	Mean	Std. dev.	Min	Max
ugds	1,480	5503.583	8560.997	27	119248

Figure 6.4 Mean Of College Size

There are many times when a researcher may want to examine the mean or the median for subgroups. For example, we can generate the average and median college tuition cost for three categories of college types in Figure 6.5 by using the code **table inst_type, stat(mean costt4_a) stat(median costt4_a) nformat(%6.0fc)** as illustrated in Figure 6.5.

```
. table inst_type, stat(mean costt4_a) ///
> stat(median costt4_a) nformat(%6.0fc)
```

	Mean	Median
Type of Institution		
Public	27,019	23,095
Private nonprofit	27,857	23,470
Private for-profit	36,758	33,115
Total	27,637	23,310

Figure 6.5 Means And Medians For Subcategories

The three forward slashes "///" in the command are used in a do-file to let Stata know that the command continues on the next line and can only be used in do-files. The last command, nformat(%6.0fc), indicates that you want to format the numeric output so that there are six characters in total (including the comma) and zero numbers to the right of the decimal point. The "fc" stands for fixed numeric variable with commas. If you wanted to add two numbers to the right of the decimal place, the command would be **nformat(%9.2fc)**. Notice that the "6" changes to "9" to allow six characters to the left of the decimal place, the decimal point, and two digits to the right of the decimal place. Using menus to generate the same result, you would click on the sequence: Statistics → Summaries, tables, and tests → Other tablese → Flexible table of summary statistics

Overall, the output shows that private nonprofit colleges have the highest average and median tuition costs.

6.5.2 Variance and Standard Deviation

Variance is a measure of how spread out the values of one variable are from their mean. Table 6.3 shows the formulas to calculate the variance for a population and the variance for a sample of a population.

TABLE 6.3 ■ Population and Sample Variance Formulas

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ where N = number of units in the population x = value of each individual observation of the variable μ = the population mean	$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ where n = number of units in the sample x = value of each individual observation of the variable $\bar{X}$ = the sample mean

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ where N = number of units in the population x = value of each individual observation of the variable μ = the population mean	$s^2 = \frac{\sum_{i=1}^{n}(X_i-X)^2}{n-1}$ where n = number of units in the sample x = value of each individual observation of the variable X = the sample mean

In the numerator for both measures, the average is subtracted from each value in the variable. The differences are then squared and added together. In other words, it is measuring how far each value falls from the mean. Notice the difference in the denominators. For the population variance, the denominator is the number of units in the population. When you work with a sample, you are estimating the variation in the population. Because the sample will not be a perfect representation of the population, the measure adjusts for this difference by dividing by "n − 1." The ***standard deviation*** is simply the square root of the variance.

To show how the variance and standard deviation work, let's suppose that we have two variables with three observations that represent a sample of a population. Table 6.4 shows the three observations for each variable, the calculation of the variance, and the resulting variance and standard deviation. As you can see, the first variable is made of three observations that are identical: 50, 50, and 50. There is no variance in these numbers, and the resulting variance and standard deviation are zero. In Variable B, however, there is a large variation in the three observations and thus a very large variance of 2,500 and a standard deviation of 50.

TABLE 6.4 ■ Calculation Of Variance And Standard Deviation

Variable	Observations	Variance Calculation	Variance	Standard Deviation
A	50, 50, 50	$\frac{(50-50)^2+(50-50)^2+(50-50)^2}{3-1}$	0	0
B	0, 50, 100	$\frac{(0-50)^2+(50-50)^2+(100-50)^2}{3-1}$	2,500	50

Variable	Observations	Variance Calculation	Variance	Standard Deviation
A	50, 50, 50	$\frac{(50-50)^2+(50-50)^2+(50-50)^2}{3-1}$	0	0
B	0, 50, 100	$\frac{(0-50)^2+(50-50)^2+(100-50)^2}{3-1}$	2,500	50

Using the College Scorecard April 23 – USNews data set, we can examine college debt with the Stata command **table inst_type, stat(mean grad_debt_mdn) stat(sd grad_debt_mdn)** or using the following sequence in the menus: Statistics → Summaries, tables, and tests → Other tables → Flexible table of summary statistics

The results in Figure 6.6 show the mean and standard deviation for colleges in the three categories of public, private nonprofit, and private for-profit universities. As you can see from the output, the debt at private for-profit universities has the highest median student debt upon graduation, but the private nonprofit universities have a higher standard deviation of the median debt upon graduation. To compare how much they vary relative to their mean, the coefficient of variation is often used.

```
. table inst_type, stat(mean grad_debt_mdn) ///
> stat(sd grad_debt_mdn) nformat(%6.0fc)
```

	Mean	Standard deviation
Type of Institution		
Public	15,459	7,912
Private nonprofit	16,369	8,022
Private for-profit	16,724	7,613
Total	15,999	7,976

Figure 6.6 Mean And Standard Deviation Of College Debt By Type Of Institution

6.5.3 Coefficient of Variation

The ***coefficient of variation***, or CV, is calculated as the standard deviation divided by the absolute value of the mean and multiplied by 100 as shown in Equation 6.3. In other words, it tells us how much variation there is in a variable relative to its mean. Using the data from Figure 6.6, the CVs would be 51, 49, and 46 for public, private nonprofit, and private for-profit universities, respectively. Thus, we could say that the standard deviation for public universities is 51% of its mean and has the largest variation among the three categories.

$$CV = \frac{S}{|\overline{X}|} * 100$$

(6.3)

$$CV = \frac{s}{|X|} * 100$$

In a recent news story about fantasy basketball, the CV of basketball players' performance (points per game) is compared to see which players are "safer." In other words, a low CV would imply that a player consistently scores close to his or her average, whereas a high CV would suggest that the player's points per game vary widely (Daily Fantasy Sports Rankings, 2018).

In general, the CV is useful because the size of the standard deviation depends on the units used to measure a variable. For example, if a variable that asked for someone's age is measured both in years (e.g., 4 years and 2 months old) and in total months (50 months), the standard deviation will differ as illustrated in Table 6.5. Notice that the standard deviation for age in months is exactly 12 times the standard deviation for the age in years. If you just looked at the standard deviation, it would look like age in months has much greater variation than age in years. But the CVs show that they have the exact same variation relative to their mean. The CV is also useful since it allows you to compare two variables with different measurements, such as years of education and income in dollars, to determine which one has greater variation relative to its mean.

TABLE 6.5 ■ A Comparison of the Standard Deviation and Coefficient of Variation

	Age in Years	Age in Months
Observations	2	24
	4	48
	6	72
Mean	4	48
Standard deviation	2	24
Coefficient of variation	0.5	0.5

	Age in Years	Age in Months
Observations	2	24
	4	48
	6	72
Mean	4	48

	Age in Years	Age in Months
Standard deviation	2	24
Coefficient of variation	0.5	0.5

6.6 DESCRIPTIVE STATISTICS FOR CATEGORICAL VARIABLES MEASURED ON A NOMINAL OR ORDINAL SCALE: CROSS TABULATION

In addition to frequency tables and modes, categorical variables can be examined with cross-tabulation, which is also referred to as a crosstab or a contingency table. This is defined and illustrated next.

A ***cross-tabulation*** allows you to combine two categorical variables to learn more about their relationship or their joint distribution. For example, to show the percentage of colleges that require or recommend the SAT by each type of college as illustrated in Figure 6.7, we would use the commands **tab inst_type admon7, row** or the following sequence if using menus: Statistics → Summaries, tables, and tests → Frequency table → Two-way table with measures of association

```
. tab inst_type admcon7, row
```

Key
frequency *row percentage*

Type of Institution	Admissions Test Score Policy				Total
	Required	Recommend	Neither r	Considere	
Public	86	48	59	250	443
	19.41	10.84	13.32	56.43	100.00
Private nonprofit	54	116	115	518	803
	6.72	14.45	14.32	64.51	100.00
Private for-profit	0	6	3	4	13
	0.00	46.15	23.08	30.77	100.00
Total	140	170	177	772	1,259
	11.12	13.50	14.06	61.32	100.00

Figure 6.7 Combining Two Categorical Variables Using The Tabulate Command

Within each cell, the numbers on top are the actual number of colleges in that category and the number below is the percentage. By including the command **row** in the Stata command, the percentages add up across the rows to 100%. For example, we can see on the first row that 59 public universities neither recommend nor require the SAT or ACT, which represents 13.32% of all public universities.

If we changed Figure 6.7 and wrote the Stata commands to add up over the columns, we would use the commands **tab inst_type admcon7, col**. The output from these commands would not make sense, as shown in Figure 6.8. Notice that the private nonprofit universities represent the largest percentage in each column, with the exception of the first column. This is because they are the largest group. Instead, you would want to know what percentage of colleges within each type require or do not require the SAT or ACT.

```
. tab inst_type admcon7, col
```

Key
frequency *column percentage*

Type of Institution	Admissions Test Score Policy Required	Recommend	Neither r	Considere	Total
Public	86 61.43	48 28.24	59 33.33	250 32.38	443 35.19
Private nonprofit	54 38.57	116 68.24	115 64.97	518 67.10	803 63.78
Private for-profit	0 0.00	6 3.53	3 1.69	4 0.52	13 1.03
Total	140 100.00	170 100.00	177 100.00	772 100.00	1,259 100.00

These numbers only tell us private not-for-profit schools represent the largest group in the sample and nothing more.

Figure 6.8 Incorrectly Adding Up A Cross Tabulation Over The Wrong Variable

There are times when you may want to examine both row and column percentages for the same set of two variables. In many cases, however, you would only examine percentages that add up across the independent variable. As described in Chapter 1, an ***independent variable*** is defined as a variable whose variation is not influenced by other variables, whereas a dependent variable is influenced by other variables. In this example, large public universities are likely to require the SAT, in contrast to small liberal arts colleges that may not. So the SAT policy (dependent variable) is dependent on the type of college (independent variable). We would therefore add up across the type of college. So if the type of college is the row variable, we would generate a row percentage. If the type of college was placed in the columns, we would generate a column percentage.

A second rule of thumb that is sometimes followed is to place the dependent variable in the rows and the independent variable in the columns. If, however, there are far more categories in the independent variable and the table can't fit on a page without wrapping around, it is fine to switch the placement of the two types of variables but always add up over the independent variable.

One last word on cross tabulation is that you can use a continuous variable in a cross-tabulation if you transform it into categories using the **recode** command discussed in Chapter 5. For example, if one of your continuous variables is income in exact dollar amounts, you could change this into a categorical variable by making the categories of below 50,000, 50,000 to 99,999.99, 100,000 and above, and so on.

6.7 APPLYING SAMPLING WEIGHTS

In Chapter 2, we described how sampling weights are used to extrapolate information from the sample to the population. If a sample is stratified or is drawn by using a multistage process, then some types of units will be overrepresented and others underrepresented in the sample. Returning to the example from Chapter 2, suppose that the population is 10% urban and 90% rural, but the sample is split 50–50

between urban and rural households. This means that urban households are overrepresented in the sample, so any statistics calculated from the sample will be disproportionately affected by the urban households. Sampling weights compensate for the distortions introduced by sampling, allowing us to calculate means and percentages from the sample that are unbiased estimates of the population statistics.

We demonstrate the use of sampling weights with data from the 2021 General Social Survey (GSS). In particular, we can explore the relationship between health (health) and happiness (happy). Without weights, we would use the commands **tab health happy, row nofreq**. With weights, we would change the code to **tab health happy [aw=wtssps], row nofreq**, where "aw" stands for analytical weights and "wtssps" is the name of the variable that contains the weights for each observation and "nofreq" means that we only want to see percentages and not the frequencies or counts.

First, comparing Figure 6.9 with no weights and Figure 6.10 with weights shows that the weights do not have a large impact on the results in this data set. This suggests that any over- or undersampling of different groups in the United States was only minor. Nonetheless, if you are generating descriptive statistics from a survey for which sampling weights are available, you should make use of those weights.

```
. tab health happy, row nofreq
```

CONDITION OF HEALTH	GENERAL HAPPINESS very happ	pretty ha	not too h	Total
excellent	38.19	48.55	13.25	100.00
good	17.33	62.50	20.17	100.00
fair	8.31	56.49	35.19	100.00
poor	7.33	35.33	57.33	100.00
Total	19.55	57.44	23.02	100.00

Figure 6.9 Relationship Between Health And Happiness Without Sampling Weights

```
. tab health happy [aw=wtssps], row nofreq
```

CONDITION OF HEALTH	GENERAL HAPPINESS very happ	pretty ha	not too h	Total
excellent	38.84	48.59	12.58	100.00
good	17.44	61.60	20.96	100.00
fair	8.81	55.16	36.03	100.00
poor	7.12	35.90	56.97	100.00
Total	19.41	56.73	23.86	100.00

Figure 6.10 Relationship Between Health And Happiness With Sampling Weights

The output in both tables also shows that there appears to be a correlation between health and happiness. As illustrated in Figure 6.10, among people with poor health, the largest percentage report that they are not too happy (57%) compared to respondents in excellent health, where only 13% report being not too happy.

6.8 FORMATTING OUTPUT FOR USE IN A DOCUMENT (WORD, GOOGLE DOCS, ETC.)

As you will notice in the output in the previous sections, the formatting of the tables is not always ideal. The value labels start with lowercase letters, some labels are too long and are cut off, other labels have abbreviations that might not be clear to the reader, and so on. For this reason, you may want to edit the table before placing it in a Word or Google document. In addition, if you are publishing your work in a journal, many journals prefer only horizontal lines. All of this editing can be done by highlighting the table in the Stata results screen and then clicking edit/copy table. You can then copy this into an Excel file where you could format each part of the table as needed. Finally, the Excel table can be copied into Word or a Google Doc. To do this, you need to select the table, right-click, and then select “copy table.”

6.9 GRAPHS TO DESCRIBE DATA

In addition to tables with data, charts or graphs are often useful to display information. In some cases, it is easier to see a pattern with a graph. Although there are many types of graphs, we illustrate four of the most common—bar graphs, box plots, histograms, and pie charts.

6.9.1 Bar Graphs

Using the same data from the previous sections, College Scorecard April 23 – USNews, we can generate a bar graph of the average tuition rate by the type of university (Figure 6.11) using the Stata command or following this sequence using the menus: Graphics → Bar chart → Other tables → Flexible table of summary statistics. Notice that the Stata command asks for a bar graph of the mean value of the continuous variable spread out “over” the type of institution.

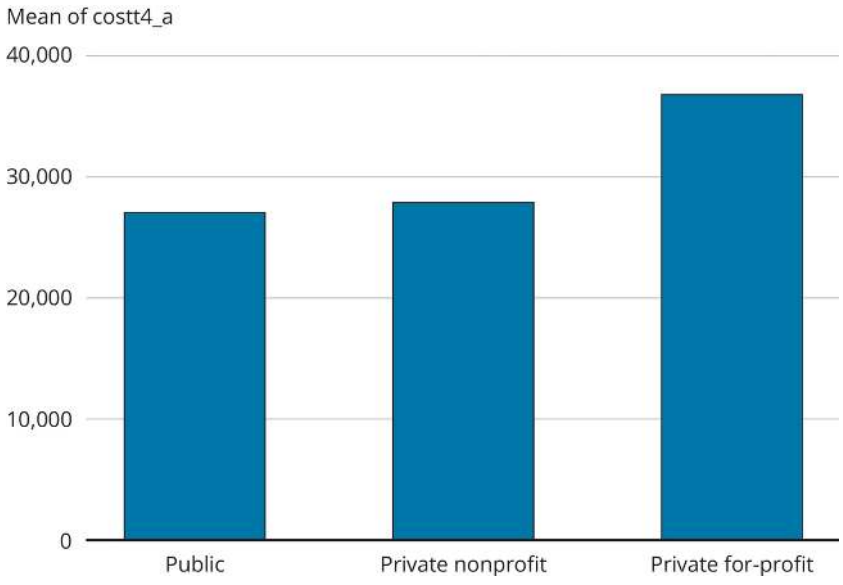


Figure 6.11 Bar Graph of Average Tuition By Type of College

```
graph bar (mean) costt4_a, over(inst_type)
```

6.9.2 Box Plots

We can also use a box plot for the same data displayed in the bar graph (Figure 6.12). With a box plot, in addition to comparing means, we can see the dispersion of a variable. The Stata command to create a box plot and the output are illustrated next, or we could use the following sequence in the menus: Graphics → Box plot

```
graph box costt4_a, over(inst_type)
```

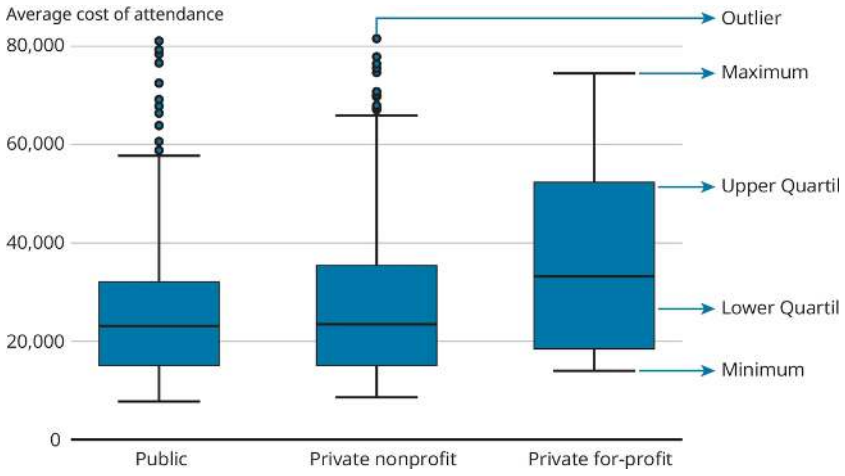


Figure 6.12 Box Plot of Average Tuition by Type of Institution

The line inside the shaded box represents the median value. The upper and lower borders of the box represent the upper and lower quartiles. In other words, the upper border is the 75th percentile and the lower border is the 25th percentile so that 50% of the observations fall within the range represented by the box. The “whiskers” or horizontal lines at the top and bottom of the graph extend out to the last value that is less than or equal to 1.5 times the interquartile range value. Finally, the outliers are extreme values that fall outside of 1.5 times the interquartile range value.

From the box plot, it is easy to see that the private for-profit institutions have the highest median value, and the public universities have the largest spread of extreme outliers.

6.9.3 Histograms

While a bar graph or box plot is useful for a limited number of categories, as in Figures 6.11 and 6.12, a histogram is a better choice for a continuous variable with numerous values. For example, the median debt among college students who graduate, which is a continuous variable, can be illustrated as a histogram (Figure 6.13). The Stata command to generate a histogram is shown next, or this can be done using the menus with the sequence Graphics → Histogram. The commands illustrated in the Stata code ask Stata to generate a histogram of the continuous variable. The **bin(10)** command lets Stata know to use 10 bars, and the term **frequency** indicates that the vertical axis should show the number of times or frequency that the range of values represented by the bar appears.

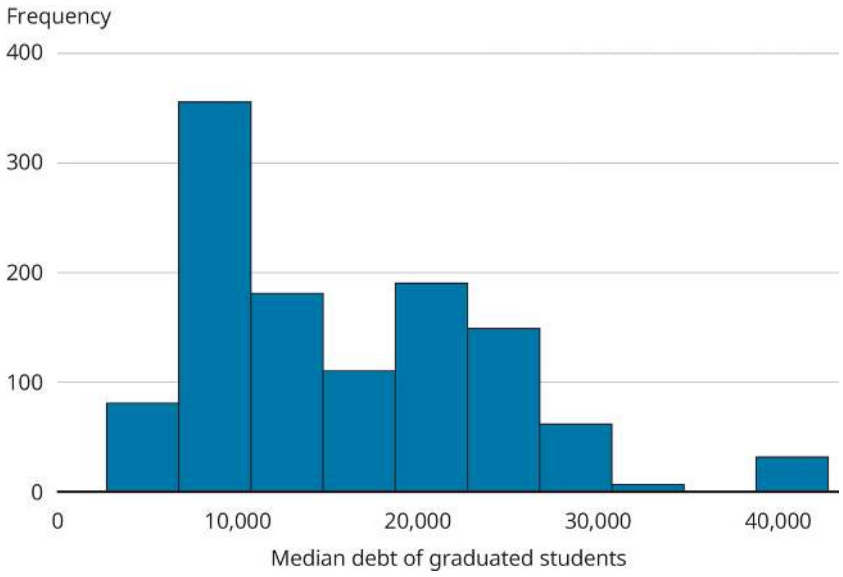


Figure 6.13 Histogram of the Median Debt Owed by College Graduates Based on the College Scorecard Data From April 23 – Usnews

```
hist grad_debt_mdn, bin(10) frequency
```

The frequency on the vertical axis shows the number students who report each level of debt. Then, looking at the widest of the bars, you can see how many students fall into each range. For example, around 350 students fall in the highest range of roughly $9,000.

6.9.4 Pie Charts

A pie chart is useful for a categorical variable with a limited number of categories where only one category can be selected by the respondent. For example, using the test score requirements of colleges, we can make a pie chart (Figure 6.14) that gives a visual example of the percentage of colleges that fall into each test score category. If we choose a variable with 40 possible responses (e.g., a student's major), then each slice of the pie would be too small.

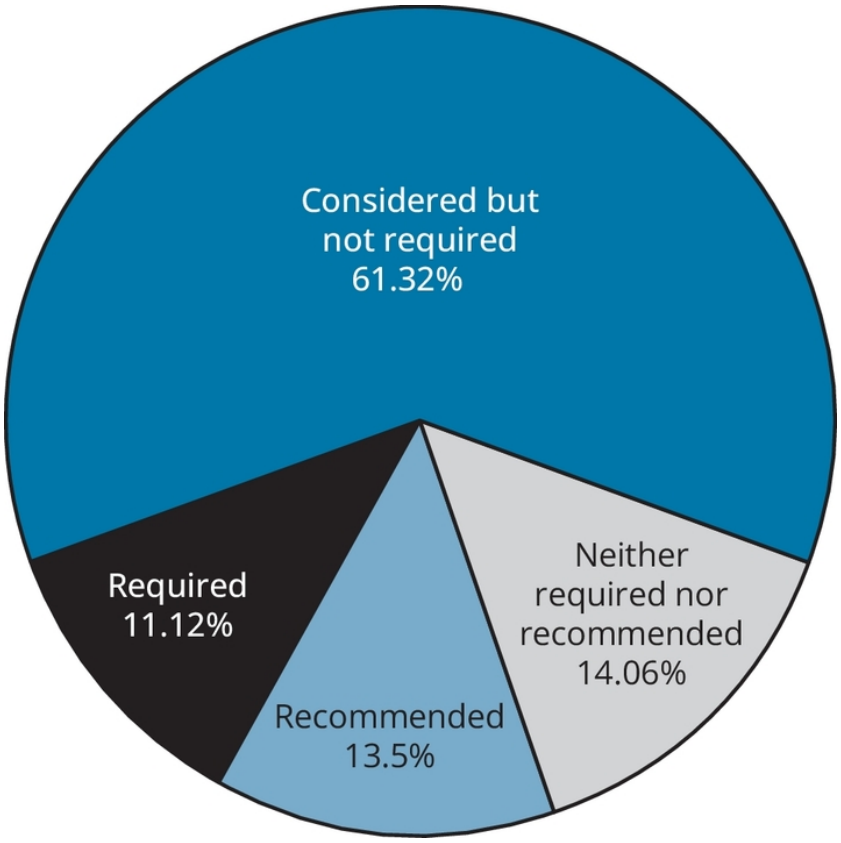


Figure 6.14 Pie Chart of College test Score Policies

To generate a pie chart, we would run the Stata command that follows or use the following sequence in the menus: Graphics → Pie chart. The command **p label(all percent)** indicates that Stata should include the percentage inside of each pie slice.

```
graph pie, over(admcon7) plabel(_all percent)
```

6.10 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes all of the Stata code used in the chapter (Table 6.6). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 6.6 ■ Code Used In Chapter 6

Function	Code
Frequency table	tab inst_type, sort tab1 region admcon7, sort
Mode	egen mode = mode(ugds) egen mode1 = mode(ugds), nummode(1) egen mode2 = mode(ugds), nummode(2)
Summary table	sum ugds, detail

Tables with mean, median, and standard deviation	table inst_type, stat(mean costt4_a) ///
	stat(median costt4_a) nformat(%6.0fc)
	table inst_type, stat(mean grad_debt_mdn) ///
	stat(sd grad_debt_mdn) nformat(%6.0fc)

Tables with row percentages	tab inst_type admon7, row
Tables with and without sample weights	tab health happy, row nofreq tab health happy [aw=wtss], row nofreq
Bar graphs, box plots, histograms, and pie charts	graph bar (mean) costt4_a, over(inst_type) graph box costt4_a, over(inst_type) hist grad_debt_mdn, bin(10) frequency graph pie, over(admcon7) plabel(_all percent)

Function	Code
Frequency table	tab inst_type, sort tab1 region admcon7, sort
Mode	egen mode = mode(ugds) egen mode1 = mode(ugds), nummode(1) egen mode2 = mode(ugds), nummode(2)
Summary table	sum ugds, detail

Tables with mean, median, and standard deviation	table inst_type, stat(mean costt4_a) ///
	stat(median costt4_a) nformat(%6.0fc)
	table inst_type, stat(mean grad_debt_mdn) ///
	stat(sd grad_debt_mdn) nformat(%6.0fc)

Tables with row percentages	tab inst_type admon7, row
Tables with and without sample weights	tab health happy, row nofreq tab health happy [aw=wtss], row nofreq

Function	Code
Bar graphs, box plots, histograms, and pie charts	graph bar (mean) costt4_a, over(inst_type) graph box costt4_a, over(inst_type) hist grad_debt_mdn, bin(10) frequency graph pie, over(admcon7) plabel(_all percent)

EXERCISES

1. For each of the following variables, indicate the type of variable (categorical or continuous) and its level of measurement (nominal, ordinal, interval, or ratio).
    a. Favorite type of cereal
    b. Car prices
    c. Total profits
    d. Level of happiness
    e. Birth date
    f. Time of birth
    g. Gender
2. Using the data set that you created from Exercise 1 in Chapter 4 (about binge-watching television), follow the instructions below.
    a. Identify both the variable type and scale of measurement for each of the four variables (TV source, Hours per week, Binge frequency, and Gender) in your data set.
    b. What is the mode for “Binge frequency”? Show your Stata command and output as part of this and for all of the following answers.
    c. What is the 25th percentile value for “Hours per week”? Explain what this means in words.
    d. What is the variance for “Hours per week”?
    e. Make a table that shows “Gender” in the rows and the mean and median of “Hours per week” in the columns. Format the table so that there are no numbers to the right of the decimal point. In other words, use only whole numbers.
    f. Calculate the coefficient of variation for “Hours per week” (use a calculator for this after obtaining the numbers that you need).
    g. Generate a cross-tabulation of “Gender” and “Binge frequency.” Be sure to think about whether the rows or columns should add up to 100%. Based on your table, what percentage of women binge watch frequently, and what percentage of men binge watch frequently?
    h. Generate a bar chart that shows the average “Hours per week” that respondents binge watch TV by gender.
    i. Generate a histogram of “Hours per week.”
    j. Generate a pie chart of “Binge frequency.” Label each slice of the pie with the percentage value.
3. Suppose there is a population of five people with height in inches as follows: 58, 62, 63, 70, and 77.
4. Calculate the population variance using a calculator. Show your work to derive the final answer.
5. Now suppose that you take a sample of three of these people who are 62, 63, and 77 inches tall. Calculate the sample variance using a calculator. Show your work to derive the final answer.
6. Using the GSS2021.dta file, answer the following questions related to political party affiliation and attitudes about gun permits.
    a. Generate a table without weights that shows the political party affiliation (partyid) in the rows and whether the respondent favors or opposes gun permits in the columns (gunlaw). Show only the percentages and decide whether to use row or column percentages.
    b. Generate the same table as in Part “a,” but apply the weights (wtssps) to the table.

7. Using the College Scorecard April 23 – USNews data set to examine annual salaries six years after graduation.
    a. Generate a table that shows the type of college (USNewsType) and the average annual median salary six years after graduation for males (md_earn_wne_male1_p6) and the average annual median salary six years after graduation for non-males (md_wne_male_p6). Format the table so that there are no digits to the right of the decimal place.
    b. Generate a box plot of average annual median salaries of non-males six years after graduation (md_earn_wne_male0_p6) by the type of university (USNewsType) by adding the commands “, over (USNewsType) at the end of your line of code. You will see that the *x*-axis labels run together. To fix this, click on the Graph Editor icon at the top of your graph. Then click directly on the *x*-axis labels. This will bring up a dialogue box called “Axis properties.” Click on label properties, and then choose 45 degrees in the “angle” box. Then click on “Ok.”

KEY TERMS

categorical variable

coefficient of variation

continuous variable

cross-tabulation

independent variable

interval scale

nominal scale

open-ended questions

ordinal scale

percentile

ratio scale

standard deviation

variable

variance

PART III TESTING HYPOTHESES

In Parts III and IV of this book, each chapter involves a research question, a null hypothesis, and a statistical test. A summary of these for each chapter is offered in Appendix 2. This summary should help you to quickly identify what type of statistical procedure or test should be used and the procedures to implement the test.

7 THE NORMAL DISTRIBUTION, HYPOTHESIS TESTING, AND STATISTICAL SIGNIFICANCE

CHAPTER PREVIEW

Steps	Example
Research question	Did 50 students who took an SAT preparatory course earn significantly higher scores on math SAT tests compared with the other students at the same high school?
Null hypothesis	There is no difference in SAT scores among those students who took a preparatory course and those who did not.
Test	Standard score or z score
When to use	Comparing a sample mean with a population mean. The population standard deviation is known.
Calculate the standard error of the mean	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
Calculate the standard or z score	$Z = \frac{(X_i - \mu)}{\sigma_{\bar{X}}}$
Compare the p value to the p critical	Use a "z score to percentile" calculator or z table.

Steps	**Example**
Research question	Did 50 students who took an SAT preparatory course earn significantly higher scores on math SAT tests compared with the other students at the same high school?
Null hypothesis	There is no difference in SAT scores among those students who took a preparatory course and those who did not.
Test	Standard score or *z* score
When to use	Comparing a sample mean with a population mean. The population standard deviation is known.
Calculate the standard error of the mean	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ $\sigma_X = \frac{}{\sqrt{n}}$

Steps	Example
Calculate the standard or *z* score	$Z = \frac{(X_i - \mu)}{\sigma_X}$
Compare the *p* value to the *p* critical	Use a "*z* score to percentile" calculator or *z* table.

7.1 INTRODUCTION

Many high school students in the United States take the Scholastic Aptitude Test (SAT) as part of the college admissions process. In addition to their individual score on a scale of 400 to 1,600, students are told their percentile or rank that allows them to compare their score with other test takers. High schools also use the SAT scores to compare their own students with national standards. In fact, a whole industry has evolved to help students improve their scores. Some organizations offer free online courses, while others offer fee-based classes that can run as high as $2,000. But do they make a difference? Are they worth it? Given the trend toward test-optional admissions policies illustrated in Figure 7.1, many people may now question the value of an SAT preparatory course.

More colleges than ever have test-optional admissions policies — and that's a good thing

Published: January 10, 2018 8.21pm EST

The number of colleges and universities with test-optional admissions policies recently topped 1,000 – a milestone that one expert says is a welcome trend.

Email
Twitter 49
Facebook 483
LinkedIn
Print

Back in the 1980s, Bates College and Bowdoin College were nearly the only liberal arts colleges not to require applicants to submit SAT or ACT test scores.

In 2018, FairTest, a Boston-based organization that has been pushing back against America's testing regime since 1985, announced that the number of colleges that are test-optional has now surpassed 1,000.

This milestone means that more than one-third of America's four-year nonprofit colleges now reject the idea that a test score should strongly determine a student's future. The ranks of test-optional institutions include hundreds of prestigious private institutions, such as George Washington, New York University, Wesleyan University and Wake Forest University. The list also includes hundreds of public universities, such as George Mason, San Francisco State and Old Dominion.

Description

Figure 7.1 Article

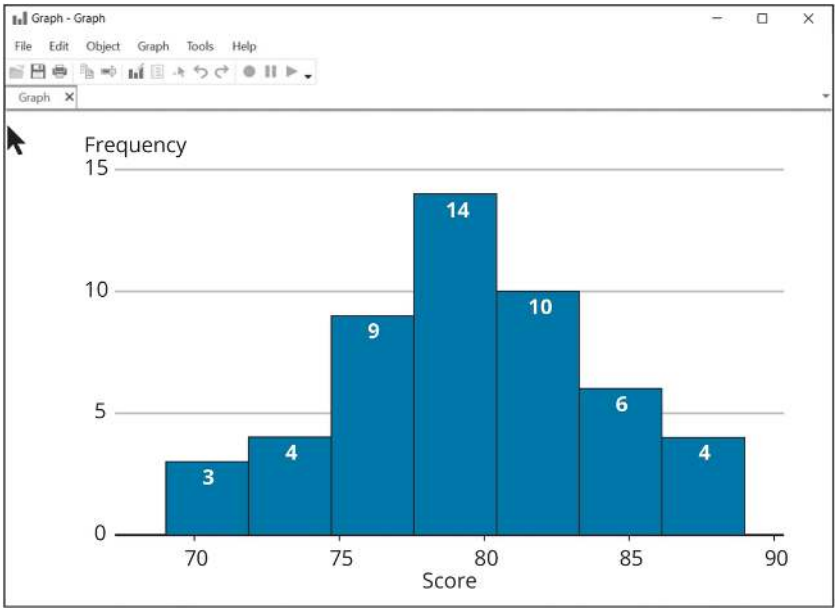
Source: More colleges than ever have test-optional admissions policies – and that's a good thing. The Conversation. January 10, 2018. https://theconversation.com/more-colleges-than-ever-have-test-optional-admissions-policies-and-thats-a-good-thing-89852

In this chapter, we will learn how to determine when something is unusually different or statistically significant. We will start by looking at exam scores and learn how a student can determine his or her rank within a class. We will then learn about sampling distributions and the standard error of the mean. Finally, we will use these concepts to test whether the average SAT scores among students who took a preparatory course are significantly different from the scores of students who did not take a preparatory course. We will then briefly turn to comparing a sample proportion to a population proportion. Rather than using descriptive statistics as we did in Chapter 6, we are now turning to *inferential* statistics, whereby we are making inferences about a population based on a sample.

7.2 THE NORMAL DISTRIBUTION AND STANDARD SCORES

Suppose you receive an exam score of 60. You may be disappointed until you learn that you earned the highest score. Alternatively, you may earn an 85 and be quite satisfied until you learn that the average was a 95. If the professor announces the average score, you only know how you did relative to the average. If, however, the professor tells you the standard deviation, you can learn what percentage of students did just as well or better. We will use the exam.dta[1] data set and the normal distribution to learn how to determine this information.

The exam.dta data set shows the exam scores for 50 students in a statistics course. The mean score for the exam is 80, and the standard deviation is 5. Figure 7.2 shows a histogram of the variable "grade," which we learned how to generate in Chapter 6 using the command **hist**. We also learned in Chapter 6 that a histogram shows ranges of values on the horizontal axis and the number of times they appear, or the proportion of each set of values, on the vertical axis. In this case, the shape of the histogram looks like a bell curve. Many continuous random variables exhibit this shape with most values clustering around the mean and fewer observations in the extremes or in the tails of the bell curve. This is known as a ***normal distribution***, and it is one of the most important concepts in statistics. It is important because not only do so many variables follow this distribution, but also we use the normal distribution to draw conclusions about the characteristics of a population based on a sample even when the underlying distribution is not normal (see Section 7.8 on the Central Limit Theorem).



Description

Figure 7.2 Histogram Of Test Scores

Figure 7.3 shows the normal distribution, which exhibits perfect symmetry. Exactly half of the area under the curve falls on each side of the mean value. As illustrated, we can also see what percentage of the area falls within 1, 2, and 3 standard deviations of the mean. Using our data from the exam data set, the second line of numbers shows 80 as the mean score and increments of 5 on either side since the standard deviation is 5. In other words, we can say that 68% of students scored between 75 and 85, or within 1 standard deviation of the mean, 95% scored within 2 standard deviations (70–90), and 99% scored within 3 standard deviations (65–95) of the mean.

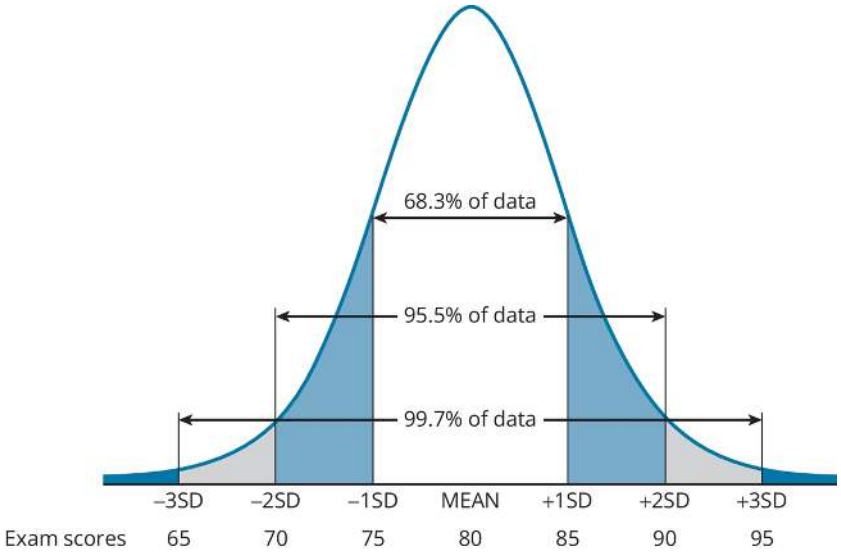


Figure 7.3 The Normal Distribution And Student Exam Scores

We can convert each number in our data set to a standard score, which is also known as a ***z score***. Essentially, every student's grade can be expressed as the number of standard deviations that it deviates from the mean. The standard score is expressed in Equation 7.1.

$$Standard\ score\ or\ z\ score = \frac{(X_i - \mu)}{\sigma}$$

(7.1)

$$Standard\ score\ or\ z\ score = \frac{(X_i - \mu)}{\sigma}$$

where

X_i = the value of one individual's score

μ = the average value of the variable X

σ = the standard deviation of the variable X

The numerator shows the difference between one student's score and the mean of all scores. If you earned an 85 and the class average was 80, for example, the numerator would be 5 points above the class average. The denominator is the value 1 standard deviation. Dividing how much your score differed from the mean by the standard deviation tells you how many standard deviations your score is above or below the mean. In this case, positive 5 divided by a standard deviation of 5 says that you are 1 standard deviation above the mean.

The next step to determine how many students did just as well or better is to find out how much of the area under the normal curve is to the right of 1 standard deviation. As we saw in Figure 7.3, 68% of the

area under a normal curve lies within 1 standard deviation (−1 to +1) of the mean, or 34%, lies between the mean and 1 standard deviation on either side of the mean. Since 50% of the area under the curve lies to the right of the mean, we can subtract 34% from 50% to determine that 16% falls to the right of 1 standard deviation. We can then say that 16% of students earned a score of 85 or higher, and 84% earned lower scores. Since 16% of 50 students is 8 students, we know that 8 students earned an equal or higher score. Because the actual distribution isn't a perfect bell curve, this can be a rough estimate. If we use the **tab** command to generate a frequency table, as we learned in Chapter 6, Figure 7.4 shows that 3 students earned an 85, and 5 students earned higher than 85.

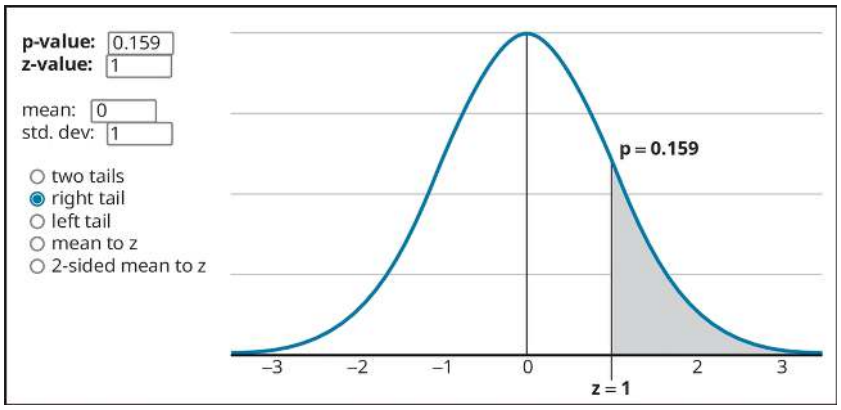
```
. tab score
```

score	Freq.	Percent	Cum.
69	1	2.00	2.00
70	1	2.00	4.00
71	1	2.00	6.00
73	2	4.00	10.00
74	2	4.00	14.00
75	2	4.00	18.00
76	4	8.00	26.00
77	3	6.00	32.00
78	6	12.00	44.00
79	3	6.00	50.00
80	5	10.00	60.00
81	3	6.00	66.00
82	2	4.00	70.00
83	5	10.00	80.00
84	2	4.00	84.00
85	3	6.00	90.00
86	1	2.00	92.00
87	1	2.00	94.00
88	2	4.00	98.00
89	1	2.00	100.00
Total	50	100.00	

Description

Figure 7.4 Frequency Distribution Of Exam Scores

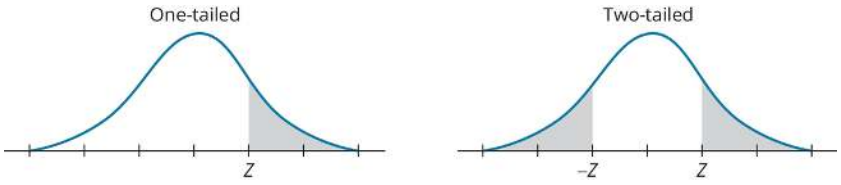
There are many online calculators that will compute the area to the right or left of a standard score. If we use the calculator at "StatDistributions" (https://www.statdistributions.com/normal), we will see the image in Figure 7.5.



Description

Figure 7.5 Area Under The Normal Curve For A Z Score Of +1

Alternatively, we could use a *z* score table, which is illustrated in Table 7.1 and included in Appendix 5. When using the table, you would look for your *z* score in the table, which is +1 in this case. Across from the +1, we see the areas for a one-tailed and a two-tailed probability. Because we only want to know how much area is above +1 and below −1, we would use the one-tailed probability. Looking at the image above the table for the one-tailed test, you can see a *z* with the area shaded to the right of the *z*. The one-tailed probability for +1 of 0.15866 represents the area that is shaded or roughly 16% of the area. Since the entire area under the curve represents 100% of the area, 100 minus 16 tells us that 84% of the area lies to the left of a *z* score of +1 (Figure 7.6).



Description

Figure 7.6 Areas Under The Normal Curve (Z Score)

TABLE 7.1 ■ Areas Under The Normal Curve (Z Score)

Z Scores	Probability	
	One-tailed	Two-tailed
0	0.50000	1.00000
0.1	0.46017	0.92034
0.2	0.42074	0.84148
0.3	0.38209	0.76418
0.4	0.34458	0.68916
0.5	0.30854	0.61708
0.6	0.27425	0.54851
0.7	0.24196	0.48393
0.8	0.21186	0.42371
0.9	0.18406	0.36812
1	0.15866	0.31731
1.1	0.13567	0.27133
1.2	0.11507	0.23014
1.3	0.09680	0.19360
1.4	0.08076	0.16151
1.5	0.06681	0.13361
1.6	0.05480	0.10960
1.7	0.04457	0.08913
1.8	0.03593	0.07186
1.9	0.02872	0.05743
2	0.02275	0.04550
2.1	0.01786	0.03573
2.2	0.01390	0.02791
2.3	0.01072	0.02145
2.4	0.00820	0.01640
2.5	0.00621	0.01242
2.6	0.00466	0.00932
2.7	0.00347	0.00693
2.8	0.00256	0.00511
2.9	0.00187	0.00373
3	0.00135	0.00270
3.1	0.00097	0.00194
3.2	0.00069	0.00137
3.3	0.00048	0.00097
3.4	0.00034	0.00067
3.5	0.00023	0.00047
3.6	0.00016	0.00032
3.7	0.00011	0.00022
3.8	0.00007	0.00014
3.9	0.00005	0.00010

Z Scores	**Probability**	
	One-tailed	**Two-tailed**
0	0.50000	1.00000
0.1	0.46017	0.92034
0.2	0.42074	0.84148
0.3	0.38209	0.76418
0.4	0.34458	0.68916
0.5	0.30854	0.61708
0.6	0.27425	0.54851
0.7	0.24196	0.48393
0.8	0.21186	0.42371
0.9	0.18406	0.36812
1	0.15866	0.31731

Z Scores	Probability	
	One-tailed	Two-tailed
1.1	0.13567	0.27133
1.2	0.11507	0.23014
1.3	0.09680	0.19360
1.4	0.08076	0.16151
1.5	0.06681	0.13361
1.6	0.05480	0.10960
1.7	0.04457	0.08913
1.8	0.03593	0.07186
1.9	0.02872	0.05743
2	0.02275	0.04550
2.1	0.01786	0.03573
2.2	0.01390	0.02781
2.3	0.01072	0.02145
2.4	0.00820	0.01640
2.5	0.00621	0.01242
2.6	0.00466	0.00932
2.7	0.00347	0.00693
2.8	0.00256	0.00511
2.9	0.00187	0.00373
3	0.00135	0.00270
3.1	0.00097	0.00194
3.2	0.00069	0.00137
3.3	0.00048	0.00097
3.4	0.00034	0.00067
3.5	0.00023	0.00047
3.6	0.00016	0.00032
3.7	0.00011	0.00022
3.8	0.00007	0.00014
3.9	0.00005	0.00010

If the *z* score is negative, we can still use the table since a normal distribution is perfectly symmetric. As described above, exactly half of the area under the curve falls on each side of the mean value. If a student scored a 75, the class average is 80, and the standard deviation is 5, their *z* score would be ((75 − 80)/5) or a negative one. We can see from the table that the area to the right of +1 is 0.15866, and therefore, we know that the area to the left of −1 is also 0.15866. Since the total area is equal to 1, 1 minus 0.15866 is 0.84134. In this case, roughly 16% of the students earned lower scores, and 84% earned equal or better scores.

Now let's suppose that the standard deviation for the class is 10 instead of 5. In that case, a student who earned an 85 when the average was 80 would have a standard score as follows:

$$Z \text{ score or Standard score} = \frac{85 - 80}{10} = 0.5$$

$$Z \text{ score or Standard score} = \frac{85 - 80}{10} = 0.5$$

Using the *z* score table, we can see that the area to the right of 0.5 is 0.30854, or roughly 31%. Therefore, 31% of students earned an equal or higher score, compared with only 16% when the standard deviation was 5.

You can also use this same information to determine your percentile rank. If, for example, 16% earned an equal or higher score in Case A, then 84% earned a lower score. You would then be at the 84th percentile. In Case B, you would be at the 69th percentile. Table 7.2 shows the results from the two examples above.

TABLE 7.2 ■ Standard Score Examples With Exam Grades

	Case A	Case B
Class mean	80	80
Standard deviation	5	10
One student's test score	85	85
Standard score or *z* score	1	0.5
One-tail probability	0.16	0.31
Number of students in class	50	50
Number of students who earned a higher score	8	16
Percentile rank	84th percentile	69th percentile

	Case A	Case B
Class mean	80	80
Standard deviation	5	10
One student's test score	85	85
Standard score or *z* score	1	0.5
One-tail probability	0.16	0.31
Number of students in class	50	50
Number of students who earned a higher score	8	16
Percentile rank	84th percentile	69th percentile

Although Stata doesn't calculate *z* scores, a user-written program is available in Stata to do this. In the Command Window, you would type in **help zscore** and scroll down until you see "Web resources from Stata and other Users." Then click on "zscore" and then on the link provided. Finally, click on "Click here to install." Once it is installed, type in the command **zscore varname**, and Stata will create a new variable called z varname in your data set. Then, if you summarize the new variable, you will see that the mean is 0 and the standard deviation is 1.

7.3 SAMPLING DISTRIBUTIONS AND STANDARD ERRORS

In the previous section, we examined one student's score compared with the rest of the class. Using the normal distribution, we were able to see the percentile rank of the student. We can also use the normal distribution to examine how one sample mean compares with a population mean. To do this, we will first need to learn about sampling distributions and standard errors. We will define these terms later as we develop an example.

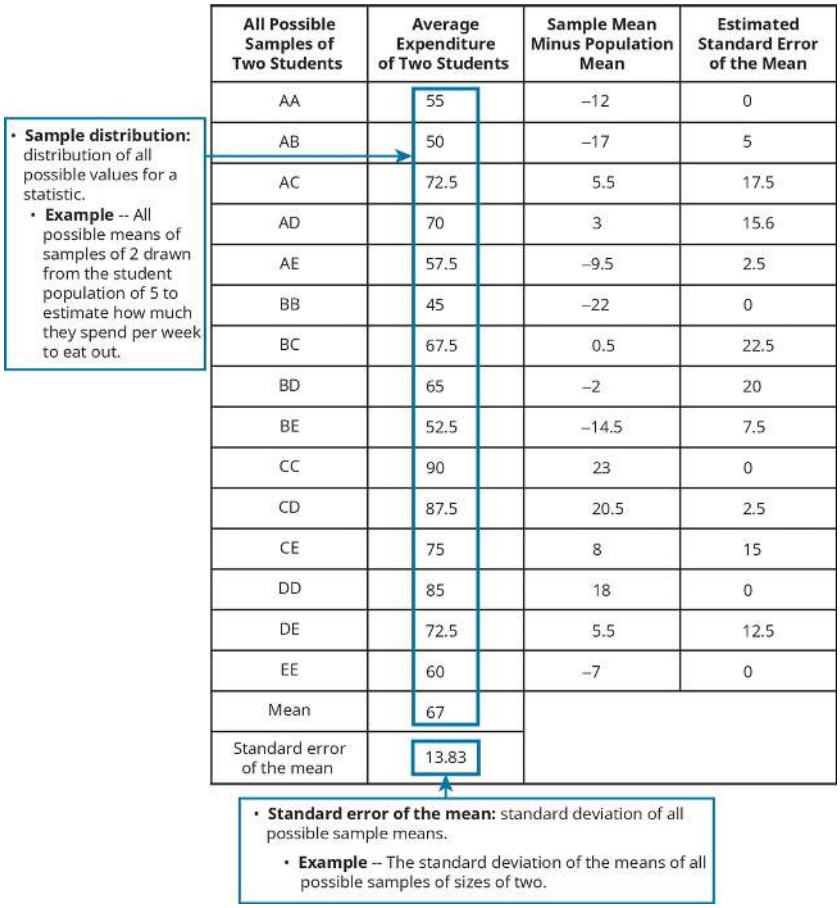
Although universities can have hundreds or thousands of students, let's suppose that only five students attend a university. The amount of money that they spend on eating out per week is shown in Table 7.3, along with the overall mean of $67 and a standard deviation of $19.60.

TABLE 7.3 ■ Weekly Expenditure On Eating Out By Five College Students

Student	Weekly Amount Spent on Eating Out in Dollars
A	55
B	45
C	90
D	85
E	60
Mean	67
Standard deviation	19.6

Student	Weekly Amount Spent on Eating Out in Dollars
A	55
B	45
C	90
D	85
E	60
Mean	67
Standard deviation	19.6

Survey designers rarely have the resources to gather information from the entire population. Instead, they take a sample to estimate the population characteristics. In this case, let's assume that we only have resources to sample two students. Figure 7.7 shows all possible combinations of two students and the average amount spent for each sample of two. Although we typically would not sample the same person twice, we have included it here to illustrate the principle of the standard error of the mean.



All Possible Samples of Two Students	Average Expenditure of Two Students	Sample Mean Minus Population Mean	Estimated Standard Error of the Mean
AA	55	–12	0
AB	50	–17	5
AC	72.5	5.5	17.5
AD	70	3	15.6
AE	57.5	–9.5	2.5
BB	45	–22	0
BC	67.5	0.5	22.5
BD	65	–2	20
BE	52.5	–14.5	7.5
CC	90	23	0
CD	87.5	20.5	2.5
CE	75	8	15
DD	85	18	0
DE	72.5	5.5	12.5
EE	60	–7	0
Mean	67		
Standard error of the mean	13.83		

Description

Figure 7.7 Sampling Distribution Of All Possible Samples Of Two Students

We know that the true mean of the population is $67. From all possible combinations of two students, we can see that the average of some samples is very close to the true mean, and others are much farther from the true mean. With larger sample sizes, we would see less variability in the sample means. Similarly, lower variation in the population values would lead to less variability in the sample means.

The distribution of all possible values for a statistic (in this case, the mean) is called a sampling distribution. When we take the standard deviation of all possible sample means, it is called the standard error of the mean, which is used extensively in statistics. Fortunately, we don't need to take all possible samples of a population to determine the standard error of the mean. Instead, we can calculate it by dividing the standard deviation of the population by the square root of the number of cases in our sample as shown in Equation 7.2.

$$Standard\ error\ of\ the\ mean = \frac{\sigma}{\sqrt{n}} = \frac{19.55671}{\sqrt{2}} = 13.83$$

(7.2)

$$Standard\ error\ of\ the\ mean = \frac{\sigma}{\sqrt{n}} = \frac{19.55671}{\sqrt{2}} = 13.83$$

.

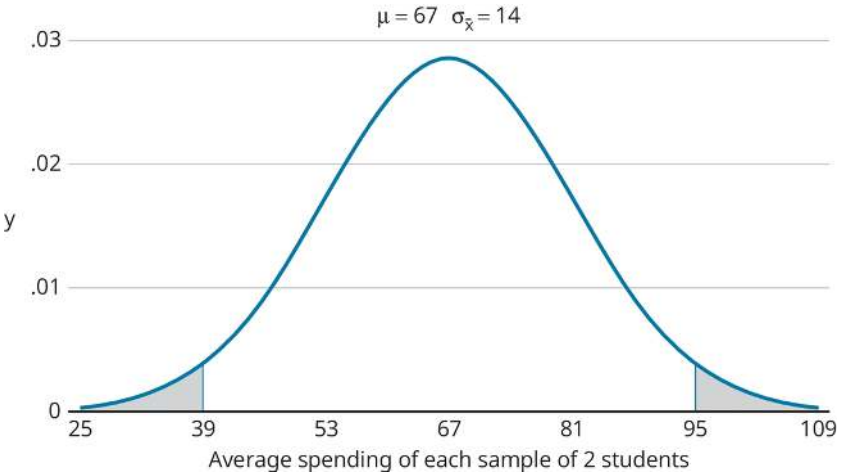
where

σ = standard deviation of the population

n = sample size

Notice that the answer to the calculation in Equation 7.2 gives us the same answer that we calculated by using the standard deviation of all possible means.

To illustrate this using a graph, Figure 7.8 shows the distribution of all possible sample means of two students where the mean is 67 and the standard deviation of all possible sample means (or the standard error of the mean) is equal to 13.83, or 14 if we round this to the nearest whole number. We can expect roughly 95 percent of all possible sample means to fall within two standard deviations of the mean or 39 (67 – 14 – 14 = 39) and 95 (67 + 14 + 14 = 95).



Description

Figure 7.8 Distribution of all possible sample means of 2 students

We can now use this information about the standard error of the mean to test a hypothesis, which we will show in the next section.

7.4 EXAMINING THE THEORY AND IDENTIFYING THE RESEARCH QUESTION AND HYPOTHESIS

With the large numbers of students who take the SAT test each year, there is an entire industry built around raising SAT scores through preparation. Many studies have shown that taking a preparatory course will raise a student's score. But are these tests biased? Are they taken by children from wealthier families or children enrolled in schools with higher achievement levels?

In this section, we will assume that the average math SAT score at High School X was 511 with a standard deviation of 120. To determine if the students could raise their scores significantly, the school randomly assigned 50 students in the high school to take a preparatory course prior to the test. The average score among the 50 students who took the course was 535. We now want to find out if the preparatory course worked, or if their average score of 535 is significantly different from 511.

In Chapter 1, we learned that part of the scientific method or the research process is to examine the theory, identify a research question, and form a hypothesis. Theory suggests that preparation for exams will lead to higher scores. In this case, our specific research question can be stated as follows: “Did students at High School X who took a preparatory course earn higher average scores on math SAT tests compared with the population of students at High School X?” As described in Chapter 1, we can then state a hypothesis, which is the answer to the question. The hypothesis could be positive or negative. For example, we could state that the students who took the preparatory course earned a higher or a lower score. When using statistical tests, however, we would define a null hypothesis, which is a testable statement indicating that there is no difference or no change. In this case, for example, the null hypothesis would be that students at High School X who took the course earned the same score as the rest of the high school population. The researcher would then use statistical techniques to test the hypothesis.

7.5 TESTING FOR STATISTICAL SIGNIFICANCE BETWEEN A SAMPLE MEAN AND A POPULATION MEAN

Now that we have identified our research question and stated our null hypothesis, we can test whether there is a statistically significant difference between the average score of the 50 students who took the preparatory course and the students who did not take the course.

Procedures

1. Calculate the standard error of the mean.

   Instead of looking at the standard deviation of the sample of 50 students, we must calculate the standard error of the mean since we are considering the distribution of possible means.

$$\text{Standard error of the mean} = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{120}{\sqrt{50}} = 16.97$$

$$\text{Standard error of the mean} = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{120}{\sqrt{50}} = 16.97$$

2. Calculate the standard score using the sample mean and the population mean in the numerator.

   In this step, our numerator shows the difference between the average score of the 50 students who took the course and the population of students at the school. When we divide by the standard error of the mean, we are essentially looking at how many standard deviation units the difference is above or below the population mean.

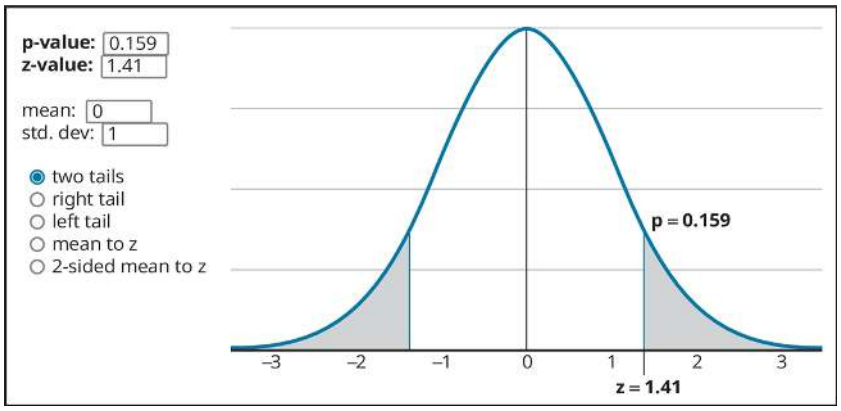
$$\text{Standard score} = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(535 - 511)}{16.97} = 1.41$$

$$\text{Standard score} = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(535 - 511)}{16.97} = 1.41$$

3. Look up the area under the normal curve for a standard score of 1.41.

In the example in Section 7.2, we were looking at one individual's score and comparing it with the class average to determine that student's percentile rank. In this case, we are examining the scores of 50 students to determine if their average score is unusual compared with the population average. Rather than looking at the area to the right of the standard score, we often want to examine both the extreme upper and lower values. It could be the case, for example, that students who take the course will earn a lower score. In fact, if you drew repeated samples of 50 students from the high school population and put them in an SAT preparatory course, many of the samples of 50 students would have a lower average score compared to the population average. For this reason, when we test hypotheses, we typically use a two-tailed test. In other words, we are testing whether our sample average is different than the population average rather than just higher or lower than the population average. In particular, we want to see how often we would see a score that is 535 or greater (24 points above the population average) or 487 or less (24 points below the population average).

If we use the calculator at "StatDistributions," (www.statdistributions.com/normal), we can plug in our standard score of 1.41 and see the image in Figure 7.9.

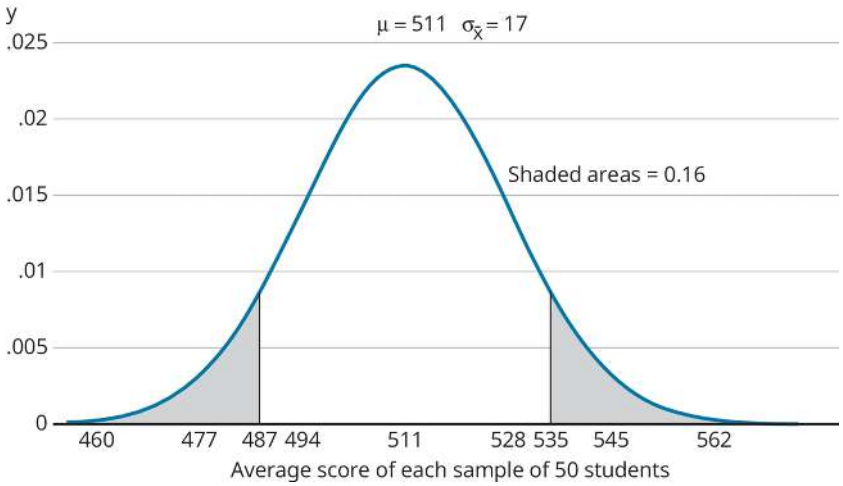


Description

Figure 7.9 Area Under The Normal Curve For A Two-Sided Standard Score of 1.41

Alternatively, we could use the areas under the normal curve in Table 7.1. Using our ***z*** score of 1.41 and the two-tailed probability, we can see that the area in the two tails is equal to 0.16151, or roughly 16%, as shown by the online calculator.

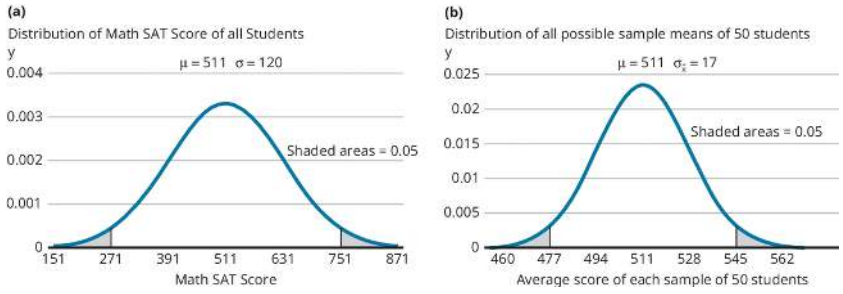
Finally, we can show this same information on a graph generated by Stata. Figure 7.10 shows the distribution of all possible sample means of 50 students where the mean is equal to 511 and the standard error of the mean (the standard deviation of all possible sample means) is 16.97, as calculated previously, or 17 if we round to the nearest whole number. The shaded areas illustrate that the area to the right or equal to 535 (or 1.41 standard deviations above the mean) combined with the area equal to or less than 487 (or 1.41 standard deviations below the mean) is 0.16. So you could say that in this is not that usual since you could expect to see 535 or higher or 487 or lower in 16% of all samples of 50 students that could be drawn from the population.



Description

Figure 7.10 Distribution of all possible sample means of 50 students

To better understand the distribution of sample means, Figure 7.11 shows two graphs. In Graph A, we see the distribution of the math SAT scores for all students in High School X. This shows the population mean (511) and the population standard deviation (120). In Graph B, we see the distribution of all sample means of 50 students with a mean of 511 and a standard deviation (which is now the standard error of the mean because it is the standard deviation of all possible sample means) of 17. Both graphs show shaded areas in the tails that represent 5% of the area under the curve. If the 50 students who had taken the SAT preparatory course had earned 545 or higher or 477 or lower, we would then say that this is unusual or significantly different than the population mean. We would only see these scores in 5% of all samples. In our example, however, the 50 students earned 535 on average, which does not fall in the shaded areas.



Description

Figure 7.11 Distribution of all SAT scores compared to distribution of all possible sample means of 50 students.

7.6 REJECTING OR NOT REJECTING THE NULL HYPOTHESIS

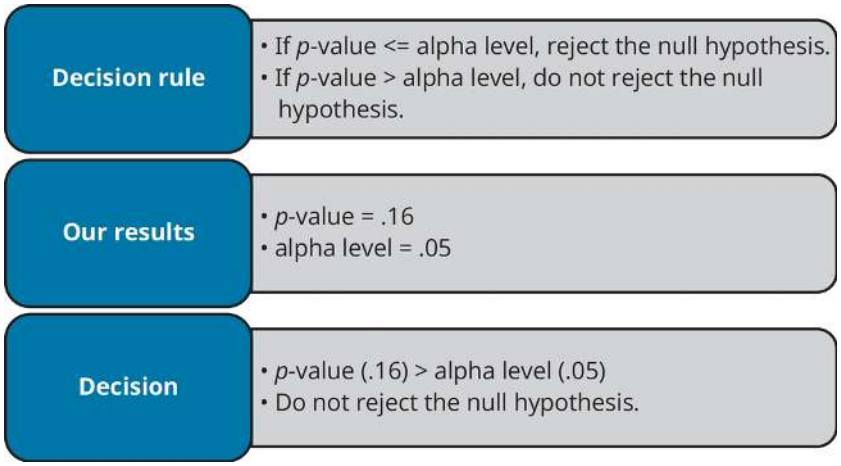
Before interpreting statistical tests, scientists or researchers set an alpha level, which is also referred to as *p* (critical). The alpha level is the probability of rejecting the null hypothesis when it is true, or a Type I error. It is typically set at 0.05, but researchers also use 0.01, 0.001, and sometimes 0.1. The larger the alpha level, the more likely you are to find statistically significant results.

Although probability is a number between 0 and 1, it is often expressed in percentage terms. For example, you could say that the probability of committing a Type I error is 0.05, or there is a 5% chance of committing a Type I error.

Using an alpha level of 0.05, we would say that the average SAT score of 535 is statistically significant if the probability of observing this value or greater (or ≤487, the opposite extreme) is 0.05.

In our example above, 16% of the samples fell in the two extremes. Our *p* value is then 0.16. The official definition of a *p* value from the American Statistical Association is given as follows: “Informally, a *p* value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value” (Wasserstein & Lazar, 2016, p. 131).

We can now compare our *p* value to the alpha level to determine whether our results are unusual or statistically significant. The rule along with our example is shown in Figure 7.12.

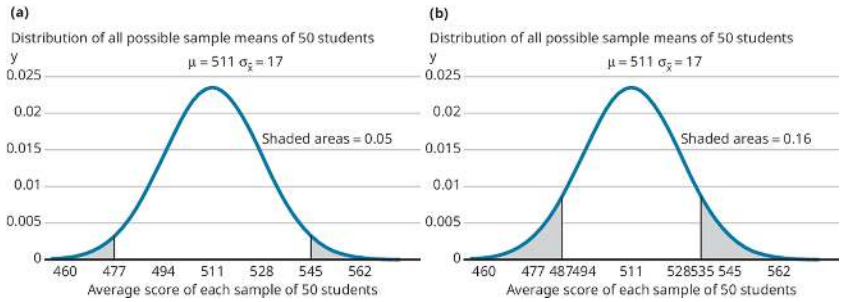


Description

Figure 7.12 Rule For Rejecting Or Not Rejecting The Null Hypothesis

In other words, you would expect to see scores of 535 or greater or 487 or less in 16% of all samples. Since this is fairly high (much higher than 5%), we would say that it is not that unusual to see this score.

Figure 7.13 shows the distribution of all possible sample means. In Figure A, we show that 5% of all sample means would be 545 or higher or 477 or lower. Given the test score results of 535, Figure B shows that 16% of all samples of 50 students would yield a test score of 535 or higher or 487 or lower. So again, we can say that our results are not that unusual since 16% of all samples of 50 students would result in 535 or something more extreme.



Description

Figure 7.13 Comparing all possible sample means of 50 students with 5% of the areas in the tails on the left and 16% on the right.

It is important to note that we would never say “we accept” the null hypothesis since there is always some chance that our samples did not accurately reflect the population. In fact, the alpha level tells us the probability of rejecting the null hypothesis when it is true. This is referred to as a Type I error, as described earlier. A Type II error occurs when we do not reject the null hypothesis when it is false. Appendix 4 offers a summary of the decision rules for statistical significance described in this chapter.

Recently, the American Statistical Association released a statement on statistical significance and *p* values to correct some of the many misuses of the concept. As they emphasize, the *p* value does not indicate if a hypothesis is true or if the data were produced by random chance. Furthermore, they emphasize that researchers should consider other factors besides the *p* values, such as the “design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.” Finally, they suggest other methods in addition to *p* values to test hypotheses, such as confidence intervals, which are discussed in later chapters (Wasserstein & Lazar, 2016).

7.7 INTERPRETING THE RESULTS

The results in Figures 7.9 and 7.13 show us that the probability of observing a standard score (or *z* score) that is greater than 1.41 or less than −1.41 is less than 0.16. As we discussed earlier, typically a *p* critical or alpha level is set at 0.05. We then compare our *p* value of 0.16 with the alpha level of 0.05. Because our *p* value is greater than 0.05, we do not reject the null hypothesis. In other words, there is not enough evidence to conclude that the students who took the preparatory course earned significantly higher or lower scores than the student population at High School X.

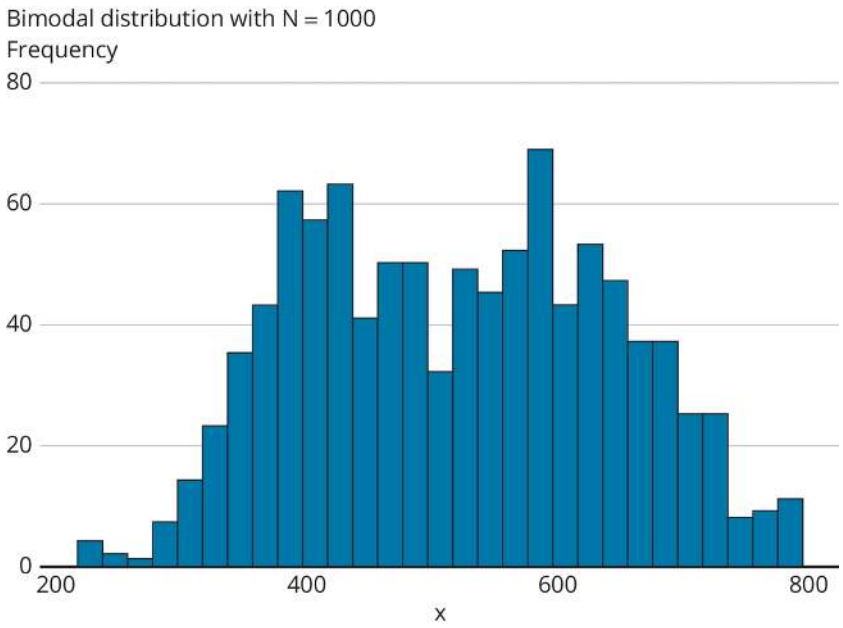
In Parts III and IV of the book, each chapter involves a research question, a null hypothesis, and a statistical test. A summary of these for each chapter is offered in Appendix 2. This summary should help you to quickly identify what type of statistical procedure or test should be used and the procedures to implement the test.

7.8 CENTRAL LIMIT THEOREM

In the previous sections, we used the normal distribution to determine one student’s percentile rank and to examine the scores of 50 students who took an SAT preparatory course compared with the population of students at that high school. As we mentioned earlier, the normal distribution is one of the most important concepts in statistics. It is used to draw conclusions about the characteristics of a population based on a sample. What is particularly unique is that the central limit theorem tells us that even if the population distribution is not normal, the sampling distribution of means from a population will

approach a normal distribution as the sampling size increases. In other words, we can still use the area under the normal curve to determine the probability of observing an equal or more extreme value of the mean observed in our sample even when the population is not normal. Next, we use an example to illustrate this point.

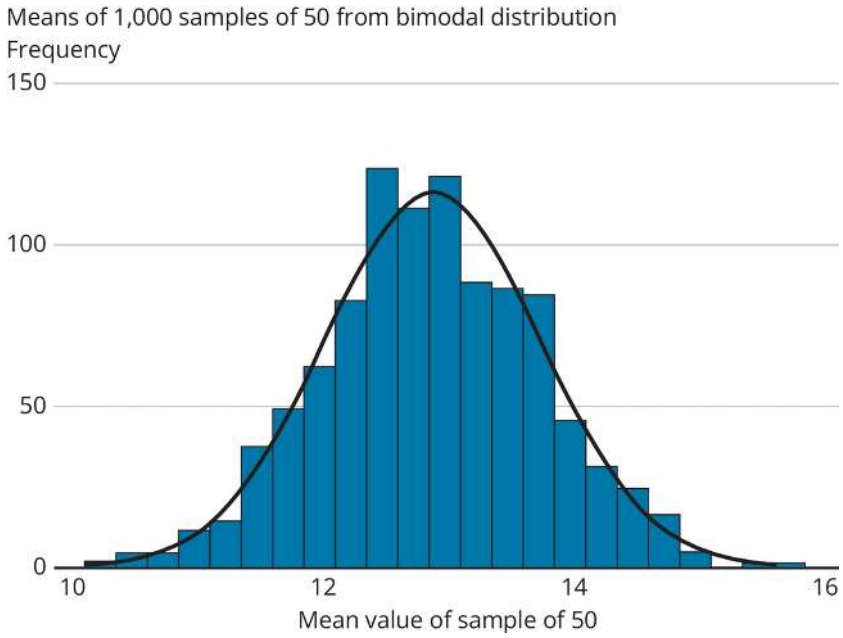
Let's suppose that there are 1,000 students at High School X. Their math SAT scores range from 200 to 800, but they are not normally distributed. Instead, as you see in Figure 7.14, there appears to be a bimodal distribution or a distribution clustered around two different values of roughly 400 and 600.



Description

Figure 7.14 Bimodal Distribution Of Sat Scores At High School X

According to the central limit theorem, if we draw many samples of 50 students, the means of each sample would form a normal distribution. Figure 7.15 shows the mean values of 1,000 samples of 50 students from the student population. As illustrated, the means form an almost perfect bell shape, which allows us to use the normal distribution to test hypotheses.[2]



Description

Figure 7.15 Means of 1,000 Samples of Math sat Scores from 50 Students

7.9 PRESENTING THE RESULTS

In addition to learning how to test hypotheses using statistics, it is important to learn how to convey the results. In particular, there may be times when you are reporting your results for a newspaper or a government report that is aimed at a nontechnical audience. You may also want to publish your results in an academic journal, which would require more details related to the statistical tests. Chapter 16, "Writing a Research Paper," offers guidelines on each type of test and how to report the results. We will also offer specific examples in each of the remaining chapters on how to address the two types of audiences based on the example used in the chapter.

Presenting the Results for a Nontechnical Audience

To present the results of the test to a nontechnical audience, we could submit the following statement:

High School X randomly chose 50 students to take part in an SAT preparatory course. They then compared their math SAT score with the rest of the high school population. Students who took the course earned 535 on average, which was higher than the high school average of 511, but it was not a statistically significant difference.

Presenting the Results in a Scholarly Journal

In a peer-reviewed journal, we would include more information. These results could be explained as follows:

Using a standard score or *z* score, we compared the math SAT scores of 50 students, who were selected randomly to take a preparatory course, with the math scores of the rest of the

population at High School X. Students who took the course earned 535 on average (*SD* = 90.25) compared with the high school average of 511 (*SD* = 120.00). This was not a statistically significant difference, *z* = 1.41, *p* = 0.16.

7.10 COMPARING A SAMPLE PROPORTION TO A POPULATION PROPORTION

The same concepts described previously can be applied to sample proportions. For example, suppose that a representative in congress won their election in the previous cycle and earned 58% of the vote (the proportion of the voting population). They are planning to run again for office and want to know if they are still able to win the same proportion of the votes. A sample of 500 voters shows that 53% of them will vote for the candidate. You now want to test if this is a statistically significant difference from the original 58% of the vote.

Similar to the distribution of sampling means, the sampling distribution of proportions will be normally distributed around the true population proportion. We therefore can use the same steps as we did above to compare a sample mean to a population mean but with slightly different formulas for the *z* score.

We begin by calculating the standard error of the proportion using the formula below:

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.58*(1-.58)}{500}} = 0.02$$

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.58*(1-.58)}{500}} = 0.02$$

Where:

$\pi = \text{the population proportion}$

$$\text{n} = \text{sample size}$$

$$\text{n} = \text{sample size}$$

In other words, 95% of sampling proportions would fall within 1.96 standard deviations of the true population proportion. In this case, that would mean within 0.0392 points (1.96 × 0.02). We would then calculate the *z* score as shown here:

$$z = \frac{p-\pi}{SE} = \frac{.53-.58}{0.02} = -2.27$$

$$z = \frac{p-\pi}{SE} = \frac{.53-.58}{0.02} = -2.27$$

where

$$p = \text{the sample proportion}$$

$$p = \text{the sample proportion}$$

We can then use a *z* table to look this up and find that the probability of observing a *z* score greater than 2.27 or less than −2.27 (a two-tailed test) is 0.02145. Using an alpha level of 0.05, we would reject the null hypothesis that the same proportion of voters would vote for the candidate in the second election. Building a confidence interval, we can be 95% sure that the proportion of voters who would vote for the candidate would be between .53 +/− (1.96 × .02) or between .4908 and .5692.

7.11 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes the hypothesis, test, procedures, and the Stata code used in the chapter (Tables 7.4 and 7.5). These are also summarized in Appendices 1 and 2.

TABLE 7.4 ■ Procedures for Chapter 7

Chapter Title	*Null Hypothesis*	*Test*	*Info Known /Type of Variables*	*Procedures/ Interpretation*
7: The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*Z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean $=(\sigma/\sqrt{n})$ 2. Standard score $[(\bar{X} - \mu)/$Standard error of mean$]$ 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.

Chapter Title	***Null Hypothesis***	***Test***	***Info Known /Type of Variables***	***Procedures/Interpretation***

Chapter Title	*Null Hypothesis*	*Test*	*Info Known /Type of Variables*	*Procedures/Interpretation*
7: The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*Z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score ((X − μ)/Standard error of mean) 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.

TABLE 7.5 ■ Code used in Chapter 7

Function	Code
Histogram	hist grade
Frequency table	tab score
Z score	help zscore zscore varname

Function	Code
Histogram	hist grade
Frequency table	tab score
Z score	help zscore zscore varname

EXERCISES

1. Your resting heart rate is 62. The average resting heart rate for the class is 72 and the variance is 25. The data are normally distributed.
   a. What percentage of the class has a lower resting heart rate relative to your own (round to the nearest whole number)?
   b. If there are 85 students in the class, how many students (the actual number, not the percentage) have a higher resting heart rate relative to your own? (Round to the nearest whole number.)
2. Many studies have confirmed that in the population, the flu lasts 21 days on average with a standard deviation of 7. The manufacturers of Tamiflu want to show that their product reduces the

length of the flu. They choose a sample of 25 people for their experiment and give them Tamiflu at the start of their flu. The average length of the flu among the 25 people taking Tamiflu was 18 days.

a. If you ran a test to determine if Tamiflu does reduce the length of a flu, what is the null hypothesis?
b. Using an alpha level or *p* critical of 0.05, use statistics to show if there is a statistically significant difference in the average length of time with and without taking Tamiflu. Show all of your work to prove this, and indicate whether you would reject your null hypothesis—why, or why not?
c. Write a paragraph to explain your results to a nontechnical audience.
d. Write a paragraph to explain your results in a scholarly journal.

3. Suppose that you have a population of three individuals and the number of times that they exercise per week is shown in the table below. We will use this information to show that there are two ways to calculate the standard error of the mean when drawing a sample.

Individual	Exercise Times per Week
A	2
B	0
C	7
Mean	3
Standard deviation	3.61

Individual	Exercise Times per Week
A	2
B	0
C	7
Mean	3
Standard deviation	3.61

a. Write out all combinations of two individuals from the three (including drawing the same person twice) and show the mean of each pair.
b. Using the formula for a standard deviation, plug in each mean in your table and use N − 1 in the denominator to calculate the standard error of the mean.
c. Instead of using information from the table that you generated, use the formula for the standard error of the mean when you know the population standard deviation ($\frac{\sigma}{\sqrt{n}}$) to determine the standard error of the mean.

4. Suppose you are a data analyst at a company that provides Internet services. You want to test whether the average download speed of a new type of modem is significantly different from the population's average download speed of 50 Mbp with a standard deviation of 8. You randomly select a sample of 16 customers and test their download speeds, finding that the sample mean download speed is 55 Mbps. Assuming an alpha level of 0.05, you want to test if there is a statistically significant difference between the new modem download speed and the population average.

a. What is your null hypothesis?
b. Show your work to test whether there is a statistically significant difference.
c. Would you reject your null hypothesis? Why, or why not?
d. Explain in words what is meant by the *p* value using statistics from this case.
e. Based on the problem above, draw a graph that illustrates the *p* value by shading in the areas that represent the *p* value. Label what is on the horizontal axis. Place the mean value and on the horizontal axis as well as the values that indicate where the *p* value areas begin.

KEY TERMS

normal distribution

z score

Descriptions of Images and Figures

Back to Figure

The number of colleges and universities with test-optional admissions policies recently topped 1,000 – a milestone that one expert says is a welcome trend.

Back in the 1980s, Bates College and Bowdoin College were nearly the only liberal arts colleges not to require applicants to submit SAT or ACT test scores.

In 2018, FairTest, a Boston-based organization that has been pushing back against America's testing regime since 1985, announced that the number of colleges that are test-optional has now surpassed 1,000.

This milestone means that more than one-third of America's four-year nonprofit colleges now reject the idea that a test score should strongly determine a student's future. The ranks of test-optional institutions include hundreds of prestigious private institutions, such as George Washington, New York University, WesleyanUniversity and Wake Forest University. The list also includes hundreds of public universities, such as George Mason, San Francisco State and Old Dominion.

Back to Figure

The x-axis represents the score, ranging from 70 to 90, and the y-axis represents the frequency, ranging from 0 to 15. The tallest bar corresponds to a score of 80, with a frequency of 14. The smallest bar corresponds to a score of 70, with a frequency of 3. The distribution is roughly bell-shaped, centered around a score of 80.

Back to Figure

Score	Freq.	Percent	Cum.
69	1	2.00	2.00
70	1	2.00	4.00
71	1	2.00	6.00
73	2	4.00	10.00
74	2	4.00	14.00
75	2	4.00	18.00
76	4	8.00	26.00
77	3	6.00	32.00
78	6	12.00	44.00
79	3	6.00	50.00
80	5	10.00	60.00
81	3	6.00	66.00
82	2	4.00	70.00
83	5	10.00	80.00

84	2	4.00	84.00
85	3	6.00	90.00
86	1	2.00	92.00
87	1	2.00	94.00
88	2	4.00	98.00
89	1	2.00	100.00
Total	50	100.00	

Back to Figure

The X-axis is labeled with z-values ranging from -3 to 3. The area under the curve to the right of z=1 is shaded, indicating the calculation of a one-tailed p-value, which is displayed as 0.159 at the top of the image and near the shaded area. The option for calculating the right tail is selected, as indicated by a radio button next to "right tail." Other options available are "two tails," "left tail," "mean to z," and "2-sided mean to z," which are not selected.

Back to Figure

In the "One-tailed" curve, there is a shaded area to the right of a specific z-value on the x-axis. In the "Two-tailed" curve, there are two symmetrically shaded areas, one to the left of a negative z-value and one to the right of a positive z-value.

Back to Figure

All Possible Samples of Two Students	**AverageExpenditureof Two Students**	**Sample MeanMinus PopulationMean**	**EstimatedStandard Errorof the Mean**
AA	55	-12	0
AB	50	-17	5
AC	72.5	5.5	17.5
AD	70	3	15.6
AE	57.5	-9.5	2.5
BB	45	-22	0
BC	67.5	0.5	22.5
BD	65	-2	20
BE	52.5	-14.5	7.5
CC	90	23	0
CD	87.5	20.5	2.5
CE	75	8	15
DD	85	18	0
DE	72.5	5.5	12.5
EE	60	-7	0
Mean	67		
Standard error of the mean	13.83		

55, 50, 72.5, 70, 57.5, 45, 67.5, 65, 52.5, 90, 87.5, 75, 85, 72.5, 60, 67 are marked and the text points to it reads "Sample distribution: distribution of all possible values for a statistic. Example -- All possible means of samples of 2 drawn from the student population of 5 to estimate how much they spend per week to eat out."

13.83 is marked and the text points to it reads "Standard error of the mean: standard deviation of all possible sample means.• Example -- The standard deviation of the means of all possible samples of sizes of two.

Back to Figure

The x-axis is labeled as "Average spending of each sample of 2 students," with values ranging from 25 to 109. The y-axis is labeled as "y" with values ranging from 0 to 0.03. There are two shaded regions under the curve, one between 25 and 39 and the other between 95 and 109, representing the tail areas of the distribution.

Back to Figure

The X-axis is labeled with z-values ranging from -3 to 3. The area under the curve to the right of z=1.41 is shaded, indicating the calculation of a two-tailed p-value, which is displayed as 0.159 at the top of the image and near the shaded area. The option for calculating the two tails is selected, as indicated by a radio button next to " two tails." Other options available are "right tail," "left tail," "mean to z," and "2-sided mean to z," which are not selected.

Back to Figure

The x-axis is labeled as "Average score of each sample of 50 students," with values ranging from 460 to 562. The y-axis is labeled as "y" with values ranging from 0 to 0.025. There are two shaded regions under the curve, one between 460 and 487 and the other between 535 and 562, representing the tail areas of the distribution.

Back to Figure

Graph (a): This represents the distribution of Math SAT scores for all students. The mean score (μ) is 511, and the standard deviation (σ) is 120. The x-axis is labeled "Math SAT Score," with values ranging from 151 to 871. The bell-shaped curve represents a normal distribution, with the shaded areas in both tails representing 5% of the distribution.

Graph (b): Represents the distribution of all possible sample means of 50 students' scores. The mean (μ) is 511, and the standard error of the sample mean ($\sigma_{\bar{x}}$) is 17. The x-axis is labeled "Average score of each sample of 50 students," with values ranging from 460 to 562. The shaded regions in the tails represent 5% of the distribution.

Back to Figure

Decision rule

• If p-value <= alpha level, reject the null hypothesis.

• If p-value > alpha level, do not reject the null hypothesis.

Our results

• p-value = .16

• alpha level = .05

Decision

• p-value (.16) > alpha level (.05)

• Do not reject the null hypothesis.

Back to Figure

Graph (a): The mean score (μ) is 511, and the standard error of the sample mean ($\sigma_{\bar{x}}$) is 17. The shaded regions in both tails represent 5% of the distribution.

Graph (b): Similarly, the mean score (μ) is 511, and the standard error of the sample mean ($\sigma_{\bar{x}}$) is 17. The shaded regions now cover a total of 16%, indicating a broader range of sample means that are more likely compared to graph (a).

Both graphs display the x-axis labeled "Average score of each sample of 50 students," with values ranging from approximately 460 to 562. The y-axis represents the probability density.

Back to Figure

The x-axis is labeled "x," with values ranging from 200 to 800. The y-axis is labeled "Frequency," with values ranging from 0 to 80. The histogram has two prominent peaks, indicating a bimodal distribution. The bars vary in height, with the frequency of values increasing near 400 and 600, which form the two modes of the distribution. The heights of the bars represent the number of occurrences (frequency) of values within each range along the x-axis.

Back to Figure

The x-axis is labeled "Mean value of sample of 50" and ranges from 10 to 16. The y-axis is labeled "Frequency," with values ranging from 0 to 150. The histogram bars cluster around two distinct peaks between 12 and 14. A smooth curve fitted to the data suggests a normal distribution despite the original bimodal nature of the data from which the samples were taken.

8 TESTING A HYPOTHESIS ABOUT A SINGLE MEAN AND A SINGLE PROPORTION

CHAPTER PREVIEW

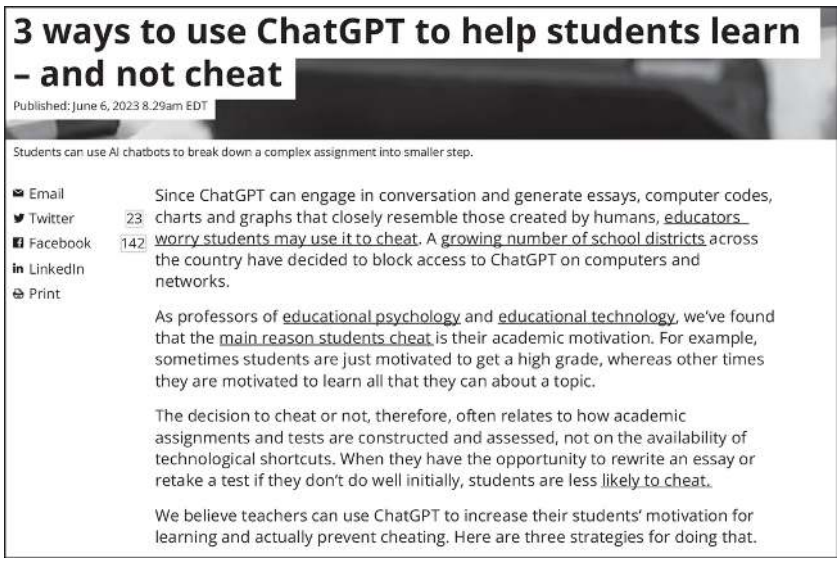
Steps	Example
Research question	Does the use of ChatGPT improve homework scores
Null hypothesis	Students who use ChatGPT to generate and practice problems earn the same average score as those who do not use it
Test	One-sample *t* test
Types of variables	One continuous variable (grade)
When to use	Comparing a sample mean with an established mean from a population The population standard deviation is not known.
Assumptions	1. The population is approximately normally distributed. 2. Sample observations are random.
Stata code: Generic	**ttest continuousvar==X** where *x* is some predetermined mean.
Stata code: Example	**ttest grade==86**

Steps	Example
Research question	Does the use of ChatGPT improve homework scores
Null hypothesis	Students who use ChatGPT to generate and practice problems earn the same average score as those who do not use it
Test	One-sample *t* test
Types of variables	One continuous variable (grade)
When to use	Comparing a sample mean with an established mean from a population The population standard deviation is not known.
Assumptions	1. The population is approximately normally distributed. 2. Sample observations are random.
Stata code: Generic	**ttest continuousvar==X** where *x* is some predetermined mean.
Stata code: Example	**ttest grade==86**

8.1 INTRODUCTION

ChatGPT (Chat Generative Pre-trained Transformer) was first introduced to the world at the end of November of 2022. Within one month, it had gained over 100 million users and had become the fastest-growing consumer software application in history. It was trained to interact intelligently and process large volumes of data. Very quickly, competitors launched competing “artificial intelligence” software.

With the birth of this new type of software, both professors and students found new uses for it (Figure 8.1). If professors are writing new assignments, for example, ChatGPT could generate endless examples of a certain type of problem. Similarly, a student could use ChatGPT to generate and solve problems for further practice. They could also cheat on homework and take-home exams by asking ChatGPT to solve problems. Thus, a great debate started on the use and value of ChatGPT in academic circles.



3 ways to use ChatGPT to help students learn – and not cheat

Published: June 6, 2023 8.29am EDT

Students can use AI chatbots to break down a complex assignment into smaller step.

Email
Twitter 23
Facebook 142
LinkedIn
Print

Since ChatGPT can engage in conversation and generate essays, computer codes, charts and graphs that closely resemble those created by humans, educators worry students may use it to cheat. A growing number of school districts across the country have decided to block access to ChatGPT on computers and networks.

As professors of educational psychology and educational technology, we've found that the main reason students cheat is their academic motivation. For example, sometimes students are just motivated to get a high grade, whereas other times they are motivated to learn all that they can about a topic.

The decision to cheat or not, therefore, often relates to how academic assignments and tests are constructed and assessed, not on the availability of technological shortcuts. When they have the opportunity to rewrite an essay or retake a test if they don't do well initially, students are less likely to cheat.

We believe teachers can use ChatGPT to increase their students' motivation for learning and actually prevent cheating. Here are three strategies for doing that.

Description

Figure 8.1 Article

Source: Xie and Anderman (2023)

https://theconversation.com/3-ways-to-use-chatgpt-to-help-students-learn-and-not-cheat-205000

In this chapter, we will learn how to test a sample mean to determine if it is significantly different from some specified value. For example, we can use homework scores from students in a data analysis course after ChatGPT was introduced and compare their average score to the average homework score from previous semesters before ChatGPT was introduced. We can then determine if there is a statistically significant difference between the average homework score for one semester to the previous semesters. We will assume that the average score in previous semesters was an 86 (considering this the population mean) and then compare this to the homework score in the first semester in the course after Chat GPT was introduced (the sample mean).

8.2 WHEN TO USE THE ONE-SAMPLE *T* TEST

Table 8.1 shows examples from different fields where the **one-sample *t* test** can be used. In each case, there is an assumed population average, but the population standard deviation is unknown. Each of these can be tested using the one-sample *t* test.

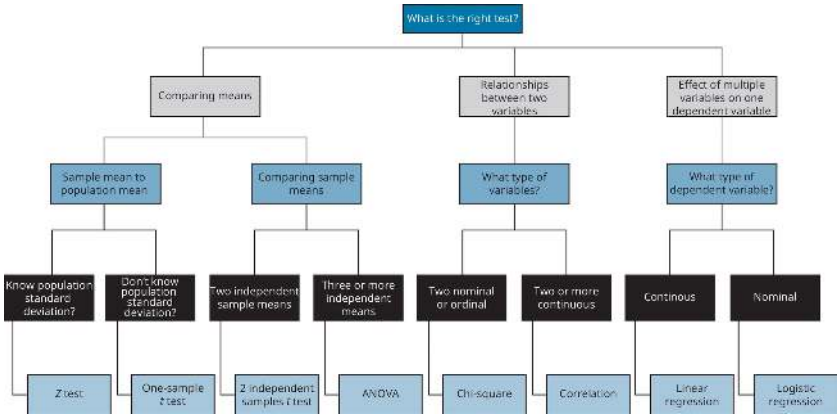
TABLE 8.1 ■ Examples of the one-Sample *t* Test

Field	Research Question	Null Hypothesis	Continuous Variable
Criminal justice	Does a judge in one district give harsher prison sentences to women convicted of child abuse?	There is no difference in the average sentence length by the judge compared with the national average.	Sentence time in months of women convicted of child abuse
Economics	Do Americans work a 40-hour workweek?	Americans work 40 hours per week.	Hours worked per week among those Americans who work full time
Political science	What is the average age of pro-life supporters?	The average age of pro-life supporters is 53.	Age of pro-life supporters
Public health	Do smokers gain weight within the first year after they stop smoking?	The average weight gain after 1 year without smoking is 0.	Weight gain after 1 year without cigarettes
Psychology	Is postpartum depression more common among mothers who do not have immediate family members living nearby than for mothers that do?	The average depression score on the "Quick Depression Assessment" test is 4.	Quick Depression Assessment test score
Sociology	What is the average age that children own their first cell phone?	Children own their first phone at the age of 9.	Ownership age of first phone

Field	Research Question	Null Hypothesis	Continuous Variable
Criminal justice	Does a judge in one district give harsher prison sentences to women convicted of child abuse?	There is no difference in the average sentence length by the judge compared with the national average.	Sentence time in months of women convicted of child abuse
Economics	Do Americans work a 40-hour workweek?	Americans work 40 hours per week.	Hours worked per week among those Americans who work full time
Political science	What is the average age of pro-life supporters?	The average age of pro-life supporters is 53.	Age of pro-life supporters
Public health	Do smokers gain weight within the first year after they stop smoking?	The average weight gain after 1 year without smoking is 0.	Weight gain after 1 year without cigarettes
Psychology	Is postpartum depression more common among mothers who do not have immediate family members living nearby than for mothers that do?	The average depression score on the “Quick Depression Assessment” test is 4.	Quick Depression Assessment test score
Sociology	What is the average age that children own their first cell phone?	Children own their first phone at the age of 9.	Ownership age of first phone

Figure 8.2 illustrates a decision tree that helps you decide which statistical test is appropriate for each type of analysis. As you can see, it will depend on whether you are comparing means or relationships between variables. In this chapter, we will be looking at a sample mean (average homework score of

students using ChatGPT) and comparing it with an assumed population mean of 86. We would follow the tree on the left-hand side from “comparing means” to “sample mean to population mean.” Because we do not know the population standard deviation, we would follow the path to the one-sample *t* test.



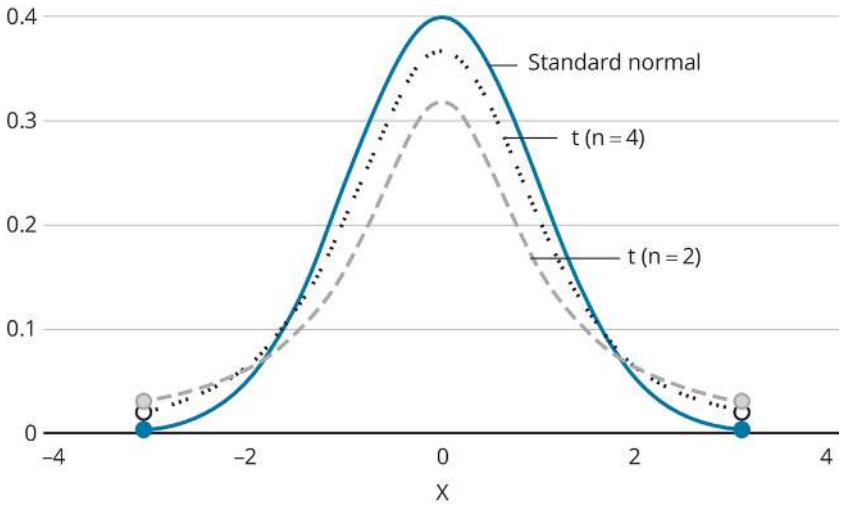
Description

Figure 8.2 Decision tree for Choosing the Right Statistic

8.3 CALCULATING THE ONE-SAMPLE *T* TEST

In Chapter 7, we learned about using the normal distribution, and we compared the math SAT scores of 50 students who took a preparatory SAT course to the average scores of the population of students at the same high school. In that case, we knew the population mean and the population standard deviation. In most cases, however, we will not know the population standard deviation. In fact, it is typical that we do not know anything about a population and therefore draw a sample to learn about the population.

In this chapter, we will assume that we know the population mean or some hypothesized value of the population mean (e.g., the average homework score in a data analysis course prior to the introduction of ChatGPT), but we do not know the population standard deviation. We will therefore use the sample standard deviation when calculating a *t* statistic, which is similar to the *z* score or standard score that we used in Chapter 7. The difference, however, is that the sample standard deviation could be larger or smaller than the population standard deviation, which introduces some uncertainty. To account for this, we use a ***t* distribution**, which looks like a normal distribution but has more area in the tails to account for the error introduced. Furthermore, its shape depends on the sample size. A larger sample size will produce a *t* distribution that is closer to the normal distribution, and eventually, with a large enough sample, the area under the *t* distribution will be close to the normal distribution, and either test can be used. Figure 8.3 shows the *t* distribution with two sample sizes and the normal distribution. The normal distribution is higher in the middle with less area in the tails. The two *t* distributions represent very small sample sizes (four and two) to emphasize the differences in the distribution. Notice that when the sample size is four, the distribution is much closer to the normal distribution than when the sample size is two.



Description

Figure 8.3 Normal and *T* Distribution for two Sample Sizes

To calculate the *t* statistic, we use the same formula that we used in the previous chapter when we calculated the standard score. In this case, however, we substitute the sample standard deviation for the population standard deviation, as shown in Equation 8.1.

$$t = \frac{(\bar{X} - \mu)}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{(X - \mu)}{\frac{s}{\sqrt{n}}}$$

where

X = sample mean

μ = population mean

s = sample standard deviation

n = sample size

The numerator shows us the difference between the sample mean and the population mean. A larger difference will lead to a larger t statistic and a greater likelihood that there is a statistically significant difference. The denominator is the standard error of the mean, which is the standard deviation of the sample means from all possible samples drawn from the population. Combined, the numerator and denominator tell you how many standard deviation units the observed sample mean is above or below the population mean.

8.4 CONDUCTING A ONE-SAMPLE *T* TEST

The "Homework" data set represents 30 students in a data analysis course. It contains only one variable, grade, which indicates the average homework grade for each of the 30 students at the end of the course in which students were allowed to use ChatGPT to generate and practice problems.

Research question

Do students who use ChatGPT to generate and practice problems earn a higher grade on their homework score compared to previous semesters?

Hypothesis

Students who use ChatGPT to generate and practice problems earn 86 on their homework score (the average from prior semesters before ChatGPT).

Null hypothesis

Students earn an 86 on their homework score when using ChatGPT

Variables

Continuous variable—the homework grade

Assumptions

In addition to using one continuous variable and one population mean, we make two assumptions to generate valid results:

---

1. ***Normal distribution:*** The continuous variable, homework score in this example, should be approximately normally distributed within each category. It only needs to be approximately normally distributed since minor violations of normality do not affect the results.
2. *Sample observations are random:* Sample data must be selected randomly. (Refer to Chapter 2 on sample selection techniques.)

Procedures using commands

Using a do-file, we would run the commands below:

```
ttest grade==86
```

Procedures using menus

Using menus in Stata, we would click on the sequence listed below that would bring us to a Dialog Box where we would select the variable “grade” and click on “One-sample.” We would also fill in “86” for the hypothesized mean as displayed.

Statistics → Summaries, tables, and tests → Classical tests of hypotheses → *t* tests (mean-comparison tests)

8.5 INTERPRETING THE OUTPUT

Figure 8.4 illustrates the results of our one-sample *t test*. We can see from the table that the average homework score in our sample of 30 students is 89.4 with a standard deviation of 5.7. When comparing these results with the hypothesized mean of 86, we turn to the *t* statistic of 3.25 and the hypotheses on the last two lines of output. The first hypothesis on the left (Ha: mean < 86) is that the mean or average is less than 86. The second hypothesis (Ha: mean != 86) is that the mean does not equal 86, and the third hypothesis (Ha: mean > 86) is that the mean is greater than 86. Since we typically want to consider the extreme values on either side of the average, as discussed in the previous chapter, we use the hypothesis that the mean is not equal to 86. Using that information, we see that when the null hypothesis is true (that the average is 86), the probability of observing a *t* statistic greater than 3.25 or less than −3.25 is less than 0.0029. Because this is less than 0.05, our alpha level, we reject the null hypothesis that the average homework score is 86.

```
. ttest grade==86

One-sample t test

Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
   grade |      30    89.36667    1.035556    5.671972    87.24872    91.48462
---------------------------------------------------------------------------
    mean = mean(grade)                                            t =   3.2511
H0: mean = 86                                     Degrees of freedom =       29

    Ha: mean < 86                 Ha: mean != 86                 Ha: mean > 86
 Pr(T < t) = 0.9985         Pr(|T| > |t|) = 0.0029          Pr(T > t) = 0.0015
```

Description

Figure 8.4 Stata Output For The One-Sample *t* Test

We can also examine the confidence interval in Figure 8.4. In Chapter 7, we learned that 95% of all sample means should fall within roughly two standard errors of the mean when the population is normally distributed. The exact number is 1.96 standard errors of the mean. To obtain the 95% confidence interval for the population mean, we would then multiply 1.96 by the standard error of the

mean and add this to the sample mean to get the upper end of the confidence interval. For the lower end, we would then multiply −1.96 by the standard error of the mean and add this to the sample mean. In this chapter, however, we don't know the population standard deviation. We therefore have to use the *t* distribution. With a sample size of 30, we first calculate the "**degrees of freedom**" as the sample size minus 1 or 29 degrees of freedom. Degrees of freedom is a statistical term that indicates the number of observations that are free to vary. In other words, we could change 29 of the values in the sample and still get the same mean of 89.4 as long as we control or set the last value so that the mean is 89.4.

We can then use Appendix 6, which shows the **critical values** of the *t* distribution, to find the exact *t* statistic that would provide the area under the curve that represents 95% of all observations. With 29 degrees of freedom and the area under the two tails adding up to .05, we see a *t* value of 2.05. This would mean that 95% of all observations fall between −2.05 and +2.05 standard errors of the mean on either side of the sample mean. The confidence interval would then be calculated as follows:

$$\text{Lower end} = 89.4 + (\text{-}2.05 \times 1.04) = 87.27$$

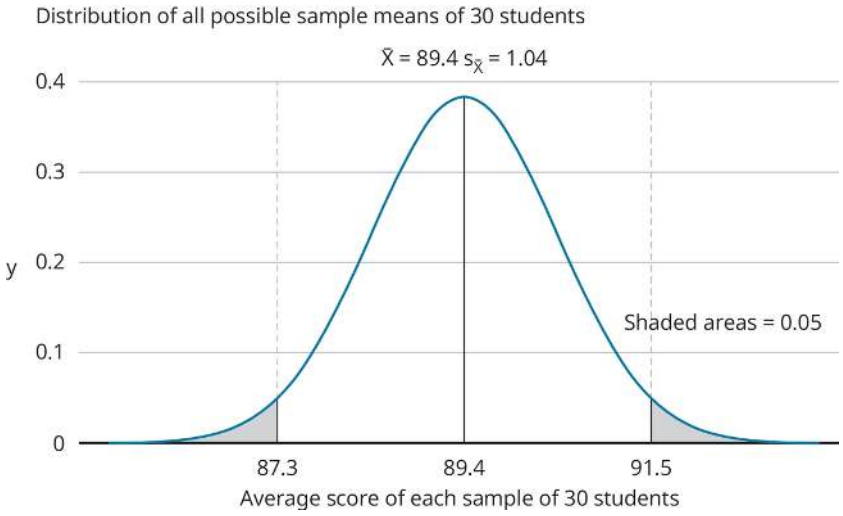
$$\text{Lower end} = 89.4 + (-2.05 \times 1.04) = 87.27$$

$$\text{Upper end} = 89.4 + (2.05 \times 1.04) = 91.53$$

$$\text{Upper end} = 89.4 + (2.05 \times 1.04) = 91.53$$

With this information, we can say that we are 95% confident that the true value of the average homework score is in the range of 87.27 to 91.53. The confidence interval can be used in place of a *p* value to test a hypothesis. In other words, if the confidence interval does not contain the null hypothesis value (86 in this case), then we can say that the results are statistically significant. As described in Chapter 7, the American Statistical Association wrote in 2016 that confidence intervals should be used in addition to *p* values. In fact, many researchers argue that the confidence interval is preferred because it offers a range of values instead of providing a cutoff where we determine if our findings are statistically significant.

We can illustrate the confidence interval in Figure 8.5 This figure shows the distribution of all possible sample means of 30 students when the average is 89.4 and the standard error of the mean (the standard deviation of all possible sample means) is 1.04, as calculated previously. The shaded areas illustrate that the areas to the right of 91.5 (or equal to 91.5) and to the left of 87.3 or equal to 87.3 is equal to 5% of all of the area under the curve.



Description

Figure 8.5 Distribution of all possible sample means of 30 students

8.6 PRESENTING THE RESULTS

Presenting the results for a nontechnical audience

To present these results to a lay audience who may not be familiar with statistical tests, we could write the following:

Based on our sample of 30 students, we found that the average homework score among 30 students who had used ChatGPT to generate and practice problems was 89. This was a statistically significant difference compared to the average homework score of 86 in previous semesters.

Presenting the results in a scholarly journal

In a peer-reviewed journal, we would include more information. These results could be explained as follows:

Using a one-sample *t test*, we examined the average homework score of 30 college students enrolled in a data analysis course where they used ChatGPT to generate and practice problems. Our results showed an average score of 89.37 (*SD* = 5.67). There was a statistically significant difference between our result and the homework score from previous semesters of 86, $t(29) = 3.25$, $p = 0.00$.

8.7 ESTIMATING A POPULATION PROPORTION FROM A SAMPLE PROPORTION

Let's suppose that instead of wanting to examine the average homework score after using ChatGPT, we want to determine if the proportion of students using ChatGPT is equal to a hypothesized proportion of

75%. In Chapter 7, we used an example where we knew the population proportion and wanted to know if a sample proportion would be the same a few years later. In particular, we saw that 58% of voters voted for a candidate and we wanted to know if the same proportion would vote for the candidate for a second term. To estimate the *z*-score, we used the following formula:

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

In many cases, we won't know the population parameter and may want to estimate it with a sample proportion. Or, as in this example, we have a hypothesized population proportion and use a sample to determine the likelihood that the hypothesized proportion is accurate. To do this analysis, we would use the same format as in the previous equation but would need to use the sample proportion in the denominator as follows:

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

As described above, we assume that 75% of students at one college are using ChatGPT to help study for exams. We could then draw a sample of 300 students that shows that 69% of students in the sample are using ChatGPT to help them study for exams. Filling in the formula, we would see that our *z*-score would be as follows:

$$Z = \frac{.69 - .75}{\sqrt{\frac{.69(1 - .69)}{300}}} = -2.22$$

$$Z = \frac{.69 - .75}{\sqrt{\frac{.69(1 - .69)}{300}}} = -2.22$$

As in Chapter 7, we would then use a *z*-table or online calculator to determine the probability of observing 0.69 if the true proportion was 0.75. Using an alpha level of 0.05, we would see that the probability of observing a value equal to or greater than 2.22 or less than or equal to −2.22 is 0.0139. We would therefore reject the null hypothesis that the proportion of students using ChatGPT is equal to .75. We could also calculate the 95 percent confidence interval as 0.69 +/− (1.96 × 0.027) or the sample proportion plus and minus 1.96 (the number of standard deviations on either side of the sample proportion that would encompass 95% of all proportions from repeated samples) multiplied by the standard error of the proportion (i.e., the denominator of the *z*-score). This gives us a confidence interval of 0.64 to 0.74. Assuming that the variable name for the whether someone used ChatGPT is “Chat,” Stata could calculate the confidence interval with the code “**ci proportions Chat**.”

8.8 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes the hypothesis, test procedures, and Stata code used in the chapter (Tables 8.2 and 8.3). They are also summarized in Appendices 1 and 2.

TABLE 8.2 ■ Procedures for Chapters 7 and 8

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/ Type of Variables*	*Procedures/Interpretation*
7: The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z*-score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score $[(\bar{X} - \mu)]$/ Standard error of mean) 3. Look up percentages for standard score using normal distribution 4. When the null hypothesis is true, the probability of observing a *z*-score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. One-sample *t* test	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know the population mean Don't know the population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score $[(\bar{X} - \mu)]$/ Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *t* statistic greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.

Chapter Title	***Null Hypothesis***	***Test***	***Info Known/ Type of Variables***	***Procedures/Interpretation***
7: The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z*-score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score ((X − μ)/Standard error of mean) 3. Look up percentages for standard score using normal distribution 4. When the null hypothesis is true, the probability of observing a *z*-score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/ Type of Variables*	*Procedures/Interpretation*
8. One-sample *t* test	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know the population mean Don't know the population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score ((X − μ)/ Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *t* statistic greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.

TABLE 8.3 ■ Code used in Chapter 8

Function	Code
One-sample *t* test	ttest grade==86
Estimating a population proportion from a sample proportion	ci proportions Chat

Function	**Code**
One-sample *t* test	ttest grade==86
Estimating a population proportion from a sample proportion	ci proportions Chat

EXERCISES

1. Your local takeout restaurant claims that their food is delivered in 20 minutes. You decide to test their claim and order food from them 36 times over the next 3 months. On average, the food is delivered in 23 minutes with a standard deviation of 5 minutes.
   a. What is the null hypothesis?
   b. Would you reject or not reject the null hypothesis assuming an alpha level or *p* critical of 0.05? Show your work to support your decision.
2. Based on your answer to Question 1, construct a 95% confidence interval for the true value of the delivery time.
3. The legal drinking age in the United States is 21 years. Many people, however, try alcohol before they turn 21. Use the National Survey on Drug Use and Health from 2015 to test whether the age when Americans first try alcohol (alctry) is 21 years. Before you run the test using Stata, use the command **tab alctry**. Notice that there are categories related to missing data—985 bad data, 991 never used alcohol, 994 don't know, 997 refused, or 998 blank. Next run the command **sum alctry**. You will notice that the mean age for first trying alcohol is 292 years, which doesn't make sense. You should also notice that the maximum value is 998. To remove these missing data from your test, include the command **if alctry < 72** at the end of your command since 71 was the oldest age reported.

   a. What is your null hypothesis?
   b. Would you reject or not reject your null hypothesis? Explain your decision using your output and add a screenshot of your output as part of your response
   c. Explain the 95% confidence interval in your output.
4. Based on your results for Question 3, write a few sentences that would explain your results to a nontechnical audience. Then write a few sentences to present your results in a scholarly journal.
5. The average college acceptance rate in 2013 according to the College Results online data set was 72.5%. Since that time, many news articles have reported on changing demographics in the United States that will lead to fewer students entering college. You want to know if acceptance rates will rise as colleges compete for a smaller pool of applicants. Use the "College Score Card April 2023 – USNews" data set to test if the acceptance rate (adm_rate) is still 72.5%. In the data set, the acceptance rate is expressed on a scale from 0 to 1. You will need to use 0.725 for your test. Show your Stata output, and then write a few sentences to explain your results for a scholarly journal.

KEY TERMS

critical values

degrees of freedom

hypothesis

normal distribution

null hypothesis

one-sample *t* test

research question

***t* distribution**

variables

Descriptions of Images and Figures

Back to Figure

Published: June 6, 2023 8.29am EDT

Students can use AI chatbots to break down a complex assignment into smaller step.

Since ChatGPT can engage in conversation and generate essays, computer codes, charts and graphs that closely resemble those created by humans, educators worry students may use it to cheat. A growing number of school districts across the country have decided to block access to ChatGPT on computers and networks.

As professors of educational psychology and educational technology, we've found that the main reason students cheat is their academic motivation. For example, sometimes students are just motivated to get a high grade, whereas other times they are motivated to learn all that they can about a topic.

The decision to cheat or not, therefore, often relates to how academic assignments and tests are constructed and assessed, not on the availability of technological shortcuts. When they have the

opportunity to rewrite an essay or take a test if they don't do well initially, students are less likely to cheat.

We believe teachers can use ChatGPT to increase their students' motivation for learning and actually prevent cheating. Here are three strategies for doing that.

Back to Figure

On the top center is a box labeled "What is the right test?" which is divided into three main branches: "Comparing means," "Relationships between two variables," and "Effect of multiple variables on one dependent variable."

Under "Comparing means", there are two sub-branches: "Sample mean to population mean" and "Comparing sample means."

"Sample mean to population mean" is further divided into "Know population standard deviation?" and "Don't Know population standard deviation?".

"Know population standard deviation?" leading to either "Z test" and "Don't Know population standard deviation?" leads to "One-sample t test."

"Comparing sample means" is divided into "Two independent sample means" and "Three or more independent means."

"Two independent sample means" leading to "2 independent samples t test" and "Three or more independent means" leading to "ANOVA."

Under "Relationships between two variables", there is a branch asking "What type of variables?": "Two nominal or ordinal" leading to "Chi-square" and "Two or more continuous" leading to "Correlation."

Under "Effect of multiple variables on one dependent variable", there is a branch asking "What type of independent variable?": "Continuous" and "Nominal."

"Continuous" leads to "Linear regression" and "Nominal" leads to "Logistic regression."

Back to Figure

The x-axis is labeled "X" and ranges from -4 to 4, while the y-axis ranges from 0 to 0.4. Normal distribution is represented by a solid line curve, centered at 0 with the highest peak. t-distribution with n=4 is depicted by a dashed line, with a slightly wider shape and lower peak than the standard normal curve. t-distribution with n=2 is shown with a dot-dashed line, having the widest spread and lowest peak, indicating higher variability. Each curve illustrates how t-distributions become more similar to the standard normal distribution as the degrees of freedom increase.

Back to Figure

.ttset grade==86

One-sample t test

Variable	obs	Mean	Std. err.	Std. dev	[95% conf. interval]	
grade	30	89.36667	1.035556	5.671972	87.24872	91.48462

mean = mean(grade)

H0: mean = 86

t = 3.2511

Degrees of freedom = 29

Ha: mean < 86

Pr(T<t) = 0.9985

Ha: mean != 86

Pr(|T|> |t|) = 0.0029

Ha: mean > 86

Pr(T>t) = 0.0015

Back to Figure

The mean score (X-bar) is 89.4 and the standard deviation (Sx) is 1.04. The curve peaks around the mean and tapers off towards both ends. Two vertical dashed lines indicate scores of 87.3 and 91.5, framing the central portion of the distribution. The areas outside these lines are shaded to represent a total area of 0.05, suggesting these regions contain 5% of the probability mass under the curve."

9 TESTING A HYPOTHESIS ABOUT TWO INDEPENDENT MEANS

CHAPTER PREVIEW

Steps	Example
Research question	Did states with Democratic electors in the 2020 presidential elections have a greater number of mask-mandated days during the coronavirus (COVID-19) pandemic from 2020 to 2022 compared to states with Republican electors?
Null hypothesis	There is no difference in the number of mask-mandated days during the COVID-19 pandemic between states with Democratic and Republican electors in the 2020 presidential election.
Test	Two independent-samples *t* test
Types of variables	One continuous variable: number of mask-mandated days (MaskDays) One categorical variable with two categories: 1 "Democrat," 2 "Republican" (ElectorParty)
When to use	Two samples Two populations Population standard deviation is unknown
Assumptions	Independence of observations Normal distribution Homogeneity of variances
Additional tests needed	Equality of variances
Stata code: generic	**ttest continuousvar, by(categoricalvar)**
Stata code: example	**ttest MaskDays, by(ElectorParty)**

Steps	**Example**
Research question	Did states with Democratic electors in the 2020 presidential elections have a greater number of mask-mandated days during the coronavirus (COVID-19) pandemic from 2020 to 2022 compared to states with Republican electors?
Null hypothesis	There is no difference in the number of mask-mandated days during the COVID-19 pandemic between states with Democratic and Republican electors in the 2020 presidential election.
Test	Two independent-samples *t* test
Types of variables	One continuous variable: number of mask-mandated days (MaskDays) One categorical variable with two categories: 1 "Democrat," 2 "Republican" (ElectorParty)

Steps	Example
When to use	Two samples Two populations Population standard deviation is unknown
Assumptions	Independence of observations Normal distribution Homogeneity of variances
Additional tests needed	Equality of variances
Stata code: generic	**ttest continuousvar, by(categoricalvar)**
Stata code: example	**ttest MaskDays, by(ElectorParty)**

9.1 INTRODUCTION

According to Figure 9.1, the response to the COVID-19 pandemic was largely based on partisan politics. Democrats preferred stricter policies while Republicans preferred fewer restrictions. These policies included the length of time when masks were mandatory, the ability to gather in groups, and the mandatory shutdown of private businesses, such as restaurants, gyms, and theaters. Using statistics, we can determine whether these policies were correlated with politics at the state level. Although all states will have Democrats, Republicans, and independent voters, one of the best ways to determine the largest proportion of voters in a given state at the time of a presidential election is to examine the party of its Electoral College votes. All states are awarded Electoral College votes during presidential elections based on their representation in the Senate and House of Representatives, which, in turn, is based on the population of each state. Once the votes are counted in each state, 48 of the 50 states award their entire slate of electoral college votes to whomever won the popular votes. (Only Maine and Nebraska allow Electoral College votes to be divided among candidates, but in the 2020 elections, Maine's votes were awarded to the Democratic Party and Nebraska's votes were awarded to the Republican Party.) So if a Republican wins the popular vote in Alabama, the nine Electoral College votes allotted to Alabama are awarded to the Republican Party.

How America's partisan divide over pandemic responses played out in the states

Published: May 12, 2021 1.46pm BST

The COVID-19 pandemic seems to have widened the partisan divide between Democrats and Republicans on health care.

Email
Twitter 170
Facebook 958
LinkedIn
Print

Throughout the COVID-19 pandemic, a partisan divide has existed over the appropriate government response to the public health crisis. Democrats have been more likely to favor stricter policies such as prolonged economic shutdowns, limits on gathering in groups and mask mandates. Republicans overall have favored less stringent policies.

As political scientists and public health scholars, we've been studying political responses to the pandemic and their impacts. In research published in the summer of 2020, we found that "sub-governments," which in the U.S. means state governments, tended to have a bigger impact on the direction of pandemic policies than the federal government. Now, as data on last year's case and death rates emerge, we're looking at whether the political party in the governor's office became a good predictor of public health outcomes as COVID-19 moved across the country.

Looking at states' COVID-19 case and death rates, researchers are finding the more stringent policies typical of Democratic governors led to lower rates of infections and deaths, compared to the the pandemic responses of the average Republican governor. In preparation for future pandemics, it may be worth considering how to address the impact that a state government's partisan leanings can have on the scope and severity of a public health crises.

Description

Figure 9.1 Article

Source: VanDusky-Allen and Shevtsova (2021).

After identifying each state as Republican or Democrat based on the party of their Electoral College votes in November of 2019, we can then proceed to examine the number of mask-mandated days in each state from the start of the COVID-19 pandemic through March 28, 2022. This will allow us to determine if there was a statistically significant difference in mask-mandated days based on political ideology. This type of test is a two independent-samples *t* test, which is described next.

In this chapter, we will learn how to test for a statistically significant difference between two independent-sample means drawn from two populations. The test uses one continuous variable (number of mask-mandated days) and one categorical variable with two categories (Democrat and Republican). We will use data from Ballotpedia on state-level mask requirements and the 2020 Electoral College results as reported in the National Archives.[1] Other examples, along with a review of assumptions, procedures, and interpretation of the output, are included later.

9.2 WHEN TO USE A TWO INDEPENDENT-SAMPLES *T* TEST

There are many situations when we may want to compare two means. In this chapter, we only consider cases where there are two independent samples. This means that individuals or objects are assigned to one of two groups. Table 9.1 offers examples from different fields and identifies the continuous variable and the categorical variable with two groups.

TABLE 9.1 ■ Examples of two Independent-Samples *t* Test

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Criminal Justice	Are men more likely to commit delinquent offenses than women?	There is no difference in the number of delinquent offenses committed by men and by women.	Number of offenses committed	1. Men 2. Women
Economics	Do men earn more than women in the same job with the same set of skills?	There is no difference between salaries of men and women in the same job with the same skill level.	Annual salary	1. Men 2. Women
Political Science	Are Democratic voters younger than Republican voters?	There is no difference in the average age of Democrats and Republicans.	Age	1. Democrats 2. Republicans
Psychology	Does multitasking while studying for an exam have an impact on a student's final score?	There is no difference in the final scores between students who multitask and those who do not.	Exam score	1. Those who multitask 2. Those who do not multitask
Public Health	Do women who smoke give birth to infants with a lower birth weight?	There is no difference in birth weight of children of pregnant mothers who smoke and those who do not.	Birth weight	1. Pregnant mothers who smoke 2. Pregnant mothers who do not smoke
Sociology	Do Catholics or Protestants spend more time volunteering for community work?	There is no difference in the number of hours per week that Catholics and Protestants spend volunteering.	Hours per week volunteering	1. Catholics 2. Protestants

Field	**Research Question**	**Null Hypothesis**	**Continuous Variable**	**Categorical Variable**
Criminal Justice	Are men more likely to commit delinquent offenses than women?	There is no difference in the number of delinquent offenses committed by men and by women.	Number of offenses committed	1. Men 2. Women
Economics	Do men earn more than women in the same job with the same set of skills?	There is no difference between salaries of men and women in the same job with the same skill level.	Annual salary	1. Men 2. Women
Political Science	Are Democratic voters younger than Republican voters?	There is no difference in the average age of Democrats and Republicans.	Age	1. Democrats 2. Republicans

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Psychology	Does multitasking while studying for an exam have an impact on a student's final score?	There is no difference in the final scores between students who multitask and those who do not.	Exam score	1. Those who multitask 2. Those who do not multitask
Public Health	Do women who smoke give birth to infants with a lower birth weight?	There is no difference in birth weight of children of pregnant mothers who smoke and those who do not.	Birth weight	1. Pregnant mothers who smoke 2. Pregnant mothers who do not smoke
Sociology	Do Catholics or Protestants spend more time volunteering for community work?	There is no difference in the number of hours per week that Catholics and Protestants spend volunteering.	Hours per week volunteering	1. Catholics 2. Protestants

We can also consult the decision tree in Figure 8.2 and Appendix 3 when we are unsure about which test to use. Since we are comparing the means, we would follow the path on the left for "comparing means." Next, we would choose "comparing sample means" since we now have two sample means in this case—the average number of mask-mandated days in states with Democratic and Republican electoral votes. Underneath the "two independent sample means" is the **two independent-samples *t* test**.

9.3 CALCULATING THE *T* STATISTIC

To test for a significant difference between the two means, we must calculate a *t* statistic. Although Stata will calculate the *t* statistic in the example that follows in this section, it is important to understand how it is calculated in order to interpret its meaning. It is expressed in Equation 9.1 below.

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{S_{\overline{X}_1 - \overline{X}_2}} \quad (9.1)$$

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{S_{\overline{X}_1 - \overline{X}_2}}$$

The numerator is simply the observed difference between the two means and how much greater it is than zero, which is the hypothesized difference. The denominator is the standard error of the mean difference. This is calculated as follows:

$$S_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (9.2)$$

$$S_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where

- S_1^2 is the variance for the first sample
- S_2^2 is the variance for the second sample
- n_1 is the sample size for the first sample
- n_2 is the sample size for the second sample

Combined, the full formula tells us how many standard error units the observed difference is from zero. As described in Chapter 7, this indicates how unusual our results are if the true difference is zero.

9.4 CONDUCTING A *T* TEST

Using the state-level mask requirements in response to the COVID-19 pandemic and the National Archives data on the 2020 Electoral College results, we will examine the number of mask-mandated days in Democratic and Republican states based on their Electoral College votes in the 2020 presidential election. We can begin by generating a summary of these values using the commands that we learned in Chapter 6, as illustrated in Figure 9.2.

```
. table ElectorParty, stat(mean MaskDays) nformat(%5.0g) stat(sd MaskDays)
```

	Mean	Standard deviation
Party of Electors from 2020 Election		
Democrat	406	213
Republican	147	140
Total	276	221

Description

Figure 9.2 Mean and standard deviation of mask-mandated days in states with democratic and republican electors in the 2020 presidential election.

Based on these results, we can see that, on average, mask-mandated days are much higher in Democratic-leaning states and they also have a higher standard deviation than Republican-leaning states. We now want to use the data to test whether this is a statistically significant difference, beginning with our research question.

Research question

Did states with Democratic electors in the 2020 presidential elections have a greater number of mask-mandated days during the COVID-19 pandemic from 2020 to 2022 compared to states with Republican electors?

Null hypothesis

There is no difference in the average number of mask-mandated days during the COVID-19 pandemic between Democratic- and Republican-leaning states.

Variables

Continuous variable—number of mask-mandated dates (MaskDays)

Categorical variable—ElectorParty (Democrat = 1, Republican = 2)

Assumptions

As described earlier, to use the two independent-samples *t* test, you must have one continuous variable and one categorical variable with two categories. We also make the following assumptions to generate valid results:

---

1. ***Independence of observations***: Each individual can appear in only one of the two groups. In addition, they can only appear once in each group.
2. ***Normal distribution***: The dependent variable—the number of mask-mandated days, in this example—should be approximately normally distributed within each category. It only needs to be approximately normally distributed since minor violations of normality do not affect the results. Normality can be tested with the Shapiro–Wilk test.
3. *Homogeneity of variances:* The variances of the two groups must be equal. This is tested with **Levene's test**. If the variances are not equal, Stata will generate output to show the results with unequal variances assumed with the command **unequal**.

Stata code for doing a t test

Using a do-file, we would run the commands below.

```
robvar MaskDays, by(ElectorParty)
ttest MaskDays, by(ElectorParty)
esize twosample MaskDays, by(ElectorParty) cohensd
```

Note that we would add the **unequal** option to the **ttest** and **esize** commands if the robust variance test indicated a significant difference in the variances of the two variables.

Using menus in Stata, we would click on the sequence listed below that would bring us to a dialog box. In the dialog box, we would select the variables “MaskDays” and “ElectorParty” in the two drop-down menus.

Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Robust equal-variance test

We would then click on the following sequence to bring us to a second dialog box. Depending on the results from the equality of variance test, we would leave the box “unequal variances” either checked or unchecked. This is explained further in Section 9.4 on interpreting the output.

Statistics → Summaries, tables, and tests → Classical tests of hypotheses → *t* test (mean-comparison test)

Finally, we could click on the following sequence and fill in the variable names in the drop-down boxes.

Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Effect size based on means comparison

9.5 INTERPRETING THE OUTPUT

The first step to determine if there is a significant difference in the number of mask-mandated dates is to check for equality of variances. As we saw in the preview to the chapter, *homogeneity of variances* is one of the assumptions for this test. If the variances are not equal, this will increase the likelihood of rejecting the null hypothesis when it is true. We, therefore, first test the assumption, and then make a correction if the variances are unequal.

To test for equality of variances, we run the robust equal-variance test, which is known as Levene’s test of equality of variances.[2] Figure 9.3 shows the results, including the *F* statistic and the probability of observing the *F* statistic when the variances are equal. In this example, we only need to interpret the *p* value (labeled as Pr) at the end of the row labeled W0. This row is testing the variance relative to the mean of the variable.[3] Because the value is greater than 0.05, we would not reject the null hypothesis that the variances are equal.

```
. robvar MaskDays, by(ElectorParty)

   Party of |
   Electors |  Summary of No. of mask mandate days
  from 2020 |            as of 8/15/22
   Election |       Mean   Std. dev.       Freq.
------------+------------------------------------
   Democrat |     405.88   212.81454          25
  Republica |     146.56   140.39174          25
------------+------------------------------------
      Total |     276.22   221.34019          50

W0  =  1.5566141   df(1, 48)     Pr > F = 0.21821431

W50 =  1.4663531   df(1, 48)     Pr > F = 0.23185058

W10 =  1.9130489   df(1, 48)     Pr > F = 0.17302516
```

Description

Figure 9.3 Istata Output for Equality of Variance Test

Once we have determined that the variances are equal or unequal, we then run the *t* test. In this example, we do not need to specify equal variances in the Stata commands since Stata assumes that they are equal. If we had rejected the null hypothesis, then we would have added “unequal” at the end of the command line. The results are illustrated in Figure 9.4. In the first column, we see the average number of mask-mandated days for Democratic states (405.88) and Republican states (146.56) and the overall average number of mask-mandated days (276.22). The difference in the average mask-mandated days by Democratic and Republican states is listed as “diff” at the bottom of the “Mean” column (259.32). To test whether this difference is statistically significant, we examine the *t* value and the **significance level**. According to the output, the probability of observing a *t* value greater than 5.1 or less than −5.1 is less than 0.05. We therefore reject the null hypothesis and say that there is a statistically significant difference in the average mask-mandated days between Democratic and Republican states.

```
. ttest MaskDays, by(ElectorParty)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
Democrat |      25      405.88    42.56291    212.8145    318.0345    493.7255
Republic |      25      146.56    28.07835    140.3917    88.60914    204.5109
---------+--------------------------------------------------------------------
Combined |      50      276.22    31.30223    221.3402    213.3158    339.1242
---------+--------------------------------------------------------------------
    diff |             259.32    50.99014                156.7974    361.8426
------------------------------------------------------------------------------
    diff = mean(Democrat) - mean(Republic)                        t =   5.0857
H0: diff = 0                                     Degrees of freedom =       48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Description

Figure 9.4 Stata Output for two-Sample *t* Test With Equal Variances

Notice that the degrees of freedom are n − 2, or 50 − 2, which is equal to 48. If the variances had been unequal, we would then use Satterthwaite's degrees of freedom, which would take into account the unequal variances. In either case, the degrees of freedom would be printed under the *t* statistic.[4]

We can also examine the confidence interval from Figure 9.4, which we learned how to generate in Chapter 8. Notice that the 95% confidence interval for the mean difference is from 156.8 to 361.8. This suggests that we are 95% confident that the true value of the difference is within that range.

In addition to examining the significance level of the difference in the two means, we may also want to examine the *effect size*, or the magnitude of the difference between the two groups. There are several measures that can estimate the effect size, but **Cohen's *d*** is commonly used. It is calculated as the difference between two means divided by the pooled standard deviation for the two independent samples. The results are illustrated in Figure 9.5. According to Cohen (1988), effect sizes are defined as small when *d* = 0.2, medium when *d* = 0.5, and large when *d* = 0.8. Since the absolute value of Cohen's *d* in Figure 9.5 is 1.4, it is a large effect.

```
. esize twosample MaskDays, by(Elector) cohensd

Effect size based on mean comparison

                                Obs per group:
                                     Democrat =         25
                                   Republican =         25

------------------------------------------------------------------------------
             Effect size |   Estimate     [95% conf. interval]
-------------------------+----------------------------------------------------
                 Cohen's d |   1.43845     .8082568    2.056732
------------------------------------------------------------------------------
```

Description

Figure 9.5 Cohen's *d*

9.6 PRESENTING THE RESULTS

Presenting the results for a nontechnical audience

To present these results to a lay audience who may not be familiar with statistical tests, we could write the following:

> On average, there were 276 mask-mandated days per state during the COVID-19 pandemic. Our results show, however, that there is a statistically significant difference between the average mask-mandated days in Democratic states and Republican states. Democratic states had 406 mask-mandated days on average compared with Republican states, which had 147 mask-mandated days on average.

Presenting the results in a scholarly journal

To present these results in a peer-reviewed scholarly journal, we would need to include more information. This could be written as follows:

To test the hypothesis that Democratic and Republican states had the same number of mask-mandated days during the COVID-19 pandemic, we used a two independent-samples *t* test. The results indicated that on average, Democratic states had 405.88 mask-mandated days (*SD* = 42.56), compared with Republican states, which had 146.56 mask-mandated days (*SD* = 28.08). This was a statistically significant difference at the 0.05 level ($t(48) = 5.09$, $p < 0.05$). Examining the effect size, or magnitude of the difference, Cohen's *d* revealed that the difference between the means is a large effect ($d = 1.44$).

9.7 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes the hypothesis, test, procedures, and the Stata code used in the chapter (Tables 9.2 and 9.3). These are also summarized in Appendices 1 and 2. In this chapter, we include the information from Chapters 7 and 8 so that you can see how the procedures differ depending on the test and information available.

TABLE 9.2 ■ Procedures For Chapter 7, 8, And 9

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/ Type of Variables*	*Procedures/ Interpretation*
7. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score $((\bar{X}-\mu)/$ Standard error of mean$)$ 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score $((\bar{X}-\mu)/$ Standard error of mean$)$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of the mean difference = $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ 2. Calculate *t* statistic = $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpretation*

Chapter Title	***Null Hypothesis***	***Test***	***Info Known/Type of Variables***	***Procedures/Interpretation***
7. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score (($\overline{X}$ − *μ*)/Standard error of mean) 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score (($\overline{X}$ − *μ*)/Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpretation*
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of the mean difference = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ 2. Calculate *t* statistic = $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.

TABLE 9.3 ■ Code Used In Chapter 9

Function	Code
Table	table ElectorParty, stat(mean MaskDays) nformat(%5.0g) stat(sd MaskDays)
Test for equal variances	robvar MaskDays, by(ElectorParty)
Two independent-means test	ttest MaskDays, by(ElectorParty)
Cohen's d effect size test	esize twosample MaskDays, by (Elector) cohensd

Function	Code
Table	table ElectorParty, stat(mean MaskDays) nformat(%5.0g) stat(sd MaskDays)
Test for equal variances	robvar MaskDays, by(ElectorParty)
Two independent-means test	ttest MaskDays, by(ElectorParty)
Cohen's d effect size test	esize twosample MaskDays, by (Elector) cohensd

EXERCISES

1. You want to determine if men and women watch the same number of hours of television per week. Assume that the robust variance test determined that there was no statistically significant difference in the variances to answer this question.

	Mean Hours of TV Watched per Week	Sample Size	Variance
Men	14	20	100
Women	6	8	88

	Mean Hours of TV Watched per Week	Sample Size	Variance
Men	14	20	100
Women	6	8	88

   a. Based on the information in the table, determine if this is a statistically significant difference. (Hint: The degrees of freedom would be equal to $n_1 + n_2 - 2$).
   b. Use the information to calculate a 95% confidence interval of the mean difference.
2. Use the National Survey on Drug Use and Health 2015 to determine if there is a difference in the average age when men and women first try alcohol by following the instructions below.
   a. Determine if there is a significant difference in the average age when men and women (irsex) first try alcohol (alctry). For each command that you use, you will need to eliminate all observations above the age of 71 since there are observations with large numeric codes that represent bad data, individuals who never used alcohol, and individuals who didn't know or refused to answer. To do this, include the code "**. if alctry < 72"** at the end of each command line.
   b. Use Cohen's *d* to examine the effect size, again using "**, if alctry < 72"** at the end of the command line.
   c. Write the results of your findings for a nontechnical audience.
   d. Write the results of your findings for a journal article.
3. One of the arguments for school uniforms is that they will deter crime and increase student safety. We can explore this by using the School Survey on Crime and Safety data set from the 2015–2016 school year, which offers data on school characteristics, crimes, practices, and policies. The data represent 2,092 public schools in the United States. In particular, we can look at the total number of disciplinary actions at schools that do and do not require uniforms. One question, however, is whether uniforms lead to fewer incidents (a negative relationship) or more incidents lead schools to require uniforms (a positive relationship). The possibility that two variables may influence each other makes it difficult to identify and measure the causal relationship, a problem called *endogeneity* that is discussed in Chapters 12 and 13.

   Using the pu_ssocs16 data set, generate a table that shows the average, the standard deviation, and the sample size of the total number of disciplinary actions (DISTOT16) among schools that require and those that do not require uniforms (C0134). Format the table so that there are two digits to the right of the decimal point.

   a. Determine if there is a significant difference in the average number of disciplinary actions between schools that require and those that do not require uniforms.
   b. What is the null hypothesis?
   c. Can you reject the null hypothesis? Use statistics to support your conclusion.
4. In this chapter, we examined whether there was a difference in the average number of mask-mandated days in states with Republican and Democratic electors in the 2020 presidential election in the United States. We can now examine the COVID-19 rate, or the number of COVID-19 cases per 100,000 people, in each state based on the party of the electors and also based on the party of

the state's governor in 2020. Keep in mind, however, that we can't assume that the COVID-19 rate can be fully explained by the number of mask-mandated days. Other factors, such as the vaccination rate, the age composition of the population, and the proportion of people who are immunocompromised, for example, would play a large role in determining the COVID-19 rate in each state.

a. Using the data set, Covid.dta, test whether there is a significant difference in the average COVID-19 rate (CovidRate) in states based on the party of the electors in the 2020 presidential election (ElectorParty). Include a test of the effect size, and write the results of your finding for a nontechnical audience.
b. Use the Covid.dta data set to test whether there is a significant difference in the average COVID-19 rate (CovidRate) in states based on the party of the governor in 2020 (Governorparty). Include a test of the effect size, and write the results of your finding for a journal article.

KEY TERMS

Cohen's *d*

independence of observations

Levene's test

normal distribution

null hypothesis

significance level

two independent-samples *t* test

variables

Descriptions of Images and Figures

Back to Figure

Published: May 12, 2021 1.46pm BST

The COVID-19 pandemic seems to have widened the partisan divide between Democrats and Republicans on health care.

Throughout the COVID-19 pandemic, a partisan divide has existed over the appropriate government response to the public health crisis. Democrats have been more likely to favor stricter policies such as prolonged economic shutdowns, limits on gathering in groups and mask mandates. Republicans overall have favored less stringent policies.

As political scientists and public health scholars, we've been studying political responses to the pandemic and their impacts. In research published in the summer of 2020, we found that "sub-governments," which in the U.S. means state governments, tended to have a bigger impact on the direction of pandemic policies than the federal government. Now, as data on last year's case and death rates emerge, we're looking at whether the political party in the governor's office became a good predictor of public health outcomes as COVID-19 moved across the country.

Looking at states' COVID-19 case and death rates, researchers are finding the more stringent policies typical of Democratic governors led to lower rates of infections and deaths, compared to the the pandemic responses of the average Republican governor. In preparation for future pandemics, it may be worth considering how to address the impact that a state government's partisan leanings can have on the scope and severity of a public health crises.

Back to Figure

.table ElectorParty, stat(mean MaskDays) nformat(%5.0g) stat(sd MaskDays)

	Mean	Standard deviation
Party of Electors from 2020 Election		
Democrat	406	213
Republican	147	140
Total	276	221

Back to Figure

. robvar MaskDays, by (ElectorParty)

Party of Electors from 2020 Election	Summary of No. of mask mandate days as of 8/15/22		
	Mean	Std. dev.	Freq.
Democrat	405.88	212.81454	25
Republica	146.56	140.39174	25
Total	276.22	221.34019	50

W0 = 1.5566141, df (1, 48), Pr F = 0.21821431

W50 = 1.4663531, df (1, 48), Pr F = 0.23185058

W10 = 1.9130489, df (1, 48), Pr F = 0.17302516

Back to Figure

.ttest MaskDays, by (ElectorParty)

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
Democrat	25	405.88	42.56291	212.8145	318.0345	493.7255
Republic	25	146.56	28.07835	140.3917	88.60914	204.5109
Combined	50	276.22	31.30223	221.3402	213.3158	339.1242
diff		259.32	50.99014		156.7974	361.8426

Diff = mean (Democrat) - mean (Republic)

t = 5.0857

H0: diff = 0

Degrees of freedom = 48

Ha: diff< 0

Pr(T< t) 1.0000

Ha: diff != 0

Pr(|T|> (|t| = 0.0000

Ha: diff > 0 Pr(Tt)=0.0000

Pr(T> t) 0.0000

Back to Figure

. esize twosample MaskDays, by (Elector) cohensd

Effect size based on mean comparison

Obs per group:

Democrat = 25

Republican = 25

Effect size	Estimate	[95% conf. interval]	
Cohen’s d	1.43845	.8082568	2.056732

10 ONE-WAY ANALYSIS OF VARIANCE

CHAPTER PREVIEW

Steps	Example
Research question	Are children from wealthier families more likely to earn higher scores on the SAT than those from lower income families?
Null hypothesis	There is no difference in SAT scores among college students from families with different levels of income.
Test	One-way analysis of variance
Types of variables	One dependent continuous variable: SAT scores (SAT) One independent categorical variable with three or more categories: 1 = "Less than $60,000," 2 = "$60,000 to $99,999," 3 = "$100,000 to $149,000"... (FAMILYINC)
When to use	Comparing three or more means
Assumptions	1. Each sample is an independent random sample. 2. Normal distribution of the continuous variable 3. Homogeneity of variances
Additional tests needed	Bartlett's test of equality of variances
Stata code: generic	**Oneway continuousvar categoricalvar**
Stata code: example	**oneway SAT FAMILYINC**

Steps	**Example**
Research question	Are children from wealthier families more likely to earn higher scores on the SAT than those from lower income families?
Null hypothesis	There is no difference in SAT scores among college students from families with different levels of income.
Test	One-way analysis of variance
Types of variables	One dependent continuous variable: SAT scores (SAT) One independent categorical variable with three or more categories: 1 = "Less than $60,000," 2 = "$60,000 to $99,999," 3 = "$100,000 to $149,000"... (FAMILYINC)
When to use	Comparing three or more means
Assumptions	1. Each sample is an independent random sample. 2. Normal distribution of the continuous variable 3. Homogeneity of variances
Additional tests needed	Bartlett's test of equality of variances
Stata code: generic	**Oneway continuousvar categoricalvar**

Steps	Example
Stata code: example	oneway SAT FAMILYINC

10.1 INTRODUCTION


The SAT and ACT are less important than you might think

Published: January 25, 2023 8.24am EST

Whether on paper or computerized, standardized tests may be in decline.

Email
Twitter 22
Facebook 123
LinkedIn
Print

College admission tests are becoming a thing of the past.

More than 80% of U.S. colleges and universities do not require applicants to take standardized tests – like the SAT or the ACT. That proportion of institutions with test-optional policies has more than doubled since the spring of 2020.

And for the fall of 2023, some 85 institutions won't even consider standardized test scores when reviewing applications. That includes the entire University of California system.

Currently, only 4% of colleges that use the Common Application system require a standardized test such as the SAT or the ACT for admission.

Description

Figure 10.1 Article

Source: Churchill (2023). "The SAT and ACT are less important than you might think." *The Conversation*, January 25, 2023.

Many studies have shown that children from higher income families earn higher scores on the Scholastic Aptitude Test (SAT). There are several reasons for this, including the fact that wealthier families can afford expensive preparation courses or private tutors for their children. This gap in test results has led to several changes. The College Board, which administers the test, has revised the exam to level the playing field by eliminating obscure vocabulary and using texts that are more typical of what students use in school. They have also offered free online tutoring. But the biggest change is that 80% of colleges no longer require the SAT or ACT tests and some won't even consider it. This change was partly due to the COVID-19 pandemic, when it became unsafe to take exams in public locations. But it was also a reconsideration of the value of the tests and their inability to accurately predict the likelihood of success in college

In this chapter, we will learn how to test for a statistically significant difference between three or more means. The test is called a one-way **analysis of variance (ANOVA)**. It is used with one continuous variable (SAT scores in the previous example) and one categorical variable with three or more categories (different family income levels). Examples from different fields are given later, along with a review of assumptions, procedures, and interpretation of the output.

10.2 WHEN TO USE ONE-WAY ANOVA

Table 10.1 shows examples from different fields where you may have three or more means. Each categorical variable must have at least three categories, and only one continuous variable is used. In all of these examples, we are testing the impact of the categorical variable on the continuous variable. This

means that the continuous variable—SAT scores, for example—is the dependent variable since its value will depend on family income levels. Family income is then the independent variable.

TABLE 10.1 ■ Examples Of One-Way Analysis Of Variance

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Criminal justice	Does birth order have an effect on the number of self-reported delinquent acts?	There is no difference in the average number of self-reported delinquent acts by birth order.	Number of self-reported delinquent acts	1. First born (or only child) 2. Middle born (if three or more children) 3. Last born
Economics	Does annual income vary across regions in the United States?	There is no difference in annual income across regions in the United States.	Annual income	1. Northeast 2. Mid-Atlantic 3. South 4. Midwest 5. West
Political science	Is voter participation affected by the type of government?	There is no difference in voter participation in countries with different types of government.	Voter turnout rate	1. Liberal democracy 2. Communist or postcommunist 3. Socialist
Psychology	Does the type of car ownership affect behavior toward bicyclists on the road?	There is no difference in behavior.	Number of feet of clearance given to bicyclists on the road	1. High-end cars 2. Medium-priced cars 3. Low-priced cars
Public health	Is there a difference in average bone density among respondents who take three levels of calcium supplement?	There is no difference in bone density.	Bone density level	1. Low calcium intake 2. Medium calcium intake 3. High calcium intake
Sociology	What is the average number of children among families from different religions?	There is no difference in the average number of children by religion.	Number of children	1. Christians 2. Muslims 3. Hindus 4. Buddhists

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Criminal justice	Does birth order have an effect on the number of self-reported delinquent acts?	There is no difference in the average number of self-reported delinquent acts by birth order.	Number of self-reported delinquent acts	1. First born (or only child) 2. Middle born (if three or more children) 3. Last born

Field	Research Question	Null Hypothesis	Continuous Variable	Categorical Variable
Economics	Does annual income vary across regions in the United States?	There is no difference in annual income across regions in the United States.	Annual income	1. Northeast 2. Mid-Atlantic 3. South 4. Midwest 5. West
Political science	Is voter participation affected by the type of government?	There is no difference in voter participation in countries with different types of government.	Voter turnout rate	1. Liberal democracy 2. Communist or postcommunist 3. Socialist
Psychology	Does the type of car ownership affect behavior toward bicyclists on the road?	There is no difference in behavior.	Number of feet of clearance given to bicyclists on the road	1. High-end cars 2. Medium-priced cars 3. Low-priced cars
Public health	Is there a difference in average bone density among respondents who take three levels of calcium supplement?	There is no difference in bone density.	Bone density level	1. Low calcium intake 2. Medium calcium intake 3. High calcium intake
Sociology	What is the average number of children among families from different religions?	There is no difference in the average number of children by religion.	Number of children	1. Christians 2. Muslims 3. Hindus 4. Buddhists

10.3 CALCULATING THE *F* RATIO

The *F* ratio, which is used to determine if there is a statistically significant difference among several means, is calculated in two parts. The first part, or numerator, estimates the between-group variability and is expressed in Equation 10.1.

$$\text{Numerator} = \frac{\sum_{i=1}^{n} n_i(\bar{X}_i - \bar{X})^2}{K-1} \qquad (10.1)$$

(10.1)

$$\text{Numerator} = \frac{\sum_{i=1}^{n} n_i (X_i - X)^2}{K - 1}$$

where

n_i = sample size for group *i*

X_i = average for group *i*

X = the overall average of all observations

K = number of groups

Notice that the numerator examines how much the mean of each individual group differs from the overall mean of all groups combined. This is then weighted by the sample size, *n*, so that larger samples are given a greater weight. The denominator is the degrees of freedom, or the number of groups (*K*) minus 1[1]. Overall, this is a measure of how much variation there is "between" the groups, or the between-groups mean square.

The second part of the *F* ratio is the within-group variability. As its name suggests, we are now looking at how much variation there is within each sample or group. This is expressed in Equation 10.2.

$$\text{Denominator} = \frac{\sum_{i=1}^{n} s_i^2 (n_i - 1)}{\sum_{i=1}^{n} (n_i - 1)} \qquad (10.2)$$

(10.2)

$$\text{Denominator} = \frac{\sum_{i=1}^{n} s_i^2 (n_i - 1)}{\sum_{i=1}^{n} (n_i - 1)}$$

where

S_i^2 = the variance of group *i*

n_i = sample size for group *i*

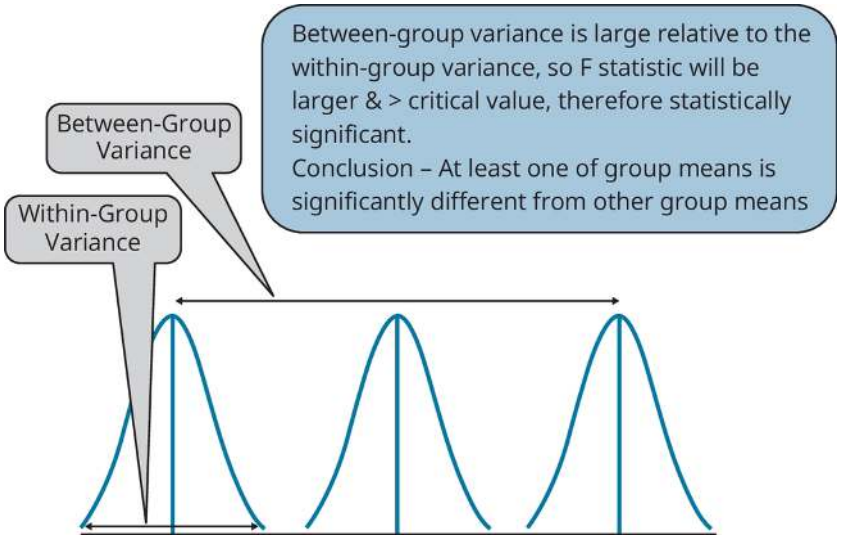
In this case, the numerator adds up the variance of each group and gives weight to each variance by multiplying by the sample size minus 1. When we then divide by the sum of the sample sizes minus 1, we are essentially getting the average variation within the groups. It is expressed as the within-groups mean square.

To calculate the *F* ratio, we then divide Equation 10.1 by Equation 10.2, which can be written as follows:

$$F = \frac{\text{Between} - \text{group variability}}{\text{Within} - \text{group variability}}$$

$$F = \frac{\text{Between} - \text{group variability}}{\text{Within} - \text{group variability}}$$

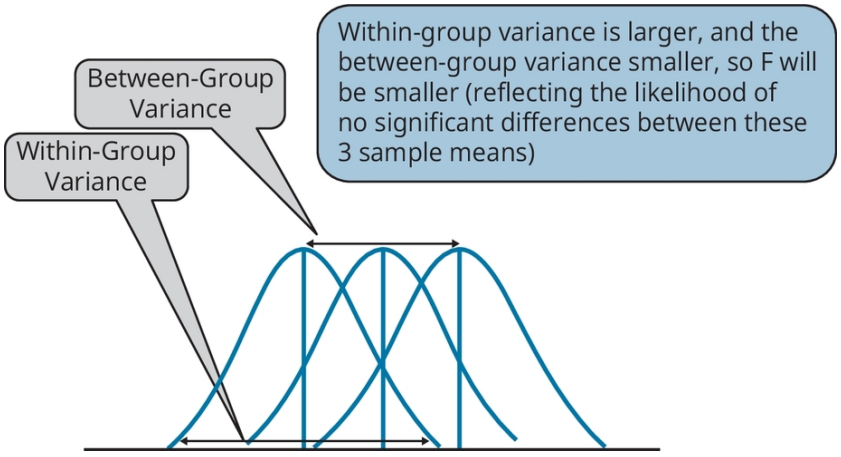
In other words, if the variability between the groups is greater than the variability within each group, you would expect a large *F* ratio. The larger the *F* ratio, the more likely you are to find a significant difference in the means. Figures 10.2 and 10.3 illustrate the concept of between- and within-group variance.



Description

Figure 10.2 Between-Group Variance Is Larger Than Within-Group Variance Illustration

Source: Khatri (2014). "Analysis of Variance (ANOVA)." *LinkedIn SlideShare*, 9 Aug. 2014, www.slideshare.net/snehkhatri/analysis-of-variance-anova.



Description

Figure 10.3 Within-Group Variance Is Larger Than Between-Group Variance Illustration

Source: Khatri (2014). "Analysis of Variance (ANOVA)." *LinkedIn SlideShare*, 9 Aug. 2014, www.slideshare.net/snehkhatri/analysis-of-variance-anova.

10.4 CONDUCTING A ONE-WAY ANOVA TEST

As described earlier, many people criticize the use of SAT scores because of their high correlation with income. Using the Admitted Student Questionnaire data set, we will examine the relationship between SAT scores and family income to see if there is a statistically significant difference among SAT scores in different income categories.

Figure 10.4 shows five income groups and the average SAT scores within each group along with the standard deviation and the number of students in each group. We can easily see that the average SAT score does increase as family incomes rise, but we cannot make the conclusion that there is a statistically significant difference until we run the one-way ANOVA test, which is described next.

```
. table FamilyInc, stat(mean SAT) stat(sd SAT) stat(n SAT) nformat (%5.0fc)
```

	Mean	Standard deviation	Number of nonmissing values
FamilyInc			
<59K	1,277	225	779
60-99K	1,312	189	641
100-149K	1,359	179	666
150-199K	1,369	176	296
>200K	1,434	143	796
Total	1,349	194	3,178

Description

Figure 10.4 Average Sat Score By Family Income

Research question

Are children from wealthier families more likely to earn higher scores on the SAT than those from lower income families?

Null hypothesis

There is no difference in SAT scores among college students from families with different levels of income.

The **alternative hypothesis** would be that at least one of the group means is not the same as the others.

Variables

Continuous variable—SAT scores of combined reading and math (SAT)

Categorical variable—family income before taxes broken into five income categories (FamilyInc)

Assumptions

As described earlier, a one-way ANOVA test is used with one categorical variable with three or more categories and one continuous variable. We also make the following assumptions to generate valid results:

1. *Independence of observations:* Each individual or observation can only appear in one of the three or more groups. In addition, they can only appear once in each group.
2. *Normal distribution:* The continuous variable, SAT score, should be approximately normally distributed within each category. It only needs to be approximately normally distributed since minor violations of normality do not affect the results.
3. *Homogeneity of variances:* The variances of the three or more groups must be equal. This is tested with **Bartlett's test**. For large samples, however, the equality of variances assumption is not required.

Procedures using code

Using a do-file, we would run the commands below:

```
oneway SAT FamilyInc, tabulate
```

Procedures using menus

Using menus in Stata, we would click on the sequence listed below that would bring us to a dialog box. In the dialog box, we would select the variables SAT and FamilyInc in the two drop-down menus as displayed.

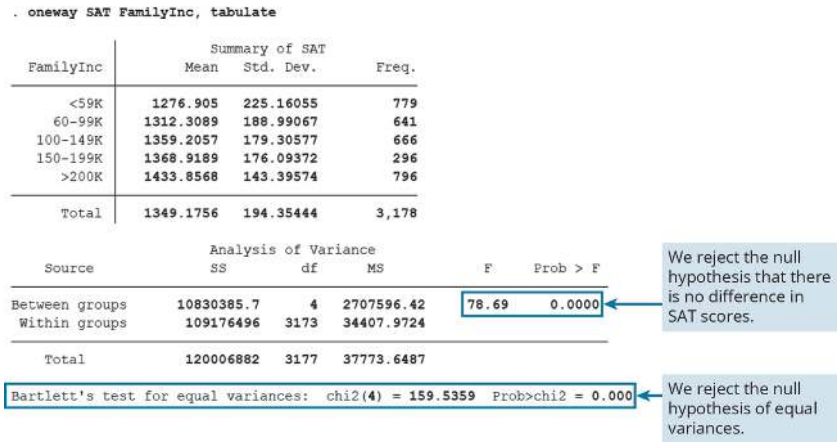
Statistics → Linear models and related → ANOVA/MANOVA → Oneway ANOVA

10.5 INTERPRETING THE OUTPUT

The first step to determine if there is a significant difference in SAT scores among different income groups is to check for equality of variances. In Stata, Bartlett's test for homogeneity of variances is automatically included in the output. The null hypothesis is that the variance for each of the five groups is equal. As illustrated in Output 10.5, the chi-square statistic is 159.5 and the significance level is 0.000. Because the significance level is less than 0.05, we reject the null hypothesis that the variances are equal. Although this means that one of the three assumptions for the one-way ANOVA is violated, the ANOVA results are typically considered acceptable if the sample size is large and in cases where the sample sizes are relatively equal for each group. You can, however, address the issue of unequal variances by using alternative tests, which can be found in more advanced statistics textbooks[2].

Examining the *F* ratio, we see that the value is 78.69 with a significance level that is less than 0.05. We then reject the null hypothesis that there is no difference in SAT scores among college students from families with different levels of income.

As we learned in Chapter 9, we may also want to examine the effect size or the magnitude of the difference between the two groups. When running an ANOVA test, we would use eta-square (η^2). This is calculated as the between-groups sum of squares divided by the total sum of squares. Using the numbers from Figure 10.5, this would be expressed as follows:



Description

Figure 10.5 Stata Output For Anova With Bartlett's Test

$$\eta^2 = \frac{\text{Between-group sum of squares}}{\text{Total sum of squares}} = \frac{10{,}830{,}385.7}{120{,}006{,}882} = 0.09$$

Turning this into a percentage, we can say that 9% of the variation in SAT scores can be explained by differences in family income. Although 9% seems low, our results did show that there was a significant difference in the mean SAT scores among income groups. Furthermore, it shows that the income accounts for a small amount of that variation and that we would need to examine other factors. It is important to keep in mind that insignificant or less dramatic results can be equally important when doing research. Not being able to show a significant difference is also a result.

10.6 IS ONE MEAN DIFFERENT, OR ARE ALL OF THEM DIFFERENT?

As mentioned earlier, the null hypothesis is that there is no difference in SAT scores among college students from families with different levels of income. The alternative hypothesis is that at least one of the means is not the same. Once we have rejected the null hypothesis, we may want to know which mean or means are different. To find out which means are different, we would run a multiple-comparison procedure. You could use multiple sets of two independent *t* tests to compare each pair of means, but the likelihood of finding a statistically significant difference in at least one pair of means increases as the number of comparisons increases, even when the means are equal. To account for the chance of this error, a multiple-comparison test adjusts the observed significance level, making it more difficult to find a statistically significant difference. The **Bonferroni test**, for example, divides the alpha level by the number of comparisons being made. In other words, with an alpha level of 0.05 and ten comparisons being made, 0.05/10 is equal to 0.005. So the **p-value** of each significance level must then be equal to or less than 0.005 in order to be considered a significant difference.

Figure 10.6 shows the commands to run a Bonferroni test and the output. The first row in the Bonferroni table compares the SAT score of students from families earning $60,000 to $99,000 with the scores of students from families earning less than $60,000. Within the first cell, 35.4039 is the average SAT score of children from families earning $60,000 to $99,000 minus the average SAT score of children from

families earning less than $60,000. The significance level is reported underneath at 0.003. With our new alpha level of 0.005, we can report that this is a statistically significant difference since 0.003 is less than 0.005. Examining all of the cells in this table, each of the 10 comparisons show a statistically significant difference.

```
. oneway SAT FamilyInc, bonferroni

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      10830385.7        4   2707596.42     78.69      0.0000
 Within groups      109176496      3173   34407.9724
------------------------------------------------------------------------
    Total           120006882      3177   37773.6487

Bartlett's test for equal variances:  chi2(4) = 159.5359  Prob>chi2 = 0.000

                    Comparison of SAT by FamilyInc
                                (Bonferroni)
Row Mean-
Col Mean        <59K     60-99K   100-149K   150-199K
---------------+--------------------------------------
  60-99K      35.4039
                0.003

100-149K      82.3007    46.8968
                0.000      0.000

150-199K      92.0139      56.61    9.71321
                0.000      0.000      1.000

   >200K      156.952    121.548    74.6511    64.9379
                0.000      0.000      0.000      0.000
```

Description

Figure 10.6 One-Way Analysis Of Variance Test With The Bonferroni Test

10.7 PRESENTING THE RESULTS

Presenting the results for a nontechnical audience

To present these results to a lay audience who may not be familiar with statistical tests, we could write the following:

Our results indicate that there is a statistically significant difference in SAT scores among children from families in five different income categories.

In the lowest category of income (less than $59,999), students earn, on average, 1,277 points on their combined reading and math SAT scores. In the wealthiest category of income (more than $200,000), students from these families earn 1,434 points on average. The results also show, however, that only 9% of the variation in SAT scores can be explained by income differences. It is therefore important to consider other factors that may affect SAT scores.

Presenting the results in a scholarly journal

In a peer-reviewed journal, we would include more information. These results could be explained as follows:

A one-way ANOVA was used to compare the combined math and reading SAT scores of students who come from families in five different income categories. The results indicate that there is a statistically significant difference at the 0.05 significance level among the SAT scores; $F(4, 3173) = 78.69$, $p < 0.001$. In particular, the scores rise as family income increases, with students from the lowest income category earning 1,276.91 on average (SD = 225.2) compared with 1,433.86 points earned on average (SD =143.3) among students from families in the wealthiest income category. Although Bartlett's test revealed unequal variances among the five income groups, the large sample sizes make the results robust. Using **eta-square** to examine the effect size, only 9% of the variation in SAT scores could be explained by income.

10.8 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes the hypothesis, test procedures, and the Stata code used in the chapter (Tables 10.2 and 10.3). They are also summarized in Appendices 1 and 2.

TABLE 10.2 ■ Procedures for Chapters 7, 8, 9, and 10

Chapter Title	Null Hypothesis	Test	Info Known/ Type of Variables	Procedures/ Interpretation
7. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	**1.** Standard error of mean $=(\sigma/\sqrt{n})$ **2.** Standard score $(\lvert\bar{X}-\mu\rvert/$ Standard error of mean$)$ **3.** Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	**1.** Standard error of mean $=(s/\sqrt{n})$ **2.** Standard score $(\lvert\bar{X}-\mu\rvert/$ Standard error of mean$)$ **3.** Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	**1.** Standard error of the mean difference $=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ **2.** Calculate *t* statistic $=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ **3.** Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		**1.** Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.
10. One-Way Analysis of Variance	There is no difference in SAT scores among college students from families with different levels of income.	One-way ANOVA	One categorical variable and more than two means	Calculate the *F* ratio by running the ANOVA test. When the null hypothesis is true, the probability of observing an *F* ratio at least as large as 78.69 is less than 0.05. Reject the null hypothesis.
	Variances of the groups are equal.	Bartlett's test for equal variances		**1.** Use the Bartlett's test from the output. When the null hypothesis is true, the probability of observing a chi-square at least as large as 159.54 is less than 0.05. Reject the null hypothesis.

Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpretation

Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpretation
1. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score ((X −μ)/Standard error of mean) 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score ((X - μ)/Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.

Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpretation
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of the mean difference= $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ 2. Calculate *t* statistic = $\frac{X_1-X_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.
10. One-Way Analysis of Variance	There is no difference in SAT scores among college students from families with different levels of income.	One-way ANOVA	One categorical variable and more than two means	Calculate the *F* ratio by running the ANOVA test When the null hypothesis is true, the probability of observing an *F* ratio at least as large as 78.69 is less than 0.05. Reject the null hypothesis.

Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpretation
	Variances of the groups are equal.	Bartlett's test for equal variances		1. Use the Bartlett's test from the output. When the null hypothesis is true, the probability of observing a chi-square at least as large as 159.54 is less than 0.05. Reject the null hypothesis.

TABLE 10.3 ■ Code used in Chapter 10

Function	Code
Table	table FamilyInc, stat(mean SAT) stat(sd SAT) stat(count SAT) nformat(%4.0f)
One-way analysis of variance	oneway SAT FamilyInc, tabulate
One-way analysis of variance with Bonferroni test	oneway SAT FamilyInc, bonferroni

Function	Code
Table	table FamilyInc, stat(mean SAT) stat(sd SAT) stat(count SAT) nformat(%4.0f)
One-way analysis of variance	oneway SAT FamilyInc, tabulate
One-way analysis of variance with Bonferroni test	oneway SAT FamilyInc, bonferroni

EXERCISES

1. Use the General Social Survey 2021 (GSS2021) data set to answer this question. You want to examine whether the number of hours that individuals work per week (hrs1) varies by education level (degree). To do this, you must first eliminate all part-time workers from the data set. This can be done by running the commands **keep if hrs1 > 31**. *When you have completed the assignment, do not save the data set since this will permanently remove part-time workers!*
   a. Generate two tables. In the first table, show the overall average of hours worked per week (hrs1) for all respondents in the sample who work full time. In a second table, show the average hours worked (hrs1), the standard deviation for hours worked, and the count for hours worked by education level (degree). Format the table so that there is one digit to the right of the decimal point.
   b. Use a one-way analysis of variance test to examine the number of hours worked per week by degree level.
   c. What is the null hypothesis?
   d. What is the alternative hypothesis?
   e. Write a paragraph that would explain your findings to a nontechnical audience.
   f. Write a paragraph that would explain your findings in a scholarly journal.
2. You want to compare the average number of hours that teenagers play video games on weeknights based on three age categories: (1) 10 to 12 years, (2) 13 to 15 years, and (3) 16 to 18 years. You are given the following information on the mean, variance, and sample size of each group. The

overall average for all individuals combined is 2. Based on this information, calculate the *F* statistic. Show all of your work (Table 10.4).

TABLE 10.4 ■ Calculate the *F* statistic

Age-Groups	Average Hours of Gaming on a Weeknight	Variance	Sample Size
10–12 years old	1	4	22
13–15 years old	3	25	31
16–18 years old	2	9	52

Age-Groups	Average Hours of Gaming on a Weeknight	Variance	Sample Size
10–12 years old	1	4	22
13–15 years old	3	25	31
16–18 years old	2	9	52

3. Use the Liberal Arts Colleges – USNews data set to determine if there are differences in the average SAT score among students in the top, middle, and bottom third of the colleges as ranked by *US News and World Report* (thirdrank).
   a. Generate a table that shows the average SAT score (sat_avg) by the ranking category of the college (thirdrank). Format the table so that it uses commas and only whole numbers.
   b. Run a one-way analysis of variance to determine if there is a statistically significant difference in the average SAT scores across ranking categories.
   c. What is the null hypothesis?
   d. What is the alternative hypothesis?
   e. What can you conclude from your results?
4. Socioeconomic mobility theories suggest that students from certain regions are more likely to go to college, earn higher incomes, or move from a low-income category to a higher income category. Use the School Survey on Crime and Safety from 2015 to 2016 (pu_ssocs16.dta) to explore this issue by answering the following questions:
   a. Generate a table that shows the average, standard deviation, and sample size for the percentage of students who are likely to go to college (C0534) by location (FR_URBAN). Format the table so that there are only whole numbers.
   b. Run a one-way analysis of variance to determine if there is a statistically significant difference in the percentage of students who are likely to go to college.
   c. Use the Bonferroni test, and explain the results.
   d. What can you conclude from your results?
5. You want to determine if the average SAT score (sat_avg) differs by type of university or region (USNewsType). Use the "College Score Card April 2023 – USNews.dta" data set to explore this issue.
   a. Generate a table that shows the average, standard deviation, and sample size of the averages SAT score (sat_avg) by type and region of college (USNewsType). Format the table so that there are only whole numbers.
   b. Run a one-way analysis of variance, including the Bonferroni test, to determine if there is a statistically significant difference in the average SAT scores by college type and region.
   c. Can you reject the null hypothesis that there is no difference in SAT scores?
   d. According to the Bonferroni test, which pairs of means show a statistically significant difference?

KEY TERMS

alternative hypothesis

analysis of variance (ANOVA)

Bartlett's test

Bonferroni test

eta-square

p-value

Descriptions of Images and Figures

Back to Figure

The information panel is titled "The SAT and ACT are less important than you might think" published on January 25, 2023, 8.24 am, EST. The content on the image is as follows.

“Whether on paper or computerized, standardized tests may be in decline.

College admission tests are becoming a thing of the past.

More than 80% of U.S. colleges and universities do not require applicants to take standardized tests – like the SAT or the ACT. That proportion of institutions with test-optional policies has more than doubled since the spring of 2020.

And for the fall of 2023, some 85 institutions won't even consider standardized test scores when reviewing applications. That includes the entire University of California system.

Currently, only 4% of colleges that use the Common Application system require a standardized test such as the SAT or the ACT for admission.”

On the left, “Email,” “Twitter,” “Facebook,” “Linkedin,” are “Print” are given.

Back to Figure

The graph has three bell curves arranged horizontally at a distance. The distance between each curve is marked as “Between-Group variance” and the length of each curve is marked as “Within-Group Variance.” The block above the graph reads “Between-group variance is large relative to the within-group variance, so F statistic will be larger & > critical value, therefore statistically significant. Conclusion – At least one of group means is significantly different from other group means.”

Back to Figure

The graph has three bell curves overlapping each other. The distance between curves is marked as “Between-Group variance” and the length of each curve is marked as “Within-Group Variance.” The block above the graph reads “Within-group variance is larger, and the between-group variance smaller, so F will be smaller (reflecting the likelihood of no significant differences between these 3 sample means).”

Back to Figure

The content of the image is given in the following table.

.table FamilyInc, stat(mean SAT) stat(sd SAT) stat(n SAT) format (%5.0fc)			

	Mean	Standard deviation	Number of nonmissing values
FamilyInc			
<59K	1,277	225	779
60-99K	1,312	189	641
100-149K	1,359	179	666
150-199K	1,369	176	296
>200K	1,434	143	796
Total	1,349	194	3,178

Back to Figure

The first table is titled “Oneway SAT FamilyInc, tabulate.” The content of the table is given in the following table.

	Summary of SAT		
FamilyInc	Mean	Std. Dev.	Freq.
<59K	1276.905	225.16055	779
60-99K	1312.3089	188.99067	641
100-149K	1359.2057	179.30577	666
150-199K	1368.9189	176.09372	296
>200K	1433.8568	143.39574	796
Total	1349.1756	194.35444	3,178

The second table is titled “Analysis of Variance.” The content of the table is given in the following table.

Source	SS	df	MS	F	Prob > F
Between groups	10830385.7	4	2707596.42	78.69	0.0000
Within groups	109176496	3173	34407.9724		
Total	120006882	3177	37773.6487		

“78.69” and “0.0000” are highlighted and marked as “We reject the null hypothesis that there is no difference in SAT scores.”

The text at the bottom reads “Bartlett’s test for equal variances: chi2 (4) = 159.5359 Prob>chi2 = 0.000.” and is marked as “We reject the null hypothesis of equal variances.”

Back to Figure

The image is titled “Oneway SAT FamilyInc, Bonferroni.” The first table is titled “Analysis of Variance.” The content of the table is given in the following table.

Source	SS	df	MS	F	Prob > F
Between groups	10830385.7	4	2707596.42	78.69	0.0000
Within groups	109176496	3173	34407.9724		
Total	120006882	3177	37773.6487		

The text below reads “Bartlett’s test for equal variances: chi2 (4) = 159.5359 Prob>chi2 = 0.000.”

The second table is titled “Comparison of SAT by FamilyInc (Bonferroni).” The content of the table is given in the following table.

Row Mean-Col mean	<599k	60-99k	100-149k	150-199k
60-99k	35.4039			
	0.003			
100-149k	82.3007	46.8968		
	0.000	0.000		
150-199k	92.0139	56.61	9.71321	
	0.000	0.000	1.000	
>200k	156.952	121.952	74.6511	64.9379
	0.000	0.000	0.000	0.000

11 COMPARING CATEGORICAL VARIABLES – THE CHI-SQUARED TEST AND PROPORTIONS

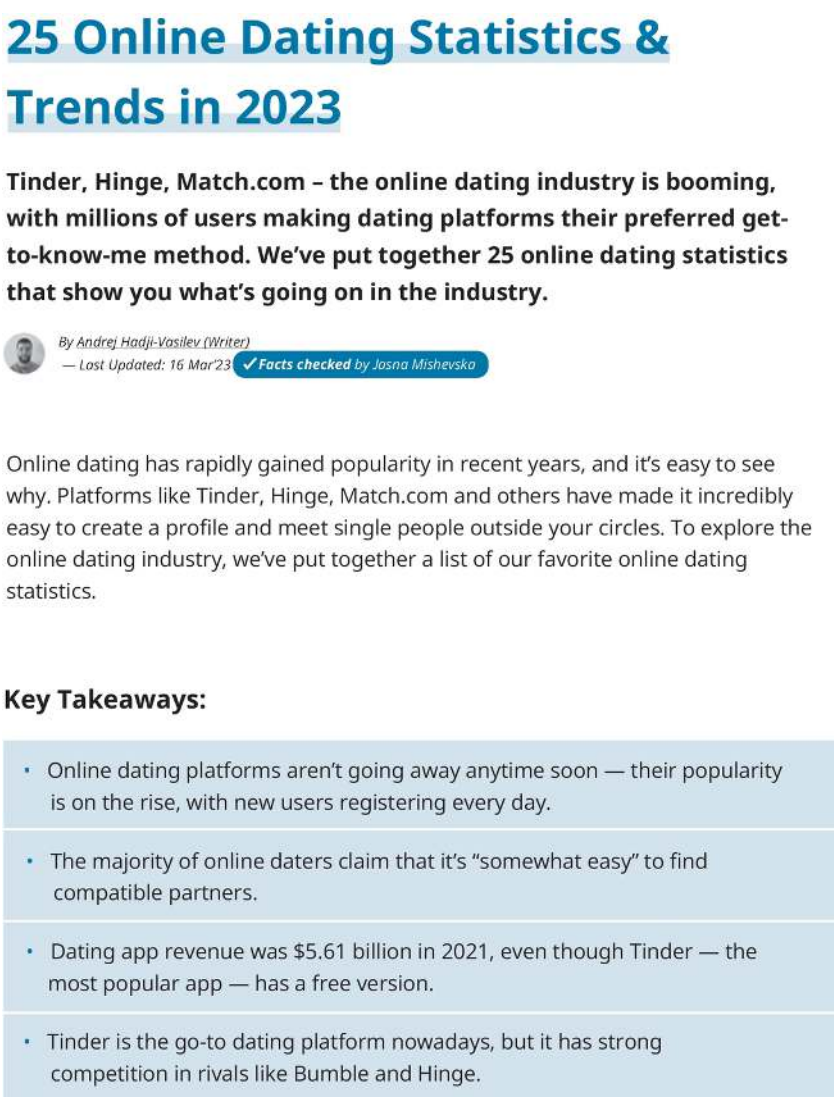
CHAPTER PREVIEW

Steps	Example
Research question	Do education levels differ between men and women who use online dating sites?
Null hypothesis	There is no difference in the education levels of men and women who use online dating sites.
Test	Chi-squared test
Types of variables	Two categorical variables with two or more categories in each: Sex—male or female Education level—high school, college, or graduate school
When to use	Comparing percentages
Assumptions	1. Independent observations 2. Minimum expected cell frequency should be 5 or greater in 80% of the cells.
Stata code: generic	**tab categoricalvar1 categoricalvar2, chi2 row** (or column if the independent variable is in the column)
Stata code: example	**tab sSex edu2, chi2 row nofre**

Steps	Example
Research question	Do education levels differ between men and women who use online dating sites?
Null hypothesis	There is no difference in the education levels of men and women who use online dating sites.
Test	Chi-squared test
Types of variables	Two categorical variables with two or more categories in each: Sex—male or female Education level—high school, college, or graduate school
When to use	Comparing percentages
Assumptions	1. Independent observations 2. Minimum expected cell frequency should be 5 or greater in 80% of the cells.
Stata code: generic	**tab categoricalvar1 categoricalvar2, chi2 row** (or column if the independent variable is in the column)
Stata code: example	**tab sSex edu2, chi2 row nofre**

11.1 INTRODUCTION

According to an article in June of 2022, 1 in 5 Americans were using an online dating app at the time and another 27% were formerly on a dating site (See Figure 11.1). Of those using a dating site at the time, 19% were talking to 11 or more people at once, and LGBTQ users were twice as likely to use a dating app. Over 13% of online dating users got engaged or married, and 24% claim they never had more than one or two dates. In terms of being matched, a much larger percentage of women (72%) think that it's essential to list the type of relationship you are looking for, compared to 53% of men. Women also report being more interested in their potential mate's occupation (27%) compared to 8% of men (Hadji-Vasilev, 2022).



25 Online Dating Statistics & Trends in 2023

Tinder, Hinge, Match.com – the online dating industry is booming, with millions of users making dating platforms their preferred get-to-know-me method. We've put together 25 online dating statistics that show you what's going on in the industry.

By Andrej Hadji-Vasilev (Writer)
— Last Updated: 16 Mar'23 ✓ Facts checked by Josna Mishevska

Online dating has rapidly gained popularity in recent years, and it's easy to see why. Platforms like Tinder, Hinge, Match.com and others have made it incredibly easy to create a profile and meet single people outside your circles. To explore the online dating industry, we've put together a list of our favorite online dating statistics.

Key Takeaways:

- Online dating platforms aren't going away anytime soon — their popularity is on the rise, with new users registering every day.
- The majority of online daters claim that it's "somewhat easy" to find compatible partners.
- Dating app revenue was $5.61 billion in 2021, even though Tinder — the most popular app — has a free version.
- Tinder is the go-to dating platform nowadays, but it has strong competition in rivals like Bumble and Hinge.

Description

Figure 11.1 Article

The rise in popularity of dating sites has also led to an increase in the number of dating sites. At the end of 2022, Tinder had the largest market share in the U.S. (30%) compared to Bumble (21%), Hinge (14%), and Plenty of Fish (13%). Worldwide, however, Badoo had the largest market share, with over 400 million users at the end of 2022.

In this chapter, we will learn how to test for a statistically significant difference in percentages using **Pearson's chi-squared test**. We use this test when there are two categorical variables with at least two categories in each variable. For example, based on the statistics just mentioned, we could test whether there was a statistically significant difference in the percentage men and women who report an interest in the occupation of a potential mate or who think that it is essential to list the type of relationship they are looking for. Since we do not have access to current worldwide trends, we will instead use the data from one particular dating app, OkCupid, that recorded the profile and general statistics for close to 60,000 users in the San Francisco area. In particular, we will examine whether the same percentage of men and women have completed high school, college, or graduate school among users of the dating app. In other words, if 60% of the overall population that uses the dating app has a college education, do 60% of women and 60% of men who use the dating app have a college education? Before turning to the the OkCupid data set to examine education levels, however, we provide examples of how the chi-squared test can be used in different fields.

11.2 WHEN TO USE THE CHI-SQUARED TEST

Table 11.1 shows examples from different fields where the chi-squared test can be used. As mentioned previously, there must be two categorical variables with at least two categories in each.

TABLE 11.1 ■ Examples Of The Chi-Squared Test

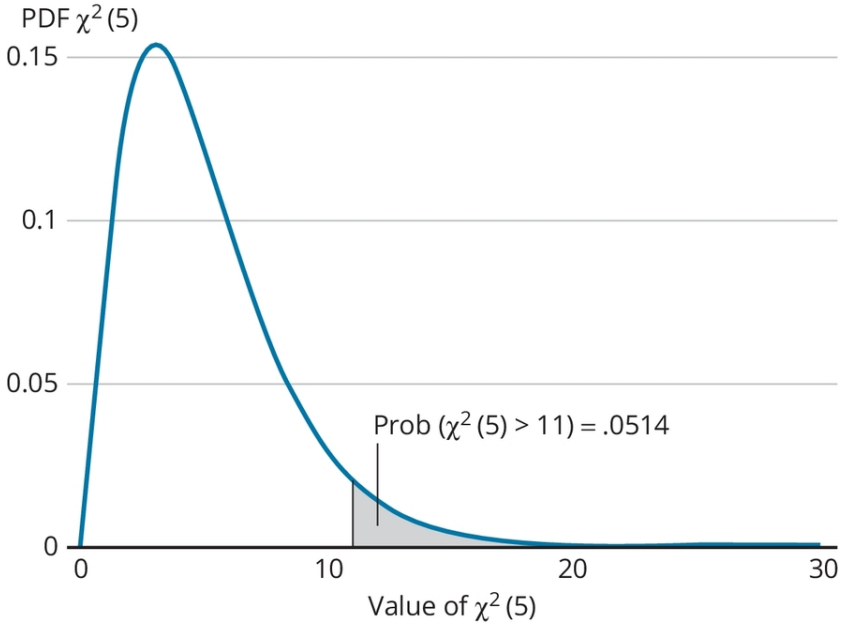
Field	Research Question	Null Hypothesis	Categorical Variables
Criminal justice	Are men and women equally likely to support decriminalization of marijuana?	Men and women are equally likely to support decriminalization of marijuana.	1. Sex 2. View on decriminalizing marijuana (yes or no)
Economics	Is income more equally distributed in developed countries?	There is no difference in income distribution between developed and developing countries.	1. Level of development (developed or developing) 2. Three levels of classification of equality based on ranges of the Gini coefficient (high equality, medium equality, and low equality)
Political science	Are men and women equally likely to vote for a Republican candidate for president?	Men and women are equally likely to vote for a Republican candidate for president.	1. Gender 2. Party they will vote for (Republican, Democrat, Green, independent)
Psychology	Does the ability to delay gratification among children lead to lower obesity?	There is no difference in obesity levels among those who were able to delay gratification and those who were not.	1. Ability to delay gratification (yes or no) 2. Obese at a later age (yes or no)
Public health	Is opioid abuse higher among men?	There is no difference in opioid abuse among men and women.	1. Opioid abuse (yes or no) 2. Gender
Sociology	Do men and women have the same reaction when a stranger invades their personal space?	There is no difference in the way men and women react when a stranger invades their personal space.	1. Gender 2. Reaction (negative, positive, or no reaction)

Field	Research Question	Null Hypothesis	Categorical Variables

Field	Research Question	Null Hypothesis	Categorical Variables
Criminal justice	Are men and women equally likely to support decriminalization of marijuana?	Men and women are equally likely to support decriminalization of marijuana.	1. Sex 2. View on decriminalizing marijuana (yes or no)
Economics	Is income more equally distributed in developed countries?	There is no difference in income distribution between developed and developing countries.	1. Level of development (developed or developing) 2. Three levels of classification of equality based on ranges of the Gini coefficient (high equality, medium equality, and low equality)
Political science	Are men and women equally likely to vote for a Republican candidate for president?	Men and women are equally likely to vote for a Republican candidate for president.	1. Gender 2. Party they will vote for (Republican, Democrat, Green, independent)
Psychology	Does the ability to delay gratification among children lead to lower obesity?	There is no difference in obesity levels among those who were able to delay gratification and those who were not.	1. Ability to delay gratification (yes or no) 2. Obese at a later age (yes or no)
Public health	Is opioid abuse higher among men?	There is no difference in opioid abuse among men and women.	1. Opioid abuse (yes or no) 2. Gender
Sociology	Do men and women have the same reaction when a stranger invades their personal space?	There is no difference in the way men and women react when a stranger invades their personal space.	1. Gender 2. Reaction (negative, positive, or no reaction)

11.3 CALCULATING THE CHI-SQUARE STATISTIC

In previous chapters, we examined differences in means and used the normal or the *t* distribution. When examining counts or percentages, we need to calculate a chi-square statistic and compare it with the chi-square distribution. Unlike the normal or *t* distributions that are bell shaped, the chi-square distribution is skewed to the right, as illustrated in Figure 11.2. Because the chi-square distribution is based on one or more squared variables, it can never be negative.



Description

Figure 11.2 Chi-Square Distribution

We use the chi-square statistic to determine the probability of observing our results when the null hypothesis is true. The formula for the chi-square statistic is illustrated in Equation 11.1.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (11.1)$$

(11.1)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where

O_i = the number of observations of type i

E_i = the expected number of type i

n = the number of cells in the table

This equation can be more easily understood with an example. Figure 11.3 uses data from the OkCupid data set to show the observed number, the expected number, and the percentage of men and women who have a high school, college, or graduate degree. In this case, sex is our independent variable, and the education level is the dependent variable. In other words, someone's education level may be influenced or depend on their sex.

```
. tab sex edu2, row exp
```

Key
frequency *expected frequency* *row percentage*

Sex	Education HS	College	Graduate	Total
Female	1,362	12,981	6,900	21,243
	1,913.9	13,100.7	6,228.4	21,243.0
	6.41	61.11	32.48	100.00
Male	3,290	18,862	8,239	30,391
	2,738.1	18,742.3	8,910.6	30,391.0
	10.83	62.06	27.11	100.00
Total	4,652	31,843	15,139	51,634
	4,652.0	31,843.0	15,139.0	51,634.0
	9.01	61.67	29.32	100.00

Description

Figure 11.3 Cross Tabulation Of Sex And Education Of Okcupid Users With With Observed And Expected Counts

The first cell shows that 1,362 women have a high school degree or less out of 21,243 women. This is the observed count. In the general population, or the "total" row, we see that 4,652 people, or 9.01%, have a high school degree or less. Our null hypothesis would suggest that 9.01% of men and 9.01% of women would have a high school degree or less. The expected count in the first cell is therefore 9.01% × 21,243, or 1,913.0. For men with a high school degree or less, the expected count is 9.01% × 30,391, or 2,738.1. After calculating the expected count for each cell, we can use Equation 11.1 to generate the chi-square statistic:

$$\chi^2 = \frac{(1{,}392 - 1{,}914)^2}{1{,}914} + \frac{(12{,}981 - 13{,}101)^2}{13{,}101} + \frac{(6{,}900 - 6{,}228)^2}{6{,}228} + \frac{(3{,}290 - 2738)^2}{2738}$$

$$\chi^2 = \frac{(1{,}392 - 1{,}914)^2}{1{,}914} + \frac{(12{,}981 - 13{,}101)^2}{13{,}101} + \frac{(6{,}900 - 6{,}228)^2}{6{,}228} + \frac{(3{,}290 - 2738)^2}{2738}$$

$$+\frac{(18{,}862 - 18{,}742)^2}{18{,}742} + \frac{(8{,}239 - 8{,}911)^2}{8{,}911} = 379$$

$$+\frac{(18{,}862 - 18{,}742)^2}{18{,}742} + \frac{(8{,}239 - 8{,}911)^2}{8{,}911} = 379$$

To determine if this is usual, we would compare this with the chi-square distribution. As with the *t* distribution, you would need to use degrees of freedom. For the chi-square statistic, the degrees of freedom are based on the number of rows and columns rather than the number of cases. In this case, the degrees of freedom are calculated as follows:

Degrees of freedom = (Number of rows in the table − 1) × (Number of columns − 1)

There are many online calculators that can determine the probability of observing a chi-square statistic at least as large as the one you observed when the null hypothesis is true. One example is the chi-square calculator by DI Management[1] (www.di-mgt.com.au/chisquare-calculator.html); it will calculate the *p* value and show you the graph.

When plugging in 379 for the chi-square value and 2 degrees of freedom, the *p* value is 0.00000. Fortunately, however, we will not need to calculate the chi-square statistic using the observed and expected counts since Stata will do this for us. This is illustrated in the next section.

11.4 CONDUCTING A CHI-SQUARED TEST

As described in the introduction, online dating has become extremely popular in the United States and around the world. Using the OkCupid data set of close to 60,000 users in the San Francisco area from 2015, we could examine many aspects of individuals who use online dating sites. In this section, we will use the same example from the previous section to assess whether education levels vary between men and women among online dating site users. We can then compare the results calculated previously with the same test generated by Stata.

Research question

Do education levels differ between men and women on online dating sites?

Null hypothesis

There is no difference in the education levels between men and women on online dating sites.

Variables

Categorical variable—education level (edu2)

Categorical variable—gender identity (sex)

Assumptions

1. *Independence of observations:* There should be only one observation for each participant.
2. *Minimum expected cell frequency:* There should be at least five observations per cell in the table in at least 80% of the cells.

Procedures using code

Using a do-file, we would run these commands:

```
tab sex edu2, nofreq row chi2 V
```

Procedures using menus

Using menus in Stata, we would click on the following sequence that would bring us to a dialog box where you would select sex and edu2 in the drop-down menus.

Statistics → Summaries, tables, and tests → Frequency tables → Two-way table with measures of association

11.5 INTERPRETING THE OUTPUT

Figure 11.4 shows the output for the chi-squared test. As illustrated, the chi-square statistic is close to the number that we calculated in Section 11.3. There is only a small difference due to rounding. Our results indicate that when the null hypothesis is true (There is no difference in the education levels between men and women on online dating sites), the probability of observing a chi-square statistic at least as large as 395 is less than 0.05. We therefore reject the null hypothesis and can state that there is a statistically significant difference in the education levels of men and women on the online dating app, OkCupid.

```
. tab sex edu2, nofre row chi2 V
```

Sex	Education HS	College	Graduate	Total
Female	6.41	61.11	32.48	100.00
Male	10.83	62.06	27.11	100.00
Total	9.01	61.67	29.32	100.00

```
Pearson chi2(2) = 395.2835   Pr = 0.000
     Cramér's V =   0.0875
```

Description

Figure 11.4 Stata Output For The Perason Chi-Squared Test

As we saw in previous chapters, we may want to examine effect size or the magnitude of the difference in education levels between men and women. This is particularly important because larger samples will often indicate a significant difference even when the difference is quite small.

There are several measures to examine the effect size, but **Cramér's *V*** is the most commonly used.[2] It is calculated as follows:

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n[\min(k-1, r-1)]}} \qquad (11.2)$$

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n[\min(k-1, r-1)]}} \qquad (11.2)$$

where

n = number of observations

k = number of columns

r = number of rows

Cramér's *V* generates a correlation coefficient that can range from 0 to 1 with 0 representing no association and +1 representing perfect correlation. In other words, a score of +1 would mean that sex can fully explain the difference in education levels among OkCupid users. The dependent variable, education levels, is dependent on sex.

For a 2 × 32 table, as in this example, a Cramér's *V* <0.07 is considered small, <0.21 is medium, and <0.35 is a large difference between the two proportions. Based on our example, the *V* of 0.0875 indicates that there is a small correlation between the two variables. In other words, sex is a significant factor in determining education levels among dating app users, but it has only a small effect[3]. Table 11.2 shows how to interpret the effect size based on degrees of freedom from 1 through 5.

TABLE 11.2 ■ Interpretation of Cramér's V

Degrees of Freedom	Small	Medium	Large
1	.2	.3	.5
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

Degrees of Freedom	Small	Medium	Large
1	.2	.3	.5
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

11.6 PRESENTING THE RESULTS

Presenting the results for a nontechnical audience

To present these results to a lay audience who may not be familiar with statistical tests, we could write the following:

Our results indicate that there is a statistically significant difference in the percentage of men and women who have a high school, college, or graduate-level education. A larger percentage of women have graduate-level education (32%) when compared with men (27%). A similar percentage have a college-level education, and a higher percentage men (11%) have a high school education compared to women (6%).

Presenting the results in a scholarly journal

In a peer-reviewed journal, we would include more information. These results could be explained as follows:

A chi-squared test for independence indicated that there is a statistically significant difference in the percentage of men and women who have a high school, college, and graduate school level of education: $\chi^2(2, n = 51{,}634) = 395$, $p = 0.00$, Cramér's $V = 0.09$. Thirty-two percent of women have a graduate-level education compared to 27% of men. At the college level, 61% of women have a college degree compared to 62% of men. In terms of high school, 11% of men have a high school degree compared to 6% of women.

11.7 COMPARING PROPORTIONS OR BINARY CATEGORICAL VARIABLES

As described previously, the chi-squared test is used to examine the association or independence between two categorical variables. In the prior example, we examined the two categorical variables of sex (male or female) and education levels (high school, college, or graduate school). The data set was based on responses among people who use OkCupid.

We can also test for an association between categorical variables when both variables are binary (for example, a success or a failure outcome or a yes/no answer). In this case, we would be comparing the proportions of two independent groups.

Let's suppose that instead of examining the educational characteristics of men and women who use dating apps, we want to compare the proportion of men and women who use dating apps. The null hypothesis would be that there is no difference in the proportion of men and women among online dating site users. Let's assume that 25% men in a sample of 400 report that they use dating apps and 18% of women in a sample of 500 report the same. We would begin by calculating the standard error as follows:

$$SE_{p1-p2} = \sqrt{\frac{\pi(1-\pi)}{n1} + \frac{\pi(1-\pi)}{n2}} = \sqrt{\frac{0.25(1-0.25)}{400} + \frac{0.18(1-0.018)}{500}} = 0.02$$

$$\boldsymbol{SE_{p1-p2} = \sqrt{\frac{\pi(1-\pi)}{n1} + \frac{\pi(1-\pi)}{n2}} = \sqrt{\frac{0.25(1-0.25)}{400} + \frac{0.18(1-0.018)}{500}} = 0.02}$$

We would then calculate a *z* score as follows:

$$\boldsymbol{z = \frac{(p1 - p2)}{SE_{p1-p2}} = \frac{0.25 - 0.18}{0.02} = \frac{0.07}{0.02} = 3.5}$$

$$z = \frac{(p1 - p2)}{SE_{p1-p2}} = \frac{0.25 - 0.18}{0.02} = \frac{0.07}{0.02} = 3.5$$

Using a *z* score calculator or the table in Appendix 5 and an alpha level of 0.05, we see that the probability of observing a *z* score when the null hypothesis is true is less than or equal to 0.00047 for a two-tailed test. We therefore reject the null hypothesis that the two proportions are equal. Assuming that we had the full data set related to this question, we could also run this test using the Stata commands prtest date, by(sex) level(95) in which "date" is the variable that asks whether someone uses dating apps (yes/no) and sex is the variable name for sex of the respondent.

11.8 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes all of the Stata code used in the chapter. In addition, all Stata code used throughout the book is summarized in Appendix 1. We also show the hypothesis, test, and procedures (Tables 11.3 and 11.4).

TABLE 11.3 ■ PROCEDURES FOR 7, 8, 9, 10, and 11

Chapter Title	Null Hypothesis	Test	Info Known/ Type of Variables	Procedures/ Interpretation
7. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = ($\sigma/\sqrt{n}$) 2. Standard score ($|\bar{X} - \mu|$/ Standard error of mean) 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = ($s/\sqrt{n}$) 2. Standard score ($|\bar{X} - \mu|$/ Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of the mean difference = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ 2. Calculate *t* statistic = $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.
10. One-way Analysis of Variance	There is no difference in SAT scores among college students from families with different levels of income.	One-way ANOVA	One categorical variable and more than two means	Calculate the *F* ratio by running the ANOVA test When the null hypothesis is true, the probability of observing an *F* ratio at least as large as 78.69 is less than 0.05. Reject the null hypothesis.
	Variances of the groups are equal.	Bartlett's test for equal variances		1. Use the Bartlett's test from the output When the null hypothesis is true, the probability of observing a chi-square at least as large as 159.54 is less than 0.05. Reject the null hypothesis.
11. Cross Tabulation and the Chi-Square Statistic	There is no difference in the education level of men and women among users of online dating sites.	Chi-square statistic	Two categorical variables Comparing percentages, not means.	Calculate the chi-square statistic by running the Pearson chi-squared test When the null hypothesis is true, the probability of observing a chi-square statistic at least as large as 395 is less than 0.05. Reject the null hypothesis.

Chapter Title	***Null Hypothesis***	***Test***	***Info Known/Type of Variables***	***Procedures/Interpretation***
7. The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score (($\overline{X}$ − *μ*)/Standard error of mean) 3. Look up percentages for standard score using normal distribution When the null hypothesis is true, the probability of observing a *z* score greater than +1.41 or less than −1.41 is less than 0.16. Do not reject the null hypothesis.
8. Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score (($\overline{X}$ − *μ*)/Standard error of mean) 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a *F* value greater than 3.25 or less than −3.25 is less than 0.0029. Reject the null hypothesis.

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpretation*
9. Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days among Democratic and Republican states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of the mean difference= $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ 2. Calculate *t* statistic = $\frac{\overline{X}_1-\overline{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothesis is true, the probability of observing a value greater than 5.1 or less than −5.1 is less than 0.01. Reject the null hypothesis.
	Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from output When the null hypothesis is true, the probability of observing an *F* value at least as large as 1.56 is less than 0.22. Do not reject the null hypothesis.
10. One-way Analysis of Variance	There is no difference in SAT scores among college students from families with different levels of income.	One-way ANOVA	One categorical variable and more than two means	Calculate the *F* ratio by running the ANOVA test When the null hypothesis is true, the probability of observing an *F* ratio at least as large as 78.69 is less than 0.05. Reject the null hypothesis.

Chapter Title	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpretation*
	Variances of the groups are equal.	Bartlett's test for equal variances		1. Use the Bartlett's test from the output When the null hypothesis is true, the probability of observing a chi-square at least as large as 159.54 is less than 0.05. Reject the null hypothesis.
11. Cross Tabulation and the Chi-Square Statistic	There is no difference in the education level of men and women among users of online dating sites.	Chi-square statistic	Two categorical variables Comparing percentages, not means.	Calculate the chi-square statistic by running the Pearson chi-squared test When the null hypothesis is true, the probability of observing a chi-square statistic at least as large as 395 is less than 0.05. Reject the null hypothesis.

TABLE 11.4 ■ Code Used In Chapter 11

Function	Code
Chi-square statistic with Cramer's V	tab sex edu2, nofreq row chi2 V
Difference in proportions of two independent samples	prtest date, by(sex) level(95)

Function	**Code**
Chi-square statistic with Cramer's V	tab sex edu2, nofreq row chi2 V
Difference in proportions of two independent samples	prtest date, by(sex) level(95)

EXERCISES

1. Misuse of prescription pain relievers has become a national crisis in the United States. Use the National Survey on Drug Use and Health data set to examine differences in prescription pain reliever abuse between men and women in the United States.
   a. Generate a table that compares the percentage of men and women (irsex) who have ever misused pain relievers (pnrnmflag). Be sure to use row or column percentages, depending on which one is appropriate. Also include Cramér's *V*.
   b. What is the null hypothesis?
   c. Based on your results, would you reject the null hypothesis?
   d. Using the appropriate statistics from your results, explain your answer to Part C.

e. Interpret Cramér's *V*. What does it mean in the context of this example?
f. Explain your results in a few sentences to a nontechnical audience.
g. Explain your results in a few sentences for a scholarly journal.

2. Use the same data set and research question from Question 1 to generate a new table that shows the observed and expected frequencies for each cell.
3. Based on your table from question 2, write out the full equation for the chi-square statistic, and calculate it using a calculator. Round each expected frequency to the nearest whole number in your equation.
4. Use Stata and the GSS2021.dta file to examine whether people with different levels of education (degree) believe in life after death (postlifev).
   a. What is the null hypothesis?
   b. Explain why you would or would not reject the null hypothesis using output from your analysis.
   c. Calculate the effect size, and interpret the number.
   d. In a few sentences, explain your results for a nontechnical audience.
   e. In a few sentences, explain your results for a scholarly journal.
5. Use the OkCupid data set to compare the percentage of men and women (sex) who like or dislike cats (likescats).
   a. What is the null hypothesis?
   b. Explain why you would or would not reject the null hypothesis using output from your analysis.
   c. Calculate the effect size and interpret the number.
   d. In a few sentences, explain your results for a nontechnical audience.
   e. In a few sentences, explain your results for a scholarly journal.

KEY TERMS

Cramér's *V*

independence of observations

null hypothesis:

research question

Pearson's chi-squared test

Descriptions of Images and Figures

Back to Figure

The article is titled "25 Online Dating Statistics & Trends in 2023." The content in the image is given as follows.

"Tinder, Hinge, Match.com – the online dating industry is booming, with millions of users making dating platforms their preferred get-to-know-me method. We've put together 25 online dating statistics that show you what's going on in the industry.

By Andrej Hadji-Vasilev (Writer)

— Last Updated: 16 Mar'23. Facts checked by Jasna Mishevska.

Online dating has rapidly gained popularity in recent years, and it's easy to see why. Platforms like Tinder, Hinge, Match.com and others have made it incredibly easy to create a profile and meet single people outside your circles. To explore the online dating industry, we've put together a list of our favorite online dating statistics.

Key Takeaways:

> Online dating platforms aren't going away anytime soon — their popularity is on the rise, with new users registering every day.
>
> The majority of online daters claim that it's "somewhat easy" to find compatible partners.
>
> Dating app revenue was $5.61 billion in 2021, even though Tinder — the most popular app — has a free version.
>
> Tinder is the go-to dating platform nowadays, but it has strong competition in rivals like Bumble and Hinge."

Back to Figure

The horizontal axis represents the "Value of Chi-square (5)" ranging from 0 to 30 in increments of 10. The vertical axis represents "PDF Chi-square (5)" ranging from 0 to 0.15 in increments of 0.05. The curve starts at (0, 0), increases to reach approximately 0.155, then decreases to reach (20, 0), and then stays constant. The area between the curve and the vertical line at 11 is shaded and marked as "Prob (Chi-square (5) > 11)=.0514."

Back to Figure

The table is titled ".tab sex edu2, row exp." The text below reads

"Key:

Frequency

Expected Frequency

Row Frequency."

The content of the table is given in the following table.

Sex	Education			
	HS	College	Graduate	Total
Female	1,362	12,981	6,900	21,243
	1,913.9	13,100.7	6,228.4	21,243.0
	6.41	61.11	32.48	100.00
Male	3,290	18,862	8,239	30,391
	2,738.1	18,742.3	8,910.6	30,391.0
	10.83	62.06	27.11	100.00
Total	4,652	31,843	15,139	51,634
	4,652.0	31,843.0	15,139.0	51,634.0
	9.01	61.67	29.32	100.00

Back to Figure

The table is titled ".tab sex edu2, nofre row chi2 V." The content of the table is given in the following table.

Sex	Education			
	HS	College	Graduate	Total
Female	6.41	61.11	32.48	100.00
Male	10.83	62.06	27.11	100.00
Total	9.01	61.67	29.32	100.00

The text below reads "Pearson chi2(2)=395.2835; Pr=0.0000; Cramer's V=0.0875."

PART IV EXPLORING RELATIONSHIPS

12 LINEAR REGRESSION ANALYSIS

CHAPTER PREVIEW

Steps	Examples
Research question	What factors influence the value of new and used cars?
Null hypothesis	Each factor has no effect on its value.
Test	*t* test of the each **coefficient** in multiple regression analysis
Types of variables	One continuous dependent variable (price) and multiple independent variables (including mileage)
When to use	To examine the relationship between one continuous dependent variable and one or more independent variables
Assumptions	Independent variables are measured without error. All relevant variables are included. The functional form is correct. The variance of the residuals is constant. The error terms are not correlated with each other. The error term is not correlated with any independent variables.
Additional tests needed	Tests for normality, omitted variables, multicollinearity, and heteroscedasticity (see Chapter 13)
Stata code: generic	**regress depvar indepvars** Where depvar is the dependent variable and indepvars is a list of one or more independent variables
Stata code: example	**regress price year mileage**

Steps	Examples
Research question	What factors influence the value of new and used cars?
Null hypothesis	Each factor has no effect on its value.
Test	*t* test of the each **coefficient** in multiple regression analysis
Types of variables	One continuous dependent variable (price) and multiple independent variables (including mileage)
When to use	To examine the relationship between one continuous dependent variable and one or more independent variables
Assumptions	Independent variables are measured without error. All relevant variables are included. The functional form is correct. The variance of the residuals is constant. The error terms are not correlated with each other. The error term is not correlated with any independent variables.
Additional tests needed	Tests for normality, omitted variables, multicollinearity, and heteroscedasticity (see Chapter 13)
Stata code: generic	**regress depvar indepvars** Where depvar is the dependent variable and indepvars is a list of one or more independent variables
Stata code: example	**regress price year mileage**

12.1 INTRODUCTION

We are often interested in exploring the effect of different factors on a variable of interest. For example, what factors influence the market price of new and used cars? Experts say that age, mileage, and condition are the main determinants of value, but other factors play a role, such as options, location, and color (D'Allegro, 2021). In this chapter, we will learn a statistical method called **regression analysis**, which is used to study the effect of one or more independent variables on one dependent variable. The ***dependent variable*** is an outcome variable that we wish to explain using a number of other variables. The ***independent variables*** are the variables used to "explain" the variation in the dependent variable; they are also called explanatory variables. Regression analysis uses data on the variables of interest to generate an equation that best describes the relationship between the dependent variable and the independent variables. Using the example from above, we can use regression analysis to generate an equation that predicts the price of cars as a function of mileage, year, and other factors.

This chapter emphasizes the use of regression analysis and the interpretation of the results. It does not look "under the hood" to explain the calculation of coefficients, standard errors, and test statistics. For additional information on regression analysis, the reader may consult Bailey (2017), Greene (2018), or Woolridge (2016), which provide much more in-depth treatments of regression analysis.

12.2 WHEN TO USE REGRESSION ANALYSIS

Regression analysis is widely used in economics, sociology, psychology, business studies, and other fields. Table 12.1 shows examples from different fields where multiple regression is used. In each case, there is a research question, a null hypothesis, a continuous dependent variable, and one or more independent variables. Each of these can be tested using **multiple regression analysis**.

TABLE 12.1 ■ Examples of Multiple Regression Analysis

Field	Research Question	Null Hypothesis	Continuous Dependent Variable	Independent Variables
Criminal Justice	Do youth sports programs predict a lower arrest rate among teenagers?	Youth sports programs are not associated with teenage arrest rate.	Number of arrests of teenagers per 100,000 teenagers in each county	Size of youth sports program and other county characteristics
Economics	How does meat demand vary with income?	Income has no effect on meat demand.	Household meat consumption from a survey	Income and other household characteristics from the survey
Political Science	How does county average education level predict county-level support for a political party?	Education level does not predict support for a political party.	Share of a county supporting a political party in a national race	Average education and other voter characteristics in each county
Psychology	How are family history characteristics associated with psychological well-being?	Family history characteristics are not associated with psychological well-being.	Indicator of psychological well-being from a survey	Family history characteristics from the survey
Public Health	Is the incidence of COVID-19 related to the percentage of the population vaccinated?	The share of people vaccinated in a county is not related to the prevalence of COVID-19.	Share of people who contract COVID-19 in a given year in each county	Share of people who have received the vaccine and other health factors
Sociology	Is the number of children a couple has affected by the parents' education?	The parents' education has no effect on the number of children a couple has.	Number of children a couple has according to a survey	Education of the father and education of the mother according to the survey

Field	Research Question	Null Hypothesis	Continuous Dependent Variable	Independent Variables
Criminal Justice	Do youth sports programs predict a lower arrest rate among teenagers?	Youth sports programs are not associated with teenage arrest rate.	Number of arrests of teenagers per 100,000 teenagers in each county	Size of youth sports program and other county characteristics
Economics	How does meat demand vary with income?	Income has no effect on meat demand.	Household meat consumption from a survey	Income and other household characteristics from the survey
Political Science	How does county average education level predict county-level support for a political party?	Education level does not predict support for a political party.	Share of a county supporting a political party in a national race	Average education and other voter characteristics in each county

Field	Research Question	Null Hypothesis	Continuous Dependent Variable	Independent Variables
Psychology	How are family history characteristics associated with psychological well-being?	Family history characteristics are not associated with psychological well-being.	Indicator of psychological well-being from a survey	Family history characteristics from the survey
Public Health	Is the incidence of COVID-19 related to the percentage of the population vaccinated?	The share of people vaccinated in a county is not related to the prevalence of COVID-19.	Share of people who contract COVID-19 in a given year in each county	Share of people who have received the vaccine and other health factors
Sociology	Is the number of children a couple has affected by the parents' education?	The parents' education has no effect on the number of children a couple has.	Number of children a couple has according to a survey	Education of the father and education of the mother according to the survey

The chapter begins with a description of correlation—a simple descriptive tool for measuring the strength of the relationship between two variables. Next, we consider simple linear regression, with a continuous dependent variable and *one* independent variable. Last, multiple linear regression is described, which has a continuous dependent variable and *multiple* independent variables. Chapter 13 describes diagnostic tools for regression analysis, including how to incorporate nonlinear relationships. Chapter 14 considers the case of regression analysis when the dependent variable is binary, rather than continuous. And Chapter 15 provides a brief overview of several advanced topics in regression analysis.

12.3 CORRELATION

Suppose we are interested in examining the relationship between two continuous variables, such as the price and mileage of a sample of cars. We can start by exploring the data visually with a scatterplot of the two variables. A scatterplot can tell us at a glance whether the two variables are positively related or negatively related. If the scatterplot shows an upward sloping pattern, the two variables are positively correlated, meaning that high values of one variable are associated with high values of the other. For example, daily temperature and ice cream sales are likely to be positively correlated. If the scatterplot shows a downward-sloping pattern, the variables are negatively correlated, meaning that high values of one variable are associated with low values of the other. For example, the number of sunny days in a month and umbrella sales are probably negatively correlated.

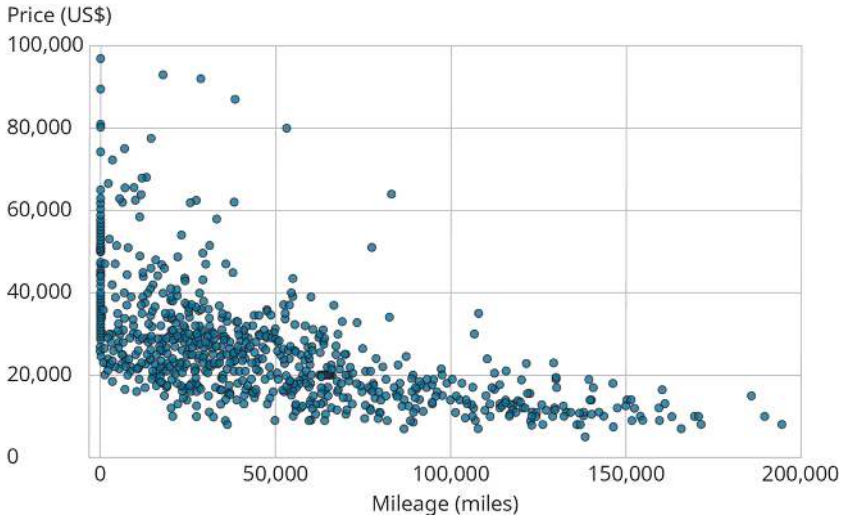
Scatterplots also tell us in a general sense how closely related the two variables are. The closer the points are to the central trend, the stronger the relationship between the two variables. Finally, the graph can let us know whether the relationship between the two variables is linear (following a straight line) or nonlinear (curved).

We have assembled a database of information on 971 new and used cars that were for sale within 20 miles of Burlington, Vermont, in mid-2023, drawing the data from the website Cars.com. For each car, we recorded information on the price, mileage, model year, fuel type, and whether it was new or used. The data are available in the file cars4sale.dta. After opening the file, we can create a scatterplot of price and mileage using the menu system or using commands. With the menu, you can use the following sequence: Graphics → Twoway graph (scatter, line, etc.) → Create → Basic plots → Scatter, then select "price" as the *y* variable and "mileage" as the *x* variable. Alternatively, you can use the following command, either in the command line or in a do-file:

```
twoway (scatter price mileage)
```

The command **twoway** means that we want to generate a graph with two variables, and **scatter** indicates the type of graph. The first variable is plotted on the vertical axis and the second on the horizontal axis.

The output in Figure 12.1 shows that the prices range up to $100,000 and the mileage up to 200,000 miles, but most of the cars have a price between $10,000 and $40,000 and have less than 100,000 miles on the odometer. As expected, price and mileage are negatively correlated, meaning that cars with high mileage tend to have low prices, and vice versa. The graph also indicates many cars are clumped in a line at zero mileage, reflecting the fact that new cars are included in the sample.



Description

Figure 12.1 Scatter Plot of Price and Mileage

How can we measure the strength of the relationship between two continuous variables? One of the most common measures is the *Pearson correlation coefficient*, or *r*. The correlation coefficient can be calculated using the following equation:[1]

$$\text{Correlation coefficient} = r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(12.1)

$$\text{Correlation coefficient} = r = \frac{\sum_{i=1}^{n}[(x_i - x)(y_i - y)]}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2 \sum_{i=1}^{n}(y_i - y)^2}}$$

where

n is the number of observations of *x* and *y*

$x_i = x_1, x_2, \ldots x_n$ are the values of x

$y_i = y_1, y_2, \ldots y_n$ are the values of y

$\bar{x}$ is the mean of x

$\bar{y}$ is the mean of y

The value of *r* varies between −1 and 1, where −1 means a perfect negative correlation, 0 means no correlation, and 1 means a perfect positive correlation. When two variables are perfectly correlated, every observation lies on a straight line if graphed on a scatterplot. When two variables have a very low correlation coefficient, the scatterplot looks like a random collection of dots with no pattern.

To calculate the Pearson correlation coefficient for price and mileage in Stata, we can use the menu system as follows: Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Pairwise correlations of variables, then select the variables price and mileage from the drop-down menu. Alternatively, we can use this command:

```
pwcorr mileage price, sig
```

It will calculate the correlation coefficient for these two variables. Adding the **sig** option will give the statistical significance of the correlation.

The results, shown in Figure 12.2, reveal that the correlation coefficient is −0.6019. The negative number indicates a negative correlation between price and mileage, meaning that as mileage increases, price declines. The magnitude suggests a relatively strong correlation. The number below the correlation coefficient, 0.0000, indicates the *p*-value of the correlation—that is, the probability of finding a correlation coefficient this large (or larger) if there were, in fact, no correlation between the two variables. The low value of *p* indicates that the probability of this occurring "by chance" is very small. The two numbers along the diagonal are 1.0 because they represent the correlation coefficient of each variable with itself.

```
. pwcorr price mileage, sig

             |    price  mileage
-------------+------------------
       price |   1.0000
             |
             |
     mileage |  -0.6019   1.0000
             |   0.0000
```

Description

Figure 12.2 Pearson Correlation Coefficient

The **pwcorr** command can be used to calculate all the correlation coefficients for each pair of variables in a list. For example, if we list five variables, Stata will display a 5 × 5 table of correlation coefficients.

A closely related measure of correlation is the ***coefficient of determination***, more commonly known as R^2. When measuring the association between two variables, R^2 can be calculated easily as the square of the Pearson correlation coefficient:

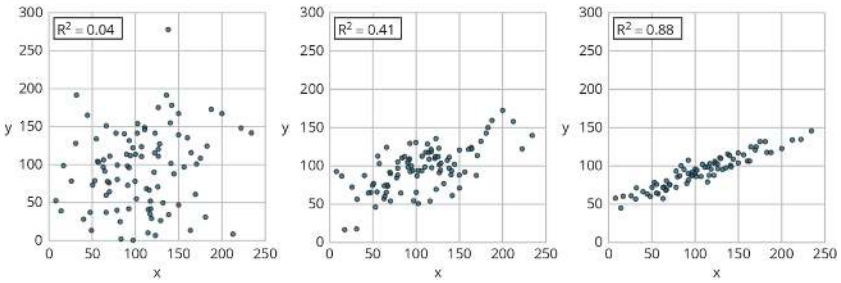
$$\text{Coefficient of determination } = R^2 = r^2$$

(12.2)

$$\text{Coefficient of determination } = R^2 = r^2$$

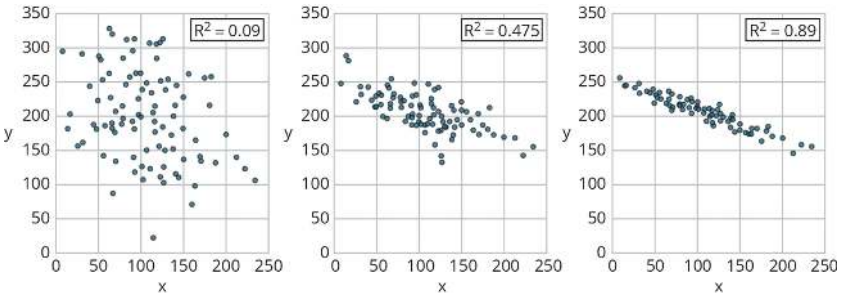
R^2 varies between 0 and 1. If R^2 = 0, the two variables are completely uncorrelated, and if R^2 = 1, they are perfectly correlated, either positively or negatively. One convenient feature of R^2 is that, under some circumstances, it represents the share of the variance in *y* that can be explained by the *x* variable.

Figures 12.3 and 12.4 provide some examples of scatterplots to give an intuitive sense of what different values of R^2 look like.



Description

Figure 12.3 Scatterplots with Different Levels of Positive Correlation



Description

Figure 12.4 Scatterplots with Different Levels of Negative Correlation

Correlation analysis has a number of limitations:

It does not tell us anything about the mathematical relationship between the two variables, such as the slope of the line or where it crosses the vertical axis.

It only considers the relationship between the two variables.

It assumes a linear relationship between the two variables.

It does not imply or confirm any causal relationship between the two variables.

As we will see in the next section, regression analysis gives an equation that describes the relationship among various variables, allows both linear and nonlinear relationships, and, subject to some assumptions, can identify causal relationships.

12.4 SIMPLE REGRESSION ANALYSIS

As mentioned earlier, regression analysis describes the relationship between a dependent variable and one or more independent variables. The distinction between dependent and independent variables is based on the assumption that the independent variables are *exogenous*, meaning that they are not affected by the dependent variable, nor are there any variables outside that model that affect both the dependent variable and the independent variables. If this assumption holds, then any relationship between *y* and *x* can be considered causal, meaning that the model describes how the independent variables *affect* the dependent variable. If these assumptions do not hold, then one or more of the independent variables are said to be *endogenous*. In this case, we cannot infer causality, but the regression analysis might still be useful as a descriptive tool. In this case, it would only describe the changes in *y* that are *associated* with changes in *x*. Chapter 13 describes in more detail the consequences of regression models that violate this or other assumptions behind regression analysis.

We start with the simple case of a linear relationship between one dependent variable and a single independent variable. Later in this chapter, we describe regression analysis with multiple independent variables. And in later chapters, we show how regression analysis can be used to describe nonlinear relationships.

The relationship between a dependent variable and one independent variable in a linear relationship can be described with the following equation[2]:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

(12.3)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

y is the dependent variable

x is the independent variable

β_0 is the constant or *y* intercept

β_1 is the coefficient on x, which is the slope of the regression line

ε is the error term

The error term, ε, reflects the fact that the relationship between *y* and *x* is not exact, but rather is subject to some error. Note that β_0 and β_1 are parameters that cannot be directly observed; we can only

estimate them using the values of *y* and *x*. Likewise, the error term, ε, cannot be directly observed.

The predicted value of *y*, written as, is defined as follows:

$$\text{Predicted value of } y = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(12.4)

$$\text{Predicted value of } y = \hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

where

$\widehat{\beta_0}$ is the estimated value of the true parameter β_0, and

$\widehat{\beta_1}$ is the estimated value of the true parameter β_1

As you can see, the "hat" indicates an estimate of a population parameter based on sample data.

The **residual** is the difference between the actual value and the predicted value:

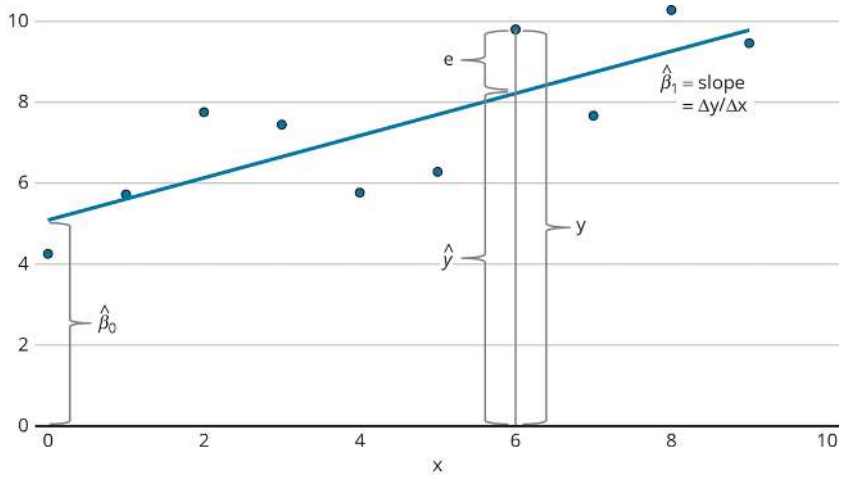
$$\text{Residual} = y - \hat{y} = e$$

(12.5)

$$\text{Residual} = y - \hat{y} = e$$

It is important not to confuse ε and *e*: ε is the unobserved error term in the "true" relationship between *y* and *x*, while *e* is the observed difference between *y* and its predicted value, $\hat{y}$, the latter based on the estimated relationship between *y* and *x*[3]. We use the distribution of the (observed) residuals to infer the distribution of the (unobserved) error term.

The relationships among these concepts is shown in a simplified example in Figure 12.5. The 10 dots represent the observations of *x* and *y*, while the line reflects the predicted values of *y* ($\hat{y}$) as a function of *x*. For each of the 10 observations, the residual (*e*) is the vertical distance between the observation (*y*) and the line representing the predicted values ($\hat{y}$), where the distance is considered negative when *y* is less than $\hat{y}$.



Description

Figure 12.5 Regression Concepts Illustrated on Hypothetical Data

Now we can ask this question: What do we mean when we say that regression analysis identifies the equation that "best describes" the relationship? In this case, regression analysis finds the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals (Σe^2) across all observations. For this reason, this type of regression analysis is also called ordinary least squares (OLS) regression.

How do we obtain the values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that best describe the data—that is, the values that minimize the sum of squared residuals? The calculation of the estimated coefficients and related statistics uses matrix algebra and is beyond the scope of this book, but interested readers will find more information in Woolridge (2016), Greene (2018), and other books dedicated to regression analysis. Fortunately, we do not need to know matrix algebra to run a regression analysis using Stata. Using the menu system, we can follow this sequence: Statistics → Linear models and related → Linear regression and then select the *y* and *x* variables from the drop-down menus. Alternatively, we can run regression analysis with the following command:

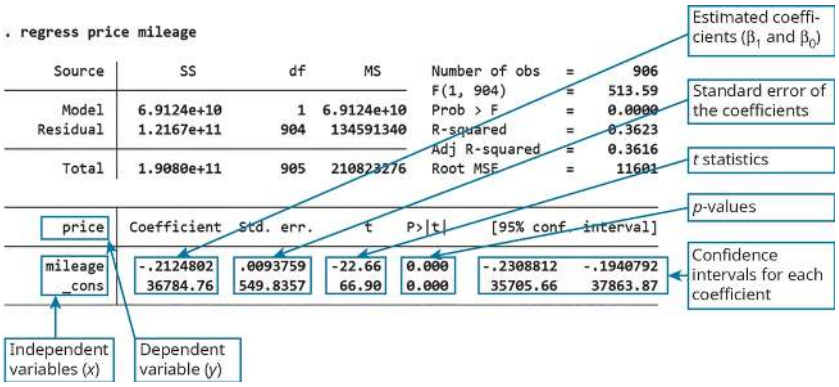
```
regress y x
```

where *y* is the dependent variable and *x* is the independent variable.

Let's see how this works in practice. Returning to the example of the effect of mileage on the price of cars, we can open the database cars4sale.dta and run the following command:

```
regress price mileage
```

The command and results are shown in Figure 12.6.



Description

Figure 12.6 Simple Regression Analysis

How do we interpret the information in Figure 12.6? In the upper right corner, we see that the number of observations (cars) is 906. The *F* statistic is a test of the null hypothesis that all coefficients (excluding the constant) are equal to 0. The "Prob > *F*" line gives the probability that an *F* statistic this large could be generated by chance if the null hypothesis were true. Since it is 0.0000, this indicates that the probability of getting this result would be very small if there were actually no linear relationship between price and mileage.

"*R*-squared" refers to R^2, the coefficient of determination of the observed values of *y* and the predicted values of *y* ($\hat{y}$). In a linear regression model with a constant, R^2can also be interpreted as the proportion of the variance in *y* that can be explained by the model. In this case, mileage explains about 36% of the variance in price across our sample of cars.

"Adj *R*-squared" refers to adjusted R^2. One limitation of R^2 is that, when you add an independent variable to the model, R^2 will always increase, even if the new variable does not help predict the dependent variable. Adjusted R^2 is adjusted for the number of independent variables, so it will increase only if the new variable increases the explanatory power of the model more than would be expected by chance. Adjusted R^2 is calculated as $1 - (1 - R^2)(n - 1)/(n - k)$, where *n* is the number of observations and *k* is the number of independent variables including the constant.

Looking at the bottom of Figure 12.6, we see a table showing a list of the variables including the constant in the first column and a list of coefficients in the second column. In this case of simple regression, there is just one independent variable plus the constant. The variables and coefficients can be rearranged to form the equation that best fits the data as follows:

$$\text{predicted price} = 36784.76 + (-0.2124802 \times \text{mileage})$$

(12.6)

$$\text{predicted price} = 36784.76 + (-0.2124802 \times \text{mileage})$$

The coefficient for mileage ($\hat{\beta}_1$) is approximately −0.212. It tells us how much *y* changes given a one-unit change in *x*. In this case, the coefficient indicates that the price declines by $0.212 or 21.2 cents for each additional mile on the car. In other words, these cars tend to depreciate $212 for each additional 1,000 miles on the odometer. Graphically, −0.212 is the slope of the line plotting predicted price against mileage.

The constant ($\hat{\beta}_0$) is 36,785. This represents the value of $\hat{y}$ (predicted price) when *x* (mileage) is 0, given this simple linear model. It is also called the **intercept** because, graphically, it indicates the value of $\hat{y}$ where the best-fit line "intercepts" the vertical (or *y*) axis.

The second column shows the standard error of the coefficient estimates. The standard error is a measure of the precision of the estimate of the coefficient. If the model fits the data well, then the residuals and the standard error will be small.

The third column gives the *t* statistic for each coefficient, calculated as the ratio of the coefficient and its standard error. As a rule of thumb, a *t* statistic greater than 2 or less than −2 indicates that the coefficient is significantly different from 0. However, the rule of thumb is redundant because Stata and other statistical software packages also report *p*-values, which are a more direct measure of statistical significance.

As described in Chapter 7, the *p*-value tells us the probability that we could get a value of $\hat{\beta}$ this large or larger (in absolute value) if the null hypothesis (that the coefficient is 0) were true. In this case, the *p*-value on the mileage variable is 0.000. This has been rounded off at three digits; it implies that there is less than 0.0005 probability (less than 0.05% probability) that we would get a result this strong (or stronger) if there were no relationship between price and mileage (that is, if $\beta_1 = 0$). Similarly, the *p*-value on the constant suggests that it is unlikely that the true intercept is 0 (that $\beta_0 = 0$). These probabilities are based on the assumption that there is a linear relationship between price and mileage, as well as other assumptions discussed in Chapter 13. By convention, if a *p*-value is less than 0.05, the coefficient is considered "significantly different from zero" or "statistically significant." If the *p*-value is less than 0.01, it is considered "statistically significant at the 1% level."

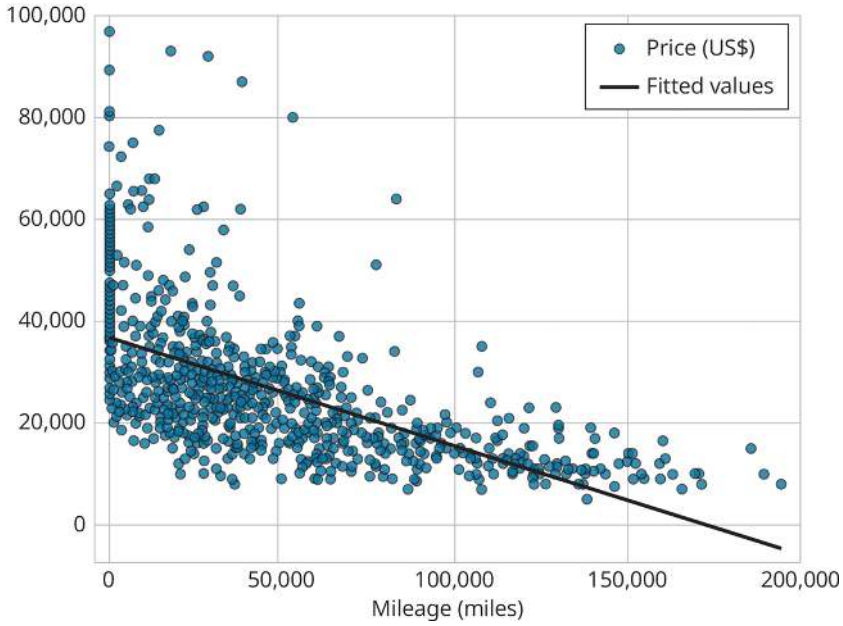
The last two columns show the lower and upper limits of the 95% confidence interval. This means that we are 95% sure that the true value of the parameter lies between these two numbers. The more precise the coefficient estimate, the smaller the standard error, the larger the *t* statistic, the smaller the *p*-value, and the narrower the confidence interval.

As discussed in Chapter 7, there is some controversy over the use of *p*-values. Sometimes, the *p*-value is misinterpreted. Some researchers argue for a stricter standard, requiring a smaller *p*-value to consider a relationship statistically significant. For example, social scientists have traditionally considered *p*-values greater than 0.05 but less than 0.10 to be "weakly significant," but recently, some have argued that it is not worth reporting coefficients with *p*-values greater than 0.05.

To see the best-fit line generated by the regression analysis, we can return to the graphing command. We can show both the scatter plot and the regression line with the following Stata command:

```
twoway (scatter price mileage) (lfit price mileage)
```

The first set of parentheses in the command tells Stata that we want to see a scatter plot of price and mileage. The second set of parentheses indicates that we would like to add the "linear fit" line of price and mileage to the same graph. The line in Figure 12.7 corresponds to the equation described in Figure 12.6 and Equation 12.6. The constant coefficient in Figure 12.6 (36,785) is the value of $\hat{\beta}_0$ in Equation 12.6 and corresponds to the price at which the line crosses the vertical axis in Figure 12.7. Similarly, the mileage coefficient in Figure 12.6 (−0.2124802) is the value of $\hat{\beta}_1$ in Equation 12.6 and corresponds to the slope of the line in Figure 12.7.



Description

Figure 12.7 Scatterplot of Price and Mileage with Regression Line

The line in Figure 12.7 does a fairly good job in describing the pattern of prices, but we can do better. First, the predicted price is clearly incorrect on the right side of the graph where it turns negative. No matter how many miles are on the odometer, a working car will have a positive price! Second, we know that there are numerous other factors that affect the value of a car. For example, a Mercedes will be worth more than a Honda even if they both have the same mileage. The next section shows how regression analysis can incorporate multiple explanatory variables.

12.5 MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis refers to the case where the analysis predicts the value of a dependent variable based on multiple independent variables in addition to the constant. A linear multiple regression model assumes that the data follow a pattern like this:

$$y = \beta_0 + \sum_{i=1}^{k-1} \beta_i x_i + \varepsilon$$

(12.7)

$$y = \beta_0 + \sum_{i=1}^{k-1} \beta_i x_i + \varepsilon$$

where

k is the number of independent variables (including the constant)

β_0 is the constant or y intercept

β_i is the coefficient on x_i where i = 1 to k − 1

x_i is one of the k − 1 independent variables

ε is the error term

As with **simple regression analysis**, the true values of the βs in Equation 12.7 are unknown, but we can estimate them from the observed values of the dependent variable (y) and the independent variables (x_i). The estimated coefficients (denoted by $\hat{\beta}_0$, $\hat{\beta}_2$,…, $\hat{\beta}_{k-1}$) are those that minimize the sum of squared residuals ($\sum e^2$). Each estimated coefficient $\hat{\beta}_i$ is interpreted as the effect of a one-unit increase in the corresponding independent variable, x_i on the dependent variable while holding constant all other independent variables.

Let's return to the model of car prices. We know that mileage is not the only characteristic that affects the price of cars. For example, the price is also influenced by the fuel type—that is, whether the car has a gasoline engine, a hybrid gas-electric system, or an electric motor. In the cars4sale.dta database, the variable fueltype has three values: 1 for a gasoline car, 2 for a hybrid, and 3 for an electric car. The fueltype variable is a nominal categorical variable, meaning that there is no natural order and we cannot assume that the intervals between them are the same. For example, we cannot assume that the difference in value between a gas car and a hybrid is the same as the difference in value between a hybrid and an electric car.

Independent variables that are categorical (nominal or ordinal) should be represented in regression models by one or more dummy variables, each taking a value of 0 or 1. Dummy variables are also called dichotomous, binary, or indicator variables. For example, the hybrid dummy variable will be equal to 1 for hybrid cars and 0 for other cars (gas and electric).

However, the number of dummy variables included in the regression analysis must be equal to the number of categories minus one[4]. In other words, one category is omitted from the regression. The coefficients of the included dummy variables represent the effect on the dependent variable of being in that category rather than the omitted category, as illustrated in Table 12.2. For this reason, it is sometimes called the reference category

TABLE 12.2 ■ Examples of Using Dummy Variables to Represent a Categorical Variable

Categorical Variable	Categories	Number of Categories	Number of Dummy Variables Needed	Example of Dummy Variables to Include	Interpretation of Coefficients
Income quintile	Poorest, Lower-middle, Middle, Upper-middle, Richest	5	4	Lower-middle, Middle, Upper-middle, Richest	Effect of being in this category relative to being in the Poorest category
Marital status	Single, married, divorced, widowed	4	3	Single, divorced, widowed	Effect of having each status relative to being married
Region	North, South, Central, East, West	5	4	North, South, East, West	Effect of living in each region relative to living in the Central region

Categorical Variable	Categories	Number of Categories	Number of Dummy Variables Needed	Example of Dummy Variables to Include	Interpretation of Coefficients
Income quintile	Poorest, Lower-middle, Middle, Upper-middle, Richest	5	4	Lower-middle, Middle, Upper-middle, Richest	Effect of being in this category relative to being in the Poorest category
Marital status	Single, married, divorced, widowed	4	3	Single, divorced, widowed	Effect of having each status relative to being married
Region	North, South, Central, East, West	5	4	North, South, East, West	Effect of living in each region relative to living in the Central region

If we need to exclude one category when creating a set of dummy variables, how do we decide which one to omit? One convention is to omit the dummy associated with the category with the largest number of observations. Another convention is to omit the category associated with the lowest values of the dependent variable so that the coefficients on the dummy variables will be positive. But it does not really matter. The R^2 and the coefficients and *p*-values of all other variables will be the same. The decision of which category to omit will only affect the constant and the coefficients on the dummy variables representing the categorical variable, but even these differences do not affect the predicted values of the dependent variable.

In the case of fueltype, there are three categories, so we need two dummy variables. We will use gasoline as the omitted category both because it is the most common type of car and because it is associated with a lower price. Thus, we need to define dummy variables for hybrid and electric cars. One option is to use **gen** and **replace** commands:

```
gen   hybrid = 0 if fueltype==1 | fueltype==3
replace hybrid = 1 if fueltype==2
gen   electric = 0 if fueltype==1 | fueltype==2
replace electric = 1 if fueltype==3
```

The first two lines create a new variable, hybrid, equal to 0 if fueltype is 1 or 3 and equal to 1 if fueltype is 2. The second two lines define a new variable, electric, in a similar way. It is not necessary to line up the commands as we did here, but it is good practice to make it easier to check for errors.

Using **gen** and **replace** to create dummy variables is effective but somewhat cumbersome. We can streamline the code by using the **recode … gen** command. The **recode** command was described in Chapter 5, but adding the **gen** option creates a new variable rather than changing the values of the original variable. With this command, we can create the two dummy variables with two commands. The first line that follows specifies that if fueltype is 1 or 3, the new hybrid variable will be 0, while if fueltype is 2, the new hybrid variable will be 1. The second line defines the new variable, electric, in a similar way.

```
recode fueltype (1 3=0) (2=1), gen(hybrid)
recode fueltype (1 2=0) (3=1), gen(electric)
```

Finally, the most streamlined approach to converting a categorical variable, such as fueltype, into a set of dummy variables is to use what Stata calls “factor variables” by attaching an “i.” prefix to the categorical variable in the **regress** command itself:

```
regress price mileage i.fueltype
```

Instead of the four **gen** and **replace** commands or the two **recode … gen** commands, the factor variable approach requires just two characters! For now, we will use the **recode … gen** method to calculate dummy variables because it is more transparent.

The commands to define hybrid and electric dummy variables and run the new regression model are shown in Figure 12.8, along with the regression results. The value of R^2 is 0.3972, indicating that the three independent variables explain about 40% of the variation in price. Not surprisingly, three independent variables plus the constant explain a larger share of the variation in price than mileage and the constant alone. The adjusted R is also higher than in the earlier model, indicating that the new variables are contributing significantly to the explanatory power of the model. This is confirmed by the fact that the *p*-values of the hybrid and electric coefficients are less than 0.01, indicating that both are statistically significant at the 1% level.

```
. regress price mileage hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(3, 898)	=	198.91
Model	7.5857e+10	3	2.5286e+10	Prob > F	=	0.0000
Residual	1.1415e+11	898	127118801	R-squared	=	0.3992
				Adj R-squared	=	0.3972
Total	1.9001e+11	901	210887932	Root MSE	=	11275

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.198785	.0093256	-21.32	0.000	-.2170875	-.1804825
hybrid	6318.549	1600.939	3.95	0.000	3176.532	9460.567
electric	12981.94	1905.485	6.81	0.000	9242.221	16721.67
_cons	35306.36	568.7436	62.08	0.000	34190.14	36422.59

Description

Figure 12.8 Multiple Regression (Version 1)

The coefficients for each variable give us information on the linear equation that best fits our data:

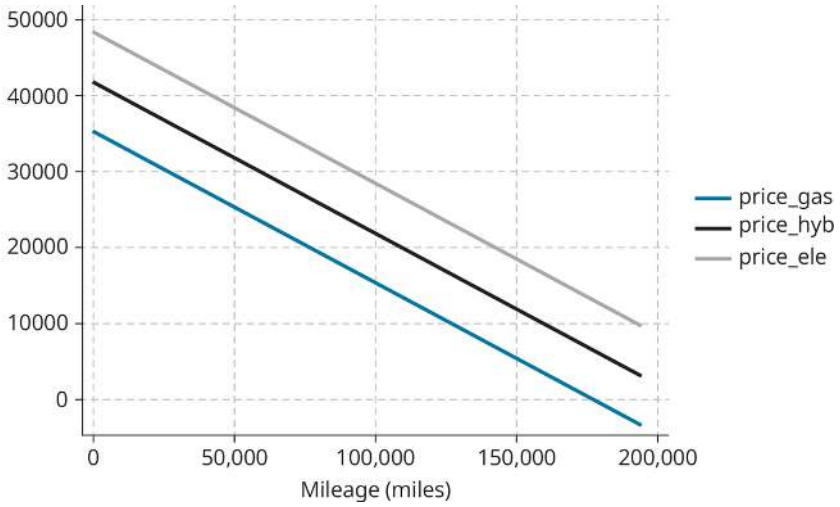
price = 35306.36 + (−0.198785 × mileage) + (6318.549 × hybrid) + (12981.94 × electric)

The coefficient on mileage is about −0.20, meaning that each additional mile on the car is associated with a price reduction of $0.20 or 20 cents. This is similar but not identical to the coefficient in the previous regression model that did not include the dummy variables to represent fuel types (see Figure 12.6). The coefficient on the hybrid variable indicates that the price of a hybrid car is $6,319 more than a gasoline car (the omitted category) given the same mileage. Similarly, an electric car has a price $12,981 more than gasoline car with the same mileage.

We can graph the best-fit line for the gas, hybrid, and electric models separately using the fact that Stata stores the coefficients from the most recent model as **_b[*varname*]**, where *varname* is the name of the independent variable for that coefficient. Thus, we can calculate predicted values for each model as follows:

```
gen price_gas = _b[_cons] + _b[mileage]*mileage
gen price_hyb = _b[_cons] + _b[mileage]*mileage + _b[hybrid]
gen price_ele = _b[_cons]+ _b[mileage]*mileage + _b[electric]
twoway (line price_gas mileage) ///
(line price_hyb mileage) ///
(line price_ele mileage)
```

Note that the three forward slashes (“///”) indicates that the command continues on the next line. The output of these commands is shown in Figure 12.9:



Description

Figure 12.9 Predicted Value of Each Type of Car

The decision whether to include the set of dummies should be based on a joint test, where the null hypothesis is that all the coefficients on the dummy variables representing the categorical variable are equal to zero. The joint Wald test will give the same result regardless of which category is omitted. We can do a joint test of the hypothesis that all the coefficients on the dummy variables are equal to zero using the **test** command. This command can be used to test a variety of null hypotheses, but if the command is followed by a list of variables, it will test the hypothesis that all the corresponding coefficients are equal to zero.

The menu procedure would be as follows: Statistics → Postestimation → Tests, contrasts, and comparisons of parameter estimates → Linear tests of parameter estimates → Create, then select the dummy variables from the drop-down menu. In this case, we would select hybrid and electric. The **test** command and its output are shown in Figure 12.10.

```
. test hybrid electric

 ( 1)  hybrid = 0
 ( 2)  electric = 0

       F(  2,   898) =   29.45
            Prob > F =    0.0000
```

Figure 12.10 Testing Joint Hypotheses

The first two lines of Figure 12.10 show the null hypothesis that the hybrid and electric coefficients are both equal to zero. The last line of the output gives the *p*-value, which suggests that we can reject the null hypothesis that the two coefficients are both equal to zero at the 1% confidence level.

In summary, when using a set of dummy variables to represent a categorical independent variable, one of the set must be omitted from the regression model. The choice of which dummy to omit has no effect

on the **predicted values** of the dependent variable. Likewise, the choice of which dummy to omit will have no effect on the joint test of the statistical significance of the set of dummy variables.

As a final exercise, let's add another variable to the regression model: a dummy to distinguish new from used cars. We can create a dummy variable called *new* from the existing variable newused. Since newused is coded 1 for new cars and 2 for used cars, we can create the *new* variable with the **recode… gen** command as shown in the first line of Figure 12.11. The regression command with one dependent variable followed by the four independent variables is also shown in Figure 12.11, along with the results.

```
. recode newused (1 = 1) (2 = 0), gen(new)
(763 differences between newused and new)

. regress price mileage new hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(4, 897)	=	179.45
Model	8.4462e+10	4	2.1116e+10	Prob > F	=	0.0000
Residual	1.0555e+11	897	117667695	R-squared	=	0.4445
				Adj R-squared	=	0.4420
Total	1.9001e+11	901	210887932	Root MSE	=	10847

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.1581854	.0101509	-15.58	0.000	-.1781077	-.138263
new	9425.751	1102.238	8.55	0.000	7262.486	11589.02
hybrid	3663.098	1571.266	2.33	0.020	579.313	6746.883
electric	9254.045	1884.4	4.91	0.000	5555.699	12952.39
_cons	32122.4	661.8525	48.53	0.000	30823.44	33421.36

Description

Figure 12.11 Multiple Regression (Version 2)

In this version of the model, the coefficient on mileage indicates that the value declines by about 16 cents per mile. The hybrid dummy coefficient implies that hybrid cars are priced $3,663 more than gasoline cars, while the electric dummy coefficient tells us that electric cars are valued $9,254 more than gas cars. And the coefficient on the dummy variable for new cars tells us that a new car is worth $9,426 more than a used car after controlling for mileage and fueltype. This confirms the saying that a car depreciates (almost) $10,000 as soon as it is driven off the car dealer lot.

How do we explain the differences in the coefficients between this model (Figure 12.11) and the earlier model (Figure 12.8)? The hybrid and electric coefficients are substantially smaller than in the earlier version of the model. Part of the higher price of hybrids and electric cars in the earlier model was because more of them are new compared to gas cars. The rapid growth in electric cars means that there are relatively few used electric cars on the market.[5] In statistical terms, hybrid and electric are positively correlated with the *new* dummy, so adding the *new* dummy to the model reduced the estimated effect of the hybrid and electric variables. In more intuitive terms, the early version of the model indicated that hybrid and electric cars were more expensive partly because most of them were new, not only because they were hybrid or electric. Once we control for newness, the price gap is smaller, but it is a more accurate representation of the difference in price between electric and hybrid cars compared to gas cars. To the extent that electric cars tend to be larger or have more advanced features, adding these variables to the model would further reduce the coefficient on the electric dummy variable, hence the implied price difference between gas and electric cars.

12.6 PRESENTING THE RESULTS

To describe the results, the researcher should focus on the independent variables that have a statistically significant and meaningful effect on the dependent variable. By statistically significant, we mean that the *p*-value is less than 0.05, indicating that the finding is unlikely to have occurred by chance if there is, in fact, no relationship. By meaningful, we mean that the size of the effect is important enough to affect policy or other decisions related to the topic of the study. With a large sample size, it is quite possible that a relationship is statistically significant (measured with little error) but too small to be of practical importance.

In addition, it may be worth identifying independent variables that did not have a statistically significant relationship with the dependent variable if this contradicts or challenges widely held beliefs. However, the size of the coefficient should not be discussed unless it is statistically significant.

For a newspaper or magazine targeting a **nontechnical audience**, we might summarize the car price regression results as follows:

Statistical analysis reveals that the price of cars is influenced by mileage, whether it is new or used, and whether it has a gasoline car, a hybrid, or an all-electric car. For example, each additional mile on the odometer is associated with a reduction in value of 16 cents. In addition, on average a new car loses $9,246 in value as soon as it is purchased. Finally, after controlling for mileage, age, and newness, electric cars are priced at about $9,254 more than a gas car, while hybrids go for about $3,663 more than a gas car.

For an academic audience, we can assume some familiarity with regression analysis and provide some additional details. We can use the **etable** command to generate a table of regression results in Word format. In its simplest form, the **etable** command uses the most recent regression analysis and sends the output to a Word file.

```
etable, export(car price model.docx)
```

The **export** option is used to indicate the name and type of the file to be created. There are numerous options that allow you to specify the statistics to be included and the layout of the table, some of which are discussed in Chapter 14. If no options are specified, the default is a simple table showing the coefficients, standard errors, and the number of observations, as shown in Figure 12.12.

	Price
Mileage (miles)	-0.199
	(0.009)
Hybrid	6318.549
	(1600.939)
Electric	12981.944
	(1905.485)
Intercept	35306.365
	(568.744)
Number of observations	902

Description

Figure 12.12 Regression Output In Word Format Using etable

In writing up the results, it is important to consult the journal for which you are writing. For example, some journals encourage the use of confidence intervals (CIs), while others prefer authors to give *p*-

value or the statistical significance of the coefficients. Below is a possible write-up for a **technical audience**:

We used regression analysis to explore the determinants of the price of 902 new and used cars advertised on cars.com in the vicinity of Burlington, Vermont, in July 2023. The independent variables were mileage and dummy variables for new cars, electric cars, and hybrid cars, with used internal combustion engine (ICE) cars being the omitted reference category. All four coefficients are statistically significant and of the expected sign. The coefficient on mileage is −0.158 ($p < 0.001$), indicating that each additional mile reduces the value of the car by 16 cents. In addition, the coefficient on the dummy variable representing the hybrid cars is 3,663 ($p < 0.05$), which implies that hybrid models are priced an average of $3,663 more than ICE cars, the omitted category, after controlling for mileage and newness. The coefficient on the electric car dummy variable is 9,254 ($p < 0.001$), suggesting that the price of an electric car is about $9,254 more than an ICE car, holding mileage and newness constant. A Wald test of the joint significance of the hybrid and electric dummy variables rejects the null hypothesis that both coefficients are equal to zero at the 1% **confidence level**.

These results would normally be followed by a discussion that places the findings in the context of previous research, identifying areas of agreement and areas where these results differ from previous studies. Finally, it is often useful to identify questions that remain unanswered and suggest future areas for research. Chapter 16 provides more information on organizing the research paper.

12.7 SUMMARY OF COMMANDS USED IN THIS CHAPTER

This section summarizes the Stata code used in the chapter (Table 12.3). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 12.3 ■ Summary of Commands Used in this Chapter

Function	Stata command(s)
Scatter plots and line of best fit	twoway (scatter price mileage) (lfit price mileage)
Correlation	pwcorr price mileage
Regression analysis	regress price mileage new hybrid electric
Generating dummy variables (method 1)	gen hybrid = 0 if fueltype==1 | fueltype==3 replace hybrid = 1 if fueltype==2
Generating dummy variables (method 2)	recode fueltype (1 3=0) (2=1), gen(hybrid)
Calculate and graph predicted values	gen price_hyb = _b[_cons] + _b[mileage]*mileage + _b[hybrid]
Test the joint hypothesis that a set of coefficients are all equal to zero	test hybrid electric
Generate a table of regression results in Word format	etable, export(car price model.docx)

Function	Stata command(s)
Scatter plots and line of best fit	twoway (scatter price mileage) (lfit price mileage)
Correlation	pwcorr price mileage
Regression analysis	regress price mileage new hybrid electric

Function	Stata command(s)
Generating dummy variables (method 1)	gen hybrid = 0 if fueltype==1 | fueltype==3 replace hybrid = 1 if fueltype==2
Generating dummy variables (method 2)	recode fueltype (1 3=0) (2=1), gen(hybrid)
Calculate and graph predicted values	gen price_hyb = _b[_cons] + _b[mileage]*mileage + _b[hybrid]
Test the joint hypothesis that a set of coefficients are all equal to zero	test hybrid electric
Generate a table of regression results in Word format	etable, export(car price model.docx)

EXERCISES

1. You are interested in whether the price of cars differs across different makes. In particular, you want to know whether being a Cadillac is associated with having a higher price, after controlling for mileage and age. Create a dummy variable called *caddy* that is equal to 1 for Cadillacs and 0 for other brands, and carry out a regression analysis.
   a. What is the coefficient on the Cadillac variable, and is it statistically significant?
   b. How would you describe this finding for a newspaper article?
   c. How would you describe the finding for an academic journal article?
2. You want to examine the relationship between a college's rank according to *U.S. News and World Report* and factors that may influence it such as their acceptance rate and the amount of the college's endowment per full-time student. Using the "Liberal Arts Colleges - USNews" data set, run a regression with the college rank (USNewsRank) as the dependent variable and the acceptance rate (adm_rate) and endowment per student (endowpp) as the independent variables.
   a. Write out the equation for the model using the coefficients from your results.
   b. What percentage of the variation in rank is explained by these two independent variables?
   c. What is the null hypothesis for the *F* value in the model?
   d. How would you interpret the coefficient for pct_adm in the model? In other words, write out a full sentence that explains the meaning of the coefficient. Is it statistically significant?
3. Suppose you have sample of 200 college students who are economics majors, business majors, and math majors. You run a regression to determine how absences affect their grade point average (GPA) and generate the equation below. All of the coefficients are statistically significant at the 5% level.

   GPA = 3.5 − 0.2 × Absences − 0.3 × Economics_major − 0.1 × Business_major

   where

   GPA = the cumulative grade point average of a student

   Absences = the number of times that a student skips a class per term on average

   Economics_major = 1 if the student is an economics major and 0 if not

   Business_major = 1 if the student is a business major and 0 if not

   Math major is the omitted category for major

a. Draw a graph of the GPA as a function of absences with separate lines for math, economics, and business majors.
b. What is the slope of your line or lines in your graph?
c. Explain in words the meaning of the coefficient on absences.
d. Explain in words the meaning of the coefficient on economics major.

4. Violence in public schools has led to an increase in the total number of full-time security guards and other sworn law enforcement officers on school campuses. Use the School Survey on Crime and Safety (pu_ssocs16.dta) to examine the relation between the total number of disciplinary actions taken by a school (DISTOT16—the dependent variable) and the number of full-time security guards or officers on campus (SEC_FT16). Include the school size by creating three dummy variables for size based on the variable FR_SIZE.
   a. Based on your results, interpret the coefficients on your dummy variables related to size.
   b. Interpret your results for the number of full-time security guards or officers on campus.
   c. Comment on the endogeneity problem of this regression equation related to disciplinary actions and security guards or law officers.
5. Various factors affect household income. We can use the 2021 General Social Survey to explore some of these relationships. Using the file GSS2021.dta, run a regression of income (“realinc”) as a function of age (“age”) and a dummy variable for female respondents. You will need to create the female dummy variable from the “sex” variable.
   a. Is the education variable statistically significant as a predictor of household income? How would you interpret the size of the coefficient?
   b. How would you interpret the coefficient on the dummy variable representing females?
   c. How would you describe the effect of age on household income? Why do you think age is not statistically significant?
   d. Suppose you wanted to estimate the effect of having a college education compared to not having one. How would you create a dummy variable to estimate this? What is the effect on household income of having a college education compared to not having one?

KEY TERMS

coefficient of determination

confidence level

dependent variable

independent variables

intercept

multiple regression analysis

non-technical audience

predicted values

regression analysis

residual

simple regression analysis

technical audience

Descriptions of Images and Figures

Back to Figure

The horizontal axis represents “mileage (miles)” ranging from 0 to 200,000 in increments of 50,000. The vertical axis represents “Price (US$)” ranging from 0 to 100,000 in increments of 20,000. The cluster is denser near the origin.

Back to Figure

The table is titled “. pwcorr price mileage, sig.” The content of the image is given in the following table.

	Price	Mileage
Price	1.0000	-
Mileage	negative 0.6019	1.0000
	0.0000	-

Back to Figure

The scatterplot for “R square =0.04” is drawn on the left (loosely packed), “R square =0.41” in the middle (dense in the middle), and “R square =0.88” on the right (dense with an increasing trend). In all three graphs, the x-axis ranges from 0 to 250 and the y-axis ranges from 0 to 300, both in increments of 50.

Back to Figure

The scatterplot for “R square =0.09” is drawn on the left (loosely packed), “R square =0.475” in the middle (dense in the middle), and “R square =0.89” on the right (dense with an decreasing trend). In all three graphs, the x-axis ranges from 0 to 250 and the y-axis ranges from 0 to 350, both in increments of 50.

Back to Figure

Both axes range from 0 to 10 in increments of 2. There are 10 plots scattered across the graph. The regression line passing through the plots is increasing. The line approximately starts at (0, 5.1), passes through (2, 6.1), (4, 7.2), (6, 8.2), (8, 9.3) and ends at (9, 9.8). The regression line is marked as “beta 1 cap = slope = delta y over delta x.” The distance between the x-axis and the regression line at 0 and 6 are marked as “beta cap 0” and “y cap.” The distance between the regression line and the plot at (6, 9.8) is marked as “e.”

Back to Figure

The table is titled “regress price mileage.” The content of the table on the top is given as follows.

Source	SS	df	MS
Model	6.9124e+10	1	6.9124e+10
Residual	1.2167e+11	904	134591340
Total	1.9080e+11	905	210823276

The list of values to the right of the table is given as follows.

“Number of obs = 906

F(1, 904) = 513.59

Prob>F = 0.0000

R-squared = 0.3623

Adj R-squared = 0.3616

Root MSE = 11601”

The table at the bottom is given as follows.

price	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
mileage	negative .2124802	.0093759	negative 22.66	0.000	negative .2308812	negative .1940792
_cons	36784.76	549.8357	66.90	0.000	35705.66	37863.87

“Price” and “mileage; _cons” are highlighted and marked as “Dependent variable (y)” and “independent variable (x)” respectively. The values on columns 2 through 5 are marked as “Estimated coefficients (beta 1 and beta 0),” “Standard error of the coefficients),” “t statistics,” “p-values,” and “Confidence intervals for each coefficient” respectively.

Back to Figure

The horizontal axis represents “mileage (miles)” ranging from 0 to 200,000 in increments of 50,000. The vertical axis ranges from 0 to 100,000 in increments of 20,000. The plots represent “Price (US$)” and the line represents “fitted values.” The cluster is denser near the origin. The line approximately starts at (0,36000), passes through (50000,25800), (100000,15500), (150000,5200), and then ends at (195000,-4000).

Back to Figure

The table is titled “regress price mileage hybrid electric.” The content of the table on the top is given as follows.

Source	SS	df	MS
Model	7.5857e+10	3	2.5286e+10
Residual	1.1415e+11	898	127118801
Total	1.9001e+11	901	210887932

The list of values to the right of the table is given as follows.

“Number of obs = 902

F(3, 898) = 198.91

Prob>F = 0.0000

R-squared = 0.3992

Adj R-squared = 0.3972

Root MSE = 11275”

The table at the bottom is given as follows.

price	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
mileage	Negative	.0093256	Negative	0.000	Negative	Negative

	.198785		21.32		.2170875	.1804825
hybrid	6318.549	1600.939	3.95	0.000	3176.532	9460.567
electric	12981.94	1905.485	6.81	0.000	9242.221	16721.67
_cons	35306.36	568.7436	62.08	0.000	34190.14	36422.59

Back to Figure

The horizontal axis represents “mileage (miles)” ranging from 0 to 200,000 in increments of 50,000. The vertical axis represents “Price (US$)” ranging from 0 to 100,000 in increments of 20,000. The three decreasing lines represent “Price_gas,” “Price_hyb,” and “Price_ele.” The approximate data on the graph are given in the following table.

Mileage (miles)	Price_gas	Price_hyb	Price_ele
0	35000	42000	48000
50,000	25250	32250	38250
100000	15500	22500	28500
1,50,000	5750	12750	18750
200000	-4000	3000	9000

Back to Figure

The text on the top reads “. recode newused (1 = 1) (2 = 0), gen (new)

(763 differences between newused and new)”

The table is titled “regress price mileage new hybrid electric.” The content of the table on the top is given as follows.

Source	SS	df	MS
Model	8.4462e+10	4	2.1116e+10
Residual	1.0555e+11	897	117667695
Total	1.9001e+11	901	210887932

The list of values to the right of the table is given as follows.

“Number of obs = 902

F(4, 897) = 179.45

Prob>F = 0.0000

R-squared = 0.4445

Adj R-squared = 0.4420

Root MSE = 10847”

The table at the bottom is given as follows.

price	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
mileage	Negative .1581854	.0101509	Negative 15.58	0.000	Negative .1781077	Negative .138263

new	9425.751	1102.238	8.55	0.000	7262.486	11589.02
hybrid	3663.098	1571.266	2.33	0.020	579.313	6746.883
electric	9254.045	1884.4	4.91	0.000	5555.699	12952.39
_cons	32122.4	661.8525	48.53	0.000	30823.44	33421.36

Back to Figure

The content of the image is given in the following table.

	Price
Mileage (miles)	Negative 0.199 (0.009)
Hybrid	6318.549 (1600.939)
Electric	12981.944 (1905.485)
Intercept	35306.265 (568.744)
Number of observations	902

13 REGRESSION DIAGNOSTICS

CHAPTER PREVIEW

Topics	Explanation
Background	Linear regression analysis generates the best equation to describe the relationship between one dependent variable and one or more independent variables, but it depends on several assumptions about the data. This chapter discusses ways to test these assumptions and remedy the problem if it is found.
Measurement error	*Assumption:* Regression analysis assumes the independent variables are measured without error.
	Diagnosis: **sum... detail**, **predict... resid**, **predict... cooksd**
	Remedies: Minimize errors in data collection. Clean data of obvious errors. Try alternative indicators. Take into account in interpretation.
Specification error	*Assumption:* Functional form is correct and all relevant independent variables are included.
	Diagnosis: **rvpplot**, **rvfplot**, **ovtest**, test significance of new variables, quadratic terms, and interaction terms
	Remedy: Include new variables, quadratic terms, or interaction terms if statistically significant.
Multicollinearity	*Assumption:* Independent variables are not highly correlated with one another.
	Diagnosis: **correl**, **vif** test
	Remedy: Test joint significance of correlated variables and explain in text.
Heteroscedasticity	*Assumption:* Variance of residuals is constant.
	Diagnosis: **rvpplot**, **rvfplot**, **hettest**
	Remedy: **vce(robust)** option, generalized least squares
Nonnormality	*Assumption:* Residuals are normally distributed.
	Diagnosis: **sktest**
	Remedy: Transform variables, take into account in interpretation.
Endogeneity	*Assumption:* Independent variables are exogenous.
	Diagnosis: Largely based on theory and experience rather than statistical tests
	Remedy: Instrumental variables regression, panel data regression, and experimental methods

Topics	Explanation
Background	Linear regression analysis generates the best equation to describe the relationship between one dependent variable and one or more independent variables, but it depends on several assumptions about the data. This chapter discusses ways to test these assumptions and remedy the problem if it is found.

Topics	Explanation
Measurement error	*Assumption:* Regression analysis assumes the independent variables are measured without error. *Diagnosis:* **sum... detail**, **predict... resid**, **predict... cooksd** *Remedies:* Minimize errors in data collection. Clean data of obvious errors. Try alternative indicators. Take into account in interpretation.
Specification error	*Assumption:* Functional form is correct and all relevant independent variables are included. *Diagnosis:* **rvpplot**, **rvfplot**, **ovtest**, test significance of new variables, quadratic terms, and interaction terms *Remedy:* Include new variables, quadratic terms, or interaction terms if statistically significant.
Multicollinearity	*Assumption:* Independent variables are not highly correlated with one another. *Diagnosis:* **correl**, **vif** test *Remedy:* Test joint significance of correlated variables and explain in text.
Heteroscedasticity	*Assumption:* Variance of residuals is constant. *Diagnosis:* **rvpplot**, **rvfplot**, **hettest** *Remedy:* **vce(robust)** option, generalized least squares
Nonnormality	*Assumption:* Residuals are normally distributed. *Diagnosis:* **sktest** *Remedy:* Transform variables, take into account in interpretation.
Endogeneity	*Assumption:* Independent variables are exogenous. *Diagnosis:* Largely based on theory and experience rather than statistical tests *Remedy:* Instrumental variables regression, panel data regression, and experimental methods

13.1 INTRODUCTION

In Chapter 12, we said that ordinary least squares (OLS) regression analysis gives us the equation that best fits the data, in the sense that it is the equation that minimizes the sum of squared residuals ($\sum e^2$). Under certain conditions, OLS gives us the best linear unbiased estimates (BLUE) of the coefficients.

Best means the lowest variance of the error terms.

Linear means that the dependent variable is a linear function of the independent variables.

Unbiased means that the estimated coefficients will not be systematically higher or lower than the true coefficients across different samples.

What are the conditions needed for OLS results to be BLUE?

The independent variables are measured without error.

The regression equation is correctly specified, meaning there are no omitted variables and it uses the right functional form (e.g., linear, quadratic, logarithmic, etc.).

None of the independent variables is perfectly correlated with any other independent variable.

The variance of the errors is constant.

The error terms are not correlated with each other.

The independent variables are exogenous.

One additional condition is convenient for the interpretation of the OLS results but not necessary for BLUE: that the error terms are normally distributed.

What happens to OLS regressions if these conditions do not hold? What follows is a list of potential problems associated with violations of these assumptions:

Measurement error: The independent variables are measured with error.

Specification error: The equation in the model is missing important variables or has the wrong functional form.

Multicollinearity: Two or more independent variables are perfectly or closely correlated with each other.

Heteroscedasticity: The variance of the error term is not constant.

Autocorrelation: The error terms are correlated with each other.

Endogeneity: The “independent” variables are influenced by the dependent variable or both dependent and independent variables are influenced by factors omitted from the model.

Nonnormality: The error terms in the regression model are not normally distributed.

This chapter considers the consequences of violating each assumption, how to test to see if the assumption is valid, and how to improve the analysis if the assumption is not valid. One of these issues, autocorrelation, is relevant primarily in time-series data, so this topic will be reserved for Chapter 15 when we give a brief review of some advanced topics, including time-series analysis.

13.2 MEASUREMENT ERROR

Regression analysis assumes that the dependent variable is measured with some error but that the independent variables are measured without error. In actual research, particularly social science research, the independent variables are almost always subject to some measurement error, meaning that values of the variable in the database differ from the true values of the variable due to errors or deception by the respondent, mistakes by the enumerator, or data entry errors. In general, measurement error in an independent variable will cause its regression coefficient to be biased toward zero[1]. This will underestimate the size of the effect of the independent variable and will reduce the likelihood of detecting a real effect. (This is Type II error, the error of not rejecting a null hypothesis when it is false.)

We can demonstrate this by adding a random number to one of the independent variables in our car price data to simulate measurement error. Then we compare the results with and without the simulated "error." The Stata function **rnormal(*m,s*)** generates a normally distributed random variable with mean *m* and standard deviation *s*. In the example in Figure 13.1, we add some artificial "measurement error" to the variable mileage by adding a random number with mean 0 and standard deviation 16,000, roughly one-half of the mean value of mileage. The **set seed** command ensures that anyone running these commands will get the same random numbers and the same results[2].

```
. set seed 2314

. gen mileage_err = mileage + rnormal(0,18000)
(65 missing values generated)

. regress price mileage_err new hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(4, 897)	=	149.72
Model	7.6072e+10	4	1.9018e+10	Prob > F	=	0.0000
Residual	1.1394e+11	897	127021531	R-squared	=	0.4004
				Adj R-squared	=	0.3977
Total	1.9001e+11	901	210887932	Root MSE	=	11270

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage_err	-.1206191	.0095686	-12.61	0.000	-.1393986	-.1018395
new	11226.23	1126.55	9.97	0.000	9015.251	13437.21
hybrid	3799.866	1633.813	2.33	0.020	593.3254	7006.407
electric	9576.098	1957.576	4.89	0.000	5734.135	13418.06
_cons	30116.18	644.2236	46.75	0.000	28851.82	31380.54

Figure 13.1 Multiple Regression with Additional Measurement Error

The results in Figure 13.1 show that the coefficient on the age variable is now −0.121 rather than −0.158 in Figure 12.11. By adding some "measurement error" to the mileage variable, the estimated coefficient is now smaller in absolute value, reflecting the bias toward zero.

How do we find measurement errors? One approach is to look at the extreme values of individual variables. The **summarize** command with the **detail** option will show us some useful information about

the distribution of a variable (Figure 13.2).

```
. sum mileage, detail

                    Mileage (miles)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           1              0
 5%           3              0
10%           3              0       Obs                1,035
25%          18              1       Sum of wgt.        1,035

50%       25972                      Mean           37708.54
                         Largest     Std. dev.      40659.5
75%       58137          171343
90%      101337          185757      Variance       1.65e+09
95%      125851          189423      Skewness       1.236314
99%      159520          194362      Kurtosis       3.962702
```

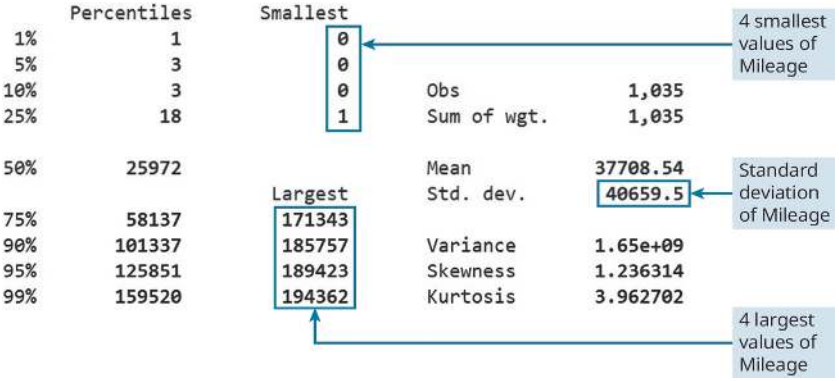


Figure 13.2 Detailed Statistics on the Mileage Variable

A rule of thumb is to check observations for which the value is more than 3 standard deviations below or above the mean. In this case, it would be 37,709 ± (3 × 40,660) or between −84,271 and 159,689. Of course, any negative numbers for mileage would raise a red flag. We can use the **browse… if** command to inspect high-mileage cars[3]:

```
browse make model year mileage if mileage>160000 & mileage!=.
```

The results show there are 10 cars in the database with between 160,000 and 200,000 miles on the odometer, which is high but not difficult to believe. Outliers are not necessarily errors but should be checked, if possible.

A second approach to finding suspicious data is to examine observations that are outliers in the relationship between dependent and independent variables. In other words, we check cases with large residuals (*e*), defined as the observed value of the dependent variable (*y*) minus the predicted value ($\hat{y}$). To calculate the residuals from the most recent regression analysis, we use the command **predict *newvar*, resid**, where *newvar* is the name we wish to assign to the residual. The first command below will calculate the residual and give it the name *e*. The second command will give us various statistics about the residual, including the five largest and the five smallest (negative) values:

```
predict e, resid
sum e, detail
```

If we run these commands after the regression model in Figure 12.11, the results (not shown) reveal that the largest outlier is 62,584, meaning the actual price is $62,584 greater than the predicted price. We can look at the data for all observations with residuals greater than 50,000 with the following command:

```
browse if e>50000 & e!=.
```

There are five observations (cars) with residuals greater than 50,000, four of which are Cadillac Escalades. This indicates that Escalades are more expensive than other cars given their age, mileage, newness, and fuel type. This is not surprising considering that the Escalade is a large, luxury sport utility vehicle. The residual reflects the effect of characteristics not included in the model, like horsepower, size, and features. The large residual does not suggest an error in this case.

The third approach is to look for observations that have the greatest influence, or *leverage*, on the coefficients. An observation will have a lot of leverage if the value of an independent variable is far from its mean. Cook's distance indicator, or Cook's *D*, measures the effect of removing an observation on the estimated coefficients. Some researchers define an outlier as an observation with a Cook's *D* value greater than 1. Others look for observations where the Cook's *D* is at least 3 times greater than the mean value of Cook's *D*. Cook's *D* can be calculated using the menu system as follows: Statistics → Postestimation → Predictions → Predictions and their SEs, leverage statistics, distance statistics, etc. → Cooks D, then type in a new variable name and click on "Submit."

Alternatively, we can calculate Cook's *D* for the most recent regression analysis with the command **predict *newvar*, cooksd**, where *newvar* is the name we want to give to the new variable. Then, we can examine the outliers with the browse commands, as shown here.

```
predict CooksD, cooksd
browse if CooksD>1 & CooksD!=.
```

In the car data set, there are no observations for which the Cook's *D* is greater than 1.

It is important to note that that not all outliers are caused by measurement error. One car in the database is recorded as being 47 years old, but looking at the original advertisement, the photo confirms that it is a 1976 Honda Civic, one of the first generations imported into the United States. In addition, not all errors are outliers: It is quite possible for a measurement error not to be an extreme value and to have a low value of Cook's *D*. Nonetheless, the tests for measurement error often depend on errors being outliers.

Below are a few guidelines to reduce measurement error:

The best way to minimize measurement errors is to avoid them in the first place with careful data collection and data entry. Software for online surveys or for electronic data collection often allow the researcher to set upper and lower limits. If one tries to enter a number outside this range, the program can be designed so that the user is either warned that it is an extreme value or blocked from entering an extreme value.

In cleaning the data, the researcher should replace a number that is impossible (e.g., age = 140) with a missing value. However, it is not good practice to replace numbers that are merely unlikely (e.g., age = 100). As mentioned before, not all outliers are data errors.

Some researchers use rules for "trimming" data, such as eliminating values that are more than 3 standard deviations from the mean. This practice should be used conservatively and must be disclosed in writing up the results.

Some researchers choose to use alternative regression methods that are less sensitive to outliers. These methods are briefly described at the end of Section 13.6.

Finally, it is always useful to take into account possible effects of measurement error in interpreting the results of regression analysis. As mentioned earlier, measurement errors tend to bias regression coefficients toward zero.

Researchers must take care that data cleaning follows transparent and consistent rules and that the procedures are documented when describing the results. Furthermore, the cleaning should never be driven by an effort to achieve a certain outcome. Aggressive data cleaning to achieve a desired finding has resulted in several high-profile cases of research fraud, with severe professional consequences.

13.3 SPECIFICATION ERROR

Ordinary least squares (OLS) regression assumes that the specification of the regression equation is correct, meaning it has the right independent variables and uses the right function to describe the relationship between dependent and independent variables.

13.3.1 Types of Specification Errors

Three common ways that the specification of the model may be incorrect are as follows: (1) a relevant variable is missing from the equation, (2) it fails to take into account nonlinearity in the relationship between the dependent variable and one or more independent variables, and (3) it does not incorporate interaction between independent variables.

13.3.1.1 Omitted Variables

The first type of specification error is omitted variables. Suppose a relevant independent variable is omitted from a linear regression model, where “relevant” means that it influences the dependent variable. If the omitted variable is uncorrelated with the other independent variables, then the estimated coefficients in the model are unbiased. In this case, the only problem associated with omitting the variable is that the explanatory power of the model (measured by the adjusted R^2) is not as good as it could be. On the other hand, if the omitted variable *is* correlated with one or more independent variables in the model, it creates another problem: The estimated coefficients of those variables will be biased. An example of this will be given later in this chapter.

13.3.1.2 Incorrect Functional Form

The second type of specification error is incorrect functional form, meaning the use of an incorrect function describing the relationship between dependent and independent variables. The most common problem with functional form is that the model is linear but the data follow a nonlinear pattern. For example, in Figure 13.3, the data clearly follow a nonlinear U-shaped relationship. If we estimate the relationship as if it were linear, we will get the best straight line that fits the data (the upward-sloping red line), but it will not be a good description of the data. The predicted values of the dependent variable would overestimate the observed values over some range of the independent variable and underestimate them over another range. Modeling a nonlinear relationship as linear could also result in heteroscedasticity, which could invalidate the estimated standard errors (as discussed later). Finally, using a linear model on a nonlinear relationship reduces the explanatory power of the model. If the nonlinearity is strong (as in Figure 13.3), the predictions from a linear model are seriously flawed.

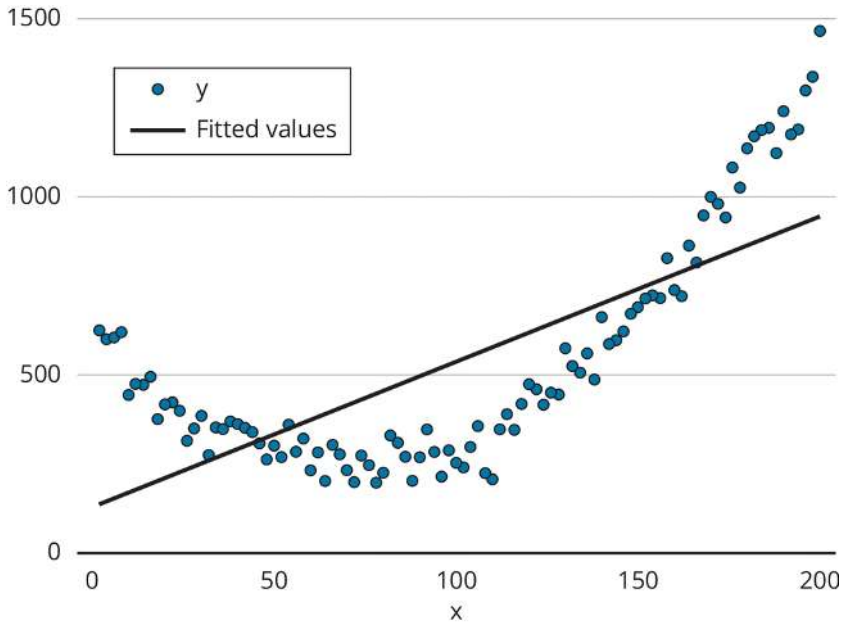


Figure 13.3 Modeling a Nonlinear Relationship with a Linear Model

In the car price regression model, we use a linear equation to represent the relationship between price and mileage, meaning that each additional mile has the same effect on the value of the car (−16 cents/mile in the most recent version) regardless of the mileage on the car. But what if the actual relationship is nonlinear? For example, it is certainly possible that the per-mile depreciation would be greater for relatively new cars with low mileage than for high-mileage cars. Later, we will test this hypothesis.

13.3.1.3 Missing Interaction Terms

The third type of specification error is that interaction between the independent variables is ignored in the model. In the regression models we have considered so far, the effect of each independent variable on the dependent variable is not affected by the other independent variables. However, the effect of one independent variable on the dependent variable may depend on one or more other independent variables.

In the model of car prices, we assume that the effect of mileage on price is not affected by whether the car is gas, hybrid, or electric. In other words, the model assumes that an additional 1,000 miles has the same effect on the price of a gas car as it does on the price of a hybrid or electric car. This is reflected in Figure 12.9, where the three lines are parallel. The reason these lines are parallel is *not* because the data tell us that all three types of cars depreciate at the same rate. Instead, they are parallel because the functional form forces them to be parallel. Specifically, there is only one coefficient to represent the effect of mileage on price.

Interaction between two independent variables can be represented by a term with the product of the two independent variables, as shown at the end of the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

(13.1)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

In Section 13.3.3, we show how an interaction term can be used to test whether the effect of mileage on price differs across models.

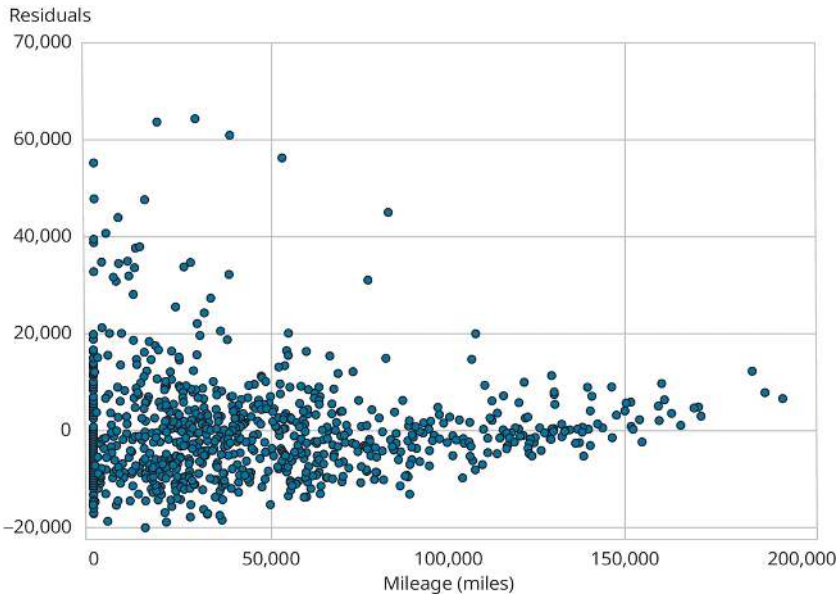
13.3.2 Diagnosing Specification Error

How do we diagnose specification error? It is useful to start by examining patterns in the residuals. A scatterplot of the residuals as a function of each of the continuous independent variables may show a U pattern or inverted-U pattern, indicating a nonlinear relationship.

As mentioned previously, we suspect that the relationship between price and mileage might be nonlinear, with per-mile depreciation being stronger (more negative) when the mileage is low and weaker (less negative) as the mileage increases. In Stata, we can generate such a graph with the menu system or using a command. Using the menus, we would follow this sequence: Graphics → Regression diagnostics plots → Residual-vs-predictor plot, then select the explanatory variable from the list. Alternatively, we can use the **rvpplot** command, where *rvp* is short for residual vs predictor:

```
rvpplot mileage
```

As shown in Figure 13.4, the scatterplot seems to show more positive residuals at low mileage and high mileage, but the pattern is not as obvious as in our hypothetical example in Figure 13.3. Later, we show how to test for nonlinearity.



Description

Figure 13.4 Scatter Plot of Residuals Against Mileage

Another method to check for specification error is to apply the Ramsey Regression Equation Specification Error Test (RESET). In Stata, the Ramsey test can be implemented using the menu system as follows: Statistics → Postestimation → Specification, diagnostics, and goodness-of-fit analysis → Ramsey regression specification-error test for omitted variables. Alternatively, the Ramsey test can be run with **estat ovtest**, which adds powers of the predicted dependent variable ($\hat{y}$) to the

original list of independent variables. An alternative version (**estat ovtest, rhs**) adds powers of the independent variables as explanatory variables. (The initials *rhs* refers to right-hand side variables.) If the coefficients on the new variables are jointly significant, we reject the null hypothesis of no specification error.

To demonstrate the omitted variable test, we run the default version after the regression command, as shown in Figure 13.5. The null hypothesis is that the model has no omitted variables, and the result in Figure 13.5 indicates that we can reject the null hypothesis that there are no omitted variables. In other words, these tests suggest there is evidence of omitted variables, one type of specification error.

```
. estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of price

H0: Model has no omitted variables

F(3, 894) =  11.04
 Prob > F = 0.0000
```

Description

Figure 13.5 Omitted Variable Test

13.3.3 Correcting Specification Error

The remedy for specification error is relatively straightforward if we have additional relevant variables in our data set: We add to the model any omitted variables that are statistically significant and experiment with alternative functional forms to find a better fit. We also consider ways to check for specification error due to (a) omitted variables, (b) nonlinear relationships, and (c) interaction between independent variables.

13.3.3.1 Correcting Omitted Variables

First, we consider the case of specification error due to omitted variables. As discussed earlier, any correlation between an omitted variable and an included variable will bias the estimate of the coefficient on the included variable.

For example, the age of the car is not taken into account in our analysis. Because age is likely to be correlated with mileage, it is important to add an age variable to get an unbiased estimate of the coefficient for mileage. We first calculate the age from the model year of the car, then run the regression[4].

The results in Figure 13.6 show that the coefficient on the age variable is negative and statistically significant. With each additional year, the price of a car declines by about $677 after controlling for mileage, newness, and type of fuel. Also, the coefficient on mileage changed from −0.158 in Figure 12.11 to −0.113 in this version. Mileage and age are positively correlated, so in the earlier regression, the variable mileage was picking up the effect of both mileage and age. When age is included as a variable, the coefficient on mileage is smaller and probably more accurate.

```
. gen age = 2023-year

. regress price mileage age new hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(5, 896)	=	161.42
Model	9.0047e+10	5	1.8009e+10	Prob > F	=	0.0000
Residual	9.9963e+10	896	111565629	R-squared	=	0.4739
				Adj R-squared	=	0.4710
Total	1.9001e+11	901	210887932	Root MSE	=	10562

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.1131618	.0117554	-9.63	0.000	-.1362332	-.0900904
age	-677.205	95.71261	-7.08	0.000	-865.052	-489.358
new	8146.808	1088.392	7.49	0.000	6010.714	10282.9
hybrid	3715.131	1529.999	2.43	0.015	712.3313	6717.931
electric	9130.816	1834.971	4.98	0.000	5529.473	12732.16
_cons	33326.24	666.5443	50.00	0.000	32018.07	34634.41

Figure 13.6 Regression Correcting for Specification Error

13.3.3.2 Correcting the Functional Form

Theory or inspection of the data may lead us to believe that the relationship between the dependent variable and one or more independent variables is nonlinear. One way to represent a nonlinear relationship between the dependent variable (y) and an independent variable (x) is to add variables representing powers of the independent variable, usually x^2 and occasionally higher power such as x^3. Although the relationship between y and x is now nonlinear, it is still considered a linear regression model because it is *linear in the parameters*, meaning that (a) the left side of the equation is the dependent variable (y) or some transformation of y and (b) the right side of the equation is a linear combination of independent variables (x) and/or transformed versions of those x variables.

In the analysis of the prices of cars, we considered whether the per-year depreciation might be greater for new cars than for older cars. In other words, is the relationship between price and age nonlinear? We can test this directly by calculating a quadratic term (age2) and adding it to the regression model, as shown in Figure 13.7.

```
. gen age2     = age^2

. regress price mileage age age2 new hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(6, 895)	=	155.91
Model	9.7106e+10	6	1.6184e+10	Prob > F	=	0.0000
Residual	9.2904e+10	895	103803543	R-squared	=	0.5111
				Adj R-squared	=	0.5078
Total	1.9001e+11	901	210887932	Root MSE	=	10188

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.0444093	.0140744	-3.16	0.002	-.072032	-.0167865
age	-2435.388	232.3414	-10.48	0.000	-2891.385	-1979.39
age2	50.75703	6.155197	8.25	0.000	38.67673	62.83733
new	4863.531	1122.813	4.33	0.000	2659.879	7067.183
hybrid	3994.29	1476.204	2.71	0.007	1097.066	6891.514
electric	9075.872	1770	5.13	0.000	5602.038	12549.71
_cons	36346.64	739.9529	49.12	0.000	34894.4	37798.89

Description

Figure 13.7 Multiple Regression With Quadratic Term

The *p*-value on age2 is 0.000, indicating that the quadratic term is statistically significant at the 1% level. This confirms that our intuition that the per-year depreciations is high in new cars and declines over time. Using calculus, the effect of age on price is ∂price/∂age = β_{age} + 2(β_{age2})(age)= −2435 + 2(50.8)(age). This implies that a one-year-old car (age = 1) depreciates at $2,333 per year, while a five-year old car depreciates at $1,927 per year.

Adding a quadratic term is only one way to represent nonlinear relationships in regression analysis. Other widely used transformations include logarithms of *y* and/or the *x* variables:

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

$$y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

Because these examples are linear in the parameters, they can be estimated using OLS and implemented by the **regress** command in Stata. Appendix 8 illustrates graphically examples of these nonlinear functional forms and explains how to calculate the marginal effect of *x* on *y* for each one.

13.3.3.3 Correcting for Missing Interaction Terms

The third type of specification error is missing interaction terms. Now we can return to the question of whether the per-mile depreciation rates differ by fuel type. Because of the small number of used electric cars in the sample (just 14), we cannot test the depreciation of hybrid and electric cars separately, but we can compare the depreciation of gas cars with hybrid and electric cars combined. In other words, is

the mileage coefficient the same for gas cars and other cars? To answer this question, we calculate a new variable that is the product of the mileage variable and a dummy for alternative fuel cars (hybrid or electric). Because a car cannot be both hybrid and electric, hybrid+electric is a dummy variable equal to 0 for gas cars and 1 for hybrid or electric cars.

```
gen mileage_alt = mileage*(hybrid+electric)
```

The new model can be written as follows:

$$\text{price} = \beta_0 + \beta_1 \text{mileage} + \beta_2(\text{mileage\_alt}) + \beta_3 \text{new} + \beta_4 \text{hybrid} + \beta_5 \text{electric} + \varepsilon$$

$$\text{price} = \beta_0 + \beta_1 \text{mileage} + \beta_2(\text{mileage\_alt}) + \beta_3 \text{new} + \beta_4 \text{hybrid} + \beta_5 \text{electric} + \varepsilon$$

The third term on the right side of the equation is the interaction term. The effect of mileage on price for gas cars is β_1, while the effect of mileage on hybrid and electric vehicles is $\beta_1+\beta_2$.

If β_2 is significantly different from zero, it implies that gas cars and alternative fuel cars depreciate at a different rate than gas cars.

Figure 13.8 shows the calculation of the interaction term, the **regress** command, and the output. The results in Figure 13.8 indicate that mileage_alt is not statistically significant at the 5% level. In other words, there is no evidence that the per-mile depreciation rate differs between gas cars and alternate fuel cars. We should note that the mileage_alt coefficient is fairly close to the threshold for statistical significance (p = 0.070). It is possible that with a larger sample of hybrid and electric cars, this coefficient would be statistically significant.

```
. gen mile_alt = mileage*(hybrid+electric)
(69 missing values generated)

. regress price mileage mile_alt age age2 new hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(7, 894)	=	134.18
Model	9.7352e+10	7	1.3907e+10	Prob > F	=	0.0000
Residual	9.2658e+10	894	103644179	R-squared	=	0.5124
				Adj R-squared	=	0.5085
Total	1.9001e+11	901	210887932	Root MSE	=	10181

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.0421777	.0141379	-2.98	0.003	-.0699251	-.0144303
mile_alt	-.0475126	.0308227	-1.54	0.124	-.108006	.0129807
age	-2418.973	232.407	-10.41	0.000	-2875.1	-1962.846
age2	50.37	6.155593	8.18	0.000	38.2889	62.45109
new	4661.284	1129.596	4.13	0.000	2444.315	6878.253
hybrid	5725.588	1853.989	3.09	0.002	2086.909	9364.266
electric	9572.561	1797.752	5.32	0.000	6044.255	13100.87
_cons	36223.6	743.6806	48.71	0.000	34764.04	37683.17

Description

Figure 13.8 Multiple Regression with an Interaction Term

In summary, we attempted to address the problem of specification error in three ways. First, we added the age variable, which was statistically significant though it reduced the coefficient on mileage. Second, we tried including an age-squared term, which was statistically significant, implying that the relationship between price and age is nonlinear. Last, we tested the interaction between mileage and an alternate

fuel dummy variable, but the coefficient was not statistically significant. Unfortunately, making these changes to the car price model did not change the results of the Ramsey RESET test (not shown). Clearly, numerous other factors affect the price of cars, including horsepower, ride quality, passenger space, cargo capacity, features, and even color. However, the changes did improve the predictive power of the model, raising the value of R^2 from 0.44 to 0.51.

13.4 MULTICOLLINEARITY

Multicollinearity refers to a condition in which two or more independent variables are closely correlated with one another. Perfect multicollinearity refers to the case where there is a linear combination of independent variables that is exactly equal to zero. For example, if one mistakenly includes variables for the male population, the female population, and the total population, then there exists a linear combination (males + females - total) that is equal to zero for all observations. If there is perfect multicollinearity, the model cannot be estimated. In this case, Stata will simply omit one of the collinear variables and report results for the rest of the model. This explains why we must omit one of the dummy variables representing a categorical variable if there is a constant. For example, if we include dummy variables for gas, hybrid, and electric cars in our price regression model, then the sum of the three dummies minus the variable associated with the constant (1) would be zero for all observations.

Imperfect multicollinearity refers to a case where there is a linear combination of independent variables that is close to zero. Because perfect multicollinearity is rare (and often the result of a mistake in coding), imperfect multicollinearity is often referred to simply as multicollinearity. Multicollinearity can occur if the model includes multiple variables that are measuring similar concepts, such as household income and expenditure or two measures of intelligence. It results in large standard errors of the coefficients and thus large confidence intervals. This is because the data do not allow us to estimate the effect of each variable independently with much accuracy. However, (imperfect) multicollinearity is not a violation of the assumptions behind OLS regression, so OLS results are still the best linear unbiased estimates (BLUE).

A simple way to identify multicollinearity is by creating a correlation matrix with the independent variables. If a pair of independent variables has a correlation coefficient greater than 0.8 or 0.9, there may be a problem of multicollinearity. The correlation matrix of independent variables in the latest model (Figure 13.7) is shown in Figure 13.9. The results indicate that only one pair of independent variables is highly correlated: age and age2 have a correlation of $r = 0.833$.

```
. correl mileage age age2 new hybrid electric
(obs=1,031)
```

	mileage	age	age2	new	hybrid	electric
mileage	1.0000					
age	0.6673	1.0000				
age2	0.2859	0.8330	1.0000			
new	-0.5535	-0.4878	-0.1605	1.0000		
hybrid	-0.0093	-0.0205	-0.0070	0.1062	1.0000	
electric	-0.1897	-0.1564	-0.0603	0.2594	-0.0558	1.0000

Description

Figure 13.9 Correlation Among Independent Variables

A more advanced test is the variance inflation factor, or VIF, which is calculated for each independent variable. The VIF for independent variable *i* is calculated as follows:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is obtained from regressing X_i on all the other independent variables. If an independent variable is closely correlated with other independent variables, the R^2 of this regression will be close to 1.0 and the VIF factor will be large. If an independent variable is not correlated with any other independent variable, R^2 will be close to 0 and the VIF will be close to 1. There is no consensus on the VIF threshold for considering multicollinearity a problem. One rule of thumb is that a VIF greater than 10 deserves attention, but some researchers prefer a threshold of 4 (O'Brien, 2007).

In Stata, this VIF test can be implemented with the menu system as follows: Statistics → Postestimation → Specification, diagnostics, and goodness-of-fit analysis → Variance inflation factors. Alternatively, the VIF test can be carried out with the command **estat vif**. Like other postestimation commands, it uses the results from the most recent regression model. The results of the VIF test for our model are shown in Figure 13.10. The table suggests multicollinearity in age and age2.

```
. estat vif
```

Variable	VIF	1/VIF
age	11.04	0.090548
age2	6.84	0.146159
mileage	2.90	0.345096
new	1.70	0.588885
electric	1.10	0.910280
hybrid	1.05	0.954866
Mean VIF	4.10	

Description

Figure 13.10 Test for Multicollinearity

In our model, both the age and age2 variables are statistically significant. Since multicollinearity does not result in bias in the coefficients or in the standard errors, and since both age and age2 are significant, we don't have to worry about multicollinearity in this case and can leave the model as is.

What is the remedy if we find strong multicollinearity and insignificant coefficients? Ideally, the researcher would collect a larger sample of data so that each coefficient can be estimated with greater precision despite the multicollinearity. In cases where this is not possible, some researchers propose dropping one of the correlated variables so that the remaining one becomes statistically significant. However, this "solution" just introduces omitted variable bias, because the remaining variable picks up some of the effect of the omitted variable. In other words, the results are misleading because they exaggerate the effect of the included variable and ignore the effect of the excluded variable. For this reason, O'Brien (2007) cautions that some of the remedies for multicollinearity may be worse than the original problem.

A better approach is to test the combined effect of the two variables by running an ***F* test** of the joint hypothesis that both coefficients are equal to zero. In Stata, this is implemented with the command **test x1 x2**, where x1 and x2 are the two correlated independent variables. If the null hypothesis is rejected, the researcher can report that two variables are jointly significant, but the effects of each variable cannot be independently measured because of the close correlation between the two.

13.5 HETEROSCEDASTICITY

An important assumption behind OLS regression is that the variance of the error term is constant. In other words, OLS assumes that the dispersion of the errors is the same throughout the sample of observations. This is called **homoscedasticity**. However, in practice, the variance of the errors may differ, a condition called heteroscedasticity. Figure 13.11 shows what heteroscedasticity looks like on a graph. The residuals (*e*) are small (in absolute value) for low values of *x* but are much larger for large values of *x*, suggesting that the variance of the (unobserved) errors is not constant.

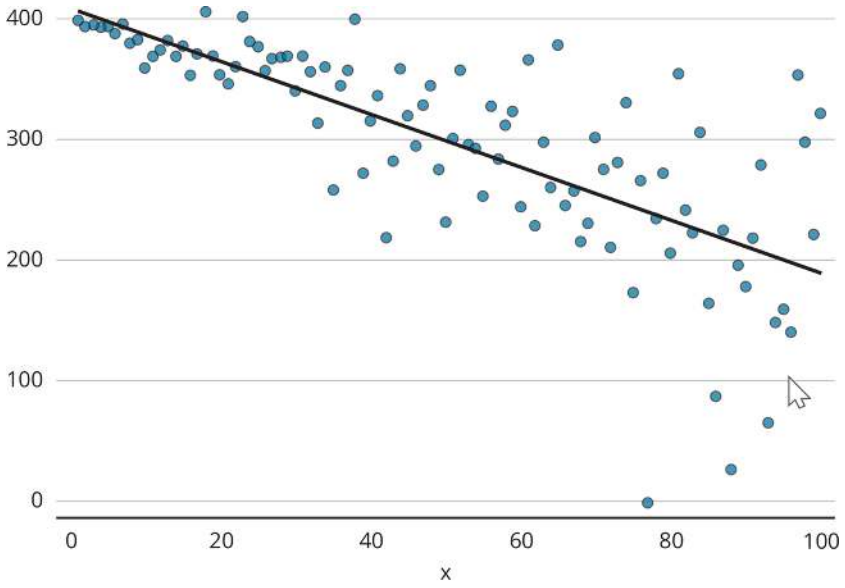
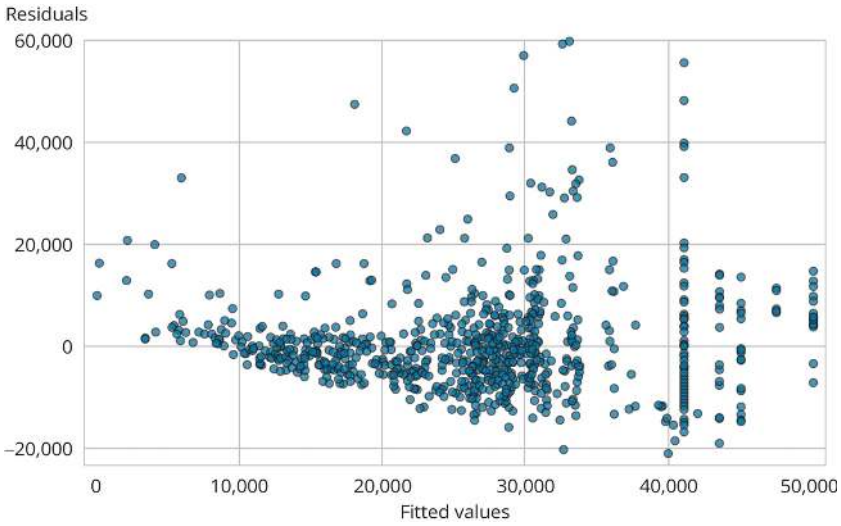


Figure 13.11 Example of Heteroscedasticity

In a regression model of food spending as a function of income, the errors might be larger among high-income households than among low-income households. This is an example of heteroscedasticity being a function of an independent variable. Or in our car price model, the errors might be greater for more expensive cars. In this case, heteroscedasticity would be a function of the dependent variable.

What are the consequences of using OLS regression when heteroscedasticity is present? The estimated coefficients are still unbiased but (a) the estimated coefficients are not *efficient*, meaning that they do not make use of all available information in the data, and (b) the standard errors of the coefficients are incorrectly measured.

We can visually check for heteroscedasticity by plotting the residuals ($e = y - \hat{y}$) against the fitted (or predicted) values of the dependent variable ($\hat{y}$). In Stata, we can do this with the menu system or a command. Using the menu system, the sequence is as follows: Graphics → Regression diagnostic plots → Residual-versus-fitted. Or we can run the command **rvfplot**. In both cases, it will use the most recent regression model. If we request this plot after our car price regression analysis, we get the plot shown in Figure 13.12.



Description

Figure 13.12 Scatterplot of Residuals Against Predicted Prices

In this graph, the variance of the residuals is represented by the degree of vertical dispersion of the dots around the horizontal center line. The degree of dispersion seems smaller at low values of $\hat{y}$ (near the left side of the graph) than at high values (toward the right side), suggesting some heteroscedasticity. As an aside, the three vertical lines of dots represent the new cars with zero mileage. All new gas cars have the same predicted value (the first line), and likewise with the new hybrid and electric cars (the second and third lines, respectively).

We can also check for heteroscedasticity by plotting the residuals against each of the continuous independent variables using the command **rvpplot**. In our car price model, we would run **rvpplot mileage** and **rvpplot age** to see if the variance of the residuals differs across values of these variables. This command was demonstrated in Section 13.3 as a tool to check for nonlinearity.

The main statistical test for heteroscedasticity is the **Breusch–Pagan/Cook–Weisberg test**, which assumes that the variance of the error term is a function of either the predicted value of the dependent variable ($\hat{y}$) or some set of independent variables (x). Multiple versions of the Breusch–Pagan/Cook–Weisberg test can be implemented with the Stata command **estat hettest**. Here are some of the more common options:

estat hettest tests whether the variance of the residuals is a function of the predicted values of the dependent variable and that the errors are normally distributed.

estat hettest *varlist* (where ***varlist*** is a list of independent variables) tests whether the variance of the residuals is a function of the independent variables listed. It also assumes that the errors are normally distributed.

estat hettest rhs tests whether the variance of the residuals is a function of all the independent variables (*rhs* refers to variables on the right-hand side of the equation).

estat hettest, iid tests for heteroscedasticity without assuming that the errors are normally distributed. It can be combined with ***varlist*** or **rhs**.

To run the Breusch–Pagan/Cook–Weisberg test for heteroscedasticity in the car price model, we can use the **estat hettest** command or the menu system as follows: Statistics → Postestimation → Specification, diagnostics, and goodness-of-fit analysis → Tests for heteroscedasticity, then select the type.

Following the model of car prices from Figure 13.7 again, Figure 13.13 shows the test for heteroscedasticity and the results. The low probability (<0.000) indicates a rejection of the null hypothesis of constant variance (homoscedasticity) in our model.

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of price

H0: Constant variance

    chi2(1) =  61.00
Prob > chi2 = 0.0000
```

Description

Figure 13.13 Test for Heteroscedasticity

When we find evidence of heteroscedasticity, there are two types of remedy. The first is to run a regression with robust standard errors, also called Huber–White or sandwich standard errors. This is the easiest approach. It uses a different method to calculate the standard error of each coefficient, resulting in a wider confidence interval and lower *p* value for each coefficient. However, the estimated coefficients are the same as the OLS estimates. In Stata, robust standard errors can be implemented by adding the **vce(robust)** option to the **regress** command, as shown here.

```
regress y x1 x2, vce(robust)
```

A second approach to dealing with heteroscedasticity is to use a generalized least squares (GLS) analysis, in which the variance of the residuals is estimated as a function of variables in the model. The variance estimates are then used to give greater weight to observations with lower variance. The advantage of GLS is that it adjusts the coefficients, making use of information about the differences in the variance of the error terms. The disadvantage is that it requires good knowledge of how the variance varies across observations, which can be difficult to obtain. A detailed description of GLS regression is, however, beyond the scope of this book. Interested readers may consult Greene (2018) and Woolridge (2016).

13.6 ENDOGENEITY

It is true that "correlation does not imply causation," but with careful use of data and methods, regression analysis *can* imply causality. The challenge is to ensure that the independent variables are *exogenous*. In statistical terms, independent variables are exogenous if they are uncorrelated with the error term (ε). If any of the independent variables are correlated with the error term, the model has an endogeneity problem. With endogeneity, the OLS coefficients are biased and may also be inconsistent, in that, as the sample size increases, the estimated coefficient will not converge toward the true value of the coefficient.

Under what conditions would an independent variable be correlated with the error terms? There are at least two situations where this may occur[5].

> The independent variable is influenced by the dependent variable. This is called reverse causation.
>
> The independent variable and the dependent variable are both affected by a variable that has been omitted from the model. The omitted variable is sometimes called a confounding factor.

Endogeneity is a serious problem in the use of regression analysis, particularly in the social sciences and other fields where it is difficult to run a controlled experiment. In our car price regression model, we can be fairly confident that fuel type, mileage, and age of the car are not influenced by the price because these characteristics predate the setting of the price. Thus, reverse causation is not likely to be an issue in our model. But there may be omitted variables that are correlated with both the dependent variable and one or more independent variables. For example, suppose hybrid car owners keep their cars cleaner or better maintained than gas or electric car owners. If cleanliness improves the resale value of a car, then the hybrid dummy coefficient will be biased upward because it captures the effect of being a hybrid *and* the effect of being cleaner or better maintained than average.

And consider the case of a study estimating the effect of the size of the police force on the crime rate (see Levitt, 1997). The research question is, "Does hiring more police officers reduce the crime rate?" If we ignore the endogeneity and use OLS regression to estimate the crime rate per 1,000 inhabitants (y) as a function of the size of the police force per 1,000 inhabitants (x) and other factors, we may get a positive coefficient. But clearly, we cannot conclude that expanding the police force increases the crime rate. The more likely explanation is that when the crime rate increases, local governments respond by increasing the size of their police force. In other words, the crime rates (y) can influence the size of the police force (x), an example of reverse causation. This demonstrates the point that endogeneity can generate biased coefficients, even changing the sign of the coefficient.

There are several techniques for addressing endogeneity that involve advanced methods. A detailed description is beyond the scope of this book, but we provide a brief introduction to some of these methods in Chapter 15.

13.7 NONNORMALITY

Normally distributed error terms are sometimes considered an "optional" assumption for OLS regression. This is because normality is not necessary for OLS, but it is convenient. It is not necessary in that, even without normally distributed errors, OLS will still generate the best linear unbiased estimates (BLUE) of the coefficients. On the other hand, normality is convenient in that it ensures that the estimated coefficients are normally distributed so that the p values and confidence interval will be correct even in small samples.

However, even if the error terms are not normally distributed, the estimated coefficients may still be normally distributed. Because OLS coefficients can be interpreted as a type of weighted average, the central limit theorem tells us that the distribution of the estimated coefficients ($\hat{\beta}_i$) becomes more normal as the sample size increases, even if the error term (ε) is not normally distributed. As few as 100 observations may be sufficient to ensure that the OLS coefficients are normally distributed, implying that the standard *p* values and confidence intervals are reliable. Thus, the problem of normality is not a serious problem for regression analysis unless the sample is quite small (Bailey, 2016).

How do we visually inspect the normality of residuals? As described in Section 13.2, we can use the **predict** command to calculate the residual for each observation (that is, each car in the database) and give it the variable name *e*. We can then compare the distribution of the residual and the normal distribution using a histogram, as described in Chapter 6.

```
lab var e Residuals
histogram e, normal width(1000) xlabel(-20000(10000)70000)
```

The **normal** option adds a line showing the normal distribution with the same mean and standard deviation. The **width** option tells Stata how wide to make each bar, while the **xlabels** option indicates where to start and end the labels and what interval to use.

The histogram in Figure 13.14 shows that the distribution of the residuals diverges from the normal distribution. There are too many residuals between −10,000 and 0 and too few between 0 and 20,000.

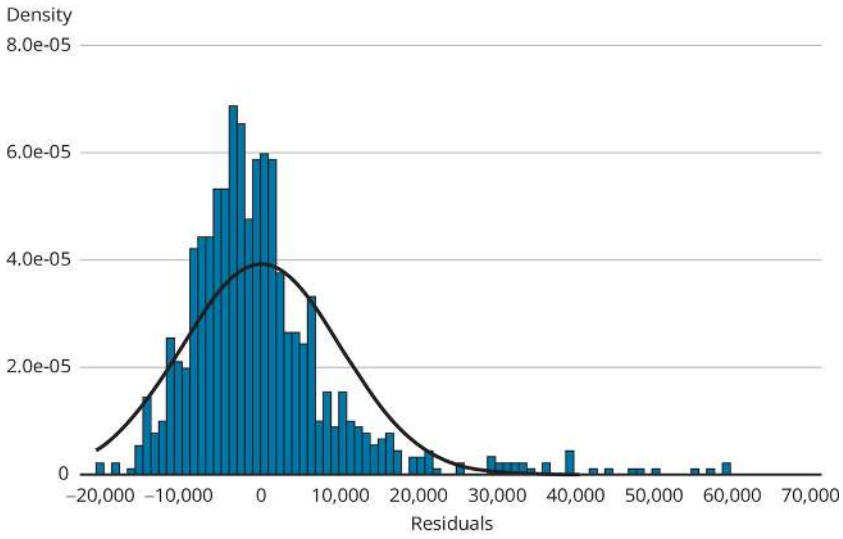


Figure 13.14 Histogram of Residuals and the Normal Distribution

We can measure the divergence from normality by calculating **skewness** and **kurtosis**. Skewness is generally considered a measure of symmetry, but this is not always the case. All symmetric distributions have a skewness of zero, but a skewness of zero does not guarantee that the distribution is symmetric. For example, a distribution may have zero skewness and still be asymmetric if the left tail of the distribution is thick and the right tail is long.

Kurtosis generally measures the presence of thick tails. The normal distribution has a kurtosis value of 3, so *excess kurtosis* is defined as kurtosis minus 3. In other words, positive excess kurtosis implies that the tails are thicker than in the normal distribution, while negative excess kurtosis means the tails are

thinner than in the normal distribution. Table 13.1 provides some guidance in interpreting skewness and kurtosis.

TABLE 13.1 ■ Interpreting Skewness and Kurtosis

	Skewness	Kurtosis
Range	−∞ to + ∞	1 to +∞
Value in the normal distribution	0	3
Interpretation of low values	Negative skewness means the distribution is skewed to the left. In other words, the left tail is longer or thicker.	Kurtosis less than 3 means a higher peak and smaller tails than a normal distribution.
Interpretation of high values	Positive skewness means the distribution is skewed to the right. In other words, the right tail is longer or thicker.	Kurtosis greater than 3 means a flatter distribution and larger tails than a normal distribution.

	Skewness	**Kurtosis**
Range	−∞ to + ∞	1 to +∞
Value in the normal distribution	0	3
Interpretation of low values	Negative skewness means the distribution is skewed to the left. In other words, the left tail is longer or thicker.	Kurtosis less than 3 means a higher peak and smaller tails than a normal distribution.
Interpretation of high values	Positive skewness means the distribution is skewed to the right. In other words, the right tail is longer or thicker.	Kurtosis greater than 3 means a flatter distribution and larger tails than a normal distribution.

In Stata, the command **sum e, detail** will calculate skewness, kurtosis, and many other statistics for the variable *e*. Alternatively, we can use the **tabstat** command to calculate these two statistics alone.

As shown in Figure 13.15, the skewness of the residual from the car price regression is 2.17, indicating that it has a positive skew. A skewness value between −2 and 2 is considered acceptable for normality, and our result is slightly outside this range. The kurtosis of the residual is 10.75, indicating that the tails are thicker or longer than in a normal distribution. This is particularly evident on the right side of the distribution.

```
. tabstat e, s(skew kurt)

    Variable |  Skewness  Kurtosis
-------------+--------------------
           e |  2.165791   10.7478
```

Figure 13.15 Skewness and Kurtosis

Is the distribution of the residual significantly different from normal? In Stata, the skewness–kurtosis test (sometimes called the D'Agostino K^2 test) checks skewness and kurtosis separately, then runs a joint test of the null hypothesis that skewness = 0 and kurtosis = 3, which would be consistent with normality.

Using the car price data, we can run the test using the menu system or using a command. To use the menu system, we follow this sequence: Statistics → Summaries, tables, and tests → Distributional plots and tests → Skewness and kurtosis normality test, and then select the variable representing the residual. Alternatively, we can use the command **sktest**. The command and results are shown in Figure 13.16.

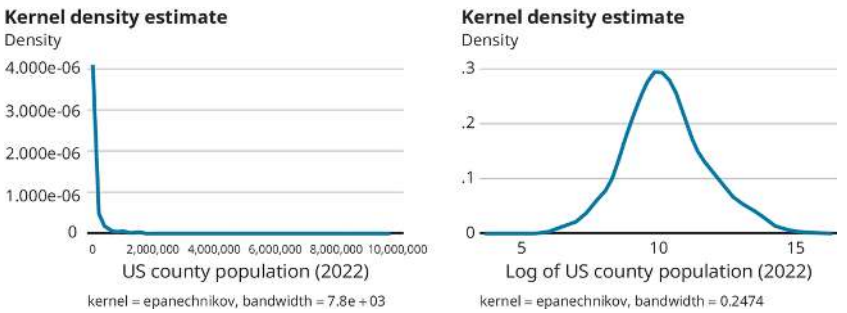
```
. sktest e

Skewness and kurtosis tests for normality
                                                    ------ Joint test ------
    Variable |        Obs   Pr(skewness)   Pr(kurtosis)   Adj chi2(2)  Prob>chi2
-------------+-------------------------------------------------------------------
           e |        902         0.0000         0.0000        336.52     0.0000
```

Figure 13.16 Test for Normality

The number under “Pr(skewness)” is the probability of getting the observed distribution of the residual if the skewness were actually zero, while the number under “Pr(kurtosis)” indicates the probability of getting this result if the kurtosis were 3. Both null hypotheses can be rejected at the 1% level. Naturally, the joint hypothesis that skewness = 0 and kurtosis = 3 is also rejected at the 1% level. In summary, the test rejects the null hypothesis that the residuals from the car price regression are normally distributed.

How do we address the problem of nonnormality in the residuals? The most common approach is to try transforming the dependent variable. For example, some variables, such as household income, individual health care spending, and population data, have many small, positive values and a few very large values. In statistical terms, these variables are positively skewed. However, if we transform the variable by taking the natural logarithm, the new variable is often much closer to a normal distribution. This makes it more likely that residuals from a regression model will be normally distributed, particularly if the independent variables are normally distributed.

We can take the example of the population of counties in the United States. The **kdensity** command gives us a “smoothed” histogram of a variable. The left side of Figure 13.17 shows the distribution of the 3,144 counties by population in 2022. Most counties have a population of less than 25,000, but a few have more than 1 million inhabitants. The right side of Figure 13.6 shows the distribution of the natural logarithm of county population, which is clearly more similar to a normal distribution. The output of the **tabstat** command (not shown) indicates that the skewness is 12.8 for the original county population but 0.26 for the logarithm of population. Recall that the skewness of the normal distribution is zero. Likewise, kurtosis is 277 in the original population and 3.35 in the log population. Kurtosis is 3.0 in the normal distribution. Thus, the natural logarithm of the population variable is much closer to normal than the original variable. The commands to generate these results are shown below:



Description

Figure 13.17 Distribution of County Population and the Logarithm of County Population

```
kdensity population
gen logpop = log(population)
kdensity logpop
tabstat population logpop, s(skewness kurtosis)
```

Transforming the dependent variable so that it is normally distributed (or closer to a normal distribution) improves the chances that the residuals will be normally distributed. This is particularly important when the sample is small.

In the case of our car price regression model, we can calculate the logarithm of price and run the new version of the analysis as follows:

```
gen lnprice = log(price)
regress lnprice mileage age new hybrid electric
```

Calculating the skewness and kurtosis using the **`tabstat`** command (not shown) confirms that this transformation was successful in reducing the skewness from 2.17 to 0.68 and reducing the kurtosis from 10.75 to 4.27, suggesting a distribution much closer to normal. However, the **`sktest`** still rejects the null hypothesis that skewness is zero and kurtosis is 3, as well as the null hypothesis of normality.

If transforming the dependent variable is not successful in producing normally distributed residuals, some researchers suggest setting a stricter criterion for statistical significance of the coefficients. For example, one might insist on a *p* value of less than 1% to conclude that the coefficient is significantly different from zero. There are also regression models that give less weight to the outliers, including robust regression (implemented with **rreg** in Stata) and quantile regression (implemented with **qreg**). However, these methods are beyond the scope of this book.

As mentioned previously, in the absence of other problems, even if the errors are not normally distributed, OLS still gives the best linear unbiased estimates (BLUE) of the coefficients. Furthermore, unless the number of observations is small, the coefficients are likely to be normally distributed if even the errors are not normally distributed. Given that we have more than 900 observations, OLS will generate reliable *p* values and confidence intervals regardless of the distribution of the error terms.

13.8 PRESENTING THE RESULTS

In this section, we normally give examples of how to write results for a nontechnical publication (e.g., a newspaper) and for a technical publication (e.g., an academic journal). However, newspapers and other nontechnical publications rarely describe regression diagnostics, so for this chapter, we will only provide the example of how to describe the results for a journal or other technical publication.

For a journal article, the write-up should include a description of the results of each test, expressed in terms of rejecting or failing to reject the null hypothesis. The value of the test statistic and the *p* value can be included in parentheses. Thus, the regression diagnostics for the car price regression could be described as follows:

> This model was subjected to various diagnostic tests. First, we ran the Ramsey RESET test of omitted variables using powers of the dependent variable. In an earlier version of the model, the test failed to reject the null hypothesis that there were no omitted variables. To address this issue, we added variables representing the age of the car, age squared, and a term for

interaction between age and the car using alternate fuel. The interaction term was not statistically significant, so it was not included. The age and age squared variables were significant ($p < 0.000$ for both). However, even after adding these explanatory variables, the RESET test still rejected the null hypothesis of no omitted variables, $F(3, 892) = 24.82$, $p < 0.000$. To the extent that the missing variables are correlated with included variables, the coefficients on the latter may be biased.

Heteroscedasticity was checked using the Breusch–Pagan/Cook–Weisberg test. It failed to reject the null hypothesis of homoscedasticity when modeling the variance of the errors as a function of the predicted values of price, $\chi^2(1) = 61.00$, $p < 0.000$. OLS estimates of the coefficients are not biased by heteroscedasticity, but the standard errors of the estimates may be misleading.

We tested for multicollinearity by calculating the variance inflation factors (VIFs) for all the independent variables. The only variables to show VIF values above 3 were age and age squared. Since the coefficients on these variables were statistically significant and multicollinearity does not result in biased coefficients or biased standard errors, we retained both variables in the model.

The D'Agostino K^2 test rejected the null hypothesis that the residuals were normally distributed, adjusted $\chi^2(2) = 336.52$, $p < 0.0000$. When the residuals are not normally distributed, caution must be taken in assigning statistical significance when p values are marginal. Fortunately, all six coefficients in our model have low p values (<0.01), providing reassurance that the coefficients are likely different from zero.

13.9 SUMMARY OF COMMANDS USED IN THIS CHAPTER

As described in Chapter 4, this last section of each chapter summarizes all the Stata code used in the chapter (Table 13.2). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 13.2 ■ Summary of Commands Used in this Chapter

Function	Stata command(s)
Add random error to a variable	gen mileage_err = mileage + rnormal(0,18000)
Looking for outliers in the residuals	predict e, resid sum e, detail
Using Cook's *D* to look for outliers	predict CooksD, cooksd browse if CooksD>1 & CooksD!=.
Adding age and age squared independent variables to a regression	gen age = 2023 – year gen age2 = age^2
Plot residuals against an independent variable to check for specification error	rvpplot mileage
Plot residuals against predicted dependent variable to check for specification error	rvfplot
Omitted variable test based on powers of predicted dependent variable	estat ovtest
Omitted variable test based on powers of the explanatory variables	estat ovtest, rhs
Calculate interaction terms for regression analysis	gen age_alt = age*(hybrid+electric)
Generate matrix of correlation coefficients	correl age age2 mileage new hybrid electric
Run variance inflation factor test of multicollinearity	estat vif
Plot regression residuals against an explanatory variable to check for heteroscedasticity	rvpplot mileage
Plot regression residuals against the predicted values of the dependent variable to check for heteroscedasticity	ryfplot
Run the Breusch–Pagan and Cook–Weisberg tests for heteroscedasticity	estat hettest
Generate histogram of residual to check for normality	predict e, resid histogram e, normal width(1000) xlabel(–20000(10000)70000)
Calculate skewness and kurtosis as part of checking for normality of residual	tabstat e, s(skew kurt)
Test for skewness, kurtosis, and normality in the residual	sktest e
Compare the skewness and kurtosis of two variables	tabstat population logpop, s(skew kurt)

Function	**Stata command(s)**
Add random error to a variable	gen mileage_err = mileage + rnormal(0,18000)
Looking for outliers in the residuals	predict e, resid sum e, detail
Using Cook's *D* to look for outliers	predict CooksD, cooksd browse if CooksD>1 & CooksD!=.

Function	Stata command(s)
Adding age and age squared independent variables to a regression	gen age = 2023 – year gen age2 = age^2
Plot residuals against an independent variable to check for specification error	rvpplot mileage
Plot residuals against predicted dependent variable to check for specification error	rvfplot
Omitted variable test based on powers of predicted dependent variable	estat ovtest
Omitted variable test based on powers of the explanatory variables	estat ovtest, rhs
Calculate interaction terms for regression analysis	gen age_alt = age*(hybrid+electric)
Generate matrix of correlation coefficients	correl age age2 mileage new hybrid electric
Run variance inflation factor test of multicollinearity	estat vif
Plot regression residuals against an explanatory variable to check for heteroscedasticity	rvpplot mileage
Plot regression residuals against the predicted values of the dependent variable to check for heteroscedasticity	ryfplot
Run the Breusch–Pagan and Cook–Weisberg tests for heteroscedasticity	estat hettest
Generate histogram of residual to check for normality	predict e, resid histogram e, normal width(1000) xlabel(−20000(10000)70000)
Calculate skewness and kurtosis as part of checking for normality of residual	tabstat e, s(skew kurt)
Test for skewness, kurtosis, and normality in the residual	sktest e
Compare the skewness and kurtosis of two variables	tabstat population logpop, s(skew kurt)

EXERCISES

1. Match the issue on the left side with the correct problem(s) that it causes with OLS regression on the correct definition on the right side (Table 13.3).

TABLE 13.3 ■ Matching Regression Problems and Their Descriptions

1. Measurement error in the independent variable	**a.** This does not create bias in the coefficient or in the standard error, but the standard errors may be large, making the coefficient statistically insignificant.
2. Omitted variable that is correlated with an included independent variable	**b.** The model is still the best linear unbiased estimate, but *p*-values associated with the standard errors are incorrect.
3. Heteroscedasticity	**c.** The coefficient on the affected variable will be biased toward zero.
4. Endogeneity due to reverse causality	**d.** The coefficient is not biased, but the standard errors are incorrect and the model is inefficient in that it does not use all available information.
5. Nonnormality of the error terms	**e.** The coefficient on the correlated independent variable is biased.
6. Multicollinearity	**f.** Biased coefficient on the affected independent variable

1. Measurement error in the independent variable	a. This does not create bias in the coefficient or in the standard error, but the standard errors may be large, making the coefficient statistically insignificant.
2. Omitted variable that is correlated with an included independent variable	b. The model is still the best linear unbiased estimate, but *p*-values associated with the standard errors are incorrect.
3. Heteroscedasticity	c. The coefficient on the affected variable will be biased toward zero.
4. Endogeneity due to reverse causality	d. The coefficient is not biased, but the standard errors are incorrect and the model is inefficient in that it does not use all available information.
5. Nonnormality of the error terms	e. The coefficient on the correlated independent variable is biased.

6. Multicollinearity	f. Biased coefficient on the affected independent variable

2. Suppose you run a regression model, but you suspect there may be heteroscedasticity.
   a. What graph would you create to check for heteroscedastity?
   b. What test would you run to check for heteroscedasticity?
   c. If the test indicates that there is heteroscedasticity, what are two strategies you could use to remedy this problem?
3. In the exercises for Chapter 12, we estimated income (realinc) as a function of age, education, and a dummy variable for female respondents. The coefficient on age was not statistically significant.
   a. What type of graph would you use to check for possible nonlinear effect of age on income? What Stata command would generate this graph?
   b. Use gen to create a new variable, age2, which is equal to age squared, and add this new variable to the model of real income. Is age2 a statistically significant variable? Has the statistical significance of age changed? Why, or why not?
   c. How would you interpret the coefficients on the age and age2 variables?
   d. Given that the derivative of a quadratic equation is $dy/dx = \beta 1 + 2 \times \beta 2 \times x$, how would you use the age coefficients to calculate the age where income peaks, according to this regression model.
4. Let's return to the model of income as a function of age, education, and a dummy for female respondents.
   a. Create a variable *e* representing the residuals, and test the residual for nonnormality. What is the result of this test? Do you accept or reject the null hypothesis of normality?
   b. If the test rejects the null hypothesis of normality, with what remedies could you try to address this problem?
5. Suppose you suspect that education has a different effect on men and women. Calculate an interaction term for education and female, and use it in the regression analysis.
   a. What is the coefficient and *t* statistic on the interaction term?
   b. How would you interpret this result?

KEY TERMS

autocorrelation

Breusch–Pagan/Cook–Weisberg test

endogeneity

***F* test**

heteroscedasticity

homoscedasticity

kurtosis

measurement error

multicollinearity

nonnormality

skewness

specification error

Descriptions of Images and Figures

Back to Figure

The x-axis ranges from 0 to 200,000 for Mileage, and the y-axis ranges from −20,000 to 70,000 for Residuals. Points follow a horizontal pattern, mostly within the Residuals range of −20,000 to 20,000 and the Mileage range of 0 to 150,000, with a concentration below 100,000 miles.

Back to Figure

The output is as follows:

. estat ovtest

Ramsey RESET test for omitted variables

Omitted: Powers of fitted values of price

Hθ: Model has no omitted variables

F(3, 894) = 11.04

Prob > F = 0.0000

Back to Figure

The table includes:

Key statistics:

Number of observations (obs): 902

F-statistic (F(6, 895)): 155.91

Prob > F: 0.0000

R-squared: 0.5111

Adjusted R-squared: 0.5075

Root MSE: 10188

ANOVA Table:

Model Sum of Squares (SS): 9.7106e+10

Residual SS: 9.2904e+10

Total SS: 1.9001e+11

Degrees of freedom (df): 6 for the model, 895 for residuals

Coefficients table:

Mileage: Coefficient = -0.0444093 (negative impact on price), statistically significant (p-value = 0.002).

Age: Coefficient = -2435.388 (negative impact on price), p-value = 0.000.

Age squared (age2): Coefficient = 50.75703 (positive impact), p-value = 0.000.

New: Coefficient = 4863.531 (positive impact), p-value = 0.000.

Hybrid: Coefficient = 3994.28 (positive impact), p-value = 0.007.

Electric: Coefficient = 9075.87 (positive impact), p-value = 0.000.

Constant (_cons): Coefficient = 36346.64, p-value = 0.000.

Back to Figure

The table includes:

Key statistics:

Number of observations (obs): 902

F-statistic (F(7, 894)): 134.18

Prob > F: 0.0000

R-squared: 0.5124

Adjusted R-squared: 0.5078

Root MSE: 10181

ANOVA Table:

Model Sum of Squares (SS): 9.7352e+10

Residual SS: 9.2658e+10

Total SS: 1.9001e+11

Degrees of freedom (df): 7 for the model, 894 for residuals

Coefficients table:

Mileage: Coefficient = −0.042177 (negative impact on price), statistically significant (p-value = 0.003).

Mile_alt: Coefficient = −0.047512 (negative impact on price), p-value = 0.003.

Age: Coefficient = -2418.973 (negative impact), p-value = 0.124.

age2: Coefficient = 50.37 (positive impact), p-value = 0.000.

New: Coefficient = 4661.284 (positive impact), p-value = 0.000.

Hybrid: Coefficient = 5725.58 (positive impact), p-value = 0.002.

Electric: Coefficient = 9572.561 (positive impact), p-value = 0.000.

Constant (_cons): Coefficient = 36223.6, p-value = 0.000.

Back to Figure

The matrix displays pairwise correlations between these variables.

Mileage: Correlation of 1.0000 with itself, 0.6673 with age, 0.2859 with age2, -0.5353 with new, -0.0093 with hybrid, and -0.1897 with electric.

Age: Correlation of 1.0000 with itself, 0.8330 with age2, -0.4878 with new, -0.0205 with hybrid, and -0.1564 with electric.

Age2: Correlation of 1.0000 with itself, -0.1665 with new, -0.0070 with hybrid, and -0.0603 with electric.

New: Correlation of 1.0000 with itself, 0.1062 with hybrid, and 0.2594 with electric.

Hybrid: Correlation of 1.0000 with itself and -0.0558 with electric.

Electric: Correlation of 1.0000 with itself.

Back to Figure

The output is as follows:

. estat vif

Variable	VIF	1/VIF
age	11.04	0.090548
Age2	6.84	0.146159
mileage	2.90	0.345096
new	1.70	0.58885
electric	1.10	0.910280
hybrid	1.05	0.954866
Mean VIF	4.10	

Back to Figure

The x-axis represents Fitted values, ranging from 0 to 50,000, while the y-axis represents Residuals, ranging from -20,000 to 60,000. The points form a distinct V-shaped pattern, spreading out from left to right. Most of the points are concentrated between Fitted values of 10,000 to 30,000 and Residuals from -20,000 to 18,000.

Back to Figure

The output is as follows:

. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroscedasticity

Assumption: Normal error terms

Variable: Fitted values of price

Hθ: Constant variance

Chi2(1) = 61.00

Prob > chi2 = 0.0000

Back to Figure

On the left graph:

The x-axis represents “US county population (2022)” with a range from 0 to 10,000,000.

The y-axis represents Density, ranging from 0 to 4.000e-06.

An L-shaped curve shows a decreasing trend, starting with a density of 4.000e-06 and approaching 0 around 10,000,000 population, then remaining constant.

The text below the graph reads “kernel = epanechnikov, bandwidth = 7.8e+03.”

On the right graph:

The x-axis represents “Log of US county population (2022)” with a range from 0 to 15.

The y-axis represents Density, ranging from 0 to 0.3.

A bell-shaped curve peaks at a value of 10 with a density of 0.3.

The text below the graph reads “Log of US county population (2022).”

14 REGRESSION ANALYSIS WITH BINARY DEPENDENT VARIABLES

CHAPTER PREVIEW

Steps	Examples
Research question	Do views on climate change vary by age and education?
Null hypothesis	Age and education have no effect on views on climate change.
Test	Logit or probit analysis and *z* test of coefficients on age and education
Types of variables	Dependent variable is binary (accept or reject human-caused climate change). Independent variables can include continuous variables, categorical variables, or a mix of both.
When to use	When dependent variable is binary (0 or 1)
Assumptions	For the logit model: The log odds is a linear function of the independent variables. For the probit model: The probability that $y = 1$ is a cumulative normal density function of the independent variables.
Stata code: generic	**logit depvar indepvars** **probit depvar indepvars**
Stata code: example	**logit humcaus age female educ** **probit humcaus age female educ**

Steps	Examples
Research question	Do views on climate change vary by age and education?
Null hypothesis	Age and education have no effect on views on climate change.
Test	Logit or probit analysis and *z* test of coefficients on age and education
Types of variables	Dependent variable is binary (accept or reject human-caused climate change). Independent variables can include continuous variables, categorical variables, or a mix of both.
When to use	When dependent variable is binary (0 or 1)
Assumptions	For the logit model: The log odds is a linear function of the independent variables. For the probit model: The probability that $y = 1$ is a cumulative normal density function of the independent variables.
Stata code: generic	**logit depvar indepvars** **probit depvar indepvars**

Steps	Examples
Stata code: example	**logit humcaus age female educ** **probit humcaus age female educ**

14.1 INTRODUCTION

Chapter 12 introduced regression analysis, which estimates the equation that best describes the relationship between a dependent variable and one or more independent variables. We focused on ordinary least squares (OLS) regression, in which the dependent variable is continuous and the model is linear in the parameters. However, in many cases, we want to estimate a relationship in which the dependent variable is binary (yes/no) rather than continuous. Suppose we want to predict whether a household will purchase a car this year, whether an adult is working, whether a student will graduate from high school, or whether a patient will survive surgery. This chapter explains how to apply regression analysis to these types of problems. In particular, we focus on two types of regression analysis used on **binary variables**: logit[1] and **probit regression**.

Let's take a concrete example. The 2021 General Social Survey found that 16% of Americans believe that the climate has not been changing or that the change is mostly due to natural causes. Another 36% say that climate change is about equally caused by natural processes and human activity. And the remaining 48% believe that climate change is mostly caused by human activity.

Suppose we want to dig deeper and analyze the factors associated with belief that climate change is mostly caused by human activity? In a **logit regression** model, the dependent variable (y) takes just two possible values, "no" and "yes," which we represent mathematically as 0 and 1. Although the observed values of y are either 0 or 1, the predicted value of y is a number between 0 and 1, which is interpreted as the probability that $y = 1$ given the values of the independent variables. In our example, regression analysis would generate an equation that predicts the probability that a person with certain characteristics (e.g., a 32-year-old college-educated female) will believe in that climate change is mostly caused by human activity. It would also allow us to test the statistical significance of each independent variable.

This chapter begins by demonstrating the relevance of this type of analysis, describing research questions from various fields where the dependent variable is binary. Then, we discuss the use of a linear OLS model to analyze data with a binary dependent variable. This leads to a more extended discussion of the logit model, including the functional form, the method used to find the solution, and the interpretation of the coefficients. We also briefly consider the closely related probit model and how it differs from the logit model.

Regression models can also handle categorical dependent variables with more than two values, such as political party affiliation or marital status. This topic is covered in Chapter 15. A more in-depth description of regression analysis for binary and categorical dependent variables can be found in Long and Freese (2006) and Greene (2018).

14.2 WHEN TO USE LOGIT OR PROBIT ANALYSIS

Table 14.1 shows examples of research questions from different fields where the dependent variable is binary. Each row gives a research question, the corresponding null hypothesis, the binary dependent variable, and one or more independent variables. The table demonstrates that binary dependent

variables are common in empirical research, highlighting the importance of statistical tools for analyzing data of this type.

TABLE 14.1 ■ Examples of Research Questions with Binary Dependent Variables

Field	Research Question	Null Hypothesis	Binary Dependent Variable	Independent Variables
Criminal justice	Does job counseling reduce the probability of ex-convicts being arrested within a year of release?	Job counseling has no effect on the rearrest rate.	Whether or not an ex-convict is rearrested within a year of release	Job counseling and personal characteristics
Economics	Is the likelihood of being employed affected by the level of education?	Level of education has no effect on the probability of being employed.	Whether or not an individual is employed	Level of education and other individual and community characteristics
Political science	Are voters who live near a polling station more likely to vote?	Distance to polling station has no effect on probability of voting.	Whether or not an individual voted in a recent election	Distance to polling station and other voter characteristics
Psychology	What factors affect an individual's likelihood of completing a 4-week therapy session?	Personal characteristics do not affect likelihood of completing therapy session.	Whether or not patients complete the therapy session	Age, sex, education, and other personal characteristics
Public health	How does the proportion of children vaccinated in a county affect the likelihood of a whooping cough outbreak over 1 year?	Proportion of children vaccinated has no effect on the probability of a whooping cough outbreak.	Whether or not there is a whooping cough outbreak in the county	Proportion of children vaccinated, demographic characteristics, and indicators of access to health care
Sociology	Is the decision to attend church affected by attendance by neighbors?	Church attendance is not affected by attendance by one's neighbors.	Whether or not a person attends church regularly	Church attendance by neighbors and other personal and social factors

Field	Research Question	Null Hypothesis	Binary Dependent Variable	Independent Variables
Criminal justice	Does job counseling reduce the probability of ex-convicts being arrested within a year of release?	Job counseling has no effect on the rearrest rate.	Whether or not an ex-convict is rearrested within a year of release	Job counseling and personal characteristics
Economics	Is the likelihood of being employed affected by the level of education?	Level of education has no effect on the probability of being employed.	Whether or not an individual is employed	Level of education and other individual and community characteristics

Field	Research Question	Null Hypothesis	Binary Dependent Variable	Independent Variables
Political science	Are voters who live near a polling station more likely to vote?	Distance to polling station has no effect on probability of voting.	Whether or not an individual voted in a recent election	Distance to polling station and other voter characteristics
Psychology	What factors affect an individual's likelihood of completing a 4-week therapy session?	Personal characteristics do not affect likelihood of completing therapy session.	Whether or not patients complete the therapy session	Age, sex, education, and other personal characteristics
Public health	How does the proportion of children vaccinated in a county affect the likelihood of a whooping cough outbreak over 1 year?	Proportion of children vaccinated has no effect on the probability of a whooping cough outbreak.	Whether or not there is a whooping cough outbreak in the county	Proportion of children vaccinated, demographic characteristics, and indicators of access to health care
Sociology	Is the decision to attend church affected by attendance by neighbors?	Church attendance is not affected by attendance by one's neighbors.	Whether or not a person attends church regularly	Church attendance by neighbors and other personal and social factors

One option is to simply apply the linear OLS model described in Chapters 12 and 13 to the case where the dependent variable is binary. This is called the linear probability model (LPM). Perhaps the main advantage of the LPM is that it is easy to interpret the coefficient(s). Each coefficient in the LPM represents the effect of a one-unit increase in the corresponding independent variable on the probability that $y = 1$. Thus, if $\beta = 0.02$, then each one-unit increase in the independent variable is associated with a 0.02 or 2 percentage point increase in the probability that $y = 1$.

The main disadvantage of the LPM is that it will generate predicted values of the dependent variable that are less than 0 or greater than 1 over some ranges of x, which are not valid as probabilities. To demonstrate this, Figure 14.1 shows 20 observations, where x takes the values between 1 and 20 and y is either 0 or 1. The LPM model generates the straight line that best fits the data, which is also shown in the figure. On the left of the graph, the LPM-predicted value of y dips below 0, while on the right, it rises above 1. In other words, the LPM model is predicting probabilities outside the 0-to-1 range.

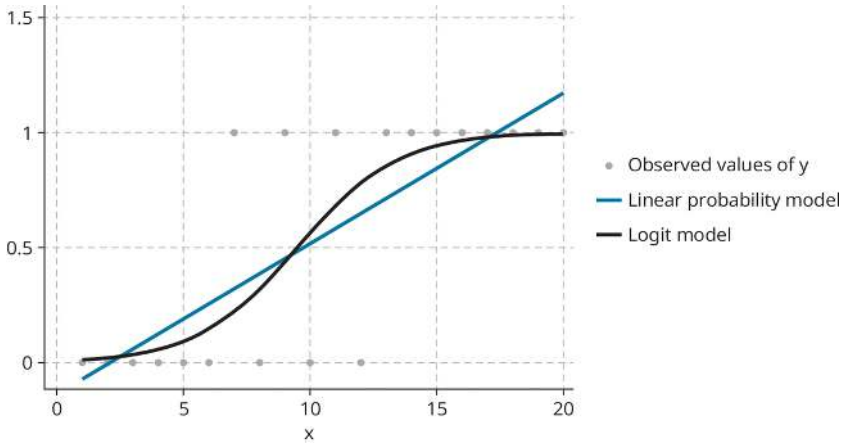


Figure 14.1 Linear Probability Model and Logit Model

On the other hand, if we apply a logit model to the same data, the predicted value of *y* is a curved line that remains greater than 0 and less than 1 throughout the range of *x* (Figure 14.1). This is only possible because the relationship between the predicted probability (*P*) and independent variable (*x*) is nonlinear. This means that the **marginal effect** (the slope on the graph) varies across observations. However, as discussed in Section 14.5, Stata can be used to calculate the marginal effects of each independent variable on the predicted probability in a logit model.

14.3 UNDERSTANDING THE LOGIT MODEL

The logit model is based on the concept of the odds, defined as the probability of an event occurring (*P*) divided by the probability of the event not occurring. Mathematically, we can express odds as follows:

$$Odds = \frac{P}{1 - P} \tag{14.1}$$

$$Odds = \frac{P}{1 - P}$$

Odds are commonly used to describe payoffs in sports gambling, but they also reflect the perceived probability of winning. For example, if a racetrack offers 3-to-1 odds on a horse, this means they will pay out three times the value of the bet if the horse wins. This implies that the perceived probability that the horse will lose is three times greater than the perceived probability that it will win.[2] In other words, the horse has a 3/(3 + 1) = 0.75 = 75% probability of losing and a 25% probability of winning.

The logit regression model expresses the natural logarithm of the odds (sometimes called the *log odds*) as a linear function of a constant and a set of independent variables:

$$In\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{k-1}\beta_i x_i$$

(14.2)

$$In\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{k-1}\beta_i x_i$$

This equation can be rewritten in terms of *P* as follows:

$$P = \frac{exp\left(\beta_0 + \sum\beta_i x_i\right)}{1 + exp\left(\beta_0 + \sum\beta_i x_i\right)}$$

(14.3)

$$P = \frac{exp(\beta_0 + \sum\beta_i x_i)}{1 + exp(\beta_0 + \sum\beta_i x_i)}$$

Because the logit regression model is nonlinear, it cannot be estimated using ordinary least squares (OLS). OLS involves a set of calculations using matrix algebra that will always generate the estimated coefficients and related statistics. In contrast, running a logit model involves maximum likelihood estimation (MLE), which uses an iterative search process to find the set of coefficients that maximizes the probability of generating the observed data. Fortunately, the calculations behind the search procedure are carried out by Stata.

Because the search procedure is computationally intensive, it takes somewhat more time to run a logit model than an OLS model. In addition, the procedure will occasionally fail to converge, meaning that it cannot “find” a set of coefficients that is better than other sets of coefficients. In graphic terms, this means that the likelihood function is “flat” over some range of coefficient estimates, making it impossible to identify a point of maximum likelihood.

The logit model relies on a set of assumptions similar to those behind OLS regression. Below are some of the more important ones:

The dependent variable takes just two possible values (0, 1).

The independent variables are measured without error.

The log odds of the event (*y* = 1) are a linear function of the independent variables.

The model includes all relevant variables.

There is no correlation between the independent variables and the error term.

14.4 RUNNING A LOGIT MODEL

Figure 14.1 demonstrates logit regression using a small hypothetical example. Now let's consider a larger, more complex, and concrete example from the 2021 General Social Survey (GSS). One of the questions asked by the GSS was this:

There has been a lot of discussion about the world's climate and the idea it has been changing in recent decades. Which of the following statements comes closest to your opinion?

1. Has not been changing.
2. Has been changing mostly due to natural processes.
3. Has been changing about equally due to natural processes and human activity.
4. Has been changing mostly due to human activity.

The results are found in the variable "clmtcaus." Suppose we are interested in the factors associated with giving the fourth response: "Has been changing mostly due to human activity." We can use the **recode** command to create a new binary variable "humcaus" equal to 1 if the respondent gives the fourth answer and 0 otherwise. Then we can recode the sex variable (1=male, 2=female) to represent a dummy variable for female respondents (0=male, 1=female). Finally, we can use **logit** command to see whether believing in human-caused climate change is associated with sex, age, or education. Using the menu system, we can run a logit model using the following: Statistics → Binary outcomes → Logistic regression, then select the dependent and independent variables. The Stata command is **logit humcaus female age educ**. Figure 14.2 shows the commands along with the output.

```
. recode clmtcaus (1/3=0) (4=1), gen(humcaus)
(1,770 differences between clmtcaus and humcaus)

. recode sex (1=0) (2=1), gen(female)
(3,940 differences between sex and female)

. logit humcaus age female educ

Iteration 0:  Log likelihood = -1155.5891
Iteration 1:  Log likelihood = -1116.9191
Iteration 2:  Log likelihood = -1116.8295
Iteration 3:  Log likelihood = -1116.8295

Logistic regression                                     Number of obs =  1,668
                                                        LR chi2(3)    =  77.52
                                                        Prob > chi2   = 0.0000
Log likelihood = -1116.8295                             Pseudo R2     = 0.0335

------------------------------------------------------------------------------
     humcaus | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -.0166404   .0029734    -5.60   0.000    -.0224682   -.0108126
      female |   .1460138   .1013562     1.44   0.150    -.0526407    .3446683
        educ |   .1223196   .0184335     6.64   0.000     .0861906    .1584485
       _cons |  -1.088241   .3250257    -3.35   0.001     -1.72528   -.4512025
------------------------------------------------------------------------------
```

Figure 14.2 Logit Model of Belief in Human-Caused Climate Change

The results indicate that the coefficient on age is negative and statistically significant, the coefficient on the female dummy variable is not statistically significant, and the coefficient on education is positive and statistically significant. In other words, belief that climate change is mainly caused by human activity is more common among young people and among people with more education, but men and women hold similar beliefs.

14.5 INTERPRETING THE RESULTS OF A LOGIT MODEL

How do we interpret the coefficients? With an OLS model, each coefficient represents the slope of the line—that is, the change in the dependent variable associated with a one-unit increase in the independent variable. But the logit function is nonlinear, so the slope changes depending on the value of the independent variable(s). One option is to calculate the marginal effect of an independent variable as $\beta(1 - P)P$, where β is the coefficient on that independent variable and P is the predicted probability that $y = 1$. From this equation, we know that the marginal effect is close to 0 when P is close to 0 or close to 1. The maximum value of $P(1 - P)$ is when $P = 0.5$, so the largest marginal effect is $\beta_i(0.5)(1 - 0.5) = \beta_i(0.25)$.

Alternatively (and more easily), we can have Stata calculate the marginal effect for us. If we want to calculate the average marginal effects of the education variable, for example, we use the command **margins, dydx(educ)**. As shown in Figure 14.3, the result is 0.02917, meaning that *on average*, each additional year of education is associated with roughly a 3 percentage point increase in the probability the respondent will agree that climate change is mainly caused by human activity.

```
. margins, dydx(educ)

Average marginal effects                                  Number of obs = 1,668
Model VCE: OIM

Expression: Pr(humcaus), predict()
dy/dx wrt:  educ
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
educ	.02917	.0041684	7.00	0.000	.0210001	.0373399

```
. margins, at(educ=(10(5)20))

Predictive margins                                        Number of obs = 1,668
Model VCE: OIM

Expression: Pr(humcaus), predict()
1._at: educ = 10
2._at: educ = 15
3._at: educ = 20
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	.3430206	.0232237	14.77	0.000	.297503	.3885383
2	.4873591	.0122656	39.73	0.000	.4633189	.5113992
3	.6337875	.0237648	26.67	0.000	.5872095	.6803656

Figure 14.3 Marginal Effects and Prediction for a Logit Model

The **margins** command without **dydx** option will give the average predicted probability for different values of an explanatory variable. For example, the command

```
margins, at(educ=10)
```

will give the average probability for respondents with 10 years of education. Similarly, the command

```
margins, at(educ=(10(5)20))
```

will produce a small table with the average probability for respondents with 10, 15, and 20 years of education (see Figure 14.3).

It is important to keep in mind several limitations of logit and probit models. First, these models are relatively sensitive to heteroscedasticity. In OLS regression, heteroscedasticity does not result in biased coefficients, but with probit and logit, it can cause the estimated coefficients to be biased.

Second, because logit models are estimated with an iterative search process, they require larger samples to achieve the same level of accuracy in estimating coefficients. A common rule of thumb is that the sample size should be at least $10k/p$, where k is the number of independent variables and p is the probability of the less likely outcome of the dependent variable. For example, if we have six independent variables and the dependent variable is zero 90% of the time, the sample size should be at least 10 × 6 / 0.1 = 600 (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996).

Finally, probit and logit models have a “zero cell” problem. Each 2 × 2 cross tabulation between the (binary) dependent variable and an independent dummy variable must have observations in all four cells of the table. Otherwise, the model cannot be estimated.

14.6 LOGIT VERSUS PROBIT REGRESSION MODELS

Another option for carrying out regression analysis with a binary dependent variable is the probit model, which uses the command **probit**. It is similar to the logit model in that both describe a function that looks like an elongated S and whose value always remains between 0 and 1. The probit function is different from the logit function: Instead of being based on the log odds, it is based on the cumulative normal probability function, denoted by Φ():

$$P = \Phi\left(\beta_0 + \sum_{i=1}^{k-1} \beta_i x_i\right)$$

(14.4)

$$P = \Phi\left(\beta_0 + \sum_{i=1}^{k-1} \beta_i x_i\right)$$

Although the equations for logit and probit look quite different, in practice, the results are almost identical. Figure 14.4 combines the logit estimation from Figure 14.1 with a probit estimation using the same data. The predicted values of y (probabilities) are virtually the same.

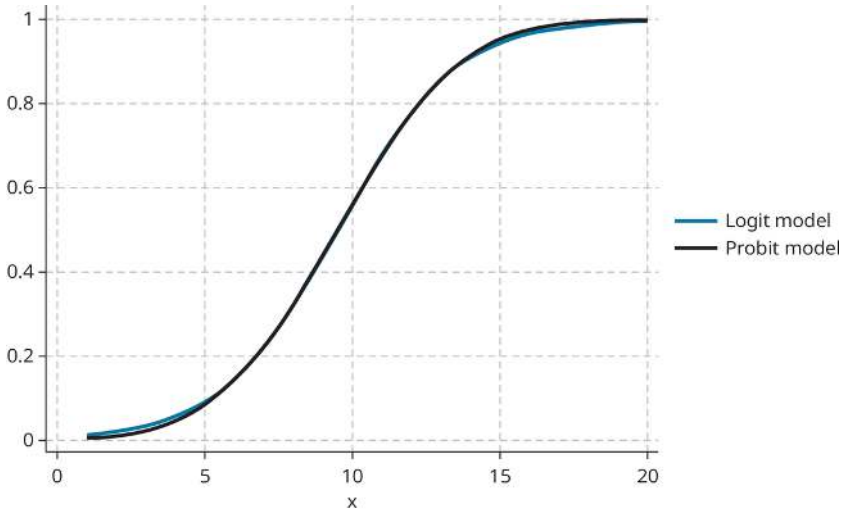


Figure 14.4 Comparison of Probit and Logit Models

Because of the similarity of results, it is not worth devoting time to determining which model gives a "better" fit. The logit model used to have an advantage because it is computationally simpler, but modern computers make this difference moot. In fields that are familiar with odds ratios, such as health and nutrition, the logit function is more common. In other fields, such as economics and political science, the probit model is the default model for regression analysis with binary dependent variables.

14.7 PRESENTING THE RESULTS

As discussed in Chapter 12, the write-up should focus on not only the statistical significance of the independent variables but also the size of the effect. It is possible to have a variable whose effect is statistically significant at the usual levels of confidence but too small to make much difference in practice. This is particularly true when analyzing databases with a large number of observations.

For a general audience, it is not necessary to provide a table of the regression results. Instead, we focus on the sign of the statistically significant results. Describing the marginal effects of the independent variables is optional. The results of the logit model could be summarized as follows:

What factors are associated with believing that human activity is the main cause of climate change in the United States? The results of the 2021 General Social Survey provide some answers. Overall, about half (48%) of respondents believe that human activity is the main cause of climate change. This belief is more common among younger people and those with more education. However, there is no significant difference between men and women in their views on the cause of climate change.

For an academic journal or a technical audience, more detail on the methods and results should be provided. Each field and each journal has different guidelines regarding the presentation of coefficients, *p*-values, and/or confidence intervals. Here is an example:

We used the 2021 General Social Survey to explore the socioeconomic correlates of views on climate change. We use a logit model and data from 1,668 respondents to estimate the effect

of age, gender, and education on the belief that human activity is the main cause of climate change. The results indicate that age has a negative and statistically significant effect on belief in human-caused climate change. On average, each additional year of age is associated with a reduction of 0.00397 in the probability ($p < 0.000$). On the other hand, education is positively and significantly related to belief in human-caused climate change, with each additional year of education being associated with a 0.0292 increase in the probability ($p < 0.000$). On the other hand, there is no statistically significant difference between men and women in the belief that human activity is the main cause of climate change ($p = 0.150$).

An academic article will also include a table showing the results of the regression analysis. As discussed in Chapter 12, the results of any regression analysis can be exported using the **etable** command. This command allows you to export the results of the most recent regression model to Word, Excel, LaTeX, or other file types. You can also use the command to specify the font, format, and layout of the results.

Figure 14.5 gives an example of an **etable** command along with some common options. The first four commands assign labels to the variables so the table will be easier for readers to understand. The **etable** command is spread over four lines; the "///" notation indicates that the command continues on the next line. The indentation is optional but helps make the code more readable. The **export** option is used to specify the title and format of the file to be created, and **replace** tells Stata to delete any previous content of the file. The **title** and **note** options allow the user to specify the title of the table and a footnote, respectively. The **showstars** option indicates that the output should use asterisks to identify the level of statistical significance of each independent variable. By convention, one asterisk is used to indicate $p < 0.10$, two asterisks are for $p < 0.05$, and three asterisks for $p < 0.01$. The **showstarsnote** adds a footnote to explain the meaning of the asterisks, and **col(dvlabel)** says that the column heading should include the label of the dependent variable ("Human-caused climate change"). Otherwise, the column heading will show the (rather uninformative) variable name "humcaus."

```
lab var humcaus "Human-caused climate change"
lab var age     "Age (years)"
lab var female  "Female"
lab var educ    "Education (years)"
etable, title(Logistic model of belief in climate change) ///
        note(Source: Analysis of 2021 GSS data) ///
        export(Logistic model.docx, replace) ///
        showstars showstarsnote col(dvlabel)
```

Figure 14.5 Exporting Regression Results

Figure 14.6 shows the contents of the Word file generated by the **etable** command. In this case, the output file gives the results from just one regression model, but the **etable** command allows users to include the results of multiple models, with the results of each model appearing in a separate column. This is often done in journal articles to compare the results of closely related models, such as models that include different sets of independent variables.

Logit model of belief in climate change		
	Human-caused climate change	
Age (years)	−0.017	**
	(0.003)	
Female	0.146	
	(0.101)	
Education (years)	0.122	**
	(0.018)	
Intercept	−1.088	**
	(0.325)	
Number of observations	1668	

** p<.01, * p<.05

Source: Analysis of 2021 GSS data

Figure 14.6 Output Of Logit Regression in Word Format

14.8 SUMMARY OF COMMANDS USED IN THIS CHAPTER

This last section summarizes all the Stata code used in the chapter (Table 14.2). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 14.2 ■ Summary of Commands Used in this Chapter

Function	Stata command(s)
Estimate a logit regression model	logit y x
Estimate a probit regression model	probit y x
Generate predicted values after regression	predict yhat
Graph y and x data with OLS estimate values and logit estimated values	twoway (scatter y x) (line y_OLS x) /// (line y_Logit x)
After non-linear regression, calculate the average marginal effect of an x variable	margins, dydx(educ)
After a non-linear regression, give the predicted value of y for a value of x	margins, at(educ=10)
After a non-linear regression, give the predicted value of y for selected values of x	margins, at(educ=(10(5)20))
Export regression results to a Word file, give it a title and footnote, include asterisks for significant coefficients, add a footnote explaining the asterisks, and use the variable label as a column heading	etable, title(Logit model of...) /// note(Source: Analysis of...) /// export(Logit model.docx, replace) /// showstars showstarsnote col(dvlabel)

Function	Stata command(s)
Estimate a logit regression model	logit y x
Estimate a probit regression model	probit y x
Generate predicted values after regression	predict yhat

Function	Stata command(s)
Graph y and x data with OLS estimate values and logit estimated values	twoway (scatter y x) (line y_OLS x) /// (line y_Logit x)
After non-linear regression, calculate the average marginal effect of an x variable	margins, dydx(educ)
After a non-linear regression, give the predicted value of y for a value of x	margins, at(educ=10)
After a non-linear regression, give the predicted value of y for selected values of x	margins, at(educ= (10(5)20))

Export regression results to a Word file, give it a title and footnote, include asterisks for significant coefficients, add a footnote explaining the asterisks, and use the variable label as a column heading	etable, title(Logit model of…) ///
	
	note(Source: Analysis of…) ///
	
	export(Logit model.docx, replace) ///
	
	showstars showstarsnote col(dvlabel)

EXERCISES

1. As the cost of college tuition rises, many politicians have called for tuition assistance for low-income students to level the playing field. Others have recommended that all community colleges should be free for anyone who wants to attend. On the flip side, some politicians argue against free college and have even called for a tax on tuition waivers and a reduction in state funding of public colleges. Using the 2021 GSS, we can explore the characteristics of those who support financial aid for college students.
    a. In the GSS2021 data set, one statement is, “The government should give financial assistance to college students from low-income families, even if it might require a tax increase to pay for it.” There were five possible responses: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree. Begin by recoding this variable (govfnaid) into two categories. Recode the first two responses as one category (1) and the other three responses as a second category (0).
    b. Run a logit regression with your new 0–1 variable as the dependent variable. The independent variables should include income (realrinc), age (age), sex (sex), and someone’s political affiliation. For the sex variable, generate a new variable in which 1 is female and 0 is male. For

the political affiliation variable (partyid), generate a dummy variable for Democrats and another for Republicans. Democrats include strong Democrat and not strong Democrat. Republicans include strong Republican and not strong Republican. Independents and others will be the omitted reference category. Be sure to examine the numeric codes before you make the new variable.

c. Use the **margins, dydx(*)** command immediately after running your logit regression in question 1b.
d. Write a paragraph for a scholarly journal that would describe the results.

2. In the 2021 GSS, respondents were asked whether they would favor or oppose a law that would require a person to obtain a police permit before he or she could buy a gun (gunlaw). Run a logit regression to examine the characteristics of people who favor or oppose the gun permit law. You can use any variables in the data set that you think would be relevant to opinions on the gun permit law. Then, write a brief report summarizing your findings. This should be three or four paragraphs: An introduction, one or two paragraphs that make up your key points, and a concluding paragraph. You can assume that you are writing this as a short article in *Newsweek* or *The Economist*.

   Hint: To quickly find variables that may be of interest, open the variable manager. In the space in the upper left corner, it says, “enter filter text here.” You can type in anything you are looking for, and it will show you all variables that have those words in the description.

KEY TERMS

binary variable

logit regression

marginal effect

probit regression

15 INTRODUCTION TO ADVANCED TOPICS IN REGRESSION ANALYSIS

CHAPTER PREVIEW

Topic	Description	Sample research question	Stata commands
Multinomial logit or probit (not ordered)	For use when the dependent variable is categorical and has no natural order	What factors predict the marital status of adults?	**mlogit** **mprobit**
Ordered multinomial logit or probit	For use when the dependent variable is categorical and has a natural order	What factors are associated with the degree of happiness reported in a survey?	**ologit** **oprobit**
Instrumental variable regression	For use when one or more explanatory variables may be affected by the dependent variable or a confounding factor	Does higher city budget for the police department translate into lower crime rates?	**ivregress**
Panel data analysis	For use when the dependent and independent variables have both time and cross-section dimensions.	What is the impact of a new program to improve literacy in a city?	**xtreg**
Time-series analysis	For use when the data describe patterns over time	Does the price of wheat in the United States affect the price of wheat in Colombia?	**var** **vec**

Topic	Description	Sample research question	Stata commands
Multinomial logit or probit (not ordered)	For use when the dependent variable is categorical and has no natural order	What factors predict the marital status of adults?	**mlogit** **mprobit**
Ordered multinomial logit or probit	For use when the dependent variable is categorical and has a natural order	What factors are associated with the degree of happiness reported in a survey?	**ologit** **oprobit**

Topic	Description	Sample research question	Stata commands
Instrumental variable regression	For use when one or more explanatory variables may be affected by the dependent variable or a confounding factor	Does higher city budget for the police department translate into lower crime rates?	**ivregress**
Panel data analysis	For use when the dependent and independent variables have both time and cross-section dimensions.	What is the impact of a new program to improve literacy in a city?	**xtreg**
Time-series analysis	For use when the data describe patterns over time	Does the price of wheat in the United States affect the price of wheat in Colombia?	**var** **vec**

15.1 INTRODUCTION

Chapter 12 introduced regression analysis, which estimates the equation that best describes the relationship between a dependent variable and one or more independent variables. We focused on ordinary least squares (OLS) regression in which the dependent variable is continuous and the model is linear in the parameters. Chapter 13 covered various problems that may occur in regression analysis and how to address them. And Chapter 14 described the logit and probit regression models, which are used when the dependent variable is binary.

However, in many cases, we want to estimate a relationship that is different in some way. This chapter provides a brief introduction to four alternative types of regression analysis:

Regression analysis when the dependent variable has multiple categories

Regression analysis when one of the explanatory variables is endogenous

Regression analysis with time-series data

Regression analysis with data that have both time-series and cross-sectional dimensions

These methods are briefly described in the next four sections of this chapter. Each of these is a large topic about which entire books have been written, so we do not attempt to cover them in any depth. Instead, the goal of this chapter is to acquaint the reader with the topic, introduce a few relevant Stata commands, and provide guidance for further reading.

15.2 REGRESSION WITH A CATEGORICAL DEPENDENT VARIABLE

In Chapter 14, we described the logit and probit models, designed for binary dependent variables, but what if the dependent variable is categorical, with three or more categories? In some cases, there is no natural order across the categories. An example is a model to predict the marital status of adults (single, married, divorced, or widowed). In this situation, we can run a multinomial logit model using the **mlogit** command or a multinomial probit model using the **mprobit** command. In other cases, there is a natural order to the categories, an example being different levels of agreement with a statement (agree, neutral,

disagree). When there is a natural order, we use an ordered logit model (**ologit**) or an ordered probit model (**oprobit**).

Suppose we want to use the 2021 GSS to address the question of whether money buys happiness. The variable “happy” gives the responses to the question:

Taken all together, how would you say things are these days—would you say that you are

1. very happy,
2. pretty happy, or
3. not too happy.

There is also a variable “realinc” which gives the real income of the household.

The ordered logit regression analysis can be implemented using the **ologit** command or by using the menu system: Statistics → Ordinal outcomes → Ordered logisitical regresson. The command and the results are shown in Figure 15.1:

```
. ologit happy realinc female age

Iteration 0:  Log likelihood = -3228.3384
Iteration 1:  Log likelihood = -3181.6284
Iteration 2:  Log likelihood =  -3181.389
Iteration 3:  Log likelihood = -3181.3889

Ordered logistic regression                       Number of obs =  3,314
                                                  LR chi2(3)    =  93.90
                                                  Prob > chi2   = 0.0000
Log likelihood = -3181.3889                       Pseudo R2     = 0.0145

------------------------------------------------------------------------------
       happy | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     realinc |  -7.91e-06   8.59e-07    -9.21   0.000    -9.59e-06   -6.22e-06
      female |   .0180778    .068961     0.26   0.793    -.1170833     .153239
         age |  -.0049096   .0020131    -2.44   0.015    -.0088551    -.000964
-------------+----------------------------------------------------------------
       /cut1 |  -2.002207   .1309775                     -2.258918   -1.745496
       /cut2 |   .6922285   .1255487                      .4461577    .9382994
------------------------------------------------------------------------------
```

Description

Figure 15.1 Ordered Logit Model of Happiness

The results from the ordered logit regression give us coefficients for each variable. The coefficient on real income is negative and statisticially significant. Since the dependent variable is coded such that happiness is a low number (1) and unhappiness is a high number (3), the negative coefficient means that higher income is associated with greater happiness. The insignificant coefficient on the female dummy variable indicates that there is no difference in happiness between men and women, after controlling for income and age. And the negative and significant coefficient on age indicates that happiness increases with age. The coefficients in the lower part of the output can be used to calculate the predicted probability of falling into each happiness category, but it is easier to use the **margins** command, as described in Chapter 14.

Of course, we should not take these results too seriously. The pseudo-R^2 indicates that these three independent variables “explain” only a tiny share of the variance in happiness in the sample, implying that there are many other factors that influence happiness, some of which may affect the size and significance of these coefficients.

For more information on using Stata for regression analysis of categorical variables, see Long and Freese (2006).

15.3 INSTRUMENTAL VARIABLES REGRESSION

In Section 13.6, we discussed the problem of **endogeneity**, where one or more explanatory variables are correlated with the unobserved error term, causing bias in the estimated coefficient. One potential solution to endogeneity is instrumental variables. The basic idea is to find one or more variables that (1) are correlated with the endogenous explanatory variable but (2) do not directly influence the dependent variable. These are called instruments. Instrumental variable (IV) regression can be considered a two-stage process. In the first stage, we use regression analysis to estimate the endogenous explanatory variable as a function of the instrument(s). In the second, we regress the dependent variable on the *estimated* value of the endogenous explanatory variable. The coefficient on the *estimated* value of the endogenous explanatory variable will be unbiased if it is a strong instrument.

This may be easier to understand with a concrete example. In Section 13.6, we discussed the example of estimating the effect of the size of the police force on crime rates. Endogeniety is an issue because a city with a high crime rate may expand the police force to address the problem, which is called reverse causation. We need an instrument that is (1) correlated with the size of the police force but (2) does not directly influence the crime rate. Levitt (1997) proposed the size of the fire department as an instrument. It is likely to be correlated with the size of the police department because both reflect the tax base and the willingness of the city government to fund municipal services. But it is not likely to directly influence the crime rate.

In Stata, we can estimate an instrumental variable regression model with the **ivregress** command or by using the menu sequence: Statistics → Endogenous covariates → Linear regression with endogenous covariates. Using the example above, suppose we have a variable “crime” for the crime rate in a sample of 500 cities, a variable “police” for the per capita city spending on the police department, and “fire” for the per capita city spending on fire departments. The command for running this would be:

```
ivregress 2sls crime x1 x2 police(fire), first
```

The **2sls** option indicates the estimation method (2-stage least squares). The first variable (crime) is the dependent variable. Exogenous independent variables are represented by x1 and x2. The phrase “police(fire)” tells Stata that “police” is an endogenous explanatory variable that should be estimated in the first stage using “fire” as the instrument. Finally, the **first** option first tells Stata to show the results of the first-stage regression.

However, finding a good instrument is often difficult. We can test statistically whether the instrument is a good predictor of the endogenous explanatory variable by looking at the goodness-of-fit in the first-stage regression. However, it is not possible to test whether the instrument has an independent effect on the dependent variable. This must be a judgement call based on theory and common sense. Weak instruments will cause biased estimates of the parameters, which may be worse than the bias caused by endogeneity. For this reason, instrumental variables are one strategy for addressing endogeneity, but there are other approaches, as discussed in the next section. For more information on regression analysis using instrumental variables, see Bailey (2020), Greene (2018), and Wooldridge (2019).

15.4 REGRESSION WITH TIME-SERIES DATA

Regression analysis can be a powerful tool for analyzing time-series data—that is, data that describe trends over time. Examples include studying the effect of an advertising campaign on weekly sales of shampoo, analyzing the relationship between the monthly prices of wheat in Chicago and Mexico City, the effect of the North American Free Trade Agreement on exports from the United States to Mexico, evaluating the impact of a literacy program on reading scores, or estimating the effect of mortgage interest rates on housing sales. Time-series data gives us new opportunities to identify causality

because of the time delay between cause and effect. However, time-series regression must be done with care because there are some additional complications that need to be taken into account. Here, we consider two complications that are specific to time-series analysis: autocorrelation and nonstationarity.

15.4.1 Autocorrelation

One of the challenges of time-series relationships is **autocorrelation** (also called serial correlation), where the error terms are correlated over time. A change in an unobserved factor often affects the dependent variable over several periods. When this occurs, the error terms are positive (or negative) for multiple periods. For example, product sales may be above expectations for several months after an unexpected endorsement by a public figure. Or crop prices may be lower than predicted over six months due to good rainfall. Or business investment may be higher than expected over several months due to a wave of optimism among investors. Because of this tendency, most time-series variables show positive autocorrelation, meaning that the error term is positively correlated with previous error terms.

A simple version of autocorrelated error terms with one lag and one explanatory variable can be represented as follows:

$$y_t = \beta_0 + \beta_1 X_t + u_t$$

(15.1)

$$y_t = \beta_0 + \beta_1 X_t + u_t$$

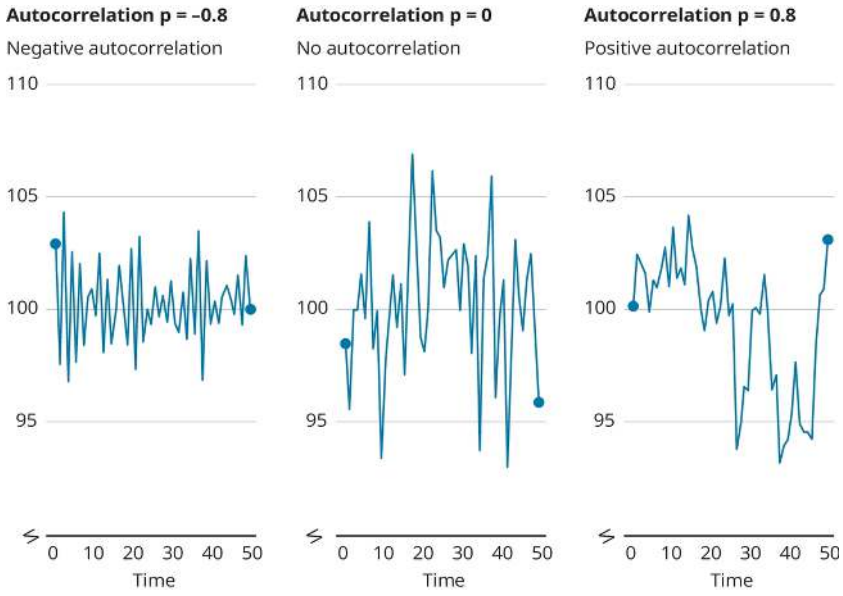
$$u_t = \rho u_{t-1} + \varepsilon_t$$

(15.2)

$$u_t = \rho u_{t-1} + \varepsilon_t$$

The error term in the first equation is u_t, where the subscript *t* refers to the time period. The second equation shows that the error term, u_t, is correlated with its previous value, u_{t-1}. The coefficient ρ (rho) indicates the direction and strength of autocorrelation. For positive autocorrelation, $0 < \rho < 1$, and for negative autocorrelation, $-1 < \rho < 0$. If $\rho = 0$, there is no autocorrelation. And ε_t is an uncorrelated random error term.

What does autocorrelation look like? We can generate hypothetical data setting $\beta_0 = 100$ and $\beta_1 = 0$, so $y_t = 100 + u_t$. We use a random number generator in Stata to create three versions of y_t, one where u_t has no autocorrelation, another where it is positively correlated, and a third which has negative correlation. In Figure 15.2, the left graph shows negative autocorrelation ($\rho = -0.8$), in which a positive residual is more likely to be followed by a negative one and vice versa, resulting in a zigzag pattern in which y_t never gets too far from the central value. The middle graph shows a variable without autocorrelation ($\rho = 0$). It varies around its mean value (100), but the residual ($u_t = y_t - 100$) in each period is unrelated to the residual in previous periods. And the right graph illustrates positive autocorrelation ($\rho = 0.8$), in which the positive (and negative) residuals are clumped together, which allows y_t to "wander" away from the center before eventually returning.



Description

Figure 15.2 Example of Data With and Without Autocorrelation

If we run an ordinary least squares (OLS) regression on data with autocorrelation, the result is similar to that of heteroscedasticity (see Section 13.5): the estimated coefficients are not biased, but the estimates of the standard error are incorrect. Specifically, if there is positive serial correlation (the more common case), then the standard errors will be underestimated, so that a coefficient may appear statistically significant when it is actually not. In addition, OLS estimates are not efficient in that they do not make use of information about the error terms.

We can test for autocorrelation in the residuals with the Durbin-Watson test, which is implemented in Stata with the **estat dwatson** command or with the menu sequence: Statistics → Time series → Tests – Time-series specification tests after regress → Durbin-Watson d test. If there is autocorrelation in the residuals, the remedy is to carry out a Cochrane-Orcutt or Prais-Winsten transformation. In simple terms, the transformation involves three steps: (1) estimate equation 15.1 using OLS, (2) use the residuals from that model to estimate equation 15.2, and (3) with the estimated value of ρ, regress ($y_t - \rho y_{t-1}$) as a function of ($x_t - \rho x_{t-1}$). As usual, we don't have to follow these manual steps because Stata will implement them as part of the **prais** command (or using the menu: Statistics → Time-series → Prais-Winsten regression). The coefficient β_1 estimated from this transformation will be unbiased, but this is not an improvement because the OLS estimate of β_1 is also unbiased. The estimated standard errors of the coefficient will generally be larger than the OLS standard errors (assuming $\rho > 0$), but they will be more accurately measured.

15.4.2 Non-stationarity

Another potential problem in the analysis of time-series data is that the value of the dependent variable depends on previous value(s) of itself, called a dynamic model. A simple version with just one lag and one independent variable can be expressed as follows:

$$y_t = \alpha Y_{t-1} + \beta_0 + \beta_1 X_t + \varepsilon_t$$

(15.3)

$$y_t = \alpha Y_{t-1} + \beta_0 + \beta_1 X_t + \varepsilon_t$$

where α is the coefficient describing the strength of autocorrelation. Generally, α varies between 0 and 1. If it is zero, the lagged dependent variable has no effect and drops out of the model. If it is greater than one, the value of y_t will explode, growing exponentially over time.

If 0 < α < 1, it is a dynamic model, which creates a number of complications in regression analysis. First, a single change in *x* has an effect on *y* that extends over time. In the same period, a one-unit increase in *x* results in a β_1 increase in the value of *y*. But since y_t affects y_{t+1}, which affects y_{t+2}, and so on, the long-run effect of *x* on *y* approaches $\beta_1/(1 - \alpha)$. If α is close to 1, the long-term effect will be much larger than the short-term effect. Another complication is that a dynamic model that also has autocorrelation in the error terms results in biased estimates of the coefficients (unlike the case with a nondynamic model discussed in section 15.4.1). Furthermore, adding a lagged dependent variable to the model when it is not justified can also result in biased coefficients.

If α = 1, y_t becomes a type of nonstationary variable called a *random walk*. The simplest random walk (with no independent variables) is expressed as:

$$y_t = y_{t-1} + \varepsilon_t$$

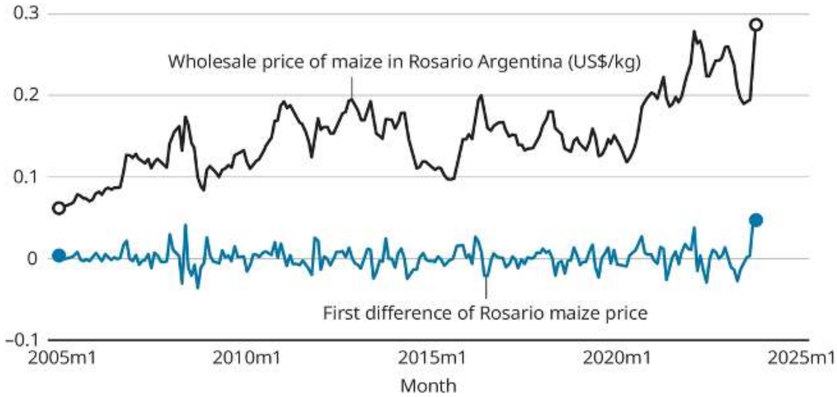
$$y_t = y_{t-1} + \varepsilon_t$$

where ε_t is a standard random normal error term. The value of *y* at time *t* is the previous value of *y* plus a random error term. Nonstationary variables have some peculiar properties. First, as the sample grows, the mean does not converge toward a specific value. The expected value of y_t is simply y_{t-1}. Second, the variance is not constant either but grows in proportion to the sample size. And most surprisingly, a simple OLS regression of two nonstationary variables will often show a “statistically significant” relationship, even if the two variables are unrelated! Granger and Newbold (1974) showed that even if the two variables are independent of each other, an OLS regression analysis will show a (false) statistically significant relationship more than half of the time.

We can test for stationarity with the augmented Dickey–Fuller test, which is implemented in Stata with the **dfuller** command or the menu sequence: Statistics → Time series → Tests → Augmented Dickey-Fuller unit-root test. The null hypothesis of this test is nonstationarity, so if the *p*-value is below 0.05 or 0.01, we reject the null hypothesis and conclude that the variable is stationary.

Nonstationarity is a characteristics of many—if not most—time-series variables, particularly those in economics, such as gross domestic product, the value of exports, the consumer price index, and the price of commodities. Fortunately, even if a variable itself is nonstationary, the first difference of the variable is often stationary. In other words, although the variable y_t is nonstationary, the transformed variable $\Delta y_t = y_t - y_{t-1}$ is often stationary.

This point can be demonstrated with data on the wholesale price of maize (corn) in Rosario, Argentina (see Figure 15.3). The black line represents the level of maize prices, while the blue line is the first difference of the maize price. The price level wanders in a random walk, while the first difference maintains a stable pattern around zero.



Description

Figure 15.3 Price of Maize in Argentina and First Difference

We can use the augmented Dicky–Fuller (ADF) procedure to test these impressions, as shown in Figure 15.4. The upper part shows the ADF test of the price level, “P_maize_Rosario”, where the *p*-value is 0.3653, implying that we cannot reject the null hypothesis that the price level is nonstationary. The lower part gives the ADF test of the first difference of the price, “DP_maize_Rosario,” where the *p*-value of 0.000 indicates that we can reject the null hypothesis of nonstationarity. These results suggest that, as is often the case, the price level is nonstationary but the first difference is stationary.

```
. dfuller P_mz_Rosario, trend

Dickey-Fuller test for unit root          Number of obs  = 226
Variable: P_mz_Rosario                    Number of lags =   0

H0: Random walk with or without drift

                                       Dickey-Fuller
                   Test        ───────── critical value ─────────
              statistic           1%           5%           10%
──────────────────────────────────────────────────────────────────
 Z(t)            -2.427       -3.998       -3.433        -3.133
──────────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.3653.

. dfuller DP_mz_Rosario, trend

Dickey-Fuller test for unit root          Number of obs  = 225
Variable: DP_mz_Rosario                   Number of lags =   0

H0: Random walk with or without drift

                                       Dickey-Fuller
                   Test        ───────── critical value ─────────
              statistic           1%           5%           10%
──────────────────────────────────────────────────────────────────
 Z(t)           -10.663       -3.998       -3.433        -3.133
──────────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.0000.
```

Description

Figure 15.4 Testing for Stationarity

If ADF tests indicate that we are working with a set of stationary variables, one common approach in time-series regression analysis is to use a vector autoregression (VAR) model. Rather than assume that one variable is dependent and the others are independent, a VAR treats all variables the same. It estimates multiple equations simultaneously, each of which consists of one variable as a function of past values of itself and past values of the other variables. The idea is to let the data determine the direction of causality rather than forcing the researcher make assumptions about causality. In Stata, the command **var** implements the vector autoregression model.

On the other hand, if we are working with a set of nonstationary variables, such as the price of maize in various countries, we need a different approach. Although nonstationary variables behave strangely, the relationship between two or more nonstationary variables may be stable. In technical terms, there may be a linear equation between two or more nonstationary variables such that the residual is stationary. This is called a cointegrated relationship, and it can be modeled with a vector error correction (VEC) model. The VEC model is similar to a VAR model of the first differences except that it also includes an expression for the long-run relationship among the levels of the original variables. This method was developed by Engle and Granger (1987) and is widely used in the analysis of prices, macroeconomic variables, and other time-series relationships. Stata has commands for testing for cointegration (**vecrank**) and running the model (**vec**). However, these are large topics and beyond the scope of this chapter. Readers interested in delving into time-series regression analysis may be interested in books by Banerjee et al. (1993) and Becketti (2013). In addition, Stata offers a 987-page manual on the commands for analysis of time-series data (StataCorp, 2023).

15.5 REGRESSION THAT COMBINES CROSS-SECTION AND TIME-SERIES DATA

Regression analysis can also be used to analyze data that combine cross-sectional units (such as households, companies, or countries) with two or more time periods (such as survey rounds, months, or years). Often the data cover the same cross-sectional units over time, such as two rounds of a survey carried out on the same households. This is called panel data, and it is particularly useful to researchers looking for evidence of causality between variables.

15.5.1 Panel Data Analysis

Why is panel data so useful? Suppose we are trying to measure the effect of income and meat prices on meat consumption using two rounds of household survey data with different samples (not a panel). And suppose vegetarians are more common in urban areas, where meat prices tend to be higher. Thus, urban households consume less meat per capita partly because prices are higher and partly because some are vegetarian. Without data on vegetarianism, the price variable will capture both effects. In the terminology of regression analysis, an unobserved variable (vegetarianism) affects the dependent variable (meat consumption) and is correlated with an independent variable of interest (price), resulting in a biased estimate of the coefficient on the variable of interest (prices).

Now suppose the household survey was carried out twice, interviewing the *same* households in each round, resulting in a panel dataset. One type of panel data regression analysis, called fixed-effects analysis, measures only the relationship between dependent and independent variables within each cross-sectional unit over time. In other words, it would estimate the effect of changes in meat prices on meat consumption between the two rounds of the survey. All household-specific characteristics, including taste, preferences, religious beliefs, dietary restrictions, and vegetarianism, would be controlled, so the coefficient on price would not be influenced by the fact that vegetarians live in areas with high meat prices.

How does fixed-effect regression work? Mathematically, it is equivalent to adding a dummy variable for each household in the sample. It is also equivalent to calculating the deviation of each observation from the mean ($y^* = y_{it} - \bar{y_i}$ and $x^* = x_{it} - \bar{x_i}$) and then regressing y* as a function of x*.

Fortunately, these calculations can be done by Stata as part of the **xtreg** command. We first need to tell Stata the variables that identify the cross-section units and the time units with the **xtset** command. Then, we run the **xtreg** command with the **fe** option for fixed effects. We can include other explanatory variables in the regression, such as per capita income and household size:

```
xtset hhid round
xtreg meatcon meatprice pcinc hhsize, fe
```

In this case, the coefficient on “meatprice” would give us a reasonable estimate of the impact of changes in the price on meat consumption without any bias caused by vegetarianism.

Panel data can also help address problems of endogeneity caused by reverse causation. Recall the example from Section 15.3 where we wanted to study the effect of the size of the police force on crime rates. The problem of reverse causation could be addressed with panel data, such as a database of crime rates and the size of police force for 500 counties over 10 years. We could estimate the crime rate as a function of the size of the police force in *previous* year(s) and other factors. By using a time lag, we could reduce the risk that we are measuring the reverse relationship—that is, the effect of crime rates on the size of the police force.

A common alternative to the fixed effect estimator (**fe**) is the the random-effects estimator (**re**), in which the coefficients are based on a weighted average of the time-series pattern (also called the within estimator) and the cross-section pattern (also called the between estimator). There are also Stata commands for analyzing panel data with instrumental variables, logit, probit, and other types of models.

15.5.2 Difference-in-Difference Analysis

One common type of analysis of combined cross-section and time-series data is the difference-in-difference approach. It is similar to the fixed-effect regression analysis described above except that the independent variable of interest is typically binary. Difference-in-difference analysis is often used to study the impact of a program using two time periods and two cross-sectional groups: a treatment group and a control group. Suppose we are interested in the effect of a literacy program on reading scores. We give a literacy test to students before and after the program is launched, and the second round of testing includes some who participated in the program (the treatment group) and some who did not (the control group). To calculate the impact of the program, we calculate the change in literacy in the treatment group minus the change in literacy in the control group. The average treatment effect on the treated group (ATET) is calculated as follows:

$$ATET = \left(\overline{y}_{T2} - \overline{y}_{T1}\right) - \left(\overline{y}_{C2} - \overline{y}_{C1}\right)$$

$$ATET = (\overline{y}_{T2} - \overline{y}_{T1}) - (\overline{y}_{C2} - \overline{y}_{C1})$$

where $\overline{y}$ refers to the average test score, subscripts *T* and *C* refer to treatment and control groups, and subscripts 1 and 2 refer to the testing before and after the program is implemented. This expression measures the increase in reading scores in the treatment group minus the increase in scores among the control group, hence the difference-in-difference label.

We could do these calculations with the command **table treat round2, stat(mean score)**, where “treat” is a dummy variable for the treatment group, “round2” is a dummy variable for the second round (after the program), and “score” represents the reading test results. However, there are two important advantages of using regression analysis. First, the regression analysis will generate tests of statistical significance, so we can say whether the impact of the literacy program is statistically significant. Second, we can add other independent variables to control for other factors, such as age, sex, and whether the person is a native English speaker. If we calculate an interaction term that is 1 for observations of the treatment group and in the second round of surveys, then we can run the regression as follows:

```
gen treat_r2 = treat*round2
regress score treat round2 treat_r2 age sex nativeEng, noconstant
```

The coefficient on the variable “treat” tells us the difference between treatment group scores and control group scores in the first round of the survey. Ideally, this is close to zero, indicating that our two groups are similar. The coefficient on the variable “round2” tells us the increase in reading scores in the second round among the control group. We expect this to be positive given that scores tend to rise with age and schooling. And the coefficient on the interaction term “treat_r2” tells us the increase in treatment group scores minus the increase in control group scores, that is, the impact of the literacy program on reading scores. This is equivalent to the average treatment effect on the treated (ATET), described above. If this coefficient is positive and statistically significant, it suggests that the literacy program was successful.

If we have panel data (the same sample of students in Round 1 and Round 2), we can use the **xtreg** command with the fixed-effect option, which will provide a better estimate of the impact of the literacy program because it will control for all time-invariant characteristics of each student. There are also

specialized Stata commands for difference-in-difference regression which provide additional options (**didregress** and **xtdidregress**).

15.5.3 Randomized Controlled Trial

The difference-in-difference strategy is a good way to measure the impact of an intervention, but it still relies on some assumptions. Continuing with the example of the literacy program, it assumes that nonparticipants will gain as much from the intervention as participants did. In practice, this type of study usually relies on schools that agree to participate in the program and students who volunteer to sign up. It is possible that students who volunteer for the literacy program are smarter or more hardworking than those who do not. In this case, participants might benefit more than nonparticipants would have, which would overstate the impact of the program if it were scaled up to include all students. Alternatively, the schools may encourage underperforming students to enroll, which may understate the impact.

The gold standard of measuring the impact of an intervention is the randomized controlled trial (RCT). Like the previous example, there is an intervention group and a control group, with two (or more) rounds of data collection on each. The RCT differs because it uses randomization to decide who is in the treatment group and who is in the control group. In the study of the literacy program, researchers would start with a list of eligible students and allocate them into the two groups using a random-number generator. If the sample is large enough, it would essentially eliminate the risk that one group will benefit more from the program than the other group.

Similarly, the study of the impact of the size of the police force on crime rates could carried out as an RCT by providing funding to (say) 100 randomly selected counties to increase the police force by 10%, while randomly selecting another 100 counties that would not get any additional funding. Fixed-effect regression analysis would compare the change in crime in the two groups of counties.

Some versions of randomized trials have been used in medical research since the 18th century, when it was used to study treatments for scurvy among British sailors. The use of RCTs in the social science research is more recent, but it has grown rapidly in the past few decades. Although RCTs are considered the gold standard for measuring causal effects, they can be costly and cannot be applied to some research topics for reasons of scale (e.g., the effect of trade policy) or ethical considerations (e.g., the effect of length of prison sentence). In the example above, providing enough funding for 100 counties to expand their police force would be quite costly.

This section has provided a brief overview of some methods for addressing the problem of endogeneity, but a detailed discussion is beyond the scope of this book. Nonetheless, it is important to be aware of the issue of endogeneity. With OLS regression in a nonexperimental setting, we must be very cautious in inferring that a relationship as causal. To do so, we must have strong reasons to believe that there is no reverse causality and that there are no confounding variables that influence both the dependent and the independent variables. Bailey (2016) describes numerous topics in regression analysis with particular emphasis on the problem of—and solutions to—endogeneity.

15.6 SUMMARY OF COMMANDS USED IN THIS CHAPTER

This last section summarizes all of the Stata code used in the chapter (Table 15.1). In addition, all Stata code used throughout the book is summarized in Appendix 1.

TABLE 15.1 ■ Summary of Commands Used in this Chapter

Function	Stata command(s)
Ordered logit model	ologit happy realinc age female
Instrumental variable regression model	ivregress 2sls crime x1 x2 police(fire), first
Test for autocorrelation	estat dwatson
Regression with Prais-Winsten transformation to address autocorrelation	prais y x1 x2
Test variable for stationarity	dfuller x1
Test for cointegration of two or more variables	vecrank x1 x2
Test for appropriate lag length	varsoc x1 x2
Run vector error-correction model	vec x1 x2
Set up panel data analysis	xtset hhid round
Run panel data analysis	xtreg meatcon meatprice pcinc hhsize, fe
Difference-in-difference with non-panel data	regress score treat round2 treat_r2
Difference-in-difference with panel data	xtreg score treat, fe

Function	Stata command(s)
Ordered logit model	ologit happy realinc age female
Instrumental variable regression model	ivregress 2sls crime x1 x2 police(fire), first
Test for autocorrelation	estat dwatson
Regression with Prais-Winsten transformation to address autocorrelation	prais y x1 x2
Test variable for stationarity	dfuller x1
Test for cointegration of two or more variables	vecrank x1 x2
Test for appropriate lag length	varsoc x1 x2
Run vector error-correction model	vec x1 x2
Set up panel data analysis	xtset hhid round
Run panel data analysis	xtreg meatcon meatprice pcinc hhsize, fe
Difference-in-difference with non-panel data	regress score treat round2 treat_r2
Difference-in-difference with panel data	xtreg score treat, fe

EXERCISES

1. Suppose you are studying the effect of youth sports programs on school attendance among teenagers. However, you realize that there is an endogeneity problem because the income of the parents could affect both their child's participation in the sport program and their child's likelihood of attending classes. To address the endogeneity problem, you decide to use distance to a youth program as an instrument for participation in a youth program.
   a. What are the two requirements for a good instrument? Which of the two can be tested, and which cannot?
   b. Assuming that you have a dataset with variables for participation in a youth program ("youthprog"), school attendance ("attendance"), and distance to a youth sports program ("distance"), what command(s) would you use to run an instrumental variable regression to

estimate income as a function of education. Assume all three variables are continuous variables.

2. Based on your analysis in Question 1, you decide that the distance to the youth program is not a good instrument for participation in the youth program. You decide a difference-in-difference model using panel data is the best way to determine whether the youth sports program is having an effect on school attendance.
   a. Describe briefly how you would collect the panel data. Explain why a control group is necessary. Explain why two rounds of data collection are needed.
   b. What Stata command would you use to analyze the panel data, using the variable names given in Question 1?
   c. How would you change the design of the study if you decided to use a randomized controlled trial to evaluate the impact of the youth sports program on school attendance?
3. You are studying the statistical properties of maize prices in Latin America using the data file “FAO maize prices.” Before starting the time-series analysis of the prices, you want to know whether they are stationary or not.
   a. What test will tell you whether a time-series variable is stationary or not?
   b. Test the price of maizes in Bogota, Colombia, and Mexico City for stationarity. What is the *p*-value of each, and what do you conclude about the stationarity of each price?
   c. Calculate a new variable of the first difference of these two prices, and test these for stationarity. What do you conclude about the stationarity of the first differences of the two prices?

Descriptions of Images and Figures

Back to Figure

In the output, the dependent variable is "happy," and the independent variables are "realinc", "female," and "age." The model includes 3,314 observations, with a log likelihood of -3181.3889 and an LR chi2(3) of 93.90. The pseudo R-squared is 0.0145.

The table shows coefficients, standard errors, z-scores, p-values, and 95% confidence intervals for the predictors:

realinc: Coefficient = -7.91e-06, Std. Err. = 8.59e-07, z = -9.21, p-value = 0.000, with a 95% confidence interval of [-9.59e-06, -6.22e-06].

female: Coefficient = 0.0180778, Std. Err. = 0.068961, z = 0.26, p-value = 0.793, with a 95% confidence interval of [-0.1170833, 0.153239].

age: Coefficient = -0.0049096, Std. Err. = 0.0020131, z = -2.44, p-value = 0.015, with a 95% confidence interval of [-0.0088551, -0.000964].

Additionally, there are two cut-off values:

/cut1: Coefficient = -2.002207, Std. Err. = 1.309775, with a 95% confidence interval of [-2.258918, -.153239].

/cut2: Coefficient = 0.6922285, Std. Err. = 1.255487, with a 95% confidence interval of [.4461577, .9382994].

Back to Figure

First graph: Autocorrelation p = –0.8 (Negative autocorrelation)

The x-axis represents "Time" ranging from 0 to 50, and the y-axis represents values between 95 and 110.

The line follows a zigzag (up-and-down) pattern, starting around (0, 103) and ending around (50, 100).

Second graph: Autocorrelation p = 0 (No autocorrelation)

The x-axis represents "Time" ranging from 0 to 50, and the y-axis represents values between 95 and 110.

The line follows a random up-and-down pattern, starting around (0, 98) and ending around (50, 96).

Third graph: Autocorrelation p = 0.8 (Positive autocorrelation)

The x-axis represents "Time" ranging from 0 to 50, and the y-axis represents values between 95 and 110.

The line follows a pattern where it initially increase, then decreases, then again increases, starting around (0, 100) and ending around (50, 103).

Back to Figure

The x-axis ranges from 2005m1 to 2025m1, and the y-axis ranges from -0.1 to 0.3. The line labeled “Wholesale price of maize in Rosario Argentina (US$/kg)” starts at approximately 0.075 in 2005m1 and rises to around 0.299 in 2025m1. The line labeled “First difference of Rosario maize price” starts at about 0 in 2005m1 and rises to around 0.05 in 2025m1.

Back to Figure

First test:

Command: dfuller P_mz_Rosario, trend

The test was conducted on the variable P_mz_Rosario with 226 observations and 0 lags.

The null hypothesis (Hθ) is: Random walk with or without drift.

The test statistic (Z(t)) is -2.427, with critical values at the 1%, 5%, and 10% levels being -3.998, -3.433, and -3.133, respectively.

The MacKinnon approximate p-value for Z(t) is 0.3653.

Second test:

Command: dfuller DP_mz_Rosario, trend

The test was conducted on the variable DP_mz_Rosario with 225 observations and 0 lags.

The null hypothesis (Hθ) is: Random walk with or without drift.

The test statistic (Z(t)) is -10.663, with critical values at the 1%, 5%, and 10% levels being -3.998, -3.433, and -3.133, respectively.

The MacKinnon approximate p-value for Z(t) is 0.0000.

PART V WRITING A RESEARCH PAPER

16 WRITING A RESEARCH PAPER

CHAPTER PREVIEW

Section	Main Points
Introduction to the research paper	Describe the general topic
	What research has been done and what are the gaps in this literature?
	Define your specific question and how it relates to the gaps or contributes to the literature
	How will you answer your question (method)?
	Overview of results
	Outline of paper
Literature review	Identify key themes in literature related to the research question
	Summarize significant sources within each theme
	Identify remaining gaps in literature that you intend to address
Theory, data, and methods	Sample size, selection method, time period
	Type of analysis
	Expected outcomes based on theory
	Measurement of variables
Results	Restate research questions
	Results related to each research question
	Do results agree or disagree with literature?
	Recommendations based on the results
	Limitations of the study
	Areas for future research
Conclusions	Summarize the key findings
	Why are the results important?

Section	Main Points

Section	Main Points
Introduction to the research paper	Describe the general topic What research has been done and what are the gaps in this literature? Define your specific question and how it relates to the gaps or contributes to the literature How will you answer your question (method)? Overview of results Outline of paper
Literature review	Identify key themes in literature related to the research question Summarize significant sources within each theme Identify remaining gaps in literature that you intend to address
Theory, data, and methods	Sample size, selection method, time period Type of analysis Expected outcomes based on theory Measurement of variables
Results	Restate research questions Results related to each research question
	Do results agree or disagree with literature? Recommendations based on the results Limitations of the study Areas for future research
Conclusions	Summarize the key findings Why are the results important?

16.1 INTRODUCTION

As described in Chapter 1, the research process begins with reading the literature, identifying gaps in the literature, and defining your research question. Once you have defined your research question, there are six parts to a typical research paper. This may vary depending on the journal or type of publication, but in general, all of the components listed previously will be included in a research paper. We will review each of these six parts.

16.2 INTRODUCTION SECTION OF A RESEARCH PAPER

The introduction to a journal article or research paper should begin by defining the general topic. In other words, what is the big picture? Why is this research important, or why should the reader be interested? Using a paper by Talan and Kalinkara (2023) “The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course,” we will illustrate each part of the introduction. Their opening paragraph is as follows:

> As scientific knowledge continues to grow exponentially, new ***technological developments*** emerge every day. These technological developments and changes ***have the potential to facilitate, transform, and improve our lives***, and can bring great benefits to the fields in which they are used. In fact, it is difficult to think of an area that is not affected by technology today. ***However, as technology continues to rapidly advance, questions are being raised about how it can be effectively used*** in various fields and what impact it will have. ***One area that has seen significant investment in recent years is education***, with virtual reality, augmented reality, metaverse, blockchain, simulation, mobile technologies, robotics and automation, and online learning environments all being implemented. Among these technological advancements, artificial intelligence (AI) stands out as one of the most successful and widely-used technologies in many sectors.

The bold and italicized parts of the text are used to illustrate that the authors first identify the big picture and why we should be interested (technological developments can transform and improve our lives). They then tell us that there are some concerns (how can it be used effectively).

The second part of the introduction is to illustrate in brief what has already been written about this topic. This is not the full literature review, but only a selection of literature that will support your rationale for choosing the topic. It will also begin to identify gaps in the literature. Talan and Kalinkara (2023) do this in these three paragraphs:

> The primary objective of AI is to enhance the comprehension of human intelligence and to augment machine intelligence to attain maximum benefit from them (Tektaş, Akbaş & Topuz, 2002). AI is a wide-ranging domain that is constantly advancing and leading the way in technological progress (Büyükgöze & Dereli, 2019). Although various research endeavors have been undertaken in AI across multiple fields, its impact on education has also been investigated. AI is applied across several disciplines, including law, science, mathematics, health, engineering, and architecture (Korkmaz & Büyükgöze, 2019; Taşçi & Çelebi, 2020). Research in this field is gaining momentum, and progress is being made through continued research and development activities.
>
> Over the years, AI technologies have evolved and taken various forms. Despite the fact that AI research has a long history, these systems have now become an essential part of human life thanks to the investments made over the years and the widespread use of technological tools such as the internet and smart devices. Currently, there are numerous AI technologies, most of which are still in the research stage. In recent years, AI technologies have been applied in diverse fields, including smart cities, smart watches, robotics, drone systems, defense industry, cybersecurity, and healthcare (Sarica, 2021; Talan, 2021).
>
> ***However, the potential use of AI in education, its contribution to education, and its impact on education are still subjects of debate, with numerous predictions and considerations. While developments in AI offer significant opportunities for the education sector, they also pose a threat at times.*** Thus, AI technology needs to be carefully considered and evaluated in many ways. AI’s potential to be one of the most

> important technologies of the future increases when the potential risks and benefits it offers are evaluated, and the necessary precautions are taken.

In these three paragraphs, Talan and Kalinkara (2023) describe articles written about the advances in AI, but they also begin to identify gaps or limitations in the studies, as illustrated in the italicized sentences.

Once it has been established that there are gaps in the literature, the next step is to define your own research question and determine how it fits into the literature and how it addresses the gaps. This is demonstrated in Talin and Kalirkara's paper as follows:

> Despite the potential benefits of ChatGPT as an educational tool, the full extent of its impact on education remains uncertain and requires further investigation (de Winter, 2023; Qadir, 2022; Zhai, 2022). It is crucial to consider both the potential advantages and risks associated with emerging technologies like ChatGPT in order to anticipate and prepare for the future of education. ***One significant concern is the possibility of students using ChatGPT to cheat, particularly on online exams, due to its ability to generate personalized and authentic responses*** (de Winter, 2023; StokelWalker, 2022; Susnjak, 2022). As online education becomes increasingly widespread, ensuring the validity and reliability of online exams is a critical issue that must be addressed. ***It is important to acknowledge that further research is required to develop effective strategies that mitigate potential risks and leverage the benefits of AI. Furthermore, there is a scarcity of literature that delves into the potential educational use of ChatGPT, a novel tool in this domain***. Hence, investigating the capabilities of this AI agent is anticipated to augment the current body of knowledge. ***This study aims to evaluate the performance of the newly launched ChatGPT on anatomy course examination questions among students enrolled in the Faculty of Health Sciences in Turkey.***

Talan and Kalinkara clearly state the concerns about GhatGPT above while also identifying the lack of literature on the potential educational use of ChatGPT. They then identify their own purpose in the last sentence.

After identifying your research question, the next section of the introduction typically describes the method used to answer the study. The method may include a variety of techniques used to do your research. These include case studies, cost–benefit analyses, regression analyses, surveys, meta-analyses, and forecasting. Some studies are simply an exhaustive literature review. Talan and Kalinkara describe their method as follows:

> This study aimed to compare the performance of ChatGPT with that of health sciences faculty students in answering anatomy course questions. A descriptive study design was employed, and 37 students from a state university in Turkey participated in the study. The students received four weeks of training on a specific anatomy topic and then took a multiple-choice test comprising 40 questions on the covered material. The same test was also given to ChatGPT, and the answers generated were compared with those of the students. The data were analyzed using descriptive statistics, including the number, percentage, and mean.

If your study is a literature review, you should not describe the method as "I used primary and secondary sources." Instead, you should be specific about how you used the sources or what you looked for in the sources. In their abstract below, Allgood, Walstad, and Siegfried (2015) provide an excellent example of

how to describe their method that uses a literature review of research on teaching economics to undergraduates:

> This survey summarizes the main research findings about teaching economics to undergraduates. After briefly reviewing the history of research on undergraduate economic education, it discusses the status of the economics major-numbers and trends, goals, coursework, outcomes, and the principles courses. Some economic theory is used to explain the likely effects of pedagogical decisions of faculty and the learning choices that students make. Major results from empirical research are reviewed from the professor's perspective on such topics as teaching methods, online technology, class size, and textbooks. Studies of student learning are discussed in relation to study time, grades, attendance, math aptitude, and cheating. The last section discusses changes in the composition of faculty who teach undergraduate economics and effects from changes in instructional technology and then presents findings from the research about measuring teaching effectiveness and the value of teacher training. (p. 285)

Following a description of the method, one paragraph is typically used to describe the results or key highlights of the paper. Finally, the last paragraph of the introduction to a journal article or report is often used to indicate what will be included in each section of the paper. This is not essential, and it is not always included in a short journal article. Longer papers or reports, however, typically do have this paragraph. Below is an example from the Allgood et al. (2015) paper on teaching economics to undergraduates:

> The article is divided into eight sections that include this short introduction and a conclusion. In the second section, we briefly review the history of research on teaching college economics to acquaint instructors with the major developments and sources of information on the subject since its origins in the 1960s. In the third section, we survey the landscape of undergraduate economics to describe enrollments, the typical curriculum of courses for undergraduate economics students, and outcomes from the economics major. In the fourth section, we use economic theory to explain the likely effects of decisions that economics professors can make either in structuring or teaching their courses, and also students' decisions about enrolling and participating in economics courses. The purpose of this theory section is to increase the understanding of empirical findings about faculty teaching decisions and selected student behaviors and decisions that are the subjects of the fifth and sixth sections. The seventh section briefly discusses how changes in the characteristics of faculty and changes in technology are likely to affect undergraduate economics instruction in the future, before turning to the issue of teaching effectiveness and examining research on the assessment of instruction from student and faculty perspectives. (p. 286)

16.3 LITERATURE REVIEW

As noted in the previous section, part of the literature review appears in the introduction. It may also be woven into other parts of the paper. For example, results from empirical studies are sometimes compared with the previous literature in the results or discussion section of the paper. The data and methods section may also review methods used in previous studies. In many papers, however, there is a separate section devoted to the literature review. We offer guidelines about this section next.

The purpose of a literature review is to summarize the most significant sources related to a research question. Rather than describing each article or source in its own paragraph like an annotated bibliography, the sources are woven into paragraphs based on themes or main points. The literature

review should identify where the sources agree or disagree and how they relate to the research question of the paper. Overall, like the introduction to the paper, the literature review should convince the reader that your topic is interesting, is important, and fills a gap in the literature.

Students often ask how many sources they should include in the literature review. There is no magic number. If the research question is well-defined, it will help narrow your search. If the question is too broad, then there could be thousands of articles on a topic. Once you have refined your question, you should include significant sources specifically related to your question. “Significant sources” would generally refer to articles that appear in peer-reviewed journals and are cited frequently by other articles or papers. Obviously, if articles are very recent, they won’t have a large number of citations. These should be reviewed to determine if they are relevant to your question.

In the example of a literature review that follows, Enfield (2013) conducted a study in which he “flipped the classroom.” He first reviewed the literature related to the benefits and drawbacks of flipping the classroom and then described the results of his own experiment.

Flipping the classroom involves providing instructional resources for students to use outside of class so that class time is freed up for other instructional activities. The Flipped Classroom Model is described and defended by Mull (2012). While not all of the principles Mull describes are utilized by all teachers who flip their classroom, all implementations include the idea that, “Students prepare for class by watching video, listening to podcasts, reading articles, or contemplating questions that access their prior knowledge.” (para. 3).

Milman (2012) explains, “the idea is that rather than taking up valuable class time for an instructor to introduce a concept (often via lecture), the instructor can create a video lecture, screencast, or podcast that teaches students the concept, freeing up valuable class time for more engaging (and often collaborative) activities typically facilitated by the instructor” (p. 85). Milman goes on to note that formative and summative assessment should be incorporated, as well as meaningful face-to-face learning activities.

Proponents of a Flipped Classroom provide many arguments for engaging students in the content outside of the class to free up time in class for other instructional activities. Milman (2012) identifies what could be considered the primary advantage: increased class time for more engaging instruction. Millard (2012) describes advantages such as increased student engagement, strengthening of team-based skills, personalized student guidance, focused classroom discussion, and creative freedom of faculty while maintaining a standardized curriculum. Fulton (2012) notes that Flipped Classrooms allow students to move at their own pace, access instruction at any time, access expertise from multiple people, benefit from better used classroom time and more.

While many educators who have flipped their classrooms tout the benefits they experienced, there are critics to this approach. Nielsen (2012) discusses concerns with accessibility to instructional resources being provided online, the growing move towards no homework, increased time requirements without improved pedagogy, lack of adapting the classroom environment to reflect the flipped classroom’s ability to support student-centered learning (allowing students to learn at their own pace), and use of lectures to provide instruction with disregard to individual student learning styles. Mull (2012) addresses several of the common concerns which, in addition to some previously mentioned, include teachers’ concerns that their role will be diminished, the students’ experience with the out-of-class instruction will not be interactive, a lack of accountability for students to complete the out-of-class instruction, and the restrictive cost and time needed to produce instructional materials. Milman (2012) also notes several concerns with the Flipped Classroom approach, including poor quality video production, conditions in which the students view the video, inability to monitor comprehension

> and provide just-in-time information when needed, and use with second-language learners or students with learning disabilities.
>
> Given all of the benefits and drawbacks of the approach, it appears that there is a place for the Flipped Classroom Model for at least some instructional contexts. “Although there are many limitations to the flipped classroom strategy and no empirical research exists to substantiate its use, anecdotal reports by many instructors maintain that it can be used as a valuable strategy at any level, depending on one’s learners, resources, and time” (Milman, 2012, p. 86). Milman notes that while the Flipped Classroom approach lends itself well to learning of procedural knowledge, it can also be used for the learning of factual, conceptual, and metacognitive learning. (Enfield, 2013, pp. 14–15)

Overall, Enfield reviews five sources in his literature review. Rather than writing about each of the five articles in separate paragraphs, he weaves them into paragraphs based on the main argument or theme of each paragraph. Note the same authors can appear in multiple paragraphs and on both sides of the argument. Overall, the structure is as follows:

Paragraph 1: Introduces the idea of a flipped classroom as presented by Mull

Paragraph 2: Discusses the benefits of a flipped classroom as presented by Milman

Paragraph 3: Lists other benefits identified by three authors: Milman, Millard, and Fulton

Paragraph 4: Identifies criticisms of a flipped classroom by three authors: Nielsen, Mull, and Milman

Paragraph 5: Acknowledges that there are benefits and drawbacks based on the literature review. Identifies the gap—no empirical research on the effectiveness of the flipped classroom

Another key question when conducting a literature review is when to use a direct quote from the literature and when to paraphrase. Direct quotes are typically used when the original passage is so unique or well stated that it can’t be easily paraphrased. They are also used if they offer a definition for an unusual word or concept. Paraphrasing, on the other hand, is used to summarize or simplify other research. Generally, direct quotes should be limited, while paraphrasing is much more common.[1]

16.4 THEORY, DATA, AND METHODS

The method used in a research paper is typically described briefly in the introduction, as discussed earlier. If the paper is based on some type of empirical research (analysis of data), then the data and methods section will go into much greater detail. It is considered the most important part of the paper since it establishes the validity of the paper. In other words, it allows the audience to judge if the results are valid, how they fit into known theories, and if they can be applied to the general population. Because of its importance, this section should be clear, precise, and detailed enough that the same study could be replicated.

For the type of study that we describe in this book (collecting data and using statistical techniques to analyze the data), the following information would be included in the data and methods section:

Data information

When and where the data were collected

Who collected the data (which organization)

The sampling method and sample size

Limitations or problems with the data

Adjustments to the data and weighting procedures

Method information

Type of analysis (regression, descriptive statistics, hypothesis testing—*t* tests, ANOVA, chi-squared tests, case studies, forecasting, etc.)

Expected outcomes or signs of variables (based on theory, hypotheses, or previous literature)

Measurement of variables used in the analysis

Although theory is often woven into the introduction and the literature review, it often appears in the methods section to position the research approach within a school of thought or to indicate the expected outcome of the research based on theory.

What follows is an example of a description of the data collection method from an article on the use of prescription stimulants among undergraduate students (Teter, McCabe, Cranford, Boyd, & Guthrie, 2005):

The Institutional Review Board at the University of Michigan approved the protocol for the present study and all participating students gave informed consent. The study was conducted during a 1-month period in March and April of 2003, drawing on a total undergraduate population of 21,294 full-time students (10,860 women and 10,434 men). Two drug-related surveys were being conducted at the same time and we did not want to burden undergraduate students with taking 2 similar surveys. Therefore, we surveyed the entire population but this study was based on a random sample of 19,278 students and the other study used the remaining students. We sent the sample group an e-mail message describing the study and inviting them to self-administer the Student Life Survey (SLS) by using a unique password and clicking on a link to access the Web survey. The Web survey was maintained on an Internet site running under the secure socket layer protocol to ensure privacy and security. We sent 3 reminder e-mails to non-respondents. By participating in the survey, students became eligible for a sweepstakes of 13 cash prizes ranging from $100 to $1,000. The final response rate was 47%, which is consistent with other college-based AOD studies. (pp. 253–254)

This same article then has additional subsections describing the questionnaire, measures used for different variables, and procedures used in data analysis. For example, in the “Measures” subsection, the authors provide the exact wording used on the questionnaire to determine how often students used illicit drugs over the past year and over their lifetime.

16.5 RESULTS

The purpose of the results section is to identify the most relevant results needed to answer the research question(s). In addition, however, summary statistics related to the variables used in the analysis are also given, such as a table showing the mean and standard deviation of each key variable.

In some journals, the results section includes an interpretation of the results and possible policy implications. In other journals, however, the results section is used strictly to state the results. Interpretation and analysis then follow in a “Discussion” or “Comment” section. Some guidelines for the results section are offered below.

16.5.1 Logical Sequence

The results section often follows the order of the research questions or hypotheses stated in the introduction and then reports on the tests related to each question or hypothesis. Typically, broader results are reported first followed by detailed analyses of each research question. For example, in the Teter et al. (2005) paper on the use of stimulant drugs among undergraduates, they state two primary objectives or questions in their introduction: (1) What is the prevalence and motive for use of stimulant drugs? (2) Is there a link between the motives for use of stimulants and the use of alcohol and other drugs? Their results section is then divided into two sections that answer these questions as follows:

Prevalence Rates and Motives for Use

More than 8.1% of the undergraduate student sample (n = 689) reported the illicit use of prescription stimulants in their lifetime, and 5.4% (n = 458) reported illicit use in the past year; undergraduate men reported significantly higher lifetime rates than did undergraduate women (9.3% vs 7.2%, $p < 0.001$). Lifetime rates were higher for White (9.5%) and Hispanic (8.9%) students than for African-American (2.7%), Asian (4.9%), or other racial student groups (5.8%), $\chi^2(4, N = 8{,}460) = 55.08$, $p < 0.001$.

The primary motives that students gave for using prescription stimulants illicitly were (1) to help with concentration, (2) to increase alertness, and (3) to get high. We observed no gender differences in motives for illicit use. The frequencies for each motive and the index describing the number of motives endorsed are presented in Table 3. Approximately half the students who endorsed the illicit use of prescription stimulants gave more than 1 motive for this behavior. On average, students reported 1.65 ± 0.91 motives (range 0–5, mode 1.0) for the illicit use of prescription stimulants. Of the 689 students who endorsed the lifetime illicit use of prescription stimulants, 31 did not provide a motive for their behavior, 19 students chose the “Refused” category, and 12 students did not provide any motive.

The proportion of each motive within a given frequency range was relatively consistent (see Figure 1). For example, using prescription stimulants “to help concentrate” accounted for approximately 30% of the motives, regardless of the number of occasions of illicit stimulant use. However, the distribution in the actual frequency range of illicit prescription stimulant use was skewed, with a steady decrease in the number of students reporting more frequent use. For example, a total of 254 students reported the illicit use of prescription stimulants on 1 to 2 occasions, compared with 45 who reported 40 or more occasions. The data in Figure 1 do not represent those 31 students who did not provide a motive and therefore consist of 658 students.

Approximately 14% (n = 97) of the illicit prescription stimulant users also reported being prescribed stimulant medication in their lifetime. We found no differences in any of the motivations endorsed by those illicit users who were also prescribed stimulants in their lifetime compared with the illicit users who had never been prescribed stimulants. For example, approximately 40% of those who endorsed the illicit use of prescription stimulants provided “to get high” as a motive, regardless of whether they had also been legitimately prescribed stimulant medications.

Alcohol and Other Drug Use Behaviors

> Chi-square analysis revealed significantly higher rates of AOD use for those students who reported the illicit use of prescription stimulants, compared with nonstimulant users (see Table 4). Furthermore, regardless of the motive or motives for the illicit use of prescription stimulants, the 689 students who endorsed these behaviors also reported significantly higher rates of AOD use in the recent past. For example, only 1.6% of those who reported no illicit prescription stimulant use had used cocaine in the past year, whereas those who reported the illicit use of prescription stimulants to help them concentrate, increase alertness, or get high had past-year cocaine prevalence rates of 28.6%, 31.1%, and 35.4%, respectively. Data in Table 4 also show that the "counteracts the effects of other drugs" and the "to get high" motives were strongly associated with cocaine and amphetamine use. Finally, AOD use was positively related to the number of motives given for the illicit use of prescription stimulants, particularly for cocaine, ecstasy, and amphetamine use (See Table 5). (Teter et al., 2005, pp. 256–257)

Notice in the first paragraph that Teter et al. (2005) begin with the larger picture—what percentage of students report stimulant drug use over their lifetime. The paragraph then continues with results that are more detailed, including drug use in the past year, drug use by men and women, and finally drug use broken down by racial and ethnic background.

16.5.2 Tables, Figures, and Numbers

Tables and figures, which include graphs and pictures, are used in the results section to display and summarize data. They should be numbered consecutively with one set of numbers for tables and a second set for figures. When referring to a specific table or figure, the word *Table* or *Figure* is always capitalized. References to specific tables or figures appear within sentences or in parentheses as illustrated in the results section of the Teter et al. (2005) paper:

> "The frequencies for each motive … are presented in Table 3."
>
> "The proportion of each motive … was relatively consistent (see Figure 1)."
>
> "The data in Figure 1 do not represent …"
>
> "Chi-square analysis revealed significantly higher … (see Table 4)."
>
> "Data in Table 4 also show that …"
>
> "Finally, AOD use was positively related to the number of motives … (See Table 5).

When referring to information from tables and figures, you should not repeat the numbers in the tables or figures since the reader can see them. Instead, you should focus on identifying patterns or highlighting the most relevant results. As one example from the Teter et al. (2005) paper, Table 3 shows the exact number of students (not percentages) who gave zero, one, two, and three or more motives for using stimulant drugs. Instead of repeating each of these numbers in the text, they write, "Approximately half the students who endorsed the illicit use of prescription stimulants gave more than 1 motive for this behavior."

All tables and figures should have complete titles and labels so that the reader can understand the table without having to read the additional text. A good rule of thumb is that if a table or figure falls out of a book and someone picks it up, they should be able to understand it fully.

Finally, there are rules for writing out numbers in academic documents. The *Publication Manual of the American Psychological Association*, which is used by the social sciences and referred to as APA style,

suggests that numbers one through nine should be spelled out and that numbers 10 and above should be written as numerals (American Psychological Association, 2009).[2] There are exceptions to these rules. Any number can be written as a numeral when referencing tables or figures, grouping numbers above and below 10 for comparison, and writing numbers that represent time, dates, and age. Alternatively, numbers that begin sentences should always be written out.

16.5.3 Reporting Results From Statistical Tests

As described in earlier chapters, the method of reporting results from statistical tests will vary depending on the publication source of the article and/or the audience. If you were writing a report for a newspaper with a wide audience, you would indicate if there was a "statistically significant difference" when comparing means or percentages. In a scholarly journal, however, you would need to include more details about the tests and the results. APA style offers specific guidelines on reporting of statistics. These rules are shown next, followed by examples of each of the statistical tests we have covered in this textbook.

16.5.3.1 APA Style Rules for Reporting the Results of Statistical Tests

Report the descriptive statistics, including means and standard deviations.

Include the test statistic, degrees of freedom, and obtained value of the test.

Round test statistics and *p*-values to two decimal places.

Italicize all statistical symbols (excluding Greek letters)—*N, n, M, SD, p, t*, and so on.

Report the *p*-value (the probability of observing the result or a more extreme value) in one of two ways:

Report the exact level ($t(40) = 2.5$, $p = 0.02$). If the *p*-value were less than 0.001, rather than rounding this to two decimal places, you would write $p < 0.001$.

Use the alpha level ($t(40) = 2.5$, $p < 0.05$), assuming that your alpha level is 0.05.

For a regression, report the R^2 value, the *F* value, the degrees of freedom in parentheses, and the *p* value.

16.5.3.2 Examples

Reporting a significant difference in a sample mean compared with the population mean

---

Students who listened to Beethoven for 1 hour before taking the Scholastic Aptitude Test scored higher ($M = 1{,}642$, $SD = 18$) than the national average of 1,250 ($SD = 88$), $t(50) = 2.47$, $p = 0.02$.

Reporting a significant difference in two means

Students who multitasked while studying for an exam scored lower (M = 82, SD = 10) than students who did not multitask (M = 88, SD = 12), $t(56) = 2.10$, $p = 0.04$.

Reporting a significant difference in more than two means

A one-way analysis of variance was conducted to examine the effect of car ownership type on behavior toward bicyclists on the road. Drivers were divided into three groups based on the cost of a new vehicle of the type that they were driving. The space between the passing car and bicyclist was then measured on a 1-mile length of road in a suburban area. There was a statistically significant difference in the average distance between the bicyclist and car among the three categories $F(2, 87) = 4.42$, $p = 0.02$. Drivers with the most expensive cars allowed 2 feet of space on average between their car and bicyclists (SD = 1.3) compared with 2.5 feet among drivers of midrange cars (SD = 1.2) and 3 feet among the drivers in the least expensive car group (SD = 1.3).

Reporting a significant difference in percentages

A higher percentage of people in the age group 20 to 40 reported that they supported gun control (85%) compared with those in the 41 to 60 age group (65%), $\chi^2(1, n = 200) = 14.6$, $p < 0.001$.

Reporting a significant correlation

Examining different regions of the world, a recent study showed a positive correlation between greater air pollution and deaths caused by respiratory disease ($r = 0.57$, $n = 42$, $p = 0.05$).

Reporting results of a regression

The hours that students studied for an exam predicted their exam score, $R^2 = .45$, $F(1, 422) = 6.88$, $p = 02$

16.5.4 Active Versus Passive Voice and the Use of First-Person Pronouns

It is important to use the active voice in writing whenever possible. One example is this:

Passive voice: It was shown that students who listen to Beethoven before an exam earn higher scores.

Active voice: Our results show that students who listen to Beethoven before an exam earn higher scores.

Regarding the first-person pronouns such as I or we, notice that in the active voice example, the sentence begins with “our results” instead of “the results.” Generally, the first person is preferred when describing tasks performed by the authors. In addition, a single author will often use “we” instead of “I.” For example, “we find that many students …” This is sometimes thought of as the “collective we,” in which you are including the audience in the plural pronoun. In other words, “we (as a group) can see that the results are interesting.”

16.6 DISCUSSION

The purpose of the discussion section is to interpret your results and place them within the context of the literature. Do your results agree or disagree with the literature or with theory? What are the possible explanations for this? Even if your results are unexpected or not significant, it is still important to discuss the implications. Generally, this section will not include any new statistics or even refer to tables or figures from the results section. Instead, it highlights the major findings and offers explanations.

In addition to interpreting and highlighting the results, the discussion section is also used to offer recommendations, identify limitations, and suggest areas for future research. The recommendations should come strictly from the results of your study and are often related to policy implications. In the paper by Talan and Kalinkara (2023) for example, they write the following:

> In the current state, ChatGPT is capable of producing accurate responses within seconds. However, it has limitations in interpreting visual aids such as diagrams, shapes, and tables, which can be easily comprehended by human students. Thus, these visual aids need to be explained in text form for ChatGPT to understand. Additionally, if a question is ambiguous or incomprehensible, ChatGPT may produce an incorrect response. To mitigate this issue, it is advisable to rephrase the question in a clear and precise manner.

Regarding limitations, all studies have them. They may arise from sampling, measurement of key variables, or a fault in the questionnaire, for example. It is important to list them clearly so that the readers can more accurately determine the quality of your work and its implications. If there are limitations that you have not listed, but the reader can identify them, they may then assume that you are not aware of these limitations and therefore question your entire paper. Or they may assume that you are aware of them and are trying to hide the limitations. Overall, all papers should list their limitations without hesitation. Below is an example of a list of limitations from the Teter et al. (2005) paper.

> We should note several limitations in this study before readers assess the implications of our findings. Our sample consists of students from a single campus, which limits the generalizability of our results. In addition, our study sample was drawn from a predominantly White student population attending a large public university. Therefore, our findings need to be compared with more diverse samples (both of students attending college and young adults who are not college students). Nonresponse may have introduced potential bias in the present study; however, these concerns were somewhat reduced because the demographic characteristics in the final obtained sample closely resembled the overall student population. In addition, the rates of drug use in the present study were comparable to rates found in other national substance use surveys of college students. We did not survey students about the quantity of prescription stimulants they used illicitly per occasion. Also, we did not collect information on the route of administration (i.e., intranasal or injection), which would have an important impact on the long-term morbidity and mortality as well as the abuse potential of stimulant medication. (p. 260)

Finally, almost all papers will suggest areas for future research. Some of these may come from the list of limitations. For example, in the Teter et al. (2005) paper, they identify one of their limitations as choosing a sample from one campus that is predominantly White. They suggest further research "in various populations, such as urban residents, those not attending college, and those with diverse racial backgrounds" (p. 261).

16.7 CONCLUSIONS

Conclusions are sometimes included as part of the discussion section of a paper, and at other times, they are presented as the final section of the paper. Regardless of their location, conclusions are always brief and frequently just one or two paragraphs. Rather than repeating information from the discussion section, the conclusion section is used to summarize the main findings of the paper, relate them back to the big ideas presented in the introduction, and emphasize their importance. An example from the Teter et al. (2005) paper is as follows:

> College students use prescription stimulants illicitly for many reasons. Our findings highlight the importance of assessing the motives for the illicit use of prescription stimulants and suggest that these motivations may be associated with greater use of alcohol and other drugs, especially if the student reports the illicit use of such stimulants to counteract the effect of other drugs or is using them to get high. In addition, those students who provide multiple motives for the illicit use of prescription stimulants may also be using excessive amounts of AODs. Although the long-term morbidity and mortality from these behaviors remain unknown, the problem of prescription stimulant abuse exists in the college population and should be addressed both clinically and experimentally. (p. 261)

Overall, this chapter on writing a research paper is a snapshot of the research process and an overview of each section of a research paper. There are many excellent sources that go into much greater detail about the research process and each part of the final paper. One source for further reading is *The Craft of Research* by Booth, Colomb, Williams, Bizup, and Fitzgerald (2016). For a concise set of rules related to writing, Weingast (2010) offers the *Caltech Rules for Writing Papers*.

EXERCISES

1. Find a newspaper or magazine article about the results of a recent study that used data. It can be in any field such as health, economics, sociology, psychology, and so on. Then find the primary source of the study in the scholarly journal where it was first published. Take one of the findings, and copy exactly how it was reported in each of the two sources. Then, point out the differences in the language used to report that finding.
2. Find two articles in a scholarly journal in your area of interest that include a literature review section. Write an outline of each of the two literature reviews where each bullet in your outline is the topic of one paragraph. How similar or different are the two literature reviews in their structure?
3. Using the same two articles from Question 2, write an outline of their data and/or methods section. How do they differ, and how are they similar?
4. Using the same two articles from Question 2, summarize what the authors have identified as the limitations of each study and the areas of future research.

APPENDIX 1: QUICK REFERENCE GUIDE TO STATA COMMANDS

This appendix summarizes the Stata commands used throughout the book, as well as providing screenshots of the output that the commands generate. Students may find this useful as they analyze data sets using multiple statistics or hypothesis tests and need a quick reference.

The first section of this appendix lists the code that does not generate output. Instead, it is used to open files, rename variables, create value labels, recode variables, and so forth. The remaining sections show the output generated by Stata code from Chapters 6 through 14. While these figures are given new numbers ("Appendix Figure A1.1," for example), they are taken directly from the chapters. At the base of each figure is a note indicating the number of the figure and the chapter where it appears. ("This is Figure 6.1 from Chapter 6", for example.) General commands are found above each output and use the following abbreviations:

contvar = continuous variable

catvar = categorical variable

varname = can be either continuous or categorical

```
. tab inst_type
```

Type of Institution	Freq.	Percent	Cum.
Public	587	39.56	39.56
Private nonprofit	866	58.36	97.91
Private for-profit	31	2.09	100.00
Total	1,484	100.00	

Figure A1.1 Frequency Table of College Types

Note: **This is Figure 6.1 from Chapter 6.**
Note: Tabulate can be abbreviated to "tab." **Tabulate** produces frequency tables, whereas **table** generates other statistics. To generate several frequency tables in a row, use "**tab1 var1 var2 var3**," and so on.

TRANSFORMING AND DEFINING VARIABLES AND VALUES PLUS GENERAL COMMANDS

TABLE A1.1 ■ Stata functions and associated codes

Function	Code
Open a file	use "c:\file location\filename"
Rename a variable	rename originalvarname newvarname rename var1 gender
Create a variable label	label variable varname "any description you like of any length" label variable gender "gender of respondent"
Create value labels	label def sexlabel 1 "female" 2 "male" label val sex sexlabel
Add new value labels	label def sexlabel 3 "other", add label val sex sexlabel
Add new value labels and change existing ones	label def sexlabel 1 "male" 2 "female" 3 "other", modify (Note: Quotation marks are only necessary in labels if there is a space in the label names.)
List values of all observations of selected variables	list varname1 varname2
Destring: Change format from string variable to numeric	destring varname, gen(numvar) ignore(",") (use the ignore command if there are commas separating numbers such as 1,000)
Recode variables	recode catvar (1 2=1 (3 4=2), gen(newcatvar) recode catvar (1/5=1) (6/10=2), gen(newcatvar) recode varname .=0, gen(newvar) (used when you want to change missing values to zero)
Generate dummy variables (If you have a categorical variable, catvar, with three or more categories, you can generate a dummy variable for each category with these commands. You can substitute any letter for "g" in parentheses.)	tab catvar, gen(g)
Generate dummy variables (If you allowed a respondent to choose more than one category, each category will appear as its own variable. To change each of these to a dummy or 0/1 variable, use this.)	recode catvar1 .=0 recode catvar2 .=0
Use only certain responses in your analysis If you want to keep certain responses, you can use the **keep if** command. This will remove the observations for all remaining analyses until you use the **clear** command. The next time you open the file, all observations will be there unless you saved the data file after using **keep if**.	keep if varname < 21 if you prefer to temporarily use some observations, but restore all observations that were removed, you can use the following comamnd: preserve keep if varname < 21 restore

Function	**Code**
Open a file	use “c:\file location\filename”
Rename a variable	rename originalvarname newvarname rename var1 gender
Create a variable label	label variable varname “any description you like of any length” label variable gender “gender of respondent”

Function	Code
Create value labels	label def sexlabel 1 “female” 2 “male” label val sex sexlabel
Add new value labels	label def sexlabel 3 “other”, add label val sex sexlabel
Add new value labels and change existing ones	label def sexlabel 1 “male” 2 “female” 3 “other”, modify (Note: Quotation marks are only necessary in labels if there is a space in the label names.)
List values of all observations of selected variables	list varname1 varname2
Destring: Change format from string variable to numeric	destring varname, gen(numvar) ignore(“,”) (use the ignore command if there are commas separating numbers such as 1,000)
Recode variables	recode catvar (1 2=1 (3 4=2), gen(newcatvar) recode catvar (1/5=1) (6/10=2), gen(newcatvar) recode varname .=0, gen(newvar) (used when you want to change missing values to zero)
Generate dummy variables (If you have a categorical variable, catvar, with three or more categories, you can generate a dummy variable for each category with these commands. You can substitute any letter for “g” in parentheses.)	tab catvar, gen(g)
Generate dummy variables (If you allowed a respondent to choose more than one category, each category will appear as its own variable. To change each of these to a dummy or 0/1 variable, use this.)	recode catvar1 .=0 recode catvar2 .=0

Function	Code
Use only certain responses in your analysis If you want to keep certain responses, you can use the **keep if** command. This will remove the observations for all remaining analyses until you use the **clear** command. The next time you open the file, all observations will be there unless you saved the data file after using **keep if**.	keep if varname < 21 if you prefer to temporarily use some observations, but restore all observations that were removed, you can use the following comamnd: preserve keep if varname < 21 restore

TABLES AND SUMMARY STATISTICS

```
tab catvar
```

```
tab1 catvar1 catvar2
```

```
. tab1 region admcon7, sort

-> tabulation of region
```

REGION	Freq.	Percent	Cum.
Southeast (AL, AR, FL, GA KY, LA, MS, N	404	27.22	27.22
Mid East (DE, DC, MD, NJ NY, PA)	278	18.73	45.96
Great Lakes (IL, IN, MI, OH, WI)	235	15.84	61.79
Plaines (IA, KS, MN, MO, NE, ND, SD)	160	10.78	72.57
New England (CT, ME, MA NH RI, VT)	129	8.69	81.27
Far West (AK, CA, HI, NV, OR, WA)	128	8.63	89.89
Southwest (AZ, NM, OK, TX)	78	5.26	95.15
Rocky Mountains (CO, ID, MT, UT, WY)	34	2.29	97.44
Outlying Areas (AS, FM, GU, MH, MP, PR,	33	2.22	99.66
U.S. Service Schools	5	0.34	100.00
Total	1,484	100.00	

```
-> tabulation of admcon7
```

Admissions Test Score Policy	Freq.	Percent	Cum.
Considered but not required	772	61.32	61.32
Neither required nor recommended	177	14.06	75.38
Recommended	170	13.50	88.88
Required	140	11.12	100.00
Total	1,259	100.00	

Figure A1.2 Multiple Frequency Tables Using The Tab1 Command

Note: **This is Figure 6.2 from Chapter 6.**

```
sum contvar, detail
```

```
. sum ugds, detail

                    Undergraduate enrollment
-------------------------------------------------------------
      Percentiles      Smallest
 1%           155             27
 5%           419             32
10%         656.5             38       Obs               1,480
25%        1151.5             67       Sum of wgt.       1,480

50%          2302                      Mean           5503.583
                        Largest       Std. dev.      8560.997
75%          6027          63963
90%         14318.5        64210       Variance       7.33e+07
95%         22690          72229       Skewness       4.204682
99%         37743         119248       Kurtosis       33.74106
```

Figure A1.3 Percentiles and Median

Note: This is Figure 6.3 from Chapter 6.

```
sum contvar
```

```
. sum ugds

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        ugds |      1,480    5503.583    8560.997        27     119248
```

Figure A1.4 Mean of College Size

Note: This is Figure 6.4 from Chapter 6.

```
table catvar, stat(mean contvar) stat(median contvar) nformat(%6.0fc)
```

```
. table inst_type, stat(mean costt4_a) ///
> stat(median costt4_a) nformat(%6.0fc)
```

	Mean	Median
Type of Institution		
Public	27,019	23,095
Private nonprofit	27,857	23,470
Private for-profit	36,758	33,115
Total	27,637	23,310

Figure A1.5 Means And Medians For Subcategories

Note: This is Figure 6.5 from Chapter 6.

```
tab catvar, stat(mean contvar) stat(sd contvar)
```

```
. table inst_type, stat(mean grad_debt_mdn) ///
> stat(sd grad_debt_mdn) nformat(%6.0fc)
```

	Mean	Standard deviation
Type of Institution		
Public	15,459	7,912
Private nonprofit	16,369	8,022
Private for-profit	16,724	7,613
Total	15,999	7,976

Figure A1.6 Mean and Standard Deviation of College Debt by Type of Institution

Note: This is Figure 6.6 from Chapter 6.

```
tab catvar1 catvar2, row
```

```
. tab inst_type admcon7, row
```

Key
frequency *row percentage*

Type of Institution	Admissions Test Score Policy Required	Recommend	Neither r	Considere	Total
Public	86 19.41	48 10.84	59 13.32	250 56.43	443 100.00
Private nonprofit	54 6.72	116 14.45	115 14.32	518 64.51	803 100.00
Private for-profit	0 0.00	6 46.15	3 23.08	4 30.77	13 100.00
Total	140 11.12	170 13.50	177 14.06	772 61.32	1,259 100.00

Figure A1.7 Combining Two Categorical Variables Using The Tabulate Command

Note: **This is Figure 6.7 from Chapter 6.**

The variable with more categories should appear first since the table will be more likely to fit on a page. When deciding on whether to generate row or column percentages, always add up over the independent variable.

GRAPHS

Pie Chart

```
graph pie, over(catvar) plabel(_all percent)
graph pie, over(Sector3) plabel(_all percent)
```

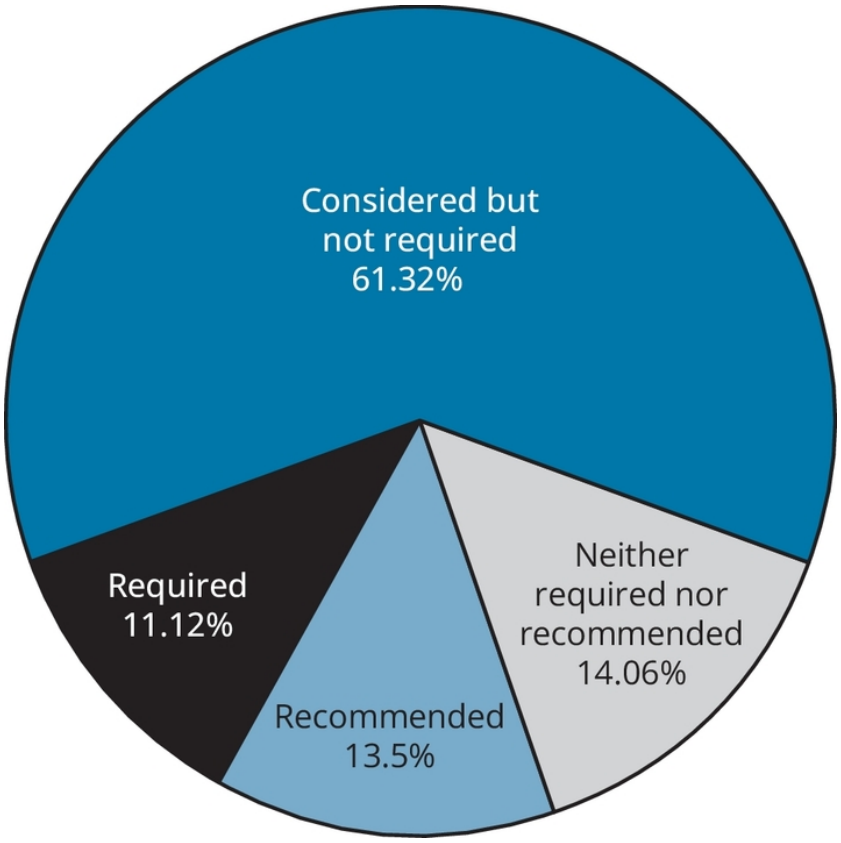


Figure A1.8 Pie Chart Of College Test Score Policies

Note: **This is Figure 6.14 from Chapter 6.**

Histogram

```
hist contvar, bin(1) frequency
hist TotalPriceInStateOnCampus, bin(10) frequency
```

(*Note:* The **bin** command is not essential since Stata will try to choose the optimal number of bins.)

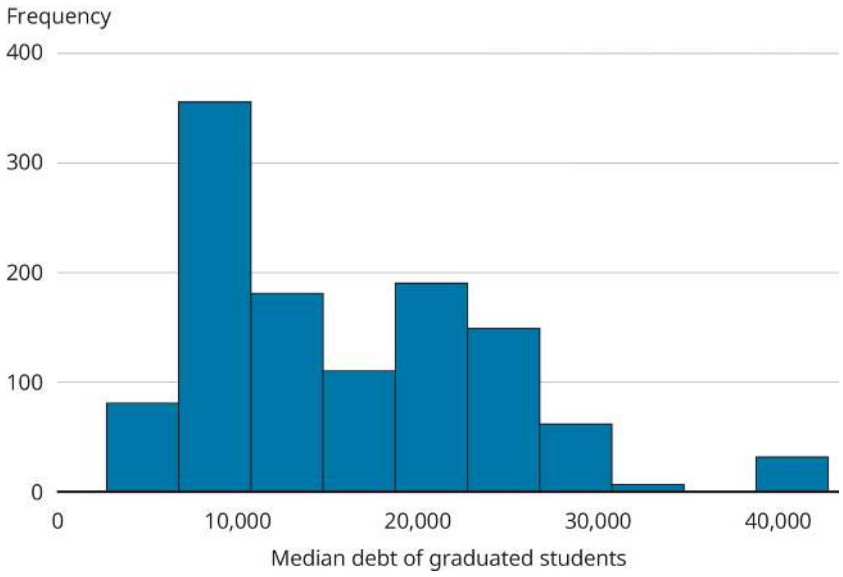


Figure A1.9 Histogram Of The Median Debt Owed By College Graduates Based On The College Scorecard Data From April 23 – Usnews

Note: **This is Figure 6.13 in Chapter 6.**

Bar Chart With Continuous Variable Summarized Over a Categorical Variable

```
graph bar(mean) contvar, over(catvar)
graph bar(mean) TotalPriceInStateOnCampus, over(Sector3)
```

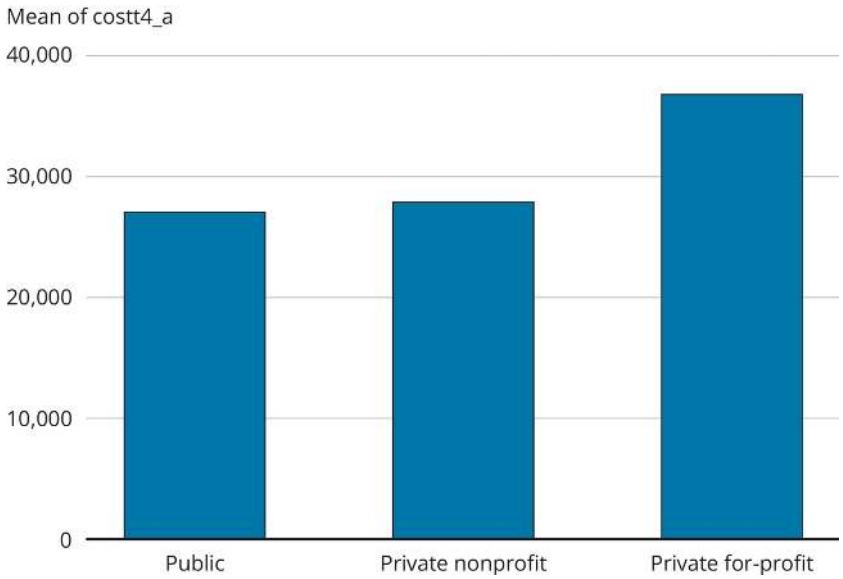


Figure A1.10 Bar Graph Of Average Tuition By Type Of College

Note: This is Figure 6.11 from Chapter 6.

Box Plot

```
graph box contvar, over(catvar)
graph box TotalPriceInStateOnCampus, over(Sector3)
```

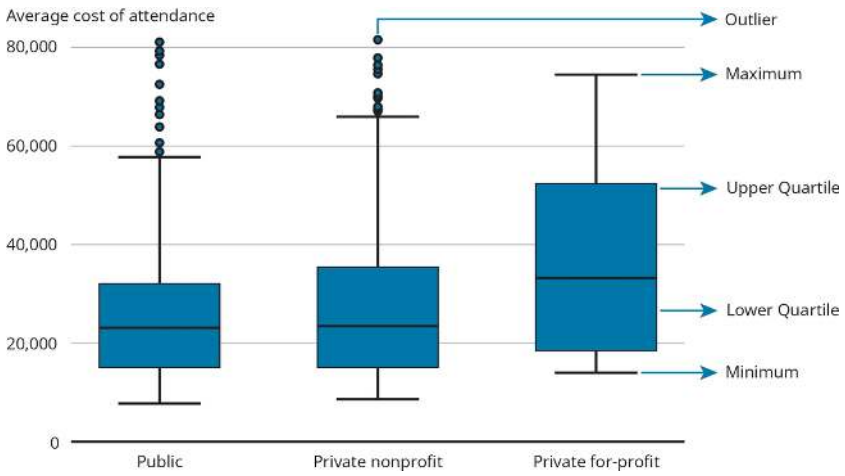


Figure A1.11 Box Plot Of Average Tuition By Type Of Institution

Note: This is Figure 6.12 from Chapter 6.

TESTING HYPOTHESES

One-Sample *t* Test

```
ttest contvar==# (where # could be any number that you are testing)
```

```
. ttest grade==86

One-sample t test

Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
   grade |      30     89.36667    1.035556    5.671972    87.24872    91.48462
---------+--------------------------------------------------------------------
    mean = mean(grade)                                            t =   3.2511
H0: mean = 86                                     Degrees of freedom =       29

    Ha: mean < 86                 Ha: mean != 86                 Ha: mean > 86
 Pr(T < t) = 0.9985         Pr(|T| > |t|) = 0.0029          Pr(T > t) = 0.0015
```

Figure A1.12 Stata Output For The One-Sample *t* Test

Note: This is Figure 8.4 from Chapter 8.

Two Independent-Samples *t* Test

```
robvar contvar, by(catvar)
ttest contvar, by(catvar)
esize twosample contvar, by(catvar) cohensd unequal (or leave out unequal if variances
are equal)
```

```
. robvar MaskDays, by(ElectorParty)
```

Party of Electors from 2020 Election	Summary of No. of mask mandate days as of 8/15/22 Mean	Std. dev.	Freq.
Democrat	405.88	212.81454	25
Republica	146.56	140.39174	25
Total	276.22	221.34019	50

```
W0  =  1.5566141   df(1, 48)     Pr > F = 0.21821431

W50 =  1.4663531   df(1, 48)     Pr > F = 0.23185058

W10 =  1.9130489   df(1, 48)     Pr > F = 0.17302516
```

Figure A1.13 Istata Output For Equality Of Variance Test

Note: **This is Figure 9.3 in Chapter 9.**

```
. ttest MaskDays, by(ElectorParty)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
Democrat	25	405.88	42.56291	212.8145	318.0345	493.7255
Republic	25	146.56	28.07835	140.3917	88.60914	204.5109
Combined	50	276.22	31.30223	221.3402	213.3158	339.1242
diff		259.32	50.99014		156.7974	361.8426

```
    diff = mean(Democrat) - mean(Republic)                    t =   5.0857
H0: diff = 0                                     Degrees of freedom =       48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Figure A1.14 Stata Output For Two-Sample *T* Test With Unequal Variances

Note: **This is Figure 9.4 from Chapter 9.**

```
. esize twosample MaskDays, by(Elector) cohensd

Effect size based on mean comparison

                                Obs per group:
                                     Democrat =          25
                                   Republican =          25
```

Effect size	Estimate	[95% conf.	interval]
Cohen's *d*	1.43845	.8082568	2.056732

Figure A1.15 Cohen's *D*

Note: **This is Figure 9.5 from Chapter 9.**

One-Way Analysis of Variance

```
oneway contvar catvar
```

```
. oneway SAT FamilyInc, tabulate
```

FamilyInc	Summary of SAT Mean	Std. Dev.	Freq.
<59K	1276.905	225.16055	779
60-99K	1312.3089	188.99067	641
100-149K	1359.2057	179.30577	666
150-199K	1368.9189	176.09372	296
>200K	1433.8568	143.39574	796
Total	1349.1756	194.35444	3,178

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	10830385.7	4	2707596.42	78.69	0.0000
Within groups	109176496	3173	34407.9724		
Total	120006882	3177	37773.6487		

We reject the null hypothesis that there is no difference in SAT scores.

```
Bartlett's test for equal variances:  chi2(4) = 159.5359  Prob>chi2 = 0.000
```

We reject the null hypothesis of equal variances.

Figure A1.16 Stata Output For Anova With Bartlett's Test

Note: **This is Figure 10.5 from Chapter 10.**

Comparing Two or More Percentages in a Cross-Tabulation

```
tab catvar1 catvar2, row V chi2
```

```
. tab sex edu2, nofre row chi2 V
```

Sex	Education HS	College	Graduate	Total
Female	6.41	61.11	32.48	100.00
Male	10.83	62.06	27.11	100.00
Total	9.01	61.67	29.32	100.00

```
          Pearson chi2(2) = 395.2835   Pr = 0.000
               Cramér's V =   0.0875
```

Figure A1.17 Stata Output For The Pearson Chisquared Test

Note: **This is Figure 11.4 from Chapter 11.**

Correlation Between Two Variables

```
pwcorr contvar1 contvar2, sig
```

```
. pwcorr price mileage, sig
```

	price	mileage
price	1.0000	
mileage	-0.6019 0.0000	1.0000

Figure A1.18 Pearson Correlation Coefficient

Note: **This is Figure 12.2 from Chapter 12.**

REGRESSION ANALYSIS

Simple Regression Analysis

```
regress contvar var1
```

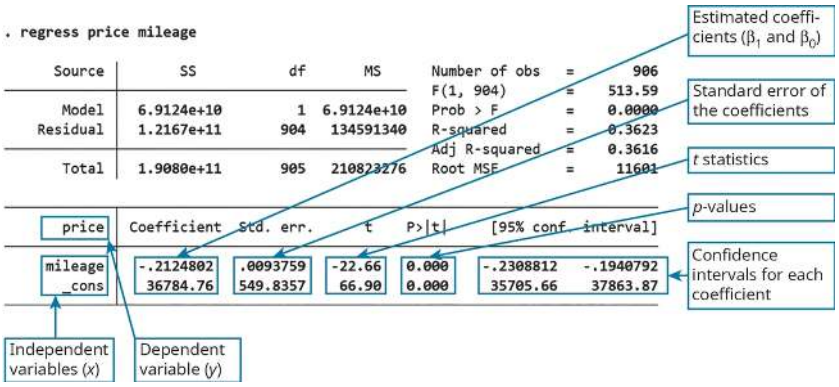


Figure A1.19 Simple Regression Analysis

Note: **This is Figure 12.6 from Chapter 12.**

Plot Data and Predicted Values From Regression Analysis

```
twoway (scatter y x) (lfit y x)
```

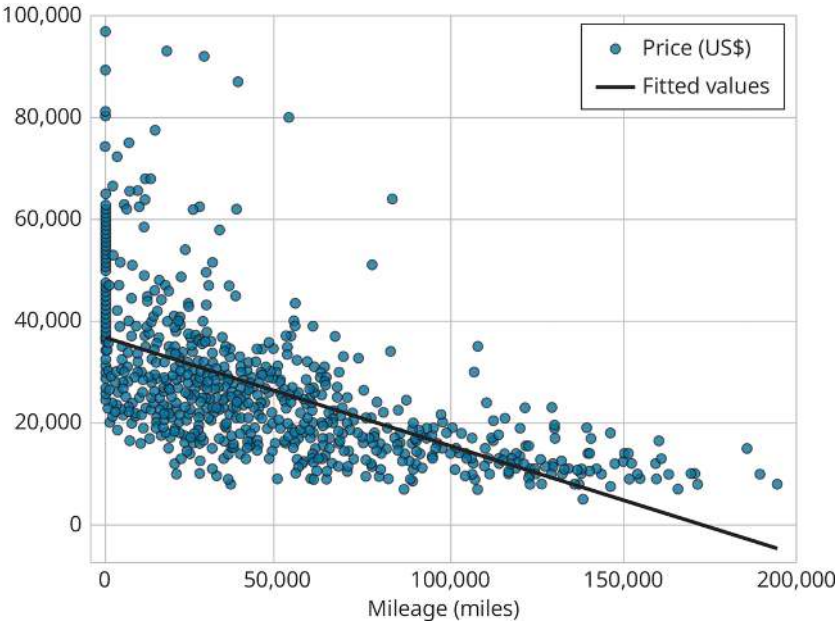


Figure A1.20 Scatter Plot Of Price And Mileage With Regression Line

Note: **This is Figure 12.7 from Chapter 12.**

Multiple Regression Analysis

```
regress contvar var1 var2 var3 var4
```

```
. regress price mileage hybrid electric
```

Source	SS	df	MS	Number of obs	=	902
				F(3, 898)	=	198.91
Model	7.5857e+10	3	2.5286e+10	Prob > F	=	0.0000
Residual	1.1415e+11	898	127118801	R-squared	=	0.3992
				Adj R-squared	=	0.3972
Total	1.9001e+11	901	210887932	Root MSE	=	11275

price	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
mileage	-.198785	.0093256	-21.32	0.000	-.2170875	-.1804825
hybrid	6318.549	1600.939	3.95	0.000	3176.532	9460.567
electric	12981.94	1905.485	6.81	0.000	9242.221	16721.67
_cons	35306.36	568.7436	62.08	0.000	34190.14	36422.59

Figure A1.21 Multiple Regression (Version 1)

Note: **This is Figure 12.8 from Chapter 12.**

Testing Joint Hypotheses After Regression Analysis

```
test x1 x2
```

```
. test hybrid electric

 ( 1)  hybrid = 0
 ( 2)  electric = 0

       F(  2,   898) =   29.45
            Prob > F =    0.0000
```

Figure A1.22 Testing Joint Hypotheses

Note: **This is Figure 12.10 from Chapter 12.**

Export Regression Results to Word

```
etable, export(filename.docx)
```

	Price
Mileage (miles)	−0.199
	(0.009)
Hybrid	6318.549
	(1600.939)
Electric	12981.944
	(1905.485)
Intercept	35306.365
	(568.744)
Number of observations	902

Figure A1.23 Regression Output In Word Format Using Etable

Note: **This is Figure 12.12 from Chapter 12.**

Plot Residuals Against an Independent Variable

```
rvpplot contvar
```

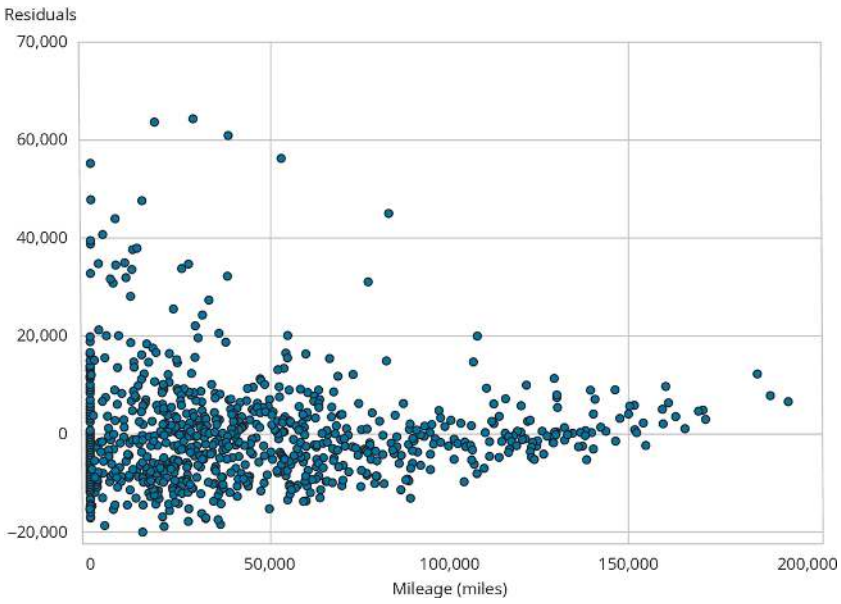


Figure A1.24 Scatterplot Of Residuals Against Mileage

Note: **This is Figure 13.4 from Chapter 13.**

Test for Omitted Variables

```
estat ovtest
```

```
. estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of price

H0: Model has no omitted variables

F(3, 894) =  11.04
 Prob > F = 0.0000
```

Figure A1.25 Omitted Variable Test

Note: This is Figure 13.5 from Chapter 13.

Test for Multicollinearity

```
estat vif
```

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
         age |     11.04    0.090548
        age2 |      6.84    0.146159
     mileage |      2.90    0.345096
         new |      1.70    0.588885
    electric |      1.10    0.910280
      hybrid |      1.05    0.954866
-------------+----------------------
    Mean VIF |      4.10
```

Figure A1.26 Test For Multicollinearity

Note: This is Figure 13.10 from Chapter 13.

Check for Heteroscedasticity With Plot of Residuals on Predicted Value

```
rvfplot
```

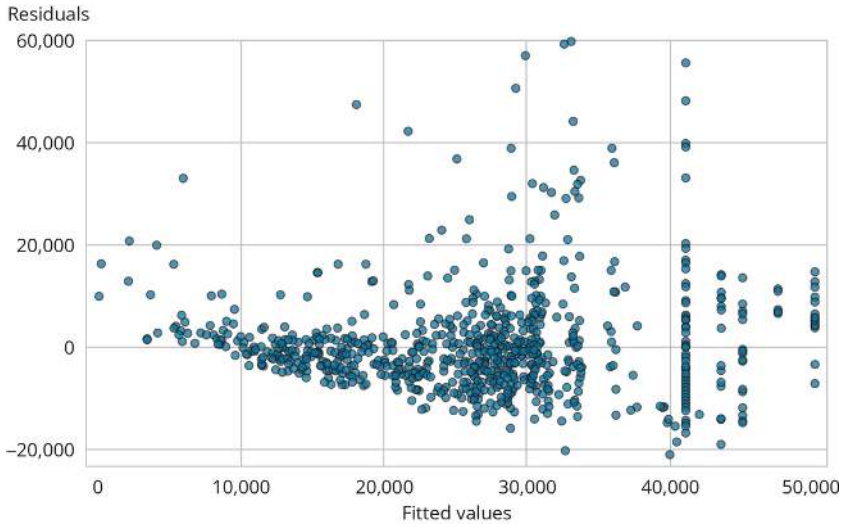


Figure A1.27 Scatter Plot Of Residuals Against Predicted Prices

Note: **This is Figure 13.12 from Chapter 13.**

Test for Heteroscedasticity

```
estat hettest
```

```
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of price

H0: Constant variance

    chi2(1) =  61.00
Prob > chi2 = 0.0000
```

Figure A1.28 Test For Heteroscedasticity

Note: **This is Figure 13.13 from Chapter 13.**

Test for Normality in Residual

```
sktest resid
```

Note: “resid” is the variable name of the residual.

```
. sktest e

Skewness and kurtosis tests for normality
                                                    ------ Joint test ------
    Variable |       Obs   Pr(skewness)   Pr(kurtosis)   Adj chi2(2)  Prob>chi2
-------------+---------------------------------------------------------------
           e |       902          0.0000          0.0000        336.52     0.0000
```

Figure A1.29 Test For Normality

Note: **This is Figure 13.16 from Chapter 13.**

Run a Logit Regression Model

```
logit binaryvar varname1 varname2 varname3
```

```
. recode clmtcaus (1/3=0) (4=1), gen(humcaus)
(1,770 differences between clmtcaus and humcaus)

. recode sex (1=0) (2=1), gen(female)
(3,940 differences between sex and female)

. logit humcaus age female educ

Iteration 0:  Log likelihood = -1155.5891
Iteration 1:  Log likelihood = -1116.9191
Iteration 2:  Log likelihood = -1116.8295
Iteration 3:  Log likelihood = -1116.8295

Logistic regression                                     Number of obs =  1,668
                                                        LR chi2(3)    =  77.52
                                                        Prob > chi2   = 0.0000
Log likelihood = -1116.8295                             Pseudo R2     = 0.0335

------------------------------------------------------------------------------
     humcaus | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -.0166404   .0029734    -5.60   0.000    -.0224682   -.0108126
      female |   .1460138   .1013562     1.44   0.150    -.0526407    .3446683
        educ |   .1223196   .0184335     6.64   0.000     .0861906    .1584485
       _cons |  -1.088241   .3250257    -3.35   0.001     -1.72528   -.4512025
------------------------------------------------------------------------------
```

Figure A1.30 Logit Model Of Belief In Human-Caused Climate Change

Note: **This is Figure 14.2 from Chapter 14.**

Run a Probit Regression Model

```
probit binaryvar varname1 varname2 varname3
```

Calculate Marginal Effects in Logit or Probit Model

```
margins, dydx(contvar)
```

```
. margins, dydx(educ)

Average marginal effects                                  Number of obs = 1,668
Model VCE: OIM

Expression: Pr(humcaus), predict()
dy/dx wrt:  educ
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
educ	.02917	.0041684	7.00	0.000	.0210001	.0373399

```
. margins, at(educ=(10(5)20))

Predictive margins                                        Number of obs = 1,668
Model VCE: OIM

Expression: Pr(humcaus), predict()
1._at: educ = 10
2._at: educ = 15
3._at: educ = 20
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	.3430206	.0232237	14.77	0.000	.297503	.3885383
2	.4873591	.0122656	39.73	0.000	.4633189	.5113992
3	.6337875	.0237648	26.67	0.000	.5872095	.6803656

Figure A1.31 Marginal Effects And Prediction For A Logit Model

Note: **This is Figure 14.3 from Chapter 14.**

Ordered Logit Model of a Categorical Dependent Variable

```
ologit catvar varname1 varname2 varname3
```

```
. ologit happy realinc female age

Iteration 0:  Log likelihood = -3228.3384
Iteration 1:  Log likelihood = -3181.6284
Iteration 2:  Log likelihood =  -3181.389
Iteration 3:  Log likelihood = -3181.3889

Ordered logistic regression                             Number of obs =  3,314
                                                        LR chi2(3)    =  93.90
                                                        Prob > chi2   = 0.0000
Log likelihood = -3181.3889                             Pseudo R2     = 0.0145

------------------------------------------------------------------------------
       happy | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     realinc |  -7.91e-06   8.59e-07    -9.21   0.000    -9.59e-06   -6.22e-06
      female |   .0180778    .068961     0.26   0.793    -.1170833    .153239
         age |  -.0049096   .0020131    -2.44   0.015    -.0088551   -.000964
-------------+----------------------------------------------------------------
       /cut1 |  -2.002207   .1309775                     -2.258918   -1.745496
       /cut2 |   .6922285   .1255487                      .4461577    .9382994
------------------------------------------------------------------------------
```

Figure A1.32 Ordered Logit Model Of Happiness

Note: **This is Figure 15.1 from Chapter 15.**

Instrumental Variable Regression

---

ivregress 2sls depvar indvar1 indvar2 endvar(instrvar)
[where depvar is the dependent variable, indvar are the independent variables, endvar is an endogenous explanatory variable, and instrvar is the instrument that predicts endvar]

Test for Autocorrelation

---

estat dwatson

Test for Stationarity

---

dfuller varname1 varname2 varname3

```
. dfuller P_mz_Rosario, trend

Dickey-Fuller test for unit root          Number of obs  = 226
Variable: P_mz_Rosario                    Number of lags =   0

H0: Random walk with or without drift

                             Dickey-Fuller
                Test      ---------- critical value ----------
            statistic           1%           5%          10%
-------------------------------------------------------------
 Z(t)          -2.427       -3.998       -3.433       -3.133
-------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3653.

. dfuller DP_mz_Rosario, trend

Dickey-Fuller test for unit root          Number of obs  = 225
Variable: DP_mz_Rosario                   Number of lags =   0

H0: Random walk with or without drift

                             Dickey-Fuller
                Test      ---------- critical value ----------
            statistic           1%           5%          10%
-------------------------------------------------------------
 Z(t)         -10.663       -3.998       -3.433       -3.133
-------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000.
```

Figure A1.33 Testing For Stationarity

Note: **This is Figure 15.4 from Chapter 15.**

APPENDIX 2: SUMMARY OF STATISTICAL TESTS BY CHAPTER

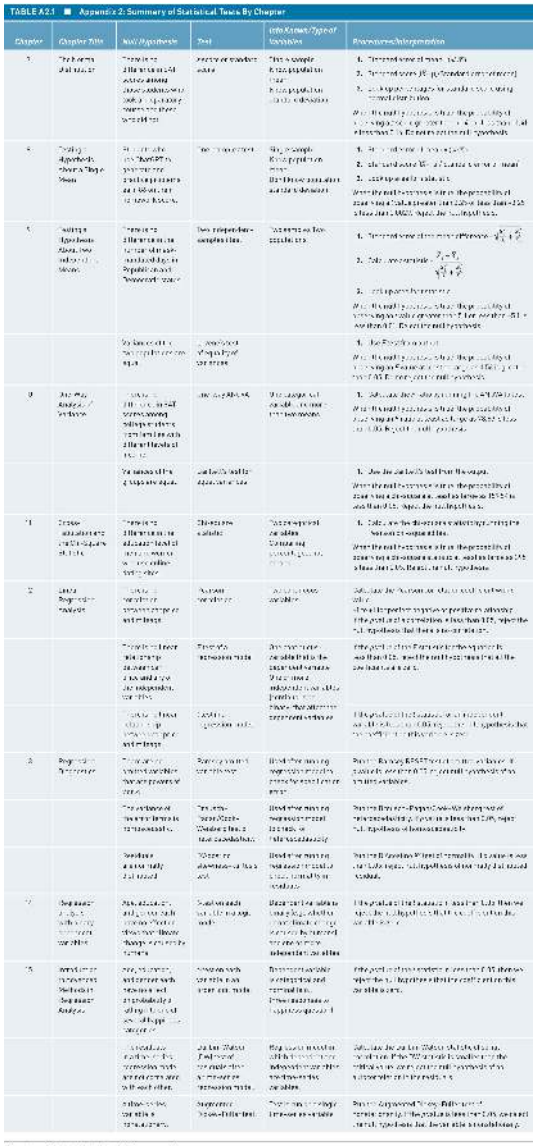


TABLE A2.1 ■ Appendix 2: Summary of Statistical Tests By Chapter

Chapter	Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpretation
5					
6					
7					
8					
11					
12					
13					
14					
15					

Chapter	Chapter Title	Null Hypothesis	Test	Info Known/Type of Variables	Procedures/Interpret

Chapter	*Chapter Title*	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpret*
7	The Normal Distribution	There is no difference in SAT scores among those students who took a preparatory course and those who did not.	*z* score or standard score	Single sample Know population mean Know population standard deviation	1. Standard error of mean = $(\sigma/\sqrt{n})$ 2. Standard score (( )/Standard error o mean) 3. Look up percenta for standard score using normal distribution When the null hypothe true, the probability of observing a *z* score gr than +1.41 or less tha −1.41 is less than 0.16 not reject the null hypothesis.
8	Testing a Hypothesis About a Single Mean	Students who use ChatGPT to generate and practice problems earn 86 on their homework score.	One-sample *t* test	Single sample Know population mean Don't know population standard deviation	1. Standard error of mean = $(s/\sqrt{n})$ 2. Standard score (( )/Standard error o mean) 3. Look up area for *t* statistic When the null hypothe true, the probability of observing a *t* value gre than 3.25 or less than −3.25 is less than 0.00 Reject the null hypothe

Chapter	*Chapter Title*	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpret*
9	Testing a Hypothesis About Two Independent Means	There is no difference in the number of mask-mandated days in Republican and Democratic states.	Two independent-samples *t* test	Two samples Two populations	1. Standard error of mean difference = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ 2. Calculate *t* statisti $\frac{X_1 - X_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ 3. Look up area for *t* statistic When the null hypothe true, the probability of observing an *t* value greater than 5.1 or les than −5.1 is less than Reject the null hypothe
		Variances of the two populations are equal.	Levene's test of equality of variances		1. Use *F* test from ou When the null hypothe true, the probability of observing an *F* value a least as large as 1.56 i greater than 0.05. Do r reject the null hypothes
10	One-Way Analysis of Variance	There is no difference in SAT scores among college students from families with different levels of income.	One-way ANOVA	One categorical variable and more than two means	1. Calculate the *F* ra by running the AN to test When the null hypothe true, the probability of observing an *F* ratio at least as large as 78.69 less than 0.05. Reject null hypothesis.
		Variances of the groups are equal.	Bartlett's test for equal variances		1. Use the Bartlett's from the output When the null hypothe true, the probability of observing a chi-square least as large as 159.5 less than 0.05. Reject null hypothesis.

Chapter	*Chapter Title*	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpret*
11	Cross-Tabulation and the Chi-Square Statistic	There is no difference in the education level of men and women who use online dating sites.	Chi-square statistic	Two categorical variables Comparing percentages, not means	1. Calculate the chi-square statistic by running the Pears chi-squared test When the null hypothe true, the probability of observing a chi-square statistic at least as larg 395 is less than 0.05. Reject the null hypothe
12	Linear Regression Analysis	There is no correlation between car price and mileage.	Pearson correlation	Two continuous variables	Calculate the Pearson correlation coefficient v *p* value −1 to +1 for perfect negative or positive relationship If the *p* value of a correlation is less than 0.05, reject the null hypothesis that there is correlation.
		There is no linear relationship between car price and any of the independent variables.	*F* test of a regression model	One continuous variable that is the dependent variable One or more independent variables (continuous or binary) that affect the dependent variables	If the *p* value of the F statistic for the equatio less than 0.05, reject t null hypothesis that all coefficients are zero.
		There is no linear relationship between car price and mileage.	*t* test in a regression model		If the *p* value of the *t* statistic for an indepen variable is less than 0. reject the null hypothes that the coefficient on t variable is zero.
13	Regression Diagnostics	There are no omitted variables that are powers of *y* or *x*.	Ramsey omitted variable test	Used after running regression model to check for specification error	Run the Ramsey RES test of omitted variable *p* value is less than 0.0 reject null hypothesis o omitted variables.

Chapter	*Chapter Title*	*Null Hypothesis*	*Test*	*Info Known/Type of Variables*	*Procedures/Interpret*
		The variance of the error terms is homoscedastic.	Breusch–Pagan/Cook–Weisberg test of heteroscedasticity	Used after running regression model to check for heteroscedasticity	Run the Breusch–Pagan/Cook–Weisberg of heteroscedasticity. I value is less than 0.05 reject null hypothesis o homoscedasticity.
		Residuals are normally distributed.	D'Agostino skewness-kurtosis test	Used after running regression model to check normality in residuals	Run the D'Agostino K^2 of normality. If *p* value less than 0.05, reject n hypothesis of normally distributed residual.
14	Regression analysis with binary dependent variables	Age, education, and gender each have no effect on views that climate change is caused by humans	*t*-test on each variable in a logit model	Dependent variable is binary (e.g., whether or not climate change is caused by humans) and one or more independent variables	If the *p* value of the *t* statistic in less than 0.(then we reject the null hypothesis that the coefficient on this varia is zero.
15	Introduction to Advanced Methods in Regression Analysis	Age, education, and gender each have no effect on probability of falling into one of several happiness categories.	*t*-test on each variable in an order logit model	Dependent variable is categorical and nominal (e.g., three responses to happiness question)	If the *p* value of the *t* statistic in less than 0.(then we reject the null hypothesis that the coefficient on this varia is zero.
		The residuals in a time-series regression model are not correlated with each other.	Durbin–Watson (DW) test of residuals after a time-series regression model.	Regression model in which dependent and independent variables are time-series variables.	Calculate the Durbin–Watson statistic of seri correlation. If the DW statistic is smaller than critical value, we reject null hypothesis of no autocorrelation in the residuals.
		A time-series variable is nonstationary.	Augmented Dickey–Fuller test.	Test is run on a single time-series variable	Run the Augmented Dickey–Fuller test of nonstationarity. If the *p* value is less than 0.05 reject the null hypothes that the variable is nonstationary.

Note: SAT = Scholastic Aptitude Test; ANOVA = analysis of variance.

APPENDIX 3: DECISION TREE FOR CHOOSING THE RIGHT STATISTIC

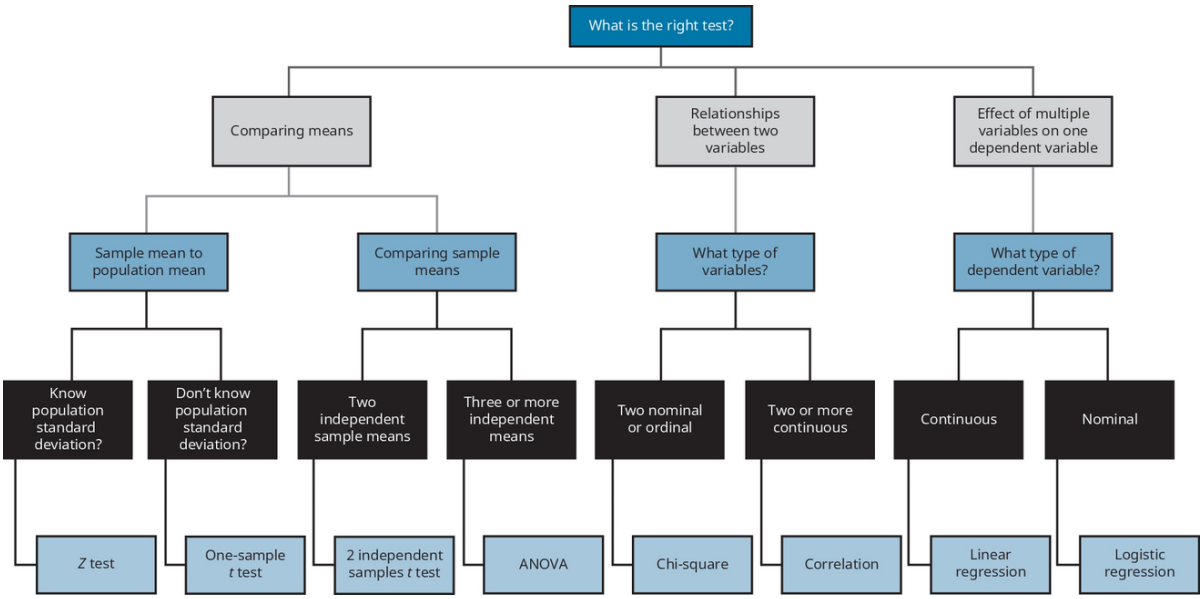


Figure A3.1 Decision Tree For Choosing the Right Statistic

Note: **This is Figure 8.2 from Chapter 8.**

APPENDIX 4: DECISION RULES FOR STATISTICAL SIGNIFICANCE

NULL HYPOTHESIS (H_O)

Hypothesis about the population

No difference among groups

It is never accepted or proved

It is rejected or not rejected

Example:

$$H_0 : \mu_1 = \mu_2$$

$H_0 : \mu_1 = \mu_2$

ALTERNATIVE HYPOTHESIS (H_1 OR H_A)

Opposite of the null hypothesis

Example

$$H_1 : \mu_1 \neq \mu_2$$

$H_1 : \mu_1 \neq \mu_2$

P VALUE OR *P* (CALCULATED)

The probability of rejecting the null hypothesis when it is true (Type I error)

The probability of obtaining a result equal to or more extreme than what was observed when the null hypothesis is true

Derived from sample results

A LEVEL OR *P* (CRITICAL)

Predetermined upper limit of the probability of making a Type I error

Significance level

Typically set at 0.05

DECISION RULE

Select α

If $p \leq \alpha$, reject the null hypothesis

If $p > \alpha$, do not reject the null hypothesis

APPENDIX 5: AREAS UNDER THE NORMAL CURVE (*Z* SCORES)

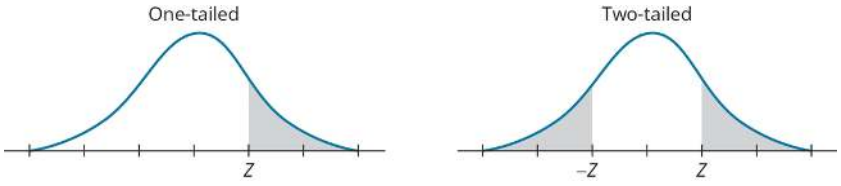


Figure A5.1 Areas Under the Normal Curve (*Z* Scores)

Note: **This is** Figure 7.6 **from Chapter 7**.

Z Scores	Probability	
	One-tailed	Two-tailed
0	0.50000	1.00000
0.1	0.46017	0.92034
0.2	0.42074	0.84148
0.3	0.38209	0.76418
0.4	0.34458	0.68916
0.5	0.30854	0.61708
0.6	0.27425	0.54851
0.7	0.24196	0.48393
0.8	0.21186	0.42371
0.9	0.18406	0.36812
1	0.15866	0.31731
1.1	0.13567	0.27133
1.2	0.11507	0.23014
1.3	0.09680	0.19360
1.4	0.08076	0.16151
1.5	0.06681	0.13361
1.6	0.05480	0.10960
1.7	0.04457	0.08913
1.8	0.03593	0.07186
1.9	0.02872	0.05743
2	0.02275	0.04550
2.1	0.01786	0.03573
2.2	0.01390	0.02781
2.3	0.01072	0.02145
2.4	0.00820	0.01640
2.5	0.00621	0.01242
2.6	0.00466	0.00932
2.7	0.00347	0.00693
2.8	0.00256	0.00511
2.9	0.00187	0.00373
3	0.00135	0.00270
3.1	0.00097	0.00194
3.2	0.00069	0.00137
3.3	0.00048	0.00097
3.4	0.00034	0.00067
3.5	0.00023	0.00047
3.6	0.00016	0.00032
3.7	0.00011	0.00022
3.8	0.00007	0.00014
3.9	0.00005	0.00010

Note: All values are computed by the authors using Excel. This is Table 7.1 in Chapter 7.

Z Scores	Probability	
	One-tailed	Two-tailed
0	0.50000	1.00000
0.1	0.46017	0.92034
0.2	0.42074	0.84148
0.3	0.38209	0.76418
0.4	0.34458	0.68916
0.5	0.30854	0.61708
0.6	0.27425	0.54851
0.7	0.24196	0.48393
0.8	0.21186	0.42371
0.9	0.18406	0.36812
1	0.15866	0.31731

Z Scores	Probability	
	One-tailed	Two-tailed
1.1	0.13567	0.27133
1.2	0.11507	0.23014
1.3	0.09680	0.19360
1.4	0.08076	0.16151
1.5	0.06681	0.13361
1.6	0.05480	0.10960
1.7	0.04457	0.08913
1.8	0.03593	0.07186
1.9	0.02872	0.05743
2	0.02275	0.04550
2.1	0.01786	0.03573
2.2	0.01390	0.02781
2.3	0.01072	0.02145
2.4	0.00820	0.01640
2.5	0.00621	0.01242
2.6	0.00466	0.00932
2.7	0.00347	0.00693
2.8	0.00256	0.00511
2.9	0.00187	0.00373
3	0.00135	0.00270
3.1	0.00097	0.00194
3.2	0.00069	0.00137
3.3	0.00048	0.00097
3.4	0.00034	0.00067
3.5	0.00023	0.00047
3.6	0.00016	0.00032
3.7	0.00011	0.00022
3.8	0.00007	0.00014
3.9	0.00005	0.00010

Note: All values are computed by the authors using Excel. This is Table 7.1 in Chapter 7.

APPENDIX 6: CRITICAL VALUES OF THE *T* DISTRIBUTION

TABLE A6.1 ■ Critical Values of the *t* Distribution

	Two-Tailed Test	
Degrees of Freedom	$\alpha = 0.05$	$\alpha = 0.01$
1	12.71	63.66
2	4.3	9.92
3	3.18	5.84
4	2.78	4.6
5	2.57	4.03
6	2.45	3.71
7	2.36	3.5
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.2	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.9
18	2.1	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.8
25	2.06	2.79
26	2.06	2.78
27	2.05	2.77
28	2.05	2.76
29	2.05	2.76
30	2.04	2.75
35	2.03	2.72
40	2.02	2.7
∞[Z]	1.96	2.58

Note: All values are computed by the authors using Excel.

	Two-Tailed Test	
Degrees of Freedom	**α = 0.05**	**α = 0.01**
1	12.71	63.66
2	4.3	9.92
3	3.18	5.84
4	2.78	4.6
5	2.57	4.03
6	2.45	3.71
7	2.36	3.5

	Two-Tailed Test	
Degrees of Freedom	**α = 0.05**	**α = 0.01**
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.2	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.9
18	2.1	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.8
25	2.06	2.79
26	2.06	2.78
27	2.05	2.77
28	2.05	2.76
29	2.05	2.76
30	2.04	2.75
35	2.03	2.72
40	2.02	2.7
∞[Z]	1.96	2.58

Note: All values are computed by the authors using Excel.

Example:

If $n = 30$ so that the degrees of freedom are 29, the positive and negative t values corresponding with 5% of the area under the tails is +2.05 and −2.05, as illustrated below.

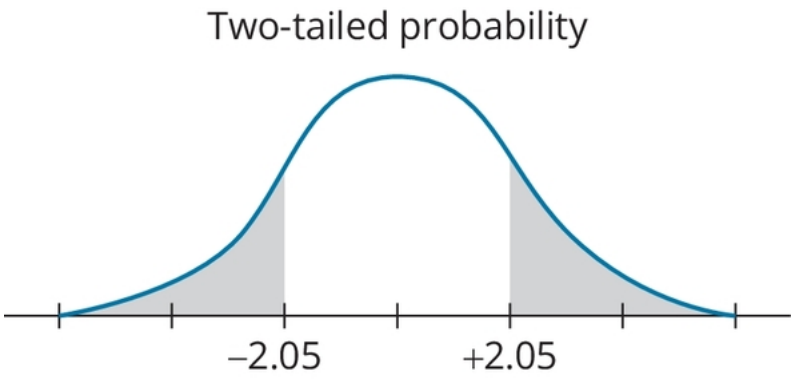


Figure A6.1 Two-tail probability when n = 30 and α = 0.05

APPENDIX 7: STATA CODE FOR RANDOM SAMPLING

This appendix provides Stata code to carry out different types of sampling. It is meant to accompany the discussion of sampling methods in Chapter 2. However, we have put this material in an appendix because it contains somewhat more advanced Stata code that will only be needed by students carrying out multistage random sample surveys.

In this appendix, we start with simple random sampling and proceed to discuss more complex types of sampling such as multistage sampling and stratification. We can use Stata or any spreadsheet software to randomly select the sample. Table A7.1 shows some useful functions in Stata.

TABLE A7.1 ■ Useful Functions In Stata For Sampling

Stata Function	Description	Examples
=runiform()	Creates a random number between 0 and 0.99999	**gen x=runiform()**
=runiform(a,b)	Creates a random number between *a* and *b* (including nonintegers)	**gen y=runiform(0.5,10.5)**
=runiformint(a,b)	Creates a random integer between *a* and *b*	**gen z=runiformint(1,150)**

Stata Function	Description	Examples
=runiform()	Creates a random number between 0 and 0.99999	**gen x=runiform()**
=runiform(a,b)	Creates a random number between *a* and *b* (including nonintegers)	**gen y=runiform(0.5,10.5)**
=runiformint(a,b)	Creates a random integer between *a* and *b*	**gen z=runiformint(1,150)**

SIMPLE RANDOM SAMPLING IN STATA

In this section, we work with the Stata file called “UScounties.dta” with a list of 3,142 counties in the United States. Initially, we wish to draw a simple random sample of 100 of them. The probability of selection for each county is 100/3142, or roughly 0.0318. We can select 100 counties randomly using the Stata code shown in Figure A7.1.

```
* Simple random sample
set seed 1234                // sets seed to ensure same sample each time
clear                        // clears data from memory
use "c:/UScounties.dta"      // opens file with list of counties
local n = 100                // defines desired sample size
gen sorter = runiform()      // defines "sorter" to be random over 0-1
sort sorter                  // sort by random number
gen select=0                 // create dummy indicating selected units
replace select=1 if _n<=`n'  // selects first `n' units randomly
list if select==1            // show list of selected counties
```

Description

Figure A7.1 Simple Random Sample

The function “**runiform()**” creates a random number uniformly distributed between 0 and 1 for each observation—that is, for each county. A uniform random number is a random number with equal

probability throughout the range. The variable "_n" is a Stata variable equal to the number of observations in the database.

The do-file in Figure A7.1 gives every county a random number between 0 and 1, then sorts by this number, so that the counties are now in random order. To pick a random sample of 100, the code just selects the first 100 counties from the newly sorted list.[1]

SYSTEMATIC RANDOM SAMPLING IN STATA

The systematic random sample can be implemented easily in Stata. We need to calculate a starting point, which identifies the first unit to be selected, and an interval, which determines the gap between subsequent units. If the population is *N* and the desired sample is *n*, then the interval between selected units is *N*/*n*, and the starting point is a random number between 1 and *N*/*n*.

Suppose we are working with the same list of 3,142 counties and wish to select 100 of them. The Stata code in Figure A7.2 will select a systematic random sample.

```
* Systematic random sample
set seed 1234                              // sets seed to ensure same sample each time
clear                                      // clears data from memory
use "c:/UScounties.dta"                    // opens file with list of counties
local n = 100                              // defines desired sample size
local interval = _N/`n'                    // defines interval between selected units
local start = runiform(0,`interval')   /   // defines random starting point
gen select = mod(_n-`start',`interval')<1  // selects `n' counties systematically
list if select==1                          // shows list of selected counties
```

Description

Figure A7.2 Single-Stage Systematic Random Sample

In this case, the interval is 3142/100 or 31.42. This means that each county (after the first one) will be 31 or 32 counties down the list from the previous one.

We identify the first unit to be selected with local 'start.' We need to use a local macro because we want just one random number for the start value. Note that when referring to local macros, the name needs to be placed inside left and right single quotation marks.

The next line defines the "select" variable to be 1 if the inequality is true and 0 if it is false. The **mod(x,y)** function calculates the modulus, defined as the difference between *x* and the largest multiple of *y* that is less than *x*. (It can also be calculated as **x-y*int(x/y)**. The expression _n is a Stata variable indicating the observation number. Whenever the modulus is less than 1, the expression (_n − start) has passed another multiple of the interval, and it is time to select the county.

Earlier, we said that if the sampling frame is sorted by a variable, a systematic random sample spreads out the sample across the values of that variable. In this case, the sampling frame is sorted by state, from Alabama to Wyoming. A systematic random sample ensures (roughly) proportional representation of each state. For example, Alabama has 67 counties, 2.1% of the counties in the country. With an interval of 31.42, a systematic random sample will always include two or possibly three counties from Alabama, corresponding to 2% or 3% of the sample. In contrast, a simple random sample might, by chance, exclude all the counties in Alabama, or it might conceivably include all 67 of them. Systematic random sampling is widely used in surveys because of this advantage and because it is only slightly more complex than simple random sampling.

MULTISTAGE SAMPLING IN STATA

To demonstrate how to carry out multistage sampling in Stata, we will use the code for systematic sampling for each stage. Figure A7.3 shows a set of Stata commands that selects 20 states and then selects five counties in each state, using systematic random sampling in each level.

```
* Multi-stage systematic sampling
* Select n1 states
set seed 1234                                  // sets seed to ensure same sample each time
clear                                          // clear data from memory
use "c:/UScounties.dta"                        // opens file with list of counties
local n1 = 20                                  // set number of states to select
local n2 = 5                                   // set number of counties/state to select
drop if state==9                               // drop DC (only 1 county)
collapse (sum) population, by(state)           // collapse to state level
local interval1 = _N/`n1'                      // calculate interval between selected states
local start1 = runiform(0,`interval1')         // generate random number for first state
gen select1 = mod(_n-`start1',`interval1')<1   // selects states
keep if select1==1                             // keep only selected states
list                                           // list selected states
gen statenbr = _n                              // create state counter for later
save "c:/SelectedStates", replace              // save list of selected states for later
* Select n2 counties in each state clear
use "c:/UScounties.dta"                        // opens file with list of counties
drop if state==9                               // drop DC
merge m:1 state using "c:/SelectedStates"      // merge counties with selected states
keep if select1==1                             // keep only counties in selected states
gen interval = .                               // define interval variable
gen select2 = .                                // define county selection variable
forvalues s = 1/`n1' {                         // loop through s=1 to n1 states
    preserve                                   // save data in memory to be restored later
    keep if statenbr==`s'                      // keep only the state numbered `s'
    local interval2 = _N/`n2'                  // calculate interval between counties
    local start2 = runiform(0,`interval2')     // generate random number for first county
    replace select2 = mod(_n-`start2',`interval2')<1  // select counties
    list state county if select2==1            // list state & selected counties
    restore                                    // restore data in memory (all selected states)}
```

Description

Figure A7.3 Multistage Sampling With Stata

The first section of the do-file selects the 20 states. It starts with the file containing the list of counties and then collapses to the state level. In other words, in the original file, each observation is a county, but after the collapse, each observation is a state. After the **collapse** command, the state selection is similar to the county selection in Figure A7.2. The do-file specifies the desired number of states, calculates the interval between states, generates a random starting point, and then uses the **mod(x,y)** function to select the states.

The second section of the code selects five counties in each of the 20 selected states. We use the **merge** command to combine the list of selected states and the full list of counties. Then we drop the counties that are in states that were not selected. The loop goes through the 20 states, carrying out a systematic random selection of five counties in each. The **preserve** command stores the data in memory, just before deleting the data for all but one of the states. After selecting and displaying the five counties in that state, the **restore** command brings back the data for all 20 states and the loop goes on to the next state.

In the end, the result is a list of 100 counties, composed of five counties from each of 20 states. Note that Delaware has only three counties, so if it were one of the selected states, the do-file would select a total of 98 counties.

STRATIFIED SAMPLING IN STATA

How do we draw a stratified sample using Stata? To keep it simple, we will consider a single-stage sampling with two strata. Suppose we decide that we want to oversample counties with large populations, either because we are particularly interested in those counties or because we believe that large counties are more diverse, so our variables of interest have greater variance in large counties. As shown in Figure A7.4, the first step is to define large counties. We generate a new variable “size,” equal

to 0 if the population is less than 500,000 and 1 if the population is greater than or equal to 500,000. We use “preserve” to store a copy of the data (with the “size” variable). Next, large counties are removed from the data, and 60 small counties are selected by systematic random sampling using the commands in Figure A7.2. After restoring the full set of counties, we remove the small counties and repeat the process, selecting 40 of the large counties.

```
* Stratification
set seed 1234                                  // saets seed to ensure same sample each time
clear                                          // clears data from memory
use "c:/UScounties.dta"                        // opens file with list of counties
gen size = (population>=500000)                // define "size" as 0 if <500k, 1 if >=500k
local n1 = 60                                  // defines sample for small counties
local n2 = 40                                  // defines sample for large counties
* Small counties
preserve                                       // save copy of data
keep if size==0                                // keep only small counties
local interval1 = _N/`n1'                      // defines interval between selected units
local start1 = runiform(0,`interval1')   // defines numbers used to start selection
gen select = mod(_n-`start1',`interval1')<1    // selects small counties
list state county population if select==1      // lists names of selected small counties
restore                                        // restore data to point of preserve
* Large counties
keep if size==1                                // keep only large counties
local interval2 = _N/`n2'                      // defines interval between selected units
local start2 = runiform(0,`interval2')         // defines rand nbrs to start selection
gen select = mod(_n-`start2',`interval2')<1    // selects large counties
list state county population if select==1      // lists names of selected large counties
```

Description

Figure A7.4 Stratified Random Sampling With Stata

There are only 134 large counties, 4.26% of the total. Because we stratified and oversampled large counties, they represent 40% of the sample. In contrast, small counties account for 95.74% of all counties but just 60% of the sample. Sampling weights can be used to calculate averages and percentages for the population that compensate for the overrepresentation of large counties in the sample.

KEY TERM

macro

Descriptions of Images and Figures

Back to Figure

* Simple random sampleset seed 1234 // sets seed to ensure same sample each time

clear // clears data from memory

use "c:/UScounties.dta" // opens file with list of counties

local n = 100 // defines desired sample size

gen sorter = runiform() // defines “sorter” to be random over 0-1

sort sorter // sort by random number

gen select=0 // create dummy indicating selected units

replace select=1 if _n<=`n’ // selects first `n’ units randomly

```
list if select==1 // show list of selected counties
```

Back to Figure

```
* Systematic random sample
set seed 1234 // sets seed to ensure same sample each time
clear // clears data from memory
use "c:/UScounties.dta" // opens file with list of counties
local n = 100 // defines desired sample size
local interval = _N/`n' // defines interval between selected units
local start = runiform(0,`interval') / // defines random starting point
gen select = mod(_n-`start',`interval')<1 // selects `n' counties systematically
list if select==1 // shows list of selected counties
```

Back to Figure

```
* Multi-stage systematic sampling
* Select n1 states
set seed 1234 // sets seed to ensure same sample each time
clear // clear data from memory
use "c:/UScounties.dta" // opens file with list of counties
local n1 = 20 // set number of states to select
local n2 = 5 // set number of counties/state to select
drop if state==9 // drop DC (only 1 county)
collapse (sum) population, by(state) // collapse to state level
local interval1 = _N/`n1' // calculate interval between selected states
local start1 = runiform(0,`interval1') // generate random number for first state
gen select1 = mod(_n-`start1',`interval1')<1 // selects states
keep if select1==1 // keep only selected states
list // list selected states
gen statenbr = _n // create state counter for later
save "c:/SelectedStates", replace // save list of selected states for later
* Select n2 counties in each state clear
```

```
use "c:/UScounties.dta" // opens file with list of counties
drop if state==9 // drop DC
merge m:1 state using "c:/SelectedStates" // merge counties with selected states
keep if select1==1 // keep only counties in selected states
gen interval = . // define interval variable
gen select2 = . // define county selection variable
forvalues s = 1/`n1' { // loop through s=1 to n1 states
preserve // save data in memory to be restored later
keep if statenbr==`s' // keep only the state numbered `s'
local interval2 = _N/`n2' // calculate interval between counties
local start2 = runiform(0,`interval2') // generate random number for first county
replace select2 = mod(_n-`start2',`interval2')<1 // select counties
list state county if select2==1 // list state & selected counties
restore // restore data in memory (all selected states)}
}
```

Back to Figure

```
* Stratification
set seed 1234 // saets seed to ensure same sample each time
clear // clears data from memory
use "c:/UScounties.dta" // opens file with list of counties
gen size = (population>=500000) // define "size" as 0 if <500k, 1 if >=500k
local n1 = 60 // defines sample for small counties
local n2 = 40 // defines sample for large counties
* Small counties
preserve // save copy of data
keep if size==0 // keep only small counties
local interval1 = _N/`n1' // defines interval between selected units
local start1 = runiform(0,`interval1') // defines numbers used to start selection
gen select = mod(_n-`start1',`interval1')<1 // selects small counties
```

```
list state county population if select==1 // lists names of selected small counties
restore // restore data to point of preserve
* Large counties
keep if size==1 // keep only large counties
local interval2 = _N/`n2' // defines interval between selected units
local start2 = runiform(0,`interval2') // defines rand nbrs to start selection
gen select = mod(_n-`start2',`interval2')<1 // selects large counties
list state county population if select==1 // lists names of selected large counties
```

APPENDIX 8: EXAMPLES OF NONLINEAR FUNCTIONS

In Chapter 13, we discuss different ways to estimate nonlinear relationships using linear regression by transforming the dependent variable (y) and/or the independent variables (x_i). To illustrate the shape of these nonlinear functions, we provide graphs of each type, along with the Stata code to generate the graphs.

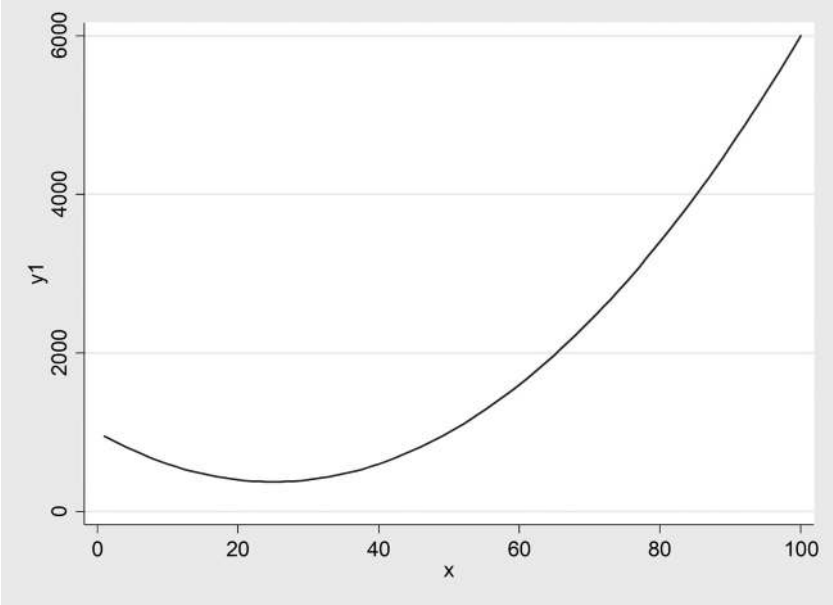
QUADRATIC FUNCTIONS

One common way to estimate a nonlinear relationship between y and x is to add powers of x, such as x^2 and x^3, to the regression equation as independent variables. Here, we consider the case of a quadratic equation, which takes the following form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

How do we graph this function using Stata? First, we generate a Stata data set with the variable x that runs from 1 to 100. The **set obs #** command (where # is a number) creates an empty data file with # observations. We then use the special Stata variable _n, which represents the row number, to create values of x from 1 to 100. Next, we define the y variable, choosing values for the three coefficients: β_0, β_1, and β_2. After defining the y variable, we graph x and y as a line graph.[1] A quadratic function with a positive value for β_2 creates a U-shaped graph, as shown in Figure A8.1.

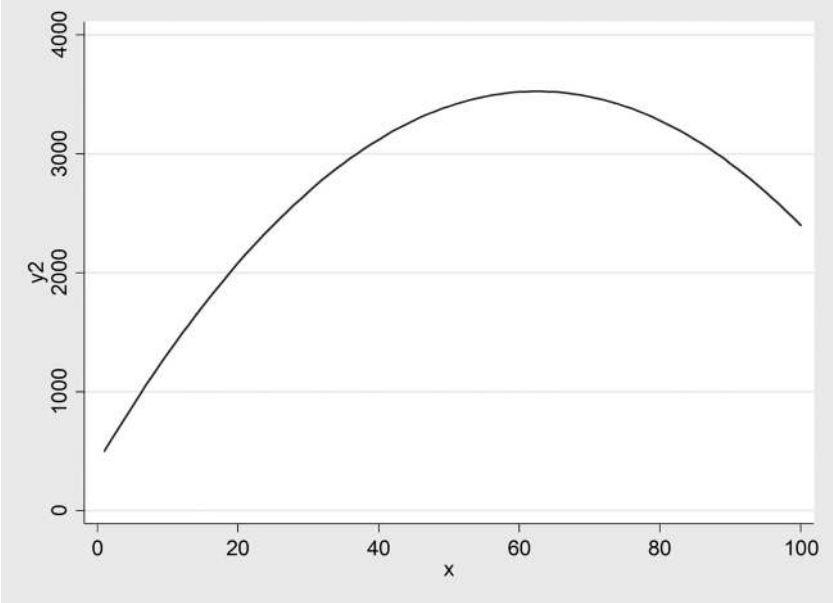


Description

Figure A8.1 Quadratic Function With Positive Quadratic Coefficient

```
clear
set obs 100
gen x = _n
gen y1 = 1000 - 50*x + x2
twoway (line y1 x)
```

If β_2 (the quadratic coefficient) is negative, as shown in the **generate** command below, the result is an inverse U shape, as shown in Figure A8.2.



Description

Figure A8.2 Quadratic Function With Negative Quadratic Coefficient

```
gen y2 = 400 + 100*x - 0.8*x2
twoway (line y2 x)
```

We are often interested in the marginal effect of *x* on *y*. In other words, what is the effect of a one-unit change in *x* on *y*. Graphically, this is the slope of the graph of *y* on the vertical axis and *x* on the horizontal axis. In a linear function, the marginal effect is simply the coefficient on the *x* variable, and it is constant. But in a nonlinear relationship, the marginal effect of *x* on *y* changes over the ranges of *x*. We can calculate the marginal effect using calculus. In the case of a quadratic equation with one independent variable,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

and the marginal effect is expressed as

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

This means the marginal effect is a function of the value of *x*. If $\beta_2 > 0$, then the marginal effect (or slope) rises as *x* increases, and the graph of *y* against *x* has a U shape. On the other hand, if $\beta_2 < 0$, the slope falls as *x* increases and the graph has an inverted U shape (∩). We can calculate the "turning point" where the slope is horizontal by setting the marginal effect to 0 and solving for *x*.

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x = 0$$

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x = 0$$

$$x = \frac{-\beta_1}{2\beta_2}$$

$$2\beta_2$$

For example, the estimated equation for the data shown in Figure A8.2 is as follows:

$$y = 400 + 100x - 0.8x^2$$

$$y = 400 + 100x - 0.8x^2$$

This means that $\beta_0 = 400$, $\beta_1 = 100$, $\beta_2 = -0.8$. Using the equation for the marginal effect,

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x = 100 + (2)(-0.8)x = 100 - 1.6x$$

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x = 100 + (2)(-0.8)x = 100 - 1.6x$$

The turning point is where $x = -\beta_1/2\beta_2 = -100/((2)(-0.8)) = 62.5$. The turning point is consistent with the curve shown in Figure A8.2.

SEMILOG FUNCTIONS

Another way to estimate a nonlinear function with linear regression is to transform the dependent variable by taking its natural logarithm. Here is the general form:

$$log(y) = \beta_0 + \beta_1 x$$

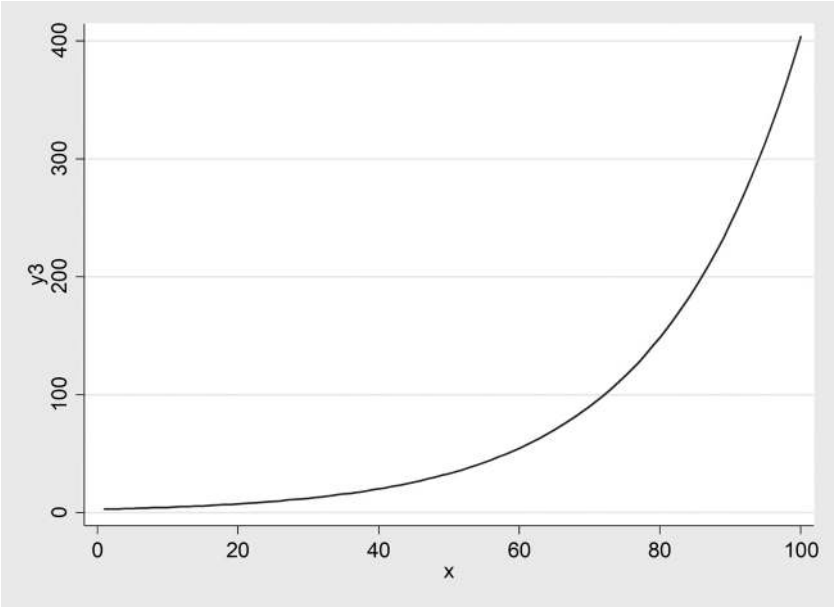
$$log(y) = \beta_0 + \beta_1 x$$

Taking the exponential function of both sides, we can isolate y as follows:

$$y = exp(\beta_0 + \beta_1 x)$$

$$y = exp(\beta_0 + \beta_1 x)$$

where exp() raises e to the power of the expression in parentheses. If β_1, the coefficient on x is positive, and the relationship will be rising at an increasing rate, as shown in Figure A8.3.

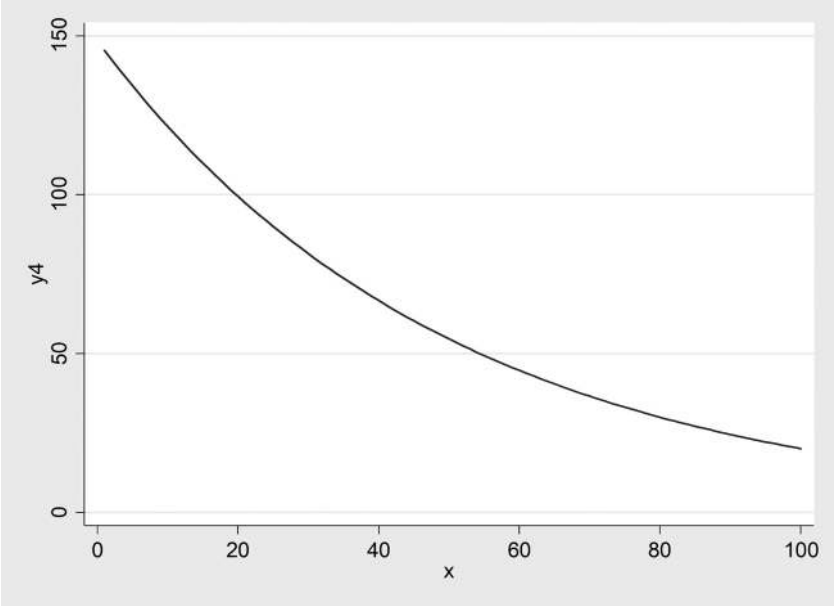


Description

Figure A8.3 Semilog Function Using Log(Y) And Positive Coefficient

```
gen y3 = exp(1 + 0.05*x)
twoway (line y3 x)
```

If we use the same functional form, but the β_1 coefficient is negative, the line slopes down but never crosses the horizontal (x) axis (see Figure A8.4). To be more precise, for each one-unit increase in x, y declines by a fixed proportion.



Description

Figure A8.4 Semilog Function Using Log(Y) And Negative Coefficient

```
gen y4 = exp(5 - 0.02*x)
twoway (line y4 x)
```

What is the marginal effect of this type of semilog function? If the function is

$$log(y) = \beta_0 + \beta_1 x$$

$$log(y) = \beta_0 + \beta_1 x$$

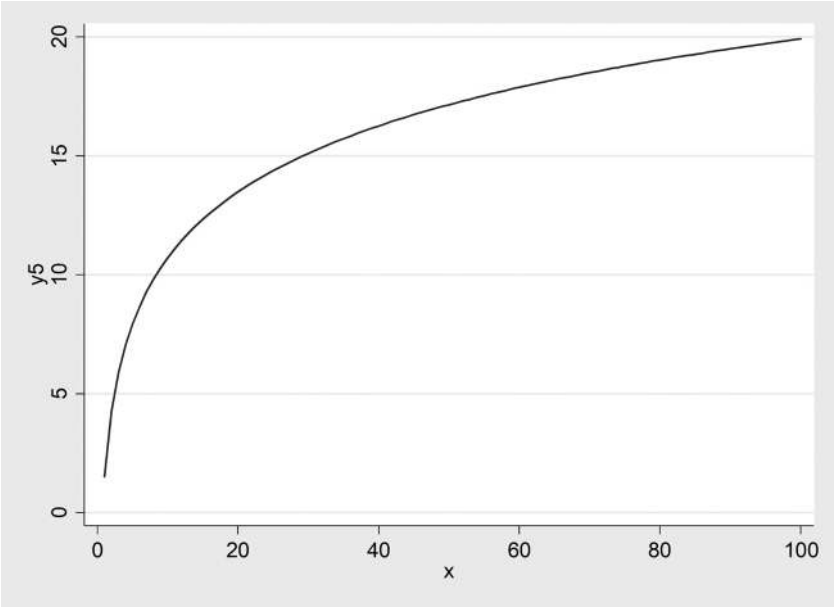
then the marginal effect is

$$\frac{\partial y}{\partial x} = \beta_1 y$$

$$\frac{\partial y}{\partial x} = \beta_1 y$$

In other words, for each unit increase in *x*, the value of *y* increases or decreases by a constant proportion, which is determined by β_1. In the example above, $\beta_1 = -0.02$, so *y* declines by about 2% for each one-unit increase in *x*. Because of compounding, the decline is actually about 1.98% per unit. In other words, as *x* changes from, say, 50 to 51, the value of *y* decreases, which in turn lowers the rate of change in *y*.

Another nonlinear function can be created by having the logarithm of *x* on the right side. With a positive coefficient on log(*x*), the curve takes the form shown in Figure A8.5. The value of *y* rises indefinitely, never reaching a maximum.

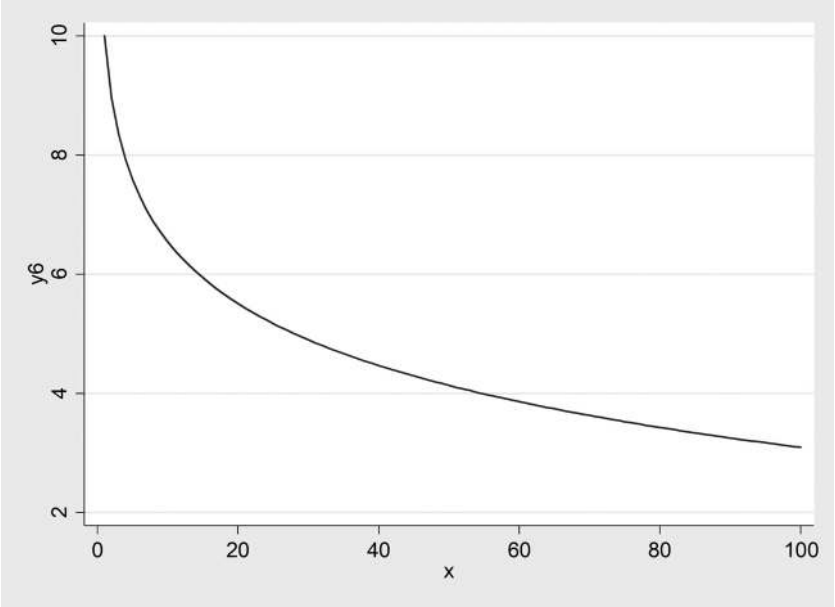


Description

Figure A8.5 Semilog Function Using Log(X) And Positive Coefficient

```
gen y5 = 1.5 + 4*log(x)
twoway (line y5 x)
```

Alternatively, if the coefficient is negative, the curve slopes down, as shown in Figure A8.6.



Description

Figure A8.6 Semilog Function Using Log(X) And Negative Coefficient

```
gen y6 = 10 - 1.5*log(x)
twoway (line y6 x)
```

The marginal effect of this type of semilog function can be calculated as follows:

$$\frac{\partial y}{\partial x} = \frac{\beta_1}{x}$$

$$\partial x \qquad x$$

In the example above, $\beta_1 = -1.5$, so the marginal effect is $-1.5/x$. If $x = 30$, then the marginal effect is -0.05.

DOUBLE-LOG FUNCTIONS

Another functional form is the double-log function, in which both *y* and *x* are transformed into logarithms:

$$log(y) = \beta_0 + \beta_1 log(x)$$

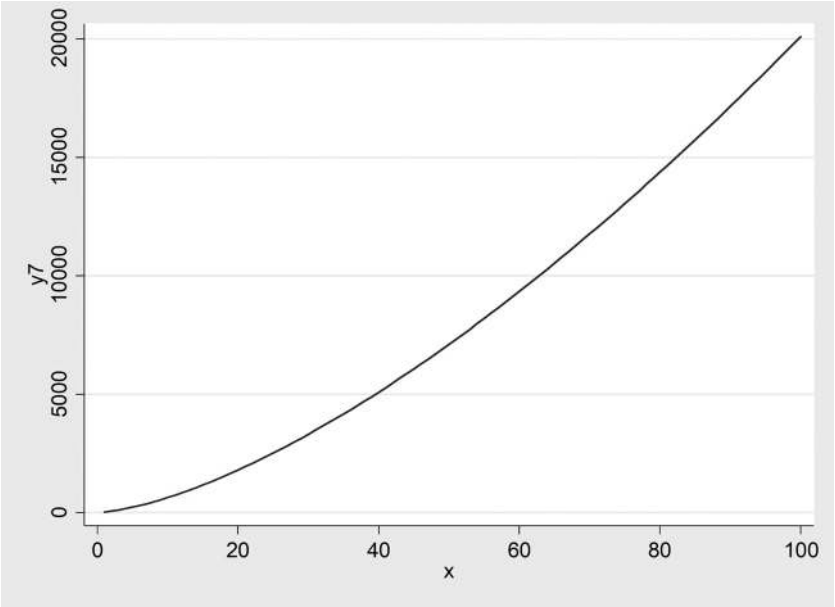
$$log(y) = \beta_0 + \beta_1 log(x)$$

As described previously, we need to take the exponential function of both sides in order to express the equation in terms of *y*.

$$y = exp\left(\beta_0 + \beta_1 log(x)\right) = exp\left(\beta_0\right) x^{\beta}{}_1 = \alpha x^{\beta}{}_1$$

$$y = exp(\beta_0 + \beta_1 log(x)) = exp(\beta_0) x^{\beta}{}_1 = \alpha x^{\beta}{}_1$$

where $\alpha = \exp(\beta_0)$. If the coefficient β_1 is greater than 1, the function will rise at an increasing rate, as shown in Figure A8.7.

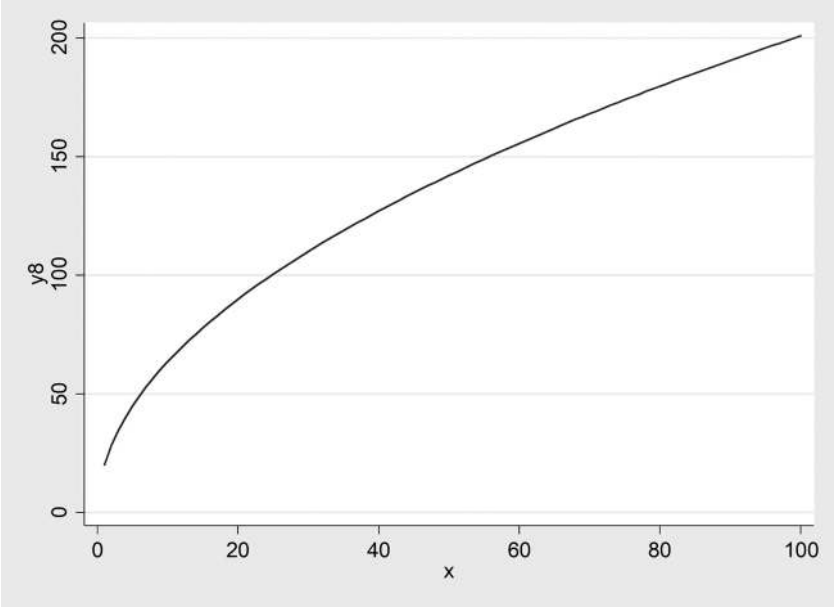


Description

Figure A8.7 Double-Log Function With Positive Coefficient Greater Than 1

```
gen y7 = exp(3 + 1.5*log(x))
twoway (line y7 x)
```

If the coefficient is positive but less than 1, the function rises but at a decreasing rate, as shown in Figure A8.8.

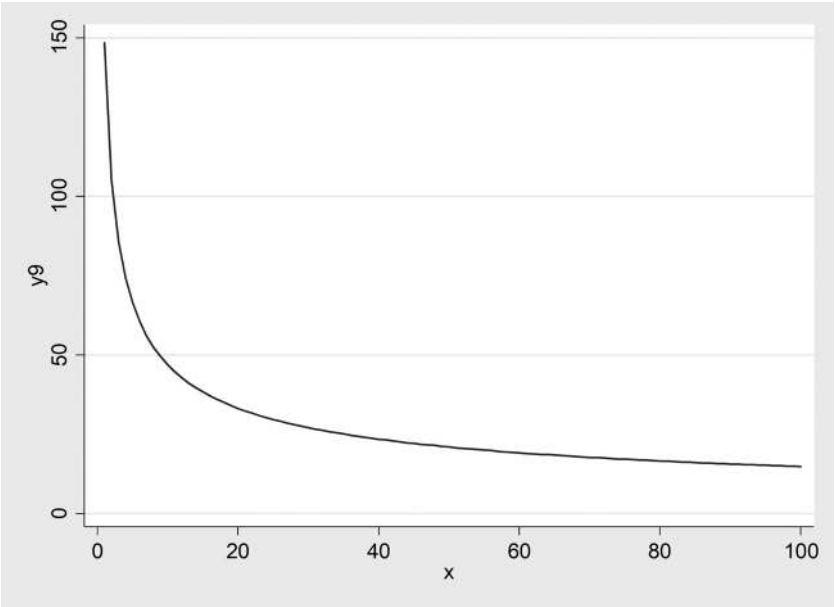


Description

Figure A8.8 Double-Log Function With Positive Coefficient Less Than 1

```
gen y8 = exp(3 + 0.5*log(x))
twoway (line y8 x)
```

On the other hand, if the coefficient is negative, then the function declines but never crosses the horizontal (*x*) axis, as shown in Figure A8.9.



Description

Figure A8.9 Double-Log Function With Negative Coefficient

```
gen y9 = exp(5 - 0.5*log(x))
twoway (line y9 x)
```

The marginal effect of the double-log functional form can be calculated as follows:

$$\frac{\partial y}{\partial x} = \beta_1 \frac{y}{x}$$

$$\frac{}{\partial x} = \beta_1 \frac{}{x}$$

This can be rewritten as follows:

$$\beta_1 = \frac{\partial y / y}{\partial x / x}$$

$$\beta_1 = \frac{\partial y / y}{\partial x / x}$$

This means that the coefficient represents the ratio of the proportional change in *y* divided by the proportional change in *x*, also called the elasticity of *y* with respect to *x*. In other words, one of the characteristics of the double-log functional form is that the elasticity is constant. In the example above, $\beta_1 = -0.5$. This means that if *x* increases by 1%, *y* will decrease by 0.5%, and this relationship holds throughout the range of *x*.

Table A8.1 summarizes the nonlinear functions discussed here and, for each one, the expression for calculating marginal effect of *x* on *y*. In each case, the marginal effect varies with different values of *x*.

TABLE A8.1 ■ Common Nonlinear Functions And The Marginal Effects Of Each

Name of Functional Form	Functional Form	Marginal Effect
Quadratic	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$
Linear with interaction of two variables	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$
Semilog [log *y*]	$log(y) = \beta_0 + \beta_1 x$	$\frac{\partial y}{\partial x} = \beta_1 y$
Semilog [log *x*]	$y = \beta_0 + \beta_1 log(x)$	$\frac{\partial y}{\partial x} = \frac{\beta_1}{x}$
Double log	$log(y) = \beta_0 + \beta_1 log(x)$	$\frac{\partial y}{\partial x} = \beta_1 \frac{y}{x}$

Name of Functional Form	Functional Form	Marginal Effect
Quadratic	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$
Linear with interaction of two variables	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$
Semilog [log *y*]	$log(y) = \beta_0 + \beta_1 x$	$\frac{\partial y}{\partial x} = \beta_1 y$
Semilog [log *x*]	$y = \beta_0 + \beta_1 log(x)$	$\frac{\partial y}{\partial x} = \frac{\beta_1}{x}$
Double log	$log(y) = \beta_0 + \beta_1 log(x)$	$\frac{\partial y}{\partial x} = \beta_1 \frac{y}{x}$

Descriptions of Images and Figures

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y1axis has values 0 to 6000 in intervals of 2000. A U-shaped graph starts from 0 at 1000, a dip at 20 at the value of 500, and reaches 100 at 6000.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y2axis has values 0 to 4000 in intervals of 1000. A U-shaped graph starts from 0 at 500, goes up with the peak value of 3500 at 60, and reaches 100 at 2500.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y3axis has values 0 to 400 in intervals of 100. An upward curve in the right direction reaches 400 at 100. The values at 20, 40, 60, and 80 are approximately 10, 25, 50, 125.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y4axis has values 0 to 150 in intervals of 50. A downward curve in the right direction reaches 20 at 100. The values at 0, 20, 40, 60, and 80 are approximately 148, 90, 60, 40, and 30.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y5axis has values 0 to 20 in intervals of 5. An upward curve in the right direction reaches 20 at 100. The values at 0, 20, 40, 60, and 80 are approximately 2, 13, 16, 18, and 19.5.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y6axis has values 0 to 10 in intervals of 2. A downward curve in the right direction reaches 3 at 100. The values at 0, 20, 40, 60, and 80 are approximately 10, 5.5, 4.5, 3.8, and 3.5.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y7axis has values 0 to 20000 in intervals of 5000. An upward curve in the right direction reaches 20000 at 100. The values at 0, 20, 40, 60, and 80 are approximately 0, 2400, 5000, 10000, and 15000.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y8axis has values 0 to 200 in intervals of 50. A downward curve in the right direction reaches 200 at 100. The values at 0, 20, 40, 60, and 80 are approximately 25, 90, 125, 150, and 180.

Back to Figure

The horizontal x-axis has values from 0 to 100 in intervals of 20. The vertical y9axis has values 0 to 150 in intervals of 50. A downward curve in the right direction reaches 18 at 100. The values at 0, 20, 40, 60, and 80 are approximately 150, 35, 25, 20, and 19.

APPENDIX 9: ESTIMATING THE MINIMUM SAMPLE SIZE

In Chapter 2, we discussed the factors that influence the necessary sample size. In that chapter, we used the example of a survey of recent college graduates designed to see whether there is a difference in salaries between men and women. In this appendix, we show how to calculate the minimum sample size needed to achieve the desired level of precision in the results. These are called power calculations.

Table A9.1 describes five factors that help determine the minimum sample size needed. On the left side, we repeat the description of the factors influencing the sample size from Chapter 2 in terms of the study of gender differences in salaries. On the right side, we present the more technical description of the five factors.

TABLE A9.1 ■ Factors Influencing the Minimum Sample Size

Intuitive Explanation	Technical Explanation
How small a difference in salaries do we want to be able measure?	Minimum detectable effect size
How much variation is there in salaries?	Standard deviation of the variable of interest
How small should the probability be of incorrectly concluding that there *is* a difference between men and women?	Alpha is the maximum probability of Type I error that we are willing to accept.
How small should the probability be of making a mistake when we state that there is *no* difference between men and women?	Beta is the maximum Type II error that we are willing to accept. The power of the test is 1−β.
How was the sample selected?	Design effect

Intuitive Explanation	Technical Explanation
How small a difference in salaries do we want to be able measure?	Minimum detectable effect size
How much variation is there in salaries?	Standard deviation of the variable of interest
How small should the probability be of incorrectly concluding that there *is* a difference between men and women?	Alpha is the maximum probability of Type I error that we are willing to accept.
How small should the probability be of making a mistake when we state that there is *no* difference between men and women?	Beta is the maximum Type II error that we are willing to accept. The power of the test is 1−β.
How was the sample selected?	Design effect

According to the National Association of Colleges and Employers (2018), the average salary of a student who graduated in 2017 was $51,022. Suppose we expect our sample of graduates to be close to the national average in salaries and want to be able to detect a gender gap of at least 5%, or (roughly) $2,500.

The standard deviation of the salaries of college graduates must be obtained from secondary data. Suppose we find that the standard deviation of these salaries is $11,000.

A Type I error is to reject the null hypothesis (no gender difference) when it is true. In this case, the Type I error would be to reject the equality of male and female wages when in fact they are equal. The

maximum acceptable probability of a Type I error is labeled alpha, or α. The convention in the social sciences is to set α = 0.05—that is, to reject the null hypothesis only if the risk of being wrong is less than 0.05.

Type II error is the risk of not rejecting the null hypothesis when it is false. In our case, it is the risk of concluding that there is no gender difference in salaries when in fact there is one. The maximum allowable probability of Type II error is labeled beta, or β. Researchers often set β at 0.20, although it depends on the "cost" of being wrong. It is worth noting that the power of the test is 1 − β, or 0.80 in this case.

The design effect is an adjustment for the sampling design, which may increase or reduce the precision of estimates relative to a simple random sample. In this case, we will assume that we are able to draw a simple random sample of recent graduates, so we do not need to take into account the design effect.

The command in Stata for carrying out power calculations is **power**. It can be used to estimate the sample size based on the size effect, standard deviation, and levels of alpha and beta. Figure A9.1 shows the command and the resulting output. Translating into English, the command says, "What is the sample size needed to detect a difference of $2,500 if the mean salary is $50,000, the standard deviation is $11,000, the maximum acceptable probability of Type I error is 0.05, and the maximum acceptable probability of Type II error is 0.20, assuming a simple random sample?"

```
. power twomeans 50000, sd(11000) a(0.05) b(0.20) diff(2500)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
         beta =    0.2000
        delta = 2500.0000
           m1 =  5.00e+04
           m2 =  5.25e+04
         diff = 2500.0000
           sd =  1.10e+04

Estimated sample sizes:

            N =       610
  N per group =       305
```

Description

Figure A9.1 Power Calculations to Derive Sample Size

The output repeats the values of the parameters being used to calculate the sample size. The result is shown at the bottom: We need a sample size of 610 graduates, including 305 men and 305 women. Note that we did not actually need to include the options **a(0.05)** and **b(0.20)** because these are the defaults. If we leave out these options, Stata will adopt α = 0.05 and β = 0.20.

Stata allows us to carry out multiple power calculations with one command by inserting number lists in **a(), b(), sd()**, and **diff()**. A number list can be a series of numbers, such as "10, 20, 30, 40, 50," or it can

be a range with step value such as “10(10)50,” which means from 10 to 50 in increments of 10. Furthermore, if number lists are put into multiple options, Stata will carry out the calculations for all combinations. In the example that follows, we check four different size effects and two levels of alpha. The output is eight sample sizes, one for each combination of the four effect sizes and the two levels of alpha.

```
. power twomeans 50000, sd(11000) a(0.05 0.01) b(0.20) diff(2000(1000)5000)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1
```

alpha	beta	N	N1	N2	delta	m1	m2
.05	.2	952	476	476	2000	50000	52000
.05	.2	426	213	213	3000	50000	53000
.05	.2	240	120	120	4000	50000	54000
.05	.2	154	77	77	5000	50000	55000
.01	.2	1418	709	709	2000	50000	52000
.01	.2	632	316	316	3000	50000	53000
.01	.2	358	179	179	4000	50000	54000
.01	.2	230	115	115	5000	50000	55000

diff	sd
2000	11000
3000	11000
4000	11000
5000	11000
2000	11000
3000	11000
4000	11000
5000	11000

Description

Figure A9.2 Power Calculations to Derive Sample Size with Multiple Parameters

The column “N” indicates the sample size needed for each value of delta (the effect size) and alpha. For example, if we want to detect salary differences down to $2,000 at the 1% confidence level, we need a sample of 1,418 graduates. At the other extreme, if we only need to detect a salary difference of $5,000 or more at the 5% confidence level, then a sample of just 154 graduates would suffice.

In this case, the power calculations were carried out to compare two sample means, hence the **twomeans** option. However, the **power** command will also carry out power calculations for other types of statistical tests:

onemean—Comparison of a sample mean with a fixed number. Example: Is the average salary for graduates from this college greater than $50,000?

oneprop—Comparison of a sample mean with a fixed proportion. Example: Is the unemployment rate for graduates from this college greater than 5%?

twoprop—Comparison of two sample proportions. Example: Is the unemployment rate different for male and female graduates?

The **power** command is quite flexible and can be used in many other ways. Take the following examples:

> It can generate graphs or tables, giving the sample size required for different values of alpha, beta, or the standard deviation.
>
> It will also calculate the minimum detectable size effect based on the sample size, the standard deviation, and alpha and beta.
>
> It can calculate the power of the test (defined as 1 − β) based on the size effect, the standard deviation, and alpha.
>
> Starting with Stata 15, it is possible to incorporate the sampling design into the power calculations, taking into account the design effect—that is, the effect of clustering and stratification on the relationship between precision, risk of Types I and II error, and sample size.

In summary, the **power** command is a useful tool in the design of surveys for examining the relationship between sample size and level of precision estimating parameters and carrying out statistical tests.

Descriptions of Images and Figures

Back to Figure

The calculation is as follows:

.power twomeans 5000, sd(11000) a(0.05) b(0.20) diff(2500)

Performing iteration…

Estimated sample sizes for a two-sample means test

t test asuming sd1 = sd2 = sd

Ho: m2 = m1 versus Ha: m2 !=m1

Study parameters:

alpha = 0.500

beta = 0.2000

delta = 2500.0000

m1 = 5.00e+04

m2 = 5.25e+04

diff = 2500.0000

sd = 1.10e+04

Estimated sample sizes:

N = 610

N per group = 305

Back to Figure

The calculation is as follows:

.power twomeans 5000, sd(11000) a(0.05 0.01) b(0.20) diff(2000 (1000) 5000)

Performing iteration…

Estimated sample sizes for a two-sample means test

t test asuming sd1 = sd2 = sd

Ho: m2 = m1 versus Ha: m2 !=m1

alpha	beta	N	N1	N2	delta	m1	m2
.05	.2	952	476	476	2000	50000	52000
.05	.2	426	213	213	3000	50000	53000
.05	.2	240	120	120	4000	50000	54000
.05	.2	154	77	77	5000	50000	55000
.01	.2	1418	709	709	2000	50000	52000
.01	.2	632	316	316	3000	50000	53000
.01	.2	358	179	179	4000	50000	54000
.01	.2	230	115	115	5000	50000	55000

diff	sd
2000	11000
3000	11000
4000	11000
5000	11000
2000	11000
3000	11000
4000	11000
5000	11000

APPENDIX 10: DESCRIPTION OF THE DATA SETS USED IN THE TEXTBOOK

2014 ASQ - ADMITTED STUDENT QUESTIONNAIRE

5,814 observations

66 variables

https://professionals.collegeboard.org/higher-ed/recruitment/asq

The Admitted Student Questionnaire (ASQ) is a market research tool offered to academic institutions by the College Board. It allows universities and colleges to explore the factors that led to student decisions about how they choose colleges. In 2014, there were 10 institutions with a total of 5,814 students that chose to use the ASQ.

COLLEGE SCORECARD APRIL 2023 – USNEWS

1,480 four-year colleges

45 variables

https://collegescorecard.ed.gov

https://www.usnews.com/best-colleges

This data set is a combination of the College Scorecard data set from April 2023 and the *U.S. News and World Report* data on best colleges from 2023. The College Scorecard data were created by the U.S. government to help students compare colleges based on various factors, including debt upon graduation, salaries six years after graduation, and size. The data include more than 6,000 technical schools, colleges, and universities (including beauty schools, massage therapy schools, etc.). The *U.S. News and World Report* data from 2023 rank 1,859 four-year colleges based on factors such as graduation rates, faculty resources, reputation, and financial resources. The type and rank of each institution based on *U.S. News and World Report* were added to the College Scorecard data, and only those colleges that are listed both in *U.S. News and World Report* and on the College Scorecard (1,480) are included in this joint data set. Some colleges in *U.S. News and World Report* are classified by type and region, but they are unranked. For these colleges, the rank is reported as missing (“.a” in Stata).

COVID

50 observations or States

13 variables

This data set was compiled by the authors from multiple websites. The number of mask mandate days was taken from Ballotpedia. The rate of COVID-19 cases and death rates from COVID-19 were taken from Statista. The party of the Electoral College votes for each state during the 2020 presidential election in the United States was taken from the National Archives. Each of these sites is listed below.

https://ballotpedia.org/State-level_mask_requirements_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020-2022

https://www.statista.com/statistics/1109004/coronavirus-covid19-cases-rate-us-americans-by-state

https://www.statista.com/statistics/1109004/coronavirus-covid19-cases-rate-us-americans-by-state

https://www.archives.gov/electoral-college/2020

EXAM

50 observations

1 variable

This is a fictitious data set of exam scores for 50 students.

FAO MAIZE PRICES

277 observations

7 variables

https://fpma.fao.org/giews/fpmat4/#/dashboard/tool/domestic

This database was extracted from the Food Price Monitoring and Analysis (FPMA) Tool maintained by the United Nations Food and Agriculture Organization (FAO). The FPMA Tool has hundreds price series from close to 100 countries for dozens of commodities covering up to 25 years of monthly data. The Stata file “FAO maize prices.dta” contains seven variables: a month variable and wholesale prices of maize from six cities in Latin America. The 227 observations cover January 2005 to November 2023.

GSS2021 – GENERAL SOCIAL SURVEY

4,032 respondents

739 variables

http://gss.norc.org

The General Social Survey (GSS) has been conducted since 1972. Operated by the National Opinion Research Center (NORC), the GSS examines trends in the attitudes and behaviors of Americans.

HOMEWORK

30 observations

1 variable

This is a fictious data set of the average homework scores of 30 students in a course who used ChatGPT to generate and practice problems related to the course material.

LIBERAL ARTS COLLEGES – USNEWS.DTA

204 four-year liberal arts colleges

49 variables

https://collegescorecard.ed.gov

https://www.usnews.com/best-colleges

This data set is a combination of the College Scorecard data set from April 2023 and the *U.S. News and World Report* data on best colleges from 2023. The College Scorecard data were created by the U.S. government to help students compare colleges based on

various factors including debt upon graduation, salaries six years after graduation, and size. It has over 6,000 technical schools, colleges, and universities (including beauty schools, massage therapy schools, etc.). The *U.S. News and World Report* data from 2023 rank 1,859 four-year colleges based on factors such as graduation rates, faculty resources, reputation, and financial resources. The type and rank of each institution based on *U.S. News and World Report* were added to the College Scorecard data, and only those colleges that are listed both in *U.S. News and World Report* and on the College Scorecard and listed as a National Liberal Arts College (204) are included in this joint data set. Some colleges in *U.S. News and World Report* are classified by type and region, but they are unranked. For these colleges, the rank is reported as missing (“.a” in Stata).

NATIONAL SURVEY ON DRUG USE AND HEALTH 2015

57,146 observations

2,666 variables

https://nsduhweb.rti.org/

The National Survey on Drug Use and Health (NSDUH) provides information on tobacco, alcohol, and drug use and mental and physical health issues in the United States. It has been conducted since 1971.

NATIONAL SURVEY ON DRUG USE AND HEALTH 2015 - TRUNCATED

57,146 observations

1,948 variables

This is the same data set as above but with many variables removed in order to use this data set with a limited version of Stata that will not accept as many variables.

CARS4SALE - NEW AND USED PRICES OF GAS AND ELECTRIC CARS

1,100 cars

7 variables

https://www.cars.com/

This data set was generated by the authors using the Cars.com website. It represents a search for all new and used cars for sale within 20 miles of Burlington, Vermont, in June of 2023.

OKCUPID MOBILE DATING APP

59,948 observations

30 variables

This data set was made available by Kim and Escobedo-Land (2015). It is data collected by OkCupid, a mobile dating app that uses a precomputed compatibility score based on optional questions users may answer. Based on users in the San Francisco area, it contains data on age, sex, and sexual orientation and responses to open-ended questions.

PU_SSOCS16 – SCHOOL SURVEY ON CRIME AND SAFETY

2,092 observations

480 variables

https://nces.ed.gov/surveys/ssocs

The 2015–2016 School Survey on Crime and Safety is a nationally representative survey of 3,500 public elementary and secondary schools in the United States. It covers topics such as security, crime, and parent involvement.

GLOSSARY

Alpha level:
The maximum acceptable probability of a Type I error (rejecting the null hypothesis when it is true), set by the researcher before starting the analysis. In social science research, alpha (α) is often 0.05, which corresponds to a 95% confidence interval.

Alternative hypothesis:
The alternative of the null hypothesis. Often, the alternative hypothesis (denoted H_1 or H_a) is that a pattern observed in data is the result of a nonrandom effect. For example, the null hypothesis is that there is no change, and the alternative hypothesis is that there is a change.

Analysis of variance (ANOVA):
A statistical method that tests for significant differences among two or more means.

Autocorrelation:
A problem in regression analysis in which the error terms are correlated (positively or negatively) with each other.

Bar graph:
A graph that uses rectangles, where the height or length of the rectangles represents the numerical values of different groups. For example, two bars could be used to represent the average wage of male and female workers.

Bartlett's test:
A test used to determine if the variances of several samples are equal.

Binary variable (also called a dichotomous or dummy variable):

A type of variable that has a value of either 0 or 1. For example, 0 = male and 1 = female.

Bonferroni test:

A method of adjusting *p*-values when multiple tests are carried out to take into account the fact that the likelihood of getting a false positive (Type I error) rises with multiple tests.

Breusch–Pagan/Cook–Weisburg test:

A test for heteroscedasticity in a linear regression model.

Categorical responses:

Answers to a question that are limited to a fixed number of options, each one defined by a group or label. The alternative is continuous responses.

Categorical variable:

A variable that allows a limited number of values, each of which represents a group and has no units (e.g., dollars or kilograms). Examples include gender, political affiliation, and religion. The alternative is continuous variables.

Chi-square distribution:

The probability distribution of the sum of squared variables, each of which has a standard normal distribution. It has one parameter, *k*, which describes the number of random variables. It is often denoted by χ^2 or $\chi^2(k)$.

Chi-square statistic:

A statistic that tests for a relationship between two categorical variables.

Cleaning data:

The process of examining data for errors and inconsistencies and then correcting or dropping errors.

Closed-ended questions:

Questions that allow only predefined categorical responses or numerical responses. The alternative is open-ended responses, which allow unlimited numerical and text responses.

Coefficient or β:

In linear regression analysis, a measure of the effect of a one-unit increase in an independent variable on the dependent variable, holding constant other independent variables.

Coefficient of determination:

A statistic that measures the strength of the relationship between two continuous variables. Denoted by R^2, the coefficient of determination varies between 0 (no relationship) and 1 (perfect correlation).

Coefficient of variation:

The ratio of the standard deviation of a variable to the mean of the variable. It is a unit-less measure of variability and is abbreviated as CV.

Cohen's *d*:

A measure of the size of the difference between two variables relative to their standard deviations. It is usually used in conjunction with measures of the statistical significance of the difference.

Confidence interval:

A pair of numbers that indicate the level of precision in measuring a number. For example, the 95% confidence interval is a range such that there is a 95% probability that the true value lies in that range.

Confidence level:

The probability that the true value of a parameter lies within a specified range.

Continuous responses:

Answers to a question that represent a numerical count, usually involving units such as hours, kilometers, or kilograms. The alternative is categorical responses.

Continuous variable:

A variable whose values represent a measurement of some quantity, usually involving units such as hours, kilometers, or kilograms. The alternative is a categorical variable.

Correlation coefficient:

A measure of strength of the relationship between two continuous variables. Denoted by *r*, it ranges from −1 (a perfect negative relationship) to 1 (a perfect positive relationship).

Cramér's *V*:

A measure of the strength of the relationship between two categorical variables. It is used in conjunction with tests of statistical significance such as the chi-squared test.

Critical value:

A threshold number that is compared with a test statistic to determine whether to reject the null hypothesis. The critical value is based on the alpha level, the type of probability distribution, and whether a one-tailed or two-tailed test is being used.

Cross tabulation:

A table that shows the number or percentage of observations in each combination of two categorical variables.

Data analysis:

The process of converting raw data into usable results such as tables, graphs, statistical tests, and regression analysis.

Degrees of freedom:

The number of independent observations in a sample minus the number of parameters that must be estimated from sample data.

Dependent variable:

In regression analysis, the variable of interest that is being explained by the independent variable(s). *See* independent variable.

Descriptive statistics:

Basic statistics that summarize a set of variables such as frequency tables for categorical variables and the mean, standard deviation, minimum, and maximum of continuous variables.

Do-file:

A type of file in Stata that contains a series of commands to be carried out. Do-files have the file extension .do.

Enumerator:

A person responsible for carrying out interviews and recording responses as part of a survey.

Endogeneity:

A problem in regression analysis where one or more "independent" variables are influenced by the dependent variable or both dependent and independent variables are influenced by factors outside the model.

Error:

In regression analysis, it is the difference between an observation and the true relationship between the dependent variable and the independent variables. It is denoted by ε.

Estimate:

In statistics, an approximation of a population parameter calculated from sample data. It may be a point estimate (the most likely value) or an interval estimate (the confidence interval around the point estimate).

Estimation:

A statistical procedure for generating one or more estimates, usually with confidence interval(s). This term is also used more narrowly to refer to generating a result that describes a population based on a sample.

Estimator:
A method for calculating the value of an estimate.

Eta-square:
A measure of the size of the relationship in an ANOVA.

Exhaustive responses:
A set of answers that covers all possible responses to a question. This is a goal in the design of a questionnaire.

Expected value:
In probability, the average result across all possible outcomes. It is calculated as the weighted average of different values of the variable, where the weights are the probabilities of getting each value.

Frequency table:
A table that shows the number and/ or percentage of observations for each value of a categorical variable.

***F* test:**
A statistical test for a variable that has an *F* distribution under the null hypothesis. For example, *F* tests can be used to compare the means of two normally distributed variables with the same variance.

General Social Survey:
A sociological survey of adults in the United States conducted regularly since 1972 by the University of Chicago.

Heteroscedasticity:

A condition in which the variance of a variable differs across the range of observations. If the error term in a regression model is heteroscedastic, this violates the assumptions behind ordinary least squares regression analysis. The alternative is homoscedasticity.

Histogram:

A graph showing the distribution of one variable, where the values of the variable are on the horizontal axis and the frequency of observations is on the vertical axis.

Homoscedasticity:

A condition in which the variance of a variable is constant across observations. It is one of the assumptions behind ordinary least squares. The alternative is heteroscedasticity.

Hypothesis:

An educated guess regarding the outcome of a test or experiment, which will be tested using data.

Imputation:

The practice of replacing missing values of a variable with estimates based on values of the same variable and/or other variables. A simple example would be replacing missing values with the mean of the variable.

Independence of observations:

The condition where each observation has no effect on other observations.

Independent variable:

A variable that causes or predicts the dependent variable. Also called the explanatory variable. *See* dependent variable.

Intercept:

In a graph, the value of Y when $X = 0$. In a regression equation, it is also called the constant.

Interval scale:

A type of measurement scale in which the difference between values is meaningful (based on measured units) but the zero is arbitrary. Examples include temperature in Fahrenheit or Celsius and year. *See also* nominal scale, ordinal scale, ratio scale.

Kurtosis:

A measure of the “thickness” of the tails in a probability distribution. The kurtosis of a normal distribution is 3.

Leading question:

A question that is phrased in a way that prompts or encourages a particular response. Leading questions should be avoided in research questionnaires.

Levene’s test:

A test of the null hypothesis that the variance in two or more groups is the same.

Likert scale:

A set of responses designed to capture the strength of agreement with a statement or an evaluation of an object or experience. Typically, a Likert scale uses five responses, with the middle one being neutral.

Linear regression:

A statistical method that identifies a linear equation that best fits the relationship between one dependent variable and one or more independent (or explanatory) variables, subject to some assumptions.

Literature:

A set of scholarly papers that describe the results of research on a topic.

Log file:

A type of file that contains both commands and the output from those commands. In Stata, log files have one of two possible extensions: .log or .smcl.

Logit regression:

A statistical method for identifying the nonlinear equation that best fits the relationship between a binary dependent variable and one or more independent (or explanatory) variables, subject to some assumptions. Also called logistic regression. It is similar to a probit regression but uses a different function.

Macro:

In Stata, a temporary variable that can be used in loops and other programming. Stata has local and global macros.

Marginal effect:

In regression analysis, the impact of a one-unit increase in an independent variable on the dependent variable. In a graph of *y* as a function of *x*, the marginal effect is the slope of the line.

Margin of error:

The maximum expected difference between the true value and a sample estimate of a parameter due to sampling for a given probability. The margin of error may be expressed at different confidence levels, most commonly 95%.

Mean:

The average value of a set of numbers or the expected value of a probability distribution.

Measurement error:

The difference between a measured value of an observation and its true value.

Median:

The middle value of a set of numbers, such that there are equal numbers of observations greater and less than this value. It is

equivalent to the 50th percentile.

Multicollinearity:
In regression analysis, a condition in which two or more independent variables are closely correlated with one another. Multicollinearity reduces the precision of coefficient estimates but does not make them biased.

Multiple regression analysis:
A statistical method for estimating the relationship between a dependent variable and two or more independent variables. *See also* regression analysis, simple regression analysis.

Nominal scale:
A scale for categorical variables in which each value describes a category or label with no natural order. The values are not measured, so the interval between values is not meaningful. Examples include sex (male and female) and region (north, center, south). *See also* interval scale, ordinal scale, ratio scale.

Nonnormality:
A condition in which a variable is not normally distributed. In regression, it refers to the situation where the error terms are not normally distributed.

Nontechnical audience:
A type of reader or listener who does not have advanced training in a topic. In the context of this book, it refers to those who do not have a background in statistical methods.

Normal distribution:
A probability distribution that occurs frequently in statistics, having two parameters: the mean and the standard deviation. The normal distribution has a symmetric bell shape with infinite tails on either side.

Null hypothesis:

The null hypothesis is a testable statement indicating that there is no significant difference in a set of observations. For example, in comparing two means, the null hypothesis is that there is no difference. In regression analysis, the null hypothesis is usually that the coefficients are zero.

Observation:

One element of a variable or a set of variables. Each observation is usually represented as a row in a database. *See* unit of observation.

One-sample *t* test:

A statistical test that compares a sample mean with a fixed (nonrandom) number.

One-tailed test:

A test of a null hypothesis in which the alternative hypothesis is expressed as an inequality (greater than or less than).

Open-ended questions:

Questions that leave room for respondents to answer in their own words. *See* closed-ended questions.

Ordinal scale:

A scale for categorical variables in which each value describes a category or label with a natural order, but they are not measured, so the interval between values is not meaningful. Examples include quality (good, better, best) and military rank. *See also* interval scale, nominal scale, ratio scale.

Outlier:

An observation that lies extremely far from the mean or median. It is sometimes defined in terms of the number of standard deviations from the mean.

Parameter:

A measurable characteristic of a population, such as the mean or the standard deviation. In contrast, a statistic is a characteristic of a sample.

Pearson's chi-square:

A statistical test applied to sets of categorical data to test the null hypothesis that there are no differences between the sets. It could be used to test whether there is a gender difference in political party affiliation (two categorical variables).

Percentile:

The percentage of observations of a variable that are below a given value. For example, 100 is the 30th percentile if 30% of the observations are below 100.

Pie graph:

A circular graph divided into slices, where each slice represents a category and the size of the slice represents the percentage of observations in this category. Also called a pie chart.

Population:

The complete set of observations that can be made. For example, the population of car dealers in the United States is the full list of all car dealers.

Predicted value:

In regression analysis, the value of the dependent variable that is expected for each observation based on the values of the independent variables and the estimated coefficients.

Probit regression:

A statistical method for identifying the nonlinear equation that best fits the relationship between a binary dependent variable and one or more independent (or explanatory) variables, subject to some assumptions. It is similar to a logit regression but uses a different function.

Purposive sampling:

A method of selecting a sample that does not rely on random selection.

***p*-Value:**

The probability that a test statistic is larger than the observed value if the null hypothesis is true and if the assumptions behind the test are valid. A low *p*-value (often <0.05) is interpreted as a rejection of the null hypothesis.

Questionnaire:

A set of questions and rules for coding the responses that is used to guide an interview and collect data for a study.

Random sampling:

A group of methods for selecting a subset of the population where the selection is made using random numbers. Types of random sampling include simple random sampling, stratified random sampling, and multistage random sampling.

Ratio scale:

A type of measurement scale for continuous variables in which the interval is meaningful (based on measured units) and there is a natural zero. Examples include income, weight, and length. *See also* interval scale, nominal scale, ordinal scale.

Regression analysis:

A statistical method for estimating the relationship between a dependent variable and one or more independent variables. *See also* multiple regression analysis, simple regression analysis.

Research question:

The query that a researcher attempts to answer in a study.

Residual:

In regression analysis, the difference between the predicted value of the dependent variable and the observed value. It is

often denoted by *e*.

Sample:
A subset of observations selected from a population to make inferences about the population. For example, a sample of 1,000 voters may be selected to make inferences about the popularity of a candidate.

Sampling distribution:
The distribution of all possible values of a statistic.

Sampling weights:
Numbers used to compensate for the under- and oversampling caused by the sampling design so that the weighted sample statistics are unbiased estimates of population parameters. The weights are calculated as the inverse of the probability of selection.

Significance level:
The probability of committing a Type I error, meaning rejecting the null hypothesis when it is actually true.

Simple random sample:
A sampling method in which the researcher starts with a full list of the population and selects a sample with each unit having an equal probability of selection.

Simple regression analysis:
A statistical method for estimating the relationship between a dependent variable and one independent variable. *See also* multiple regression analysis, regression analysis.

Skewness:
A characteristic of a probability distribution that is often used to assess the level of asymmetry. A symmetric distribution has a skewness of zero, though the reverse is not always true.

Skip patterns:

In questionnaire design, the rules for skipping over questions based on the responses to earlier questions. For example, if the respondent is single, the skip patterns indicate that one should skip over questions about his or her spouse.

Specification error:

A problem in regression analysis where the model is missing important variables or has the wrong functional form.

Standard deviation:

A statistic that measures the degree of dispersion around the mean. The standard deviation is the square root of the variance.

Standard error of the mean:

The standard deviation of the means of all possible samples from a population. An estimate of this parameter can be calculated as the standard deviation of the sample divided by the square root of the sample size.

Statistic:

A measurable characteristic of a sample, such as the sample mean or the sample standard deviation. In contrast, parameters are characteristics of the population.

Statistically significant:

A condition in which the probability of Type I error (rejecting the null hypothesis when it is true) is below the value of alpha (the highest acceptable level of Type I error). In practice, it is often defined as when the *p*-value is less than 0.05, indicating that the 95% confidence interval does not include zero.

Strata:

In sampling, groups within a population, each with their own sampling design. The singular is stratum.

Stratification:

In sampling, the process of dividing the population into groups (or strata) and having a different sampling design for each one. For example, a population may be stratified by region or by income group.

Systematic random sample:
A random sampling of units characterized by an equal interval between selected units. It is used to ensure that the sample is dispersed across the population in the dimension in which they are sorted.

***t* Distribution:**
A probability distribution that results from estimating the mean from a normal distribution with a small sample.

Technical audience:
Readers or listeners who have some training in scientific methods and statistics. The type of audience influences the appropriate writing style.

Two independent-samples *t* test:
Compares the means of two independent groups to determine whether there is statistical evidence that the population means are different from each other.

Type I error:
The error of rejecting the null hypothesis when it is true.

Type II error:
The error of accepting the null hypothesis when it is false.

Unit of observation:
An object about which information is collected. For example, in survey data, the unit of observation may be people, households, companies, or some other category of objects. *See* observation.

Variable:

A quantified characteristic or attribute of each observation that varies across observations. In a database, each variable is usually represented by a column of numbers.

Variance:

A parameter used to indicate the degree of dispersion in a variable. It is the square of the standard deviation.

***Z* score:**

The value of an observation minus the mean, divided by the standard deviation. In other words, it measures how many standard deviations above or below the mean an observation is.

1. Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.
2. Greenlaw, S. A. (2009). *Doing economics*. South-Western Cengage Learning.
3. Leedy, P. D., & Ormrod, J. E. (2001). *Practical research: Planning and design*. Merrill Prentice Hall.
4. Teter, C. J., McCabe, S. E., Cranford, J. A., Boyd, C. J., & Guthrie, S. K. (2005). Prevalence and motives for illicit use of prescription stimulants in an undergraduate student sample. *Journal of American College Health*, *53*(6), 253–262.
5. Valkenburg, P. M., Peter, J., & Schouten, A. P. (2006). Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & Behavior*, *9*(5), 584–590. doi:10.1089/cpb.2006.9.584
6. Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.
7. Daniel, J. (2011). *Sampling essentials: Practical guidelines for making sampling choices*. Sage.
8. National Constitution Center. (2012). *The five biggest polling mistakes in U.S. history*. https://perma.cc/5LFL-TSAK
9. Rea, L. M., & Parker, R. A. (2005). *Designing and conducting survey research: A comprehensive guide*. Jossey-Bass.
10. Scheaffer, R. L., Mendenhall, W., III, Ott., L, R., & Gerow, K. G. (2011). *Elementary survey sampling*. Brooks/Cole, Cengage Learning.
11. Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, *52*(1), 125–133
12. Ekinci, Y. (2015). *Designing research questionnaires for business and management students*. Sage.
13. Grosh, M., & Glewwe, P. (2000). *Designing household survey questionnaires for developing countries: Lessons from 15 years of the Living Standard Measurement Study*. World Bank. http://siteresources.worldbank.org/INTPOVRES/Resources/477227–1142020443961/2311843–1197996479165/part1DesigningHHS.pdf

14. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508cFINAL.pdf
15. Protection of Human Subjects. (2009). *Code of Federal Regulations, Title 45, Part 46*. www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html
16. Qiao, H. (2018). A brief introduction to institutional review boards in the United States. *Pediatric Investigation*, *2*(1), 46–51. https://onlinelibrary.wiley.com/doi/abs/10.1002/ped4.12023
17. Rea, L. M., & Parker, R. A. (2005). *Designing and conducting survey research: A comprehensive guide*. Jossey-Bass
18. Enders, C. K. (2010). *Applied missing data analysis: Methodology in the social sciences*.T. D. Little & Series(Eds.). Guildford Press.
19. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (2nd ed.). Wiley.
20. Minot, N. (2012). *Using Stata for data analysis*. International Food Policy Research Institute.
21. Osborne, J. W. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.
22. Sauro, J. (2015, June 2). *7 Ways to handle missing data*. https://measuringu.com/handlemissing-data/
23. Daily Fantasy Sports Rankings. (2018, February 20). *Daily fantasy NBA: Consistency, scoring and fun facts from the first half of the season*. www.dailyfantasysportsrankings.com/2018/02/20/consistency-scoring-and-fun-facts-from-the-first-half-of-the-season/
24. Enders, C. K. (2010). T. D. Little & Series (Eds.), *Applied missing data analysis: Methodology in the social sciences*. Guildford Press.
25. Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (2nd ed.). Wiley.

26. Sauro, J. (2015, June 2). *7 ways to handle missing data*. https://measuringu.com/handle-missing-data/
27. Soares, J. (2018, January 10). *“More colleges than ever have test-optional admissions policies – and that’s a good thing.”*. *The Conversation*. https://theconversation.com/more-colleges-than-ever-have-test-optional-admissions-policies-and-thats-a-good-thing-89852
28. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108
29. Whittier, N., Wildhagen, T., & Gold, H. J. (2019). *Statistics for social understanding*. Rowman & Littlefield Publishing Group https://bookshelf.vitalsource.com/books/9781538109847
30. Xie, K., & Anderman, E. (2023, June 6). *3 ways to use CHATGPT to help students learn—and not cheat* https://theconversation.com/3-ways-to-use-chatgpt-to-help-students-learn-and-not-cheat-205000
31. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
32. VanDusky-Allen, J., & Shvetsova, O. (2021, May 12). *How America’s partisan divide over pandemic responses played out in the states*. *The Conversation* 2021 https://theconversation.com/how-americas-partisan-divide-over-pandemic-responses-played-out-in-the-states-157565
33. Churchill, M. L. (2023, January 25). *The SAT and ACT are less important than you might think*. *The Conversation*. https://theconversation.com/the-sat-and-act-are-less-important-than-you-might-think-197658
34. Khatri, S. (2014, August 9). *Analysis of variance (LinkedIn SlideShare)*. www.slideshare.net/snehkhatri/analysis-of-variance-anova
35. Hadji-Vasilev, A. (2022, June 25). *25 online dating statistics & trends in 2022* Retreived from http://cloudwards.net/online-dating-statistics

36. Bailey, M. A. (2017). *Real econometrics: The right tools to answer important questions*. Oxford University Press.
37. Dallegro, J. A. (2018). *Just what factors into the value of your used car*. *Investopedia*. www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp
38. Greene, W. H. (2018). *Econometric analysis*. Pearson.
39. Woolridge, J. M. (2016). *Introductory econometrics: A modern approach*. South-Western Cengage Learning
40. Bailey, M. A. (2016). *Real econometrics the right tools to answer important questions*. Oxford University Press.
41. Greene, W. H. (2018). *Econometric analysis*. Pearson.
42. Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review*, *87*(3), 270–290.
43. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*, 673–690.
44. Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. South-Western Cengage Learning
45. De Pinto, J., Backus, F., Khanna, K., & Salvanto, A. (2017). *Marijuana legalization support at all-time high*. www.cbsnews.com/news/support-for-marijuana-legalization-at-all-time-high
46. Greene, W. H. (2018). *Econometric analysis*. Pearson.
47. Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata Press.
48. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379
49. Bailey, M. A. (2020). *Real econometrics: The right tools to answer important questions*. Oxford University Press.
50. Banerjee, A., Dolado, J. J., Galbraith, J. W., & Hendry, D. (1993). *Co-integration, error correction, and the econometric analysis of non-stationary data*. Oxford University Press.

51. Becketti, S. (2013). *Introduction to Time Series Using Stata*. Stata Press.
52. Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 251–276.
53. Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, *2*(2), 111–120.
54. Greene, W. H. (2018). *Econometric analysis*. Pearson.
55. Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata Press.
56. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379.
57. StataCorp. (2023). *Stata 18 Time-Series Reference Manual*. Stata Press. https://www.stata.com/manuals/ts.pdf
58. Wooldridge, J. M. (2019). *Introductory econometrics: A modern approach*. Cengage Learning.
59. Allgood, S., Walstad, W. B., & Siegfried, J. J. (2015). Research on teaching economics to undergraduates. *Journal of Economic Literature*, *53*(2), 285–325.
60. American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Author.
61. Booth, W. C., Colomb, G. G., Williams, J. M., Bizup, J., & Fitzgerald, W. T. (2016). *The craft of research* (4th ed.). University of Chicago Press.
62. Enfield, J. (2013). Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at CSUN. *Tech Trends*, *57*(6), 14–27.
63. Jerman, B. (2012, June). *When to quote and when to paraphrase*. https://writingcommons.org/open-text/research-methods-methodologies/integreat-evidence/summarize-paraphrase-sources/692-when-to-quote-and-when-toparaphase
64. Talan, T., & Kalinkara, Y. (2023). The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course.

International Journal of Management Information Systems and Computer Science, *7*(1), 33–40.
65. Teter, C. J., McCabe, S. E., Cranford, J., Boyd, C., & Guthrie, S. (2005). Prevalence and motives for illicit use of prescription stimulants in an undergraduate student sample. *Journal of American College Health*, *53*(6), 253–262.
66. University of Chicago Press. (2010). *Chicago manual of style* (16th ed.). Author.
67. Weingast, B. R. (2010). *Caltech rules for writing papers: How to structure your paper and write an introduction*. https://web.stanford.edu/group/mcnollgast/cgi-bin/wordpress/wp-content/uploads/2013/10/CALTECH.RUL_.pdf
68. National Association of Colleges and Employers. (2018). *Compensation*. http://www.naceweb.org/job-market/compensation
69. Kim, A., & Escobedo-Land, A. (2015). OkCupid data for introductory statistics and data science courses. *Journal of Statistics Education*, *23*. 10.1080/106981898.2015.11889737

ENDNOTES

Chapter 1

1. The home page of a journal will indicate if and how articles are peer reviewed.

2. Note that the title of this section is “Examine the Data or *Other Evidence*.” Not all research is based on data. Research can also be theoretical, conceptual, or exploratory, for example.

Chapter 4

1. Many students in a course related to Stata may not have Stata installed on their own computers. By converting the log to text, they can review their work on their own computers without Stata. If, however, you do have Stata on your computer, the smcl files within Stata are easier to read than text files outside of Stata.

2. When using the **log on** and **log off** commands, Stata keeps the log file open and ready to use. If you then want to use a new log file in the do-file, you will get an error message.

Chapter 5

1. For a complete review of outliers and advanced methods for dealing with them, see Osborne (2012).

2. A less transparent but more streamlined approach to recode and generate the new variable maritalstat would be **recode mar1 2/5=2, generate(maritalstat)**.

Chapter 6

1. To find multiple modes for “Size,” we would use the following code:

```
egen mode = mode(ugds), nummode(1)
egen mode2 = mode(ugds), nummode(2)
```

The first line of the command tells Stata to create a new variable, mode1, which is the lowest mode of the variable ugds. Similarly, the second line tells Stata to create a new variable, mode2, which is the second lowest value of the mode. If we added lines of code with 3, 4, and 5 at the end, Stata would tell us that there are only four modes when running the fifth line. Finally, to see what these modes are, we would generate a frequency table of the four new variables—**tab1 mode1 mode2 mode3 mode4**.

Chapter 7

1. The exam.dta data set is a simulated data set. It does not represent actual grades of any students.

2. There are many books written entirely about the central limit theorem. For a concise description of the central limit theorem, along with its key components and why it is useful, please refer to this Investopedia site: https://www.investopedia.com/terms/c/central_limit_theorem.asp

Chapter 9

1. Sources for these data can be found at these links:

https://ballotpedia.org/State-level_mask_requirements_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020-2022

https://www.archives.gov/electoral-college/2020

2. The calculation for Levene's test of equality of variances is as follows:

$$W = \frac{(N-K)}{(k-1)} \times \frac{\sum_{i=1}^{k} N_i \left(Z_{i.-} - Z_{...}\right)^2}{\sum_{i=1}^{k} \sum_{i=1}^{Ni} \left(Z_{ij} - Z_{i.}\right)^2}$$

$$W = \frac{(N-K)}{(k-1)} \times \frac{\sum_{i=1}^{k} N_i (Z_{i.-} - Z_{...})^2}{\sum_{i=1}^{k} \sum_{i=1}^{Ni} (Z_{ij} - Z_{i.})^2}$$

Where *k* is the number of different groups to which the sampled cases belong. N_i is the number of cases in the ith group. *N* is the total number of cases in all groups. Y_{ij} is the value of the measured variable for the jth case from the ith group.

$$Z_{ij} = \left| Y_{ij} - \overline{Y}_{i.} \right|, \overline{Y}_{i.} \text{is a mean of the ith group.}$$

$$Z_{i.} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij} \text{is the mean of the Zij for group i.}$$

$Z_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ is the mean of all Z_{ij}.

3. The W50 row tests the variance relative to the median value. The W10 uses a trimmed mean with 10% of extreme values trimmed.

4. A full description of Satterthwaite's degrees of freedom and their calculation can be found at this website: https://www.statology.org/satterthwaite-approximation

Chapter 10

1. Two other tests, the Tukey method and the Scheffe method, are similar to the Bonferroni correction. Alternative methods to address unequal variances include the Welch's ANOVA and nonparametric tests such as the Kruskal-Wallis test.

2. Refer to Chapter 8 for a discussion of degrees of freedom.

Chapter 11

1. Stata can calculate the *p* value as well using the commands **display chi2tail(1,379)**, but it will not show the graph.

2. In the case of a 2 × 2 table, the Cramér's *V* is equivalent to the Phi coefficient, which is used to test the effect size for 2 × 2 tables.

3. For a more complete explanation of Cramér's *V* and other measures of association, please refer to this website: www.statisticssolutions.com/nominal-variable-association

Chapter 12

1. This equation has an intuitive interpretation for those with some statistics background. The correlation coefficient is the ratio of (a) the covariance of *x* and *y* and (b) the product of the standard deviation of *x* and the standard deviation of *y*.

2. This equation is similar to one that may be familiar from algebra classes, $y = Mx + B$, with different notation and the addition of the error term, ε. The slope is *M* in this equation and β_1 in the regression equation, while the *y*-intercept is *B* here and β_0 in the regression equation.

3. To be consistent, we could label the residual $\hat{\varepsilon}$, since it is an estimate of the error term, but we follow the convention in statistics of labeling it *e*.

4. The intuitive explanation is that the independent variables in a regression model cannot include any redundant information. If there are three categories, they can be represented by two dummy variables. Adding a third dummy variable would not add any new information.

5. In this database, new cars represent just 27% of gas cars, 50% of hybrid cars, and 75% of electric cars.

Chapter 13

1. This is always true when there is just one independent variable and usually true when there are multiple independent variables.

2. Like all statistical software, Stata generates pseudorandom numbers starting with a "seed." By fixing the seed, we can ensure that Stata generates the same set of random numbers in multiple runs or runs by different users. The "2314" is arbitrary, but the seed must be a positive integer.

3. We add the condition "& mileage!=." to this command to exclude observations with missing mileage. Stata treats missing values as very high values so they would be included in the display without this condition.

4. In the database, year refers to the model year of the car. Typically, car companies release the new year's models toward the end of the previous year. As a result, 39 of the cars in the database are 2024 models, so the "age" of these cars is −1.

5. Technically, correlation between the independent variables and the error term is the underlying cause of biased OLS coefficients when there is measurement error in the independent variables (Section 13.2), specification error (Section 13.3), and endogeneity (Section 13.7). However, for teaching purposes, we find it useful to consider these separate topics.

Chapter 14

1. The terms *logit regression* and *logistic regression* can be used interchangeably.

2. Strictly speaking, in order to make a profit, betting companies offer payout odds slightly less than implied by the perceived probabilities. For example, if two evenly matched teams are playing, the perceived probabilities that each will win is 50%, implying payoff odds of 1-to-1. However, betting companies offer slightly lower payoffs for each team, say 9-to-10. For this reason, the sum of implied probabilities across outcomes is slightly greater than 100%, in this case, 1 − [9 / (9 + 10)] = 0.526 = 52.6% for each team.

Chapter 16

1. For a nice summary of when to quote and when to paraphrase, see Jerman (2012).

2. The humanities and some social science fields use the *Chicago Manual of Style*, which suggests spelling out numbers 1 through 99 and using numerals for all higher numbers (University of Chicago Press, 2010).

Appendix 7

1. Thanks to Bill Rising from Stata Corporation for suggesting this approach, as well as for improvements in the other do-files in this appendix.

Appendix 8

1. Stata also offers a way to generate graphs directly from the **twoway** command. The above quadratic function can be graphed with the command: **twoway function y1 = 1000 – 50*x + x^2, range(0 100)**. This approach is more concise but somewhat less transparent to the new Stata user.

INDEX

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z

ABOUT THE AUTHORS

Lisa Daniels
is the Hodson Trust Professor Emeritus of Economics at Washington College in Chestertown, Maryland. She specializes in development in Africa, where she worked for 10 years, beginning as a Peace Corps volunteer. During her time in Africa and later in Asia, she studied agricultural markets, market information systems, poverty trends, and micro- and small-scale enterprises. As part of her research on micro- and small-scale enterprises, she directed national surveys of 7,000 to 56,000 households and businesses in Bangladesh, Botswana, Kenya, Malawi, and Zimbabwe funded by the U.S. Agency for International Development. In each survey, she was responsible for the questionnaire design, sample selection, data collection and analysis, and report preparation. Her work from these surveys and other research in Africa and Asia appears in consulting reports and in peer-reviewed journals. In addition to research and fieldwork, she has taught a wide range of courses over the past 28 years, including a research methods course and a data analysis course, which she has taught over 20 times. She has also presented her work related to teaching at more than a dozen workshops.

Nicholas Minot
is a senior research fellow at the International Food Policy Research Institute (IFPRI) in Washington, D.C. Since joining IFPRI in 1997, he has carried out research on agricultural market reform, income diversification, food security, food price volatility, and the impact of policies and programs on poverty in developing countries. This research often involves carrying out surveys of farmers, cooperatives, traders, and consumers to better understand changes in food marketing systems. In addition to research, he is involved in outreach and capacity-building activities, including offering short courses on the use of Stata for survey data analysis. Before joining IFPRI, he taught at the University of Illinois Urbana–Champaign,

served as a policy adviser in Zimbabwe, and analyzed survey data in Rwanda. Overall, he has worked in more than 30 countries in Latin America, sub-Saharan Africa, North Africa, and Asia.

The authors are married, live in Annapolis, Maryland, and have two children—Andrea (26) and Alex (23)—and one dog, Kara (5).