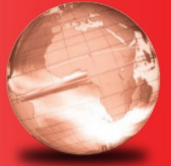


GLOBAL  
EDITION



# Statistical Methods for the Social Sciences

FIFTH EDITION

**ALAN AGRESTI**



---

# STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fifth Edition  
Global Edition

Alan Agresti  
*University of Florida*



Pearson

---

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong  
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

Director, Portfolio Management: Deirdre Lynch  
Senior Portfolio Manager: Suzanna Bainbridge  
Portfolio Management Assistant: Justin Billing  
Content Producer: Sherry Berg  
Managing Producer: Karen Wernholm  
Producer: Jean Choe  
Manager, Courseware QA: Mary Durnwald  
Manager, Content Development: Bob Carroll  
Senior Acquisitions Editor, Global Edition: Sandhya Ghoshal  
Editor, Global Edition: Punita Kaur Mann  
Content Producer, Global Edition: Isha Sachdeva  
Product Marketing Manager: Yvonne Vannatta  
Field Marketing Manager: Evan St. Cyr  
Product Marketing Assistant: Jennifer Myers  
Field Marketing Assistant: Erin Rush  
Senior Author Support/Technology Specialist: Joe Vetere  
Manager, Rights and Permissions: Gina Cheselka  
Manufacturing Buyer: Carol Melville, LSC Communications  
Senior Manufacturing Controller, Global Edition: Kay Holman  
Associate Director of Design: Blair Brown  
Production Coordination, Composition, and Illustrations: iEnergizer Aptara<sup>®</sup>, Ltd.  
Cover Image: Marina Zakharova/Shutterstock

Attributions of third party content appear on page 549, which constitutes an extension of this copyright page.

Pearson Education Limited  
KAO Two  
KAO Park  
Harlow  
CM17 9NA  
United Kingdom

and Associated Companies throughout the world

Visit us on the World Wide Web at:  
[www.pearsonglobaleditions.com](http://www.pearsonglobaleditions.com)

© Pearson Education Limited 2018

The rights of Alan Agresti to be identified as the author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

*Authorized adaptation from the United States edition, entitled Statistical Methods for the Social Sciences, 5th Edition, ISBN 978-0-13-450710-1 by Alan Agresti, published by Pearson Education © 2018.*

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a license permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

#### **British Library Cataloguing-in-Publication Data**

A catalogue record for this book is available from the British Library

Typeset in 10/12 TimesTenLTStd-Roman by iEnergizer Aptara<sup>®</sup>, Ltd.  
Printed and bound by Vivar in Malaysia

ISBN 10: 1-292-22031-7  
ISBN 13: 978-1-29-222031-4

TO MY PARENTS  
LOUIS J. AGRESTI AND MARJORIE H. AGRESTI

This page intentionally left blank

# Contents

Preface 9

Acknowledgments 11

## 1 INTRODUCTION 13

---

1.1 Introduction to Statistical Methodology 13

1.2 Descriptive Statistics and Inferential Statistics 16

1.3 The Role of Computers and Software in Statistics 18

1.4 Chapter Summary 20

## 2 SAMPLING AND MEASUREMENT 23

---

2.1 Variables and Their Measurement 23

2.2 Randomization 26

2.3 Sampling Variability and Potential Bias 29

2.4 Other Probability Sampling Methods\* 33

2.5 Chapter Summary 35

## 3 DESCRIPTIVE STATISTICS 41

---

3.1 Describing Data with Tables and Graphs 41

3.2 Describing the Center of the Data 47

3.3 Describing Variability of the Data 53

3.4 Measures of Position 58

3.5 Bivariate Descriptive Statistics 63

3.6 Sample Statistics and Population Parameters 67

3.7 Chapter Summary 67

## 4 PROBABILITY DISTRIBUTIONS 79

---

4.1 Introduction to Probability 79

4.2 Probability Distributions for Discrete and Continuous Variables 81

4.3 The Normal Probability Distribution 84

4.4 Sampling Distributions Describe How Statistics Vary 92

4.5 Sampling Distributions of Sample Means 97

4.6 Review: Population, Sample Data, and Sampling Distributions 103

4.7 Chapter Summary 106

## 5 STATISTICAL INFERENCE: ESTIMATION 115

---

5.1 Point and Interval Estimation 115

5.2 Confidence Interval for a Proportion 118

5.3 Confidence Interval for a Mean 125

5.4 Choice of Sample Size 132

5.5 Estimation Methods: Maximum Likelihood and the Bootstrap\* 138

5.6 Chapter Summary 142

## 6 STATISTICAL INFERENCE: SIGNIFICANCE TESTS 151

---

6.1 The Five Parts of a Significance Test 152

6.2 Significance Test for a Mean 155

6.3 Significance Test for a Proportion 164

6.4 Decisions and Types of Errors in Tests 167

6.5 Limitations of Significance Tests 171

6.6 Finding  $P(\text{Type II Error})^*$  175

6.7 Small-Sample Test for a Proportion—the Binomial Distribution\* 177

6.8 Chapter Summary 181

## 7 COMPARISON OF TWO GROUPS 191

---

7.1 Preliminaries for Comparing Groups 191

7.2 Categorical Data: Comparing Two Proportions 194

7.3 Quantitative Data: Comparing Two Means 199

- 7.4 Comparing Means with Dependent Samples **202**
- 7.5 Other Methods for Comparing Means\* **205**
- 7.6 Other Methods for Comparing Proportions\* **210**
- 7.7 Nonparametric Statistics for Comparing Groups\* **213**
- 7.8 Chapter Summary **216**

## 8 ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES **227**

---

- 8.1 Contingency Tables **227**
- 8.2 Chi-Squared Test of Independence **230**
- 8.3 Residuals: Detecting the Pattern of Association **237**
- 8.4 Measuring Association in Contingency Tables **239**
- 8.5 Association Between Ordinal Variables\* **245**
- 8.6 Chapter Summary **250**

## 9 LINEAR REGRESSION AND CORRELATION **259**

---

- 9.1 Linear Relationships **259**
- 9.2 Least Squares Prediction Equation **262**
- 9.3 The Linear Regression Model **268**
- 9.4 Measuring Linear Association: The Correlation **271**
- 9.5 Inferences for the Slope and Correlation **278**
- 9.6 Model Assumptions and Violations **284**
- 9.7 Chapter Summary **289**

## 10 INTRODUCTION TO MULTIVARIATE RELATIONSHIPS **299**

---

- 10.1 Association and Causality **299**
- 10.2 Controlling for Other Variables **302**
- 10.3 Types of Multivariate Relationships **306**
- 10.4 Inferential Issues in Statistical Control **311**
- 10.5 Chapter Summary **313**

## 11 MULTIPLE REGRESSION AND CORRELATION **319**

---

- 11.1 The Multiple Regression Model **319**
- 11.2 Multiple Correlation and  $R^2$  **328**
- 11.3 Inferences for Multiple Regression Coefficients **332**
- 11.4 Modeling Interaction Effects **337**
- 11.5 Comparing Regression Models **341**
- 11.6 Partial Correlation\* **343**
- 11.7 Standardized Regression Coefficients\* **346**
- 11.8 Chapter Summary **349**

## 12 REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS **363**

---

- 12.1 Regression Modeling with Dummy Variables for Categories **363**
- 12.2 Multiple Comparisons of Means **367**
- 12.3 Comparing Several Means: Analysis of Variance **370**
- 12.4 Two-Way ANOVA and Regression Modeling **374**
- 12.5 Repeated-Measures Analysis of Variance\* **381**
- 12.6 Two-Way ANOVA with Repeated Measures on a Factor\* **385**
- 12.7 Chapter Summary **390**

## 13 MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS **399**

---

- 13.1 Models with Quantitative and Categorical Explanatory Variables **399**
- 13.2 Inference for Regression with Quantitative and Categorical Predictors **406**
- 13.3 Case Studies: Using Multiple Regression in Research **409**

- 13.4** Adjusted Means\* **413**
- 13.5** The Linear Mixed Model\* **418**
- 13.6** Chapter Summary **423**

## **14** MODEL BUILDING WITH MULTIPLE REGRESSION **431**

---

- 14.1** Model Selection Procedures **431**
- 14.2** Regression Diagnostics **438**
- 14.3** Effects of Multicollinearity **445**
- 14.4** Generalized Linear Models **447**
- 14.5** Nonlinear Relationships: Polynomial Regression **451**
- 14.6** Exponential Regression and Log Transforms\* **456**
- 14.7** Robust Variances and Nonparametric Regression\* **460**
- 14.8** Chapter Summary **462**

## **15** LOGISTIC REGRESSION: MODELING CATEGORICAL RESPONSES **471**

---

- 15.1** Logistic Regression **471**
- 15.2** Multiple Logistic Regression **477**

- 15.3** Inference for Logistic Regression Models **482**
- 15.4** Logistic Regression Models for Ordinal Variables\* **484**
- 15.5** Logistic Models for Nominal Responses\* **489**
- 15.6** Loglinear Models for Categorical Variables\* **492**
- 15.7** Model Goodness-of-Fit Tests for Contingency Tables\* **496**
- 15.8** Chapter Summary **500**

**Appendix: R, Stata, SPSS, and SAS for Statistical Analyses 509**

**Bibliography 545**

**Credits 549**

**Index 551**



This page intentionally left blank

# Preface

When Barbara Finlay and I undertook the first edition of this book nearly four decades ago, our goal was to introduce statistical methods in a style that emphasized their concepts and their application to the social sciences rather than the mathematics and computational details behind them. We did this by focusing on how the methods are used and interpreted rather than their theoretical derivations.

This edition of the book continues the emphasis on concepts and applications, using examples and exercises with a variety of “real data.” This edition increases its illustrations of statistical software for computations, and takes advantage of the outstanding applets now available on the Internet for explaining key concepts such as sampling distributions and for conducting basic data analyses. I continue to downplay mathematics, in particular probability, which is all too often a stumbling block for students. On the other hand, the text is not a cookbook. Reliance on an overly simplistic recipe-based approach to statistics is not the route to good statistical practice.

## Changes in the Fifth Edition

Users of earlier editions will notice that the book no longer lists Barbara Finlay as a co-author. I am grateful to Barbara Finlay for her contributions to the first two editions of this text. Combining her sociology background with my statistics background, she very much helped me develop a book that is not only statistically sound but also relevant to the social sciences.

Since the first edition, the increase in computer power coupled with the continued improvement and accessibility of statistical software has had a major impact on the way social scientists analyze data. Because of this, this book does not cover the traditional shortcut hand-computational formulas and approximations. The presentation of computationally complex methods, such as regression, emphasizes interpretation of software output rather than the formulas for performing the analysis. The text contains numerous sample software outputs, both in chapter text and in homework exercises. In the appendix on using statistical software, this edition adds R and Stata to the material on SPSS and SAS.

Exposure to realistic but simple examples and to numerous homework exercises is vital to student learning. This edition has updated data in most of the exercises and replaced some exercises with new ones. Each chapter’s homework set is divided into two parts, straightforward exercises on the text material in *Practicing the Basics* and exercises dealing with open-ended data analyses, understanding of concepts, and advanced material in *Concepts and Applications*. The data sets in the examples and exercises are available at [www.pearsonglobaleditions.com/Agresti](http://www.pearsonglobaleditions.com/Agresti).

This edition contains some changes and additions in content, directed toward a more modern approach. The main changes are as follows:

- The text has greater integration of *statistical software*. Software output shown now uses R and Stata instead of only SAS and SPSS, although much output has a generic appearance. The text appendix provides instructions about basic use of these software packages.
- New examples and exercises in Chapters 4–6 ask students to use applets to help learn the fundamental concepts of sampling distributions, confidence

intervals, and significance tests. The text also now relies more on applets for finding tail probabilities from distributions such as the normal,  $t$ , and chi-squared. I strongly encourage instructors and students to look at the excellent applets cited at [www.pearsonglobal editions.com/Agresti](http://www.pearsonglobal editions.com/Agresti). They were prepared by Prof. Bernhard Klingenberg for the fourth edition of the text *Statistics: The Art and Science of Learning from Data*, by Agresti, Franklin, and Klingenberg (Pearson, 2017).

- Chapter 5 has a new section introducing maximum likelihood estimation and the bootstrap method.
- Chapter 12 on ANOVA has been reorganized to put more emphasis on using regression models with dummy variables to handle categorical explanatory variables.
- Chapter 13 on regression modeling with both quantitative and categorical explanatory variables has a new section using case studies to illustrate how research studies commonly use regression with both types of explanatory variables. The chapter also has a new section introducing linear mixed models.
- Chapter 14 has a new section introducing robust regression standard errors and nonparametric regression.
- The text Web site [www.pearsonglobal editions.com/Agresti](http://www.pearsonglobal editions.com/Agresti) has the data sets analyzed in the text, in generic form to copy for input into statistical software. Special directories there also have data files in Stata format and in SPSS format, so they are ready for immediate use with those packages.
- Answers to Select Odd-Numbered Exercises are available at the text Website [www.pearsonglobal editions.com/Agresti](http://www.pearsonglobal editions.com/Agresti).

## Use of Text in Introductory Statistics Courses

Like the first four editions, this edition is appropriate for introductory statistics courses at either the undergraduate or beginning graduate level, and for either a single-term or a two-term sequence. Chapters 1–9 are the basis for a single-term course. If the instructor wishes to go further than Chapter 9 or wishes to cover some material in greater depth, sections that can easily be omitted without disturbing continuity include 2.4, 5.5, 6.6–6.7, 7.5–7.7, and 8.5. Also, Chapters 7–9 are self-contained, and the instructor could move directly into any of these after covering the fundamentals in Chapters 1–6. Three possible paths for a one-term course are as follows:

- Chapters 1–9 (possibly omitting sections noted above): Standard cross-section of methods, including basic descriptive and inferential statistics, two-sample procedures, contingency tables, and linear regression.
- Chapters 1–7, 9, and 11: Emphasis on regression.
- Chapters 1–7, and 9, and Sections 11.1–11.3 and 12.1–12.3: After two-group comparisons, introduction to regression and analysis of variance.

Regardless of the type of data, my belief is that a modeling paradigm emphasizing parameter estimation is more useful than the artificial hypothesis-testing approach of many statistics texts. Thus, the basic inference chapters (5–8) explain the advantages confidence intervals have over significance testing, and the second half of this text (starting in Chapter 9) is primarily concerned with model building. The modeling material forms the basis of a second course. Instructors who focus on

observational data rather than designed experiments may prefer to cover only the first section of Chapter 12 (ANOVA), to introduce dummy variables before moving to later chapters that incorporate both categorical and quantitative explanatory variables.

Some material appears in sections, subsections, or exercises marked by asterisks. This material is optional, having lesser importance for introductory courses. The text does not attempt to present every available method, since it is meant to be a teaching tool, not an encyclopedic cookbook. It does cover the most important methods for social science research, however, and it includes topics not usually discussed in introductory statistics texts, such as

- Methods for contingency tables that are more informative than chi-squared, such as cell residuals and analyses that utilize category orderings.
- Controlling for variables, and issues dealing with causation.
- The generalized linear modeling approach, encompassing ordinary regression, analysis of variance and covariance, gamma regression for nonnegative responses with standard deviation proportional to the mean, logistic regression for categorical responses, and loglinear association models for contingency tables.
- Relatively new methods that are increasingly used in research, such as the linear mixed model approach of using both fixed effects and random effects (and related multilevel models), and multiple imputation for dealing with missing data.

I believe that the student who works through this book successfully will acquire a solid foundation in applied statistical methodology.

## Acknowledgments

I thank those who invested considerable time in helping this book to reach fruition. Thanks to Alfred DeMaris, Regina Dittrich, Susan Herring, Haeil Jung, James Lapp, Graham Lord, Brian Marx, Brian McCall, Mack Shelley, Peter Steiner, Gary Sweeten, and Henry Wakhungu for providing comments for this edition. Other individuals who provided advice or data sets include Don Hedeker, John Henretta, Glenn Pierce, and Michael Radelet. Thanks to NORC for permission to use General Social Survey data. (The GSS is a project of the independent research organization NORC at the University of Chicago, with principal funding from the National Science Foundation.) I am grateful to Stata Corp. and IBM for supplying copies of Stata and SPSS. Special thanks to Bill Rising at Stata for reviewing the book's Stata discussion and pointing out glitches and improvements.

Thanks also to the many people whose comments helped in the preparation of the first four editions, such as Jeffrey Arnold, Arne Bathke, Roslyn Brain, Beth Chance, Brent Coull, Alfred DeMaris, E. Jacquelin Dietz, Dorothy K. Davidson, Burke Grandjean, Mary Gray, Brian Gridley, Ralitzia Gueorguieva, Maureen Hallinan, John Henretta, Ira Horowitz, Youqin Huang, Harry Khamis, Bernhard Klingenberg, Michael Lacy, Norma Leyva, David Most, Michael Radelet, Paula Rausch, Susan Reiland, Euijung Ryu, Shirley Scritchfield, Paul Smith, Sarah Streett, Andrew Thomas, Robert Wilson, Jeff Witmer, Sonja Wright, Mary Sue Younger, Douglas Zahn, and Zoe Ziliak. My editors for this and the previous edition, Suzy Bainbridge at Pearson and Petra Recter and Ann Heath at Prentice Hall, provided outstanding support and encouragement.

Finally, extra special thanks to my wife, Jacki Levine, for assistance with editing and style in the third edition and with overall encouragement during the preparation of the fourth and fifth editions.

*Alan Agresti*  
Gainesville, Florida and Brookline, Massachusetts

## Global Edition Acknowledgments

Pearson would like to thank the following people for their work on the content of the Global Edition:

### **Contributors:**

MaryJane Bock, Murdoch University Dubai  
Gagari Chakrabarti, Presidency University  
Raghvi Garg, doctoral student, Ashoka University  
Pooja Thakur

### **Reviewers:**

Dave Centeno, University of the Philippines  
Chitrita Bhowmick Chakrabarti, Victoria Institution (College)  
Gagari Chakrabarti, Presidency University  
Samprit Chakrabarti, ICAI Business School  
Albert Lee Kai Chung, Nanyang Technological University  
Eric Li, The University of Hong Kong  
Francisco de los Reyes, University of the Philippines  
Pooja Sengupta  
Raymond Wong, The University of Hong Kong  
Elizabeth Wright, Murdoch University Dubai

## INTRODUCTION

**CHAPTER  
OUTLINE**

- 1.1 Introduction to Statistical Methodology
- 1.2 Descriptive Statistics and Inferential Statistics
- 1.3 The Role of Computers and Software in Statistics
- 1.4 Chapter Summary

**1.1 Introduction to Statistical Methodology**

Recent years have seen a dramatic increase in the use of statistical methods by social scientists, whether they work in academia, government, or the private sector. Social scientists study their topics of interest, such as analyzing how well a program works or investigating the factors associated with beliefs and opinions of certain types, by analyzing quantitative evidence provided by data. The growth of the Internet and computing power has resulted in an increase in the amount of readily available quantitative information. At the same time, the evolution of new statistical methodology and software makes new methods available that can more realistically address the questions that social scientists seek to answer.

This chapter introduces “statistics” as a science that deals with describing data and making predictions that have a much wider scope than merely summarizing the collected data. So, why should knowledge of statistical science be important for a student who is studying to become a social scientist?

**WHY STUDY STATISTICAL SCIENCE?**

The increased use of statistical methods is evident in the changes in the content of articles published in social science research journals and reports prepared in government and industry. A quick glance through recent issues of journals such as *American Political Science Review* and *American Sociological Review* reveals the fundamental role of statistical methodology in research. For example, to learn about which factors have the greatest impact on student performance in school or to investigate what affects people’s political beliefs or the quality of their health care or their decisions about work and home life, researchers collect information and analyze it using statistical methods. Because of this, more and more academic departments require that their majors take statistics courses.

These days, social scientists work in a wide variety of areas that use statistical methods, such as governmental agencies, business organizations, and health care facilities. Social scientists in government agencies dealing with human welfare or environmental issues or public health policy commonly need to read reports that contain statistical arguments, and perhaps use statistical methods themselves in preparing such reports. Some social scientists help managers to evaluate employee performance using quantitative benchmarks and to determine factors that help predict sales of products. Medical sociologists and physicians often must evaluate

recommendations from studies that contain statistical evaluations of new therapies or new ways of caring for the elderly. In fact, a recent issue of *The Journal of the American Medical Association* indicated that the Medical College Admissions Test has been revised to require more statistics, because doctors increasingly need to be able to evaluate quantitatively the factors that affect peoples' health.

In fact, increasingly many jobs for social scientists require a knowledge of statistical methods as a basic work tool. As the joke goes, "What did the sociologist who passed statistics say to the sociologist who failed it? 'I'll have a Big Mac, fries, and a Coke.'"

But an understanding of statistical science is important even if you never use statistical methods in your own career. Often you are exposed to communications containing statistical arguments, such as in advertising, news reporting, political campaigning, and surveys about opinions on controversial issues. Statistical science helps you to make sense of this information and evaluate which arguments are valid and which are invalid. You will find concepts from this text helpful in judging the information you encounter in your everyday life. Look at [www.youtube.com/user/ThisIsStats](http://www.youtube.com/user/ThisIsStats) to view some short testimonials with the theme that "Statistics isn't just about data analysis or numbers; it is about understanding the world around us. The diverse face of statistics means you can use your education in statistics and apply it to nearly any area you are passionate about, such as the environment, health care, human rights, sports. . . ."

We realize you are not reading this book in hopes of becoming a statistician. In addition, you may suffer from math phobia and feel fear at what lies ahead. Please be assured that you can read this book and learn the primary concepts and methods of statistics with little knowledge of mathematics. Just because you may have had difficulty in math courses before does not mean you will be at a disadvantage here. To understand this book, logical thinking and perseverance are more important than mathematics. In our experience, the most important factor in how well you do in a statistics course is how much time you spend on the course—attending class, doing homework, reading and re-reading this text, studying your class notes, working together with your fellow students, and getting help from your professor or teaching assistant—not your mathematical knowledge or your gender or your race or whether you now feel fear of statistics.

Please do not be frustrated if learning comes slowly and you need to read a chapter a few times before it makes sense. Just as you would not expect to take a single course in a foreign language and be able to speak that language fluently, the same is true with the language of statistical science. Once you have completed even a portion of this text, however, you will better understand how to make sense of statistical information.

## DATA

Information gathering is at the heart of all sciences, providing the *observations* used in statistical analyses. The observations gathered on the characteristics of interest are collectively called *data*.

For example, a study might conduct a survey of 1000 people to observe characteristics such as opinion about the legalization of same-sex marriage, political party affiliation, how often attend religious services, number of years of education, annual income, marital status, race, and gender. The data for a particular person would consist of observations such as (opinion = do not favor legalization, party = Republican, religiosity = once a week, education = 14 years, annual income in the range of 40–60 thousand dollars, marital status = married, race = white, gender = female). Looking

at the data in the right way helps us learn about how the characteristics are associated. We can then answer questions such as “Do people who attend church more often tend to be less favorable toward same-sex marriage?”

To generate data, the social sciences use a wide variety of methods, including surveys using questionnaires, experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze data already recorded for other purposes, such as police records, census materials, and hospital files. Existing archived collections of data are called *databases*. Many databases are now available on the Internet. An important database for social scientists contains results since 1972 of the *General Social Survey*.

### Example 1.1

**The General Social Survey** Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of about 2000 adults provides data about opinions and behaviors of the American public. Social scientists use it to investigate how adult Americans answer a wide diversity of questions, such as “Do you believe in life after death?” “Would you be willing to pay higher prices in order to protect the environment?” and “Do you think a preschool child is likely to suffer if his or her mother works?” Similar surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database. We’ll demonstrate, using a question it asked in one survey, “About how many good friends do you have?”

- Go to the website [sda.berkeley.edu/GSS/](http://sda.berkeley.edu/GSS/) at the Survey Documentation and Analysis site at the University of California, Berkeley.
- Click on *GSS—with NO WEIGHT VARIABLES predefined*. You will then see a “variable selection” listing in the left margin dealing with issues addressed over the years, and a menu on the right for selecting particular characteristics of interest.
- The GSS name for the question about number of good friends is NUMFRIEND. Type NUMFRIEND in the *Row* box. Click on *Run the table*.

The GSS site will then generate a table that shows the possible values for “number of good friends” and the number of people and the percentage who made each possible response. The most common responses were 2 and 3 (about 16% made each of these responses). ■

## WHAT IS STATISTICAL SCIENCE?

You already have a sense of what the word *statistics* means. You hear statistics quoted about sports events (such as the number of points scored by each player on a basketball team), statistics about the economy (such as the median income or the unemployment rate), and statistics about opinions, beliefs, and behaviors (such as the percentage of students who indulge in binge drinking). In this sense, a statistic is merely a number calculated from data. But this book uses *statistics* in a much broader sense—as a science that gives us ways of obtaining and analyzing data.



Specifically, statistical science provides methods for

1. **Design:** Planning how to gather data for a research study to investigate questions of interest to us.
2. **Description:** Summarizing the data obtained in the study.
3. **Inference:** Making predictions based on the data, to help us deal with uncertainty in an objective manner.

**Design** refers to planning a study so that the data it yields are informative. For a survey, for example, the design specifies how to select the people to interview and constructs the questionnaire to administer to those people.

**Description** refers to summarizing data, to help understand the information the data provide. For example, an analysis of the number of good friends based on the GSS data might start with a list of the number reported for each person surveyed. The raw data are then a complete listing of observations, person by person. These are not easy to comprehend, however. We get bogged down in numbers. For presentation of results, instead of listing *all* observations, we could summarize the data with a graph or table showing the percentages reporting 1 good friend, 2 good friends, 3 good friends, and so on. Or we could report the average number of good friends, which was about 5, or the most common response, which was 2. Graphs, tables, and numerical summaries such as averages and percentages are called **descriptive statistics**. We use descriptive statistics to reduce the data to a simpler and more understandable form without distorting or losing much information.

**Inference** refers to using the data to make predictions. For instance, for the GSS data on reported number of good friends, 6.1% reported having only 1 good friend. Can we use this information to predict the percentage of the 250 million adults in the United States who have only 1 good friend? A method presented in this book allows us to predict confidently that that percentage is no greater than 8%. Predictions made using data are called **statistical inferences**.

**Description** and **inference** are the two types of ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, “Is having the death penalty available for punishment associated with a reduction in violent crime?” “Does student performance in schools depend on the amount of money spent per student, the size of the classes, or the teachers’ salaries?”

## 1.2 Descriptive Statistics and Inferential Statistics

Section 1.1 explained that statistical science consists of methods for *designing* studies and *analyzing* data collected in the studies. A statistical analysis is classified as **descriptive** or **inferential**, according to whether its main purpose is to describe the data or to make predictions. To explain this distinction further, we next define the *population* and the *sample*.

### POPULATIONS AND SAMPLES

The entities on which a study makes observations are called the sample **subjects** for the study. Usually the subjects are people, such as in the General Social Survey, but they need not be. For example, subjects in social research might be families, schools, or cities. Although we obtain data for the sample subjects, our ultimate interest is in the population that the sample represents.

**Population and Sample**

The **population** is the total set of subjects of interest in a study. A **sample** is the subset of the population on which the study collects data.

In the 2014 General Social Survey, the sample was the 2538 adult Americans who participated in the survey. The population was all adult Americans at that time—about 250 million people. One person was sampled for about every 100,000 people in the population.

The goal of any study is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, survey organizations such as the Gallup Poll usually select samples of about 1000–2000 Americans to collect information about opinions and beliefs of the population of *all* Americans.

**Descriptive Statistics**

**Descriptive statistics** summarize the information in a collection of data.

Descriptive statistics consist of graphs, tables, and numbers such as averages and percentages. Descriptive statistics reduce the data to simpler and more understandable form without distorting or losing much information.

Although data are usually available only for a sample, descriptive statistics are also useful when data are available for the entire population, such as in a census. By contrast, inferential statistics apply when data are available only for a sample, but we want to make a prediction about the entire population.

**Inferential Statistics**

**Inferential statistics** provide predictions about a population, based on data from a sample of that population.

**Example  
1.2**

**How Many People Believe in Heaven?** In three of its surveys, the General Social Survey asked, “Do you believe in heaven?” The population of interest was the collection of all adults in the United States. In the most recent survey in which this was asked, 85% of the 1326 sampled subjects answered *yes*. This is a descriptive statistic. We would be interested, however, not only in those 1326 people but in the *entire population* of all adults in the United States.

Inferential statistics use the sample data to generate a prediction about the entire population. An inferential method presented in Chapter 5 predicts that the population percentage that believe in heaven falls between 83% and 87%. That is, the sample value of 85% has a “margin of error” of 2%. Even though the sample size was tiny compared to the population size, we can conclude that a large percentage of the population believed in heaven. ■

Inferential statistical analyses can predict characteristics of populations well by selecting samples that are small relative to the population size. That’s why many polls sample only about a thousand people, even if the population has millions of people. In this book, we’ll see why this works.

In recent years, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 4.

## PARAMETERS AND STATISTICS

A descriptive statistic is a numerical summary of the sample data. The corresponding numerical summary for the population is called a **parameter**.

**Parameter**

A *parameter* is a numerical summary of the population.

Example 1.2 estimated the percentage of Americans who believe in heaven. The parameter was the population percentage who believed in heaven. Its value was unknown. The inference about this parameter was based on a descriptive statistic—the percentage of the 1326 subjects interviewed in the survey who answered *yes*, namely, 85%.

In practice, our main interest is in the values of the *parameters*, not merely the values of the *statistics* for the particular sample selected. For example, in viewing results of a poll before an election, we're more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed. The sample and statistics describing it are important only insofar as they help us make inferences about unknown population parameters.

An important aspect of statistical inference involves reporting the likely *precision* of the sample statistic that estimates the population parameter. For Example 1.2 on belief in heaven, an inferential statistical method predicted how close the *sample* value of 85% was likely to be to the unknown percentage of the *population* believing in heaven. The reported margin of error was 2%.

When data exist for an entire population, such as in a census, it's possible to find the values of the parameters of interest. Then, there is no need to use inferential statistical methods.

## DEFINING POPULATIONS: ACTUAL AND CONCEPTUAL

Usually the population to which inferences apply is an actual set of subjects, such as all adult residents of the United States. Sometimes, though, the generalizations refer to a *conceptual* population—one that does not actually exist but is hypothetical.

For example, suppose a medical research team investigates a newly proposed drug for treating lung cancer by conducting a study at several medical centers. Such a medical study is called a *clinical trial*. The conditions compared in a clinical trial or other experiment are called *treatments*. Basic descriptive statistics compare lung cancer patients who are given the new treatment to other lung cancer patients who instead receive a standard treatment, using the percentages who respond positively to the two treatments. In applying inferential statistical methods, the researchers would ideally like inferences to refer to the conceptual population of *all* people suffering from lung cancer now or at some time in the future.

## 1.3 The Role of Computers and Software in Statistics

Over time, powerful and easy-to-use software has been developed for implementing statistical methods. This software provides an enormous boon to the use of statistics.

### STATISTICAL SOFTWARE

Statistical software packages include R, SPSS,<sup>1</sup> SAS,<sup>2</sup> and Stata. Appendix A explains how to use them, organized by chapter. You can refer to Appendix A for the software used in your course as you read each chapter, to learn how to implement the analyses of that chapter. It is much easier to apply statistical methods using software

<sup>1</sup> Originally, this was an acronym for *Statistical Package for the Social Sciences*.

<sup>2</sup> Originally, this was an acronym for *Statistical Analysis System*.

than using hand calculation. Moreover, many methods presented in this text are too complex to do by hand or with hand calculators. Software relieves us of computational drudgery and helps us focus on the proper application and interpretation of the statistical methods.

Many examples in this text also show output obtained by using statistical software. One purpose of this textbook is to teach you what to look for in output and how to interpret it. Knowledge of computer programming is not necessary for using statistical software.

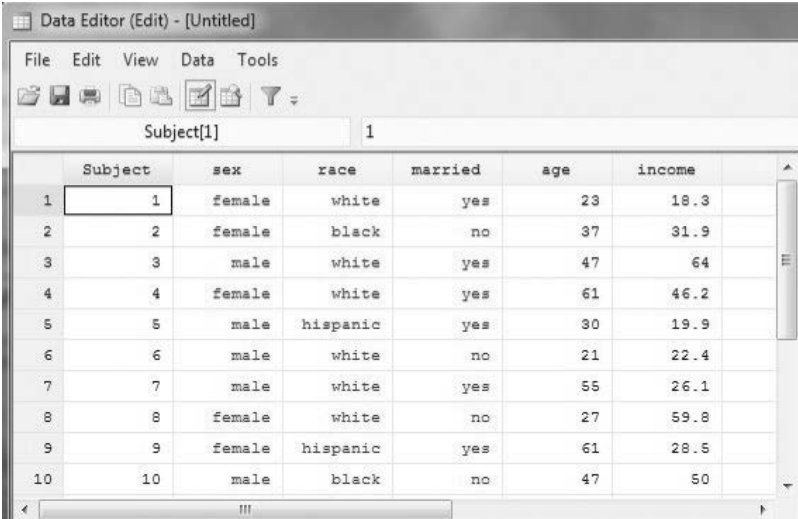
## DATA FILES

Statistical software analyzes data organized in the spreadsheet form of a *data file*:

- Any one row contains the observations for a particular subject (e.g., person) in the sample.
- Any one column contains the observations for a particular characteristic.

Figure 1.1 shows an example of a data file, in the form of a window for editing data using Stata software. It shows data for the first 10 subjects in a sample, for the characteristics sex, racial group, marital status, age, and annual income (in thousands of dollars). Some of the data are numerical, and some consist of labels. Chapter 2 introduces the types of data for data files.

**FIGURE 1.1:** Example of Part of a Stata Data File



Subject	sex	race	married	age	income
1	female	white	yes	23	18.3
2	female	black	no	37	31.9
3	male	white	yes	47	64
4	female	white	yes	61	46.2
5	male	hispanic	yes	30	19.9
6	male	white	no	21	22.4
7	male	white	yes	55	26.1
8	female	white	no	27	59.8
9	female	hispanic	yes	61	28.5
10	male	black	no	47	50

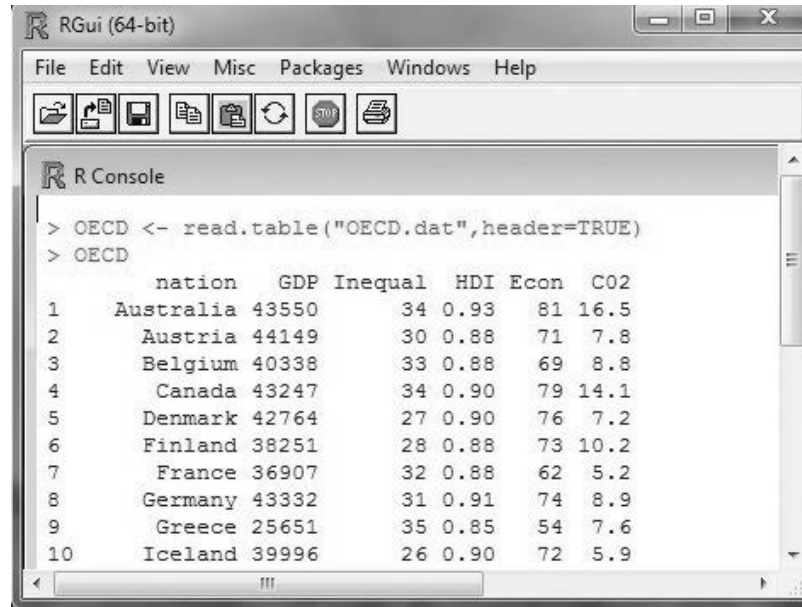
R is a software package that is increasingly popular, partly because it is available to download for free at [www.r-project.org](http://www.r-project.org). Figure 1.2 shows part of an R session for loading a data file called `OECD.dat` from a PC directory and displaying it.

## USES AND MISUSES OF STATISTICAL SOFTWARE

A note of caution: The easy access to statistical methods using software has dangers as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Incorrect analyses result when researchers take insufficient time to understand the statistical method, the assumptions for its use, or its appropriateness for the

**FIGURE 1.2:** Example of Part of an R Session for Loading and Displaying Data. The full data file is in Table 3.13 on page 70.



```

RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> OECD <- read.table("OECD.dat",header=TRUE)
> OECD
  nation  GDP Inequal  HDI Econ  CO2
1  Australia 43550    34 0.93  81 16.5
2  Austria 44149    30 0.88  71  7.8
3  Belgium 40338    33 0.88  69  8.8
4  Canada 43247    34 0.90  79 14.1
5  Denmark 42764    27 0.90  76  7.2
6  Finland 38251    28 0.88  73 10.2
7  France 36907    32 0.88  62  5.2
8  Germany 43332    31 0.91  74  8.9
9  Greece 25651    35 0.85  54  7.6
10 Iceland 39996    26 0.90  72  5.9

```

specific problem. It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis. You'll need a good background in statistics to understand which method to select, which options to choose in that method, and how to make valid conclusions from the output. The purpose of this text is to give you this background.

## I.4 Chapter Summary

The field of statistical science includes methods for

- designing research studies,
- describing the data (*descriptive statistics*), and
- making predictions using the data (*inferential statistics*).

Statistical methods apply to observations in a *sample* taken from a *population*. *Statistics* summarize sample data, while *parameters* summarize entire populations.

- *Descriptive statistics* summarize sample or population data with numbers, tables, and graphs.
- *Inferential statistics* use sample data to make predictions about population parameters.

A *data file* has a separate row of data for each subject and a separate column for each characteristic. Software applies statistical methods to data files.

## Exercises

### Practicing the Basics

**1.1.** A medical university conducts an annual national survey of cancer patients who are in remission about their lifestyle habits. In 2016, 1764 patients were surveyed. Identify the **(a)** subject, **(b)** sample, and **(c)** population.

**1.2.** In 2015, a French national survey asked adults about marriage and divorce. Of the 1754 individuals surveyed, 43% reported that they were married. Of the entire adult French population, 41% were married.

- (a)** What was the population and what was the sample?
- (b)** Identify a statistic and a parameter.