

6313

STATISTICS

THE ART AND SCIENCE OF LEARNING FROM DATA

FIFTH EDITION

Alan Agresti • Christine A. Franklin • Bernhard Klingenberg

4633

2649

2416

1810



Statistics

The Art and Science of Learning from Data

Fifth Edition
Global Edition

Alan Agresti

University of Florida

Christine Franklin

University of Georgia

Bernhard Klingenberg

Williams College and New College of Florida



Product Management: Gargi Banerjee and Paromita Banerjee
Content Strategy: Shabnam Dohutia and Bedasree Das
Product Marketing: Wendy Gordon, Ashish Jain, and Ellen Harris
Supplements: Bedasree Das
Production and Digital Studio: Vikram Medepalli, Naina Singh, and Niharika Thapa
Rights and Permissions: Rimpay Sharma and Nilofar Jahan
Cover Image: majcot / Shutterstock

Please contact <https://support.pearson.com/getsupport/s/> with any queries on this content

Pearson Education Limited
KAO Two
KAO Park
Hockham Way
Harlow
Essex
CM17 9SR
United Kingdom

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

© Pearson Education Limited 2023

The rights of Alan Agresti, Christine A. Franklin, and Bernhard Klingenberg to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Authorized adaptation from the United States edition, entitled *Statistics: The Art and Science of Learning from Data*, 5th Edition, ISBN 978-0-13-646876-9 by Alan Agresti, Christine A. Franklin, and Bernhard Klingenberg published by Pearson Education © 2021.

PEARSON, ALWAYS LEARNING, and MYLAB are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

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. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsoned.com/permissions/.

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. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights and Permissions department, please visit www.pearsoned.com/permissions/.

Attributions of third-party content appear on page 921, which constitutes an extension of this copyright page.

This eBook is a standalone product and may or may not include all assets that were part of the print version. It also does not provide access to other Pearson digital products like MyLab and Mastering. The publisher reserves the right to remove any material in this eBook at any time.

ISBN 10 (Print): 1-292-44476-2
ISBN 13 (Print): 978-1-292-44476-5
ISBN 13 (eBook): 978-1-292-44479-6

British Library Cataloguing-in-Publication Data

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

Typeset in Times NR MT Pro by Straive

Dedication

To my wife, Jacki, for her extraordinary support, including making numerous suggestions and putting up with the evenings and weekends I was working on this book.

ALAN AGRESTI

To Corey and Cody, who have shown me the joys of motherhood, and to my husband, Dale, for being a dear friend and a dedicated father to our boys. You have always been my biggest supporters.

CHRIS FRANKLIN

To my wife, Sophia, and our children, Franziska, Florentina, Maximilian, and Mattheus, who are a bunch of fun to be with, and to Jean-Luc Picard for inspiring me.

BERNHARD KLINGENBERG

Contents

Preface 10

Part One Gathering and Exploring Data

Chapter 1 Statistics: The Art and Science of Learning From Data 24

- 1.1 Using Data to Answer Statistical Questions 25
- 1.2 Sample Versus Population 30
- 1.3 Organizing Data, Statistical Software, and the New Field of Data Science 42

Chapter Summary 52

Chapter Exercises 53

Chapter 2 Exploring Data With Graphs and Numerical Summaries 57

- 2.1 Different Types of Data 58
- 2.2 Graphical Summaries of Data 64
- 2.3 Measuring the Center of Quantitative Data 82
- 2.4 Measuring the Variability of Quantitative Data 90
- 2.5 Using Measures of Position to Describe Variability 98
- 2.6 Linear Transformations and Standardizing 110
- 2.7 Recognizing and Avoiding Misuses of Graphical Summaries 117

Chapter Summary 122

Chapter Exercises 125

Chapter 3 Exploring Relationships Between Two Variables 134

- 3.1 The Association Between Two Categorical Variables 136
- 3.2 The Relationship Between Two Quantitative Variables 146
- 3.3 Linear Regression: Predicting the Outcome of a Variable 160
- 3.4 Cautions in Analyzing Associations 175

Chapter Summary 194

Chapter Exercises 196

Chapter 4 Gathering Data 204

- 4.1 Experimental and Observational Studies 205
- 4.2 Good and Poor Ways to Sample 213
- 4.3 Good and Poor Ways to Experiment 223
- 4.4 Other Ways to Conduct Experimental and Nonexperimental Studies 228

Chapter Summary 240

Chapter Exercises 241

Part Two Probability, Probability Distributions, and Sampling Distributions

Chapter 5 Probability in Our Daily Lives 252

- 5.1 How Probability Quantifies Randomness 253
- 5.2 Finding Probabilities 261

- 5.3 Conditional Probability 275
- 5.4 Applying the Probability Rules 284

Chapter Summary 298

Chapter Exercises 300

Chapter 6 Random Variables and Probability Distributions 307

- 6.1 Summarizing Possible Outcomes and Their Probabilities 308
- 6.2 Probabilities for Bell-Shaped Distributions: The Normal Distribution 321
- 6.3 Probabilities When Each Observation Has Two Possible Outcomes: The Binomial Distribution 334
- Chapter Summary 345
- Chapter Exercises 347

Chapter 7 Sampling Distributions 354

- 7.1 How Sample Proportions Vary Around the Population Proportion 355
- 7.2 How Sample Means Vary Around the Population Mean 367
- 7.3 Using the Bootstrap to Find Sampling Distributions 380
- Chapter Summary 390
- Chapter Exercises 392

Part Three Inferential Statistics

Chapter 8 Statistical Inference: Confidence Intervals 400

- 8.1 Point and Interval Estimates of Population Parameters 401
- 8.2 Confidence Interval for a Population Proportion 409
- 8.3 Confidence Interval for a Population Mean 426
- 8.4 Bootstrap Confidence Intervals 439
- Chapter Summary 447
- Chapter Exercises 449

Chapter 9 Statistical Inference: Significance Tests About Hypotheses 457

- 9.1 Steps for Performing a Significance Test 458
- 9.2 Significance Test About a Proportion 464
- 9.3 Significance Test About a Mean 481

- 9.4 Decisions and Types of Errors in Significance Tests 493
- 9.5 Limitations of Significance Tests 498
- 9.6 The Likelihood of a Type II Error and the Power of a Test 506
- Chapter Summary 513
- Chapter Exercises 515

Chapter 10 Comparing Two Groups 522

- 10.1 Categorical Response: Comparing Two Proportions 524
- 10.2 Quantitative Response: Comparing Two Means 539
- 10.3 Comparing Two Groups With Bootstrap or Permutation Resampling 554
- 10.4 Analyzing Dependent Samples 568
- 10.5 Adjusting for the Effects of Other Variables 581
- Chapter Summary 587
- Chapter Exercises 590

Part Four Extended Statistical Methods and Models for Analyzing Categorical and Quantitative Variables

Chapter 11 Categorical Data Analysis 602

- 11.1 Independence and Dependence (Association) 603
- 11.2 Testing Categorical Variables for Independence 608
- 11.3 Determining the Strength of the Association 622
- 11.4 Using Residuals to Reveal the Pattern of Association 631
- 11.5 Fisher's Exact and Permutation Tests 635

- Chapter Summary 643
- Chapter Exercises 645

Chapter 12 Regression Analysis 652

- 12.1 The Linear Regression Model 653
- 12.2 Inference About Model Parameters and the Relationship 663
- 12.3 Describing the Strength of the Relationship 670

6 Contents

- 12.4** How the Data Vary Around the Regression Line 681
- 12.5** Exponential Regression: A Model for Nonlinearity 693
- Chapter Summary 698
- Chapter Exercises 701

Chapter 13 Multiple Regression 707

- 13.1** Using Several Variables to Predict a Response 708
- 13.2** Extending the Correlation Coefficient and R^2 for Multiple Regression 714
- 13.3** Inferences Using Multiple Regression 720
- 13.4** Checking a Regression Model Using Residual Plots 731
- 13.5** Regression and Categorical Predictors 737
- 13.6** Modeling a Categorical Response: Logistic Regression 743
- Chapter Summary 752
- Chapter Exercises 754

Appendix A-833

Answers A-839

Index I-865

Index of Applications I-872

Credits C-877

Chapter 14 Comparing Groups: Analysis of Variance Methods 760

- 14.1** One-Way ANOVA: Comparing Several Means 761
- 14.2** Estimating Differences in Groups for a Single Factor 772
- 14.3** Two-Way ANOVA: Exploring Two Factors and Their Interaction 781
- Chapter Summary 795
- Chapter Exercises 797

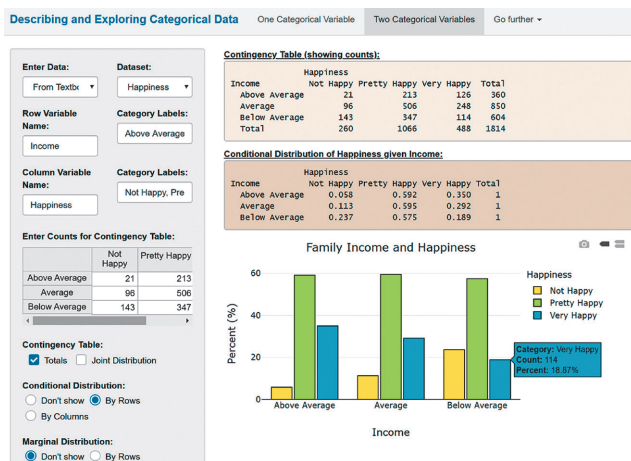
Chapter 15 Nonparametric Statistics 804

- 15.1** Compare Two Groups by Ranking 805
- 15.2** Nonparametric Methods for Several Groups and for Dependent Samples 816
- Chapter Summary 827
- Chapter Exercises 829

An Introduction to the Web Apps

The website ArtofStat.com links to several web-based apps that run in any browser. These apps allow students to carry out statistical analyses using real data and engage with statistical concepts interactively. The apps can be used to obtain graphs and summary statistics for exploratory data analysis, fit linear regression models and create residual plots, or visualize statistical distributions such as the normal distribution. The apps are especially valuable for teaching and understand sampling distributions through simulations. They can further guide students through all steps of inference based on one or two sample categorical or quantitative data. Finally, these apps can carry out computer intensive inferential methods such as the bootstrap or the permutation approach. As the following overview shows, each app is tailored to a specific task.

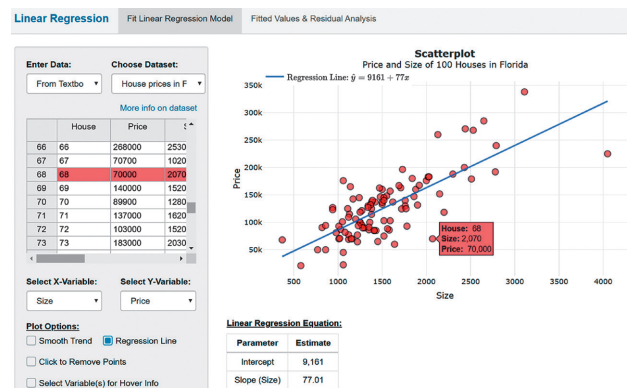
- The **Explore Categorical Data** and **Explore Quantitative Data** apps provide basic summary statistics and graphs (bar graphs, pie charts, histograms, box plots, dot plots), both for data from a single sample or for comparing two samples. Graphs can be downloaded.



Screenshot from the *Explore Categorical Data* app

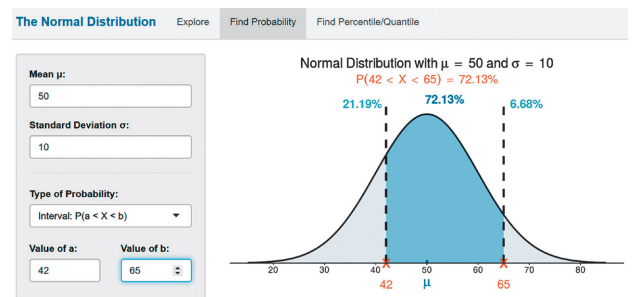
- The **Times Series** app plots a simple time series and can add a smooth or linear trend.
- The **Random Numbers** app generates random numbers (with or without replacement) and simulates flipping (potentially biased) coins.
- The **Mean vs. Median** app allows users to add or delete points by clicking in a graph and observing the effect of outliers or skew on these two statistics.
- The **Scatterplots & Correlation** and the **Explore Linear Regression** apps allow users to add or delete points from a scatterplot and observe how the correlation coefficient or the regression line are affected. The **Guess the Correlation** app lets users guess the correlation for

randomly generated data sets. The **Fit Linear Regression** app provides summary statistics, an interactive scatterplot, a smooth trend line, the regression line, predictions, and a residual analysis. Options include confidence and prediction intervals. Users can upload their own data. The **Multivariate Relationships** app explores graphically and statistically how the relationship between two quantitative variables is affected when adjusting for a third variable.



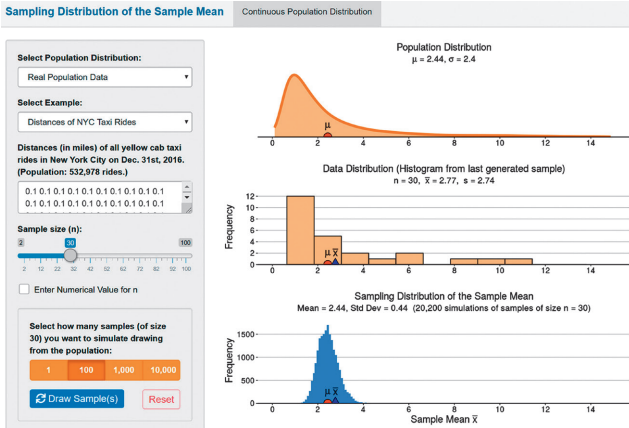
Screenshot from the *Linear Regression* app

- The **Binomial, Normal, t, Chi-Squared, and F Distribution** apps provide an interactive graph of the distribution and how it changes for different parameter values. Each app easily finds percentile values and lower, upper, or interval probabilities, clearly marking them on the graph of the distribution.



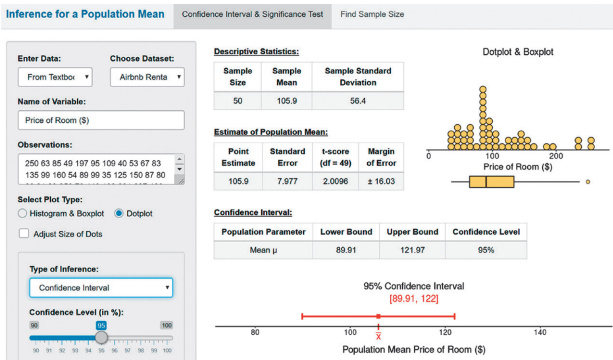
Screenshot from the *Normal Distribution* app

- The three **Sampling Distribution** apps generate sampling distributions of the sample proportion or the sample mean (for both continuous and discrete populations). Sliders for the sample size or population parameters help with discovering and exploring the Central Limit Theorem interactively. In addition to the uniform, skewed, bell-shaped, or bimodal population distributions, several real population data sets (such as the distances traveled by all taxi rides in NYC on Dec. 31) are preloaded.



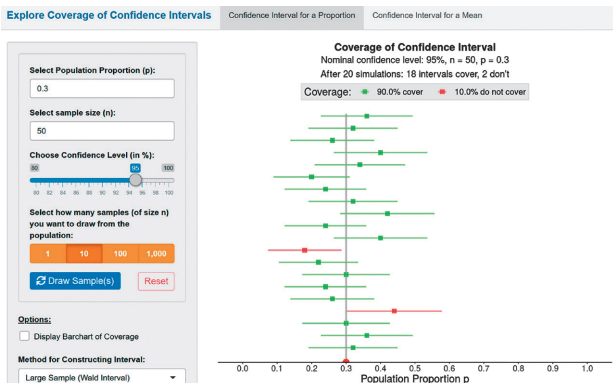
Screenshot from the *Sampling Distribution for the Sample Mean* app

- The **Inference for a Proportion** and the **Inference for a Mean** apps carry out statistical inference. They provide graphs and summary statistics from the observed data to check assumptions, and they compute margin of errors, confidence intervals, test statistics, and P-values, displaying them graphically.



Screenshot from the *Inference for a Population Mean* app

- The **Explore Coverage** app uses simulation to demonstrate the concept of the confidence coefficient, both for confidence intervals for the proportion and for the mean. Different sliders for true population parameters, the sample size, or the confidence coefficient show their effect on the coverage and width of confidence intervals. The **Errors and Power** app visually explores and provides Type I and Type II errors and power under various settings.



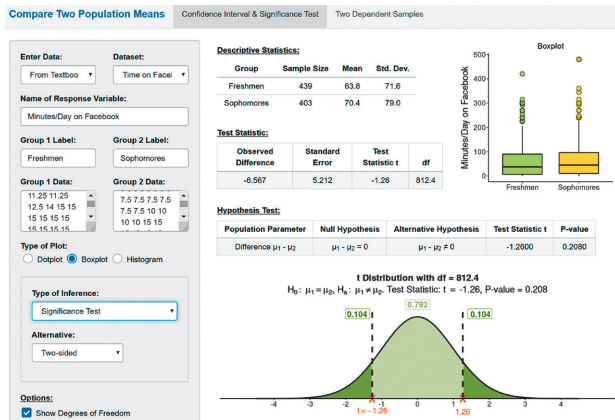
Screenshot from the *Explore Coverage* app

- The **Bootstrap for One Sample** app illustrates the idea behind the bootstrap and provides bootstrap percentile confidence intervals for the mean, median, or standard deviation.



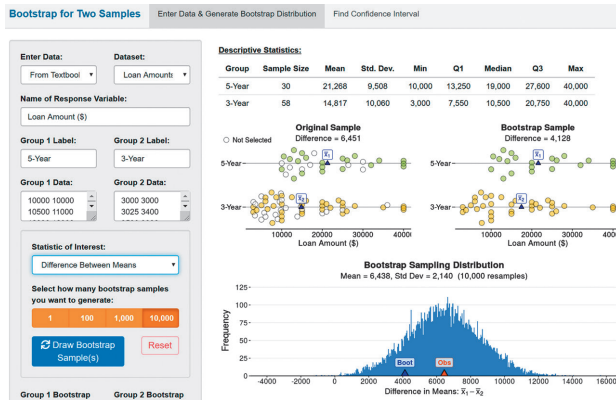
Screenshot from the *Bootstrap for One Sample* app

- The **Compare Two Proportions** and **Compare Two Means** apps provide inference based on data from two independent or dependent samples. They show graphs and summary statistics for data preloaded from the textbook or entered by the user and compute confidence intervals or P-values for the difference between two proportions or means. All results are illustrated graphically. Users can provide their own data in a variety of ways, including as summary statistics.



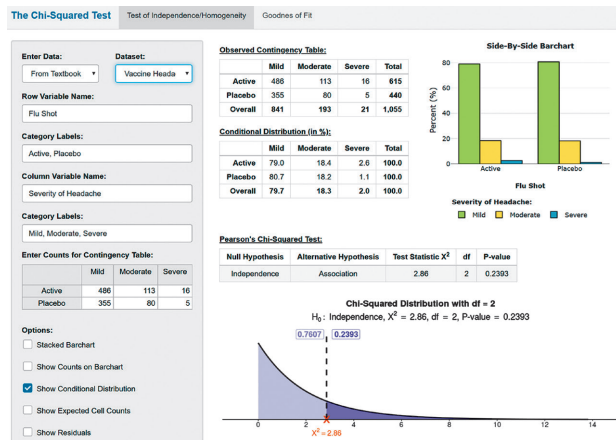
Screenshot from the *Compare Two Means* app

- The **Bootstrap for Two Samples** app illustrates how the bootstrap distribution is built via resampling and provides bootstrap percentile confidence intervals for the difference in means or medians. The **Scatterplots & Correlation** app provides resampling inference for the correlation coefficient. The **Permutation Test** app illustrates and carries out the permutation test comparing two groups with a quantitative response.



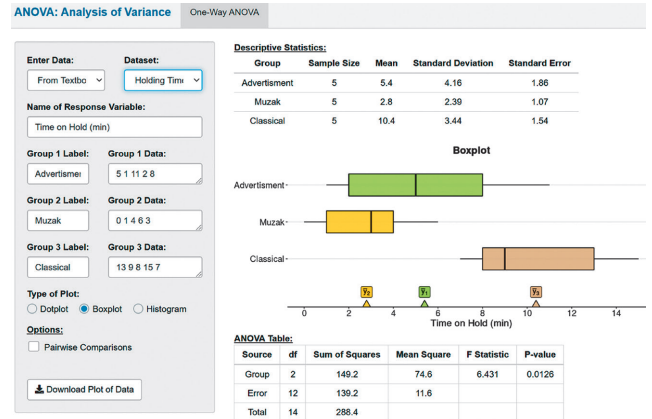
Screenshot from the *Bootstrap for Two Samples* app

- The **Chi-Squared Test** app provides the chi-squared test for independence. Results are illustrated using graphs, and users can enter data as contingency tables. The corresponding permutation test is available in the **Permutation Test for Independence (Categ. Data)** app and for 2×2 tables in the **Fisher's Exact Test** app.



Screenshot from the *Chi-Squared Test* app

- The **ANOVA (One-Way)** app compares several means and provides simultaneous confidence intervals for multiple comparisons. The **Wilcoxon Test** and the **Kruskal-Wallis Test** apps for nonparametric statistics provide the analogs to the two-sample t test and ANOVA, both for independent and dependent samples.



Screenshot from the *ANOVA (One-Way)* app

- The **Multiple Linear Regression** app fits linear models with several explanatory variables, and the **Logistic Regression** app fits a simple logistic regression model, complete with graphical output.

Preface

Each of us is impacted daily by information and decisions based on data, in our jobs or free time and on topics ranging from health and medicine, the environment and climate to business, financial or political considerations. In 2020, never was data and statistical fluency more important than with understanding the information associated with the COVID-19 pandemic and the measures taken globally to protect against this virus. Fueled by the digital revolution, data is now readily available for gaining insights not only into epidemiology but a variety of other fields. Many believe that data are the new oil; it is valuable, but only if it can be refined and used in context. This book strives to be an essential part of this refinement process.

We have each taught introductory statistics for many years, to different audiences, and we have witnessed and welcomed the evolution from the traditional formula-driven mathematical statistics course to a concept-driven approach. One of our goals in writing this book was to help make the conceptual approach more interesting and readily accessible to students. In this new edition, we take it a step further. We believe that statistical ideas and concepts come alive and are memorable when students have a chance to interact with them. In addition, those ideas stay relevant when students have a chance to use them on real data. That's why we now provide many screenshots, activities and exercises that use our online web apps to learn statistics and carry out nearly all the statistical analysis in this book. This removes the technological barrier and allows students to replicate our analysis and then apply it on their own data. At the end of the course, we want students to look back and realize that they learned practical concepts for reasoning with data that will serve them not only for their future careers, but in many aspects of their lives.

We also want students to come to appreciate that in practice, assumptions are not perfectly satisfied, models are not exactly correct, distributions are not exactly normal, and different factors should be considered in conducting a statistical analysis. The title of our book reflects the experience of statisticians and data scientists, who soon realize that statistics is an art as well as a science.

What's New in This Edition

Our goal in writing the fifth edition was to improve the student and instructor experience and provide a more accessible introduction to statistical thinking and practice. We have:

- Integrated the online apps from ArtofStat.com into every chapter through multiple screenshots and activities, inviting the user to replicate results using data from our many examples. These apps can be used live in lectures (and together with students) to visualize concepts or carry out analysis. Or, they can be recorded for asynchronous online instructions. Students can download (or screenshot) results for inclusion in assignments or projects. At the end of each chapter, a new section on statistical software describes the functionality of each app used in the chapter.
- Streamlined Chapter 1, which continues to emphasize the importance of the statistical investigative process. New in Chapter 1 is an introduction on the opportunities and challenges with Big Data and Data Science, including a discussion of ethical considerations.

- Written a new (but optional) section on linear transformations in Chapter 2.
- Emphasized in Section 3.1 the two descriptive statistics that students are most likely to encounter in the news media: differences and ratios of proportions.
- Expanded the discussion on multivariate thinking in Section 3.3.
- Explicitly stated and discussed the definition of the standard deviation of a discrete random variable in Section 6.1.
- Included the simulation of drawing random samples from *real* population distributions to illustrate the concept of a sampling distribution in Section 7.2.
- Significantly expanded the coverage of resampling methods, with a thorough discussion of the bootstrap for one- and two-sample problems and for the correlation coefficient. (New Sections 7.3, 8.3, and 10.3.)
- Continued emphasis on using interval estimation for inference and less reliance on significance testing and incorporated the 2016 American Statistical Association's statement on P-values.
- Provided commented R code in a new section on statistical software at the end of each chapter which replicates output obtained in featured examples.
- Updated or created many new examples and exercises using the most recent data available.

Our Approach

In 2005 (and updated in 2016), the American Statistical Association (ASA) endorsed guidelines and recommendations for the introductory statistics course as described in the report, “Guidelines for Assessment and Instruction in Statistics Education (GAISE) for the College Introductory Course” (www.amstat.org/education/gaise). The report states that the overarching goal of all introductory statistics courses is to produce statistically educated students, which means that students should develop statistical literacy and the ability to think statistically. The report gives six key recommendations for the college introductory course:

1. Teach statistical thinking.
 - Teach statistics as an investigative process of problem solving and decision making.
 - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

We wholeheartedly support these recommendations and our textbook takes every opportunity to implement these guidelines.

Ask and Answer Interesting Questions

In presenting concepts and methods, we encourage students to think about the data and the appropriate analyses by posing questions. Our approach, learning by framing questions, is carried out in various ways, including (1) presenting a structured approach to examples that separates the question and the analysis from the scenario presented, (2) letting the student replicate the analysis with a specific app and not worrying about calculations, (3) providing homework exercises that encourage students to think and write, and (4) asking questions in the figure captions that are answered in the Chapter Review.

Present Concepts Clearly

Students have told us that this book is more “readable” and interesting than other introductory statistics texts because of the wide variety of intriguing real data examples and exercises. We have simplified our prose wherever possible, without sacrificing any of the accuracy that instructors expect in a textbook.

A serious source of confusion for students is the multitude of inference methods that derive from the many combinations of confidence intervals and tests, means and proportions, large sample and small sample, variance known and unknown, two-sided and one-sided inference, independent and dependent samples, and so on. We emphasize the most important cases for practical application of inference: large sample, variance unknown, two-sided inference, and independent samples. The many other cases are also covered (except for known variances), but more briefly, with the exercises focusing mainly on the way inference is commonly conducted in practice. We present the traditional probability distribution–based inference but now also include inference using simulation through bootstrapping and permutation tests.

Connect Statistics to the Real World

We believe it’s important for students to be comfortable with analyzing a balance of both quantitative and categorical data so students can work with the data they most often see in the world around them. Every day in the media, we see and hear percentages and rates used to summarize results of opinion polls, outcomes of medical studies, and economic reports. As a result, we have increased the attention paid to the analysis of proportions. For example, we use contingency tables early in the text to illustrate the concept of association between two categorical variables and to show the potential influence of a lurking variable.

Organization of the Book

The statistical investigative process has the following components: (1) asking a statistical question; (2) obtaining appropriate data; (3) analyzing the data; and (4) interpreting the data and making conclusions to answer the statistical questions. With this in mind, the book is organized into four parts.

Part 1 focuses on gathering, exploring and describing data and interpreting the statistics they produce. This equates to components 1, 2, and 3, when the data is analyzed descriptively (both for one variable and the association between two variables).

Part 2 covers probability, probability distributions, and sampling distributions. This equates to component 3, when the student learns the underlying probability necessary to make the step from analyzing the data descriptively to analyzing the data inferentially (for example, understanding sampling distributions to develop the concept of a margin of error and a P-value).

Part 3 covers inferential statistics. This equates to components 3 and 4 of the statistical investigative process. The students learn how to form confidence intervals and conduct significance tests and then make appropriate conclusions answering the statistical question of interest.

Part 4 introduces statistical modeling. This part cycles through all 4 components. The chapters are written in such a way that instructors can teach out of order. For example, after Chapter 1, an instructor could easily teach Chapter 4, Chapter 2, and Chapter 3. Alternatively, an instructor may teach Chapters 5, 6, and 7 after Chapters 1 and 4.

Features of the Fifth Edition

Promoting Student Learning

To motivate students to think about the material, ask appropriate questions, and develop good problem-solving skills, we have created special features that distinguish this text.

Chapter End Summaries

Each chapter ends with a high-level summary of the material studied, a review of new notation, learning objectives, and a section on the use of the apps or the software program R.

Student Support

To draw students to important material we highlight key definitions, and provide summaries in blue boxes throughout the book. In addition, we have four types of margin notes:

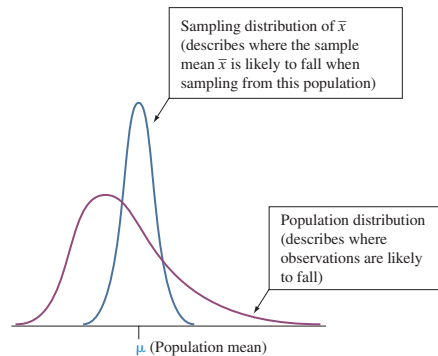
In Words

To find the **95% confidence interval**, you take the sample proportion and add and subtract 1.96 standard errors.

Recall

From Section 8.2, for using a confidence interval to estimate a proportion p , the **standard error** of a sample **proportion** \hat{p} is

$$\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$



- **In Words:** This feature explains, in plain language, the definitions and symbolic notation found in the body of the text (which, for technical accuracy, must be more formal).
- **Caution:** These margin boxes alert students to areas to which they need to pay special attention, particularly where they are prone to make mistakes or incorrect assumptions.
- **Recall:** As the student progresses through the book, concepts are presented that depend on information learned in previous chapters. The Recall margin boxes direct the reader back to a previous presentation in the text to review and reinforce concepts and methods already covered.
- **Did You Know:** These margin boxes provide information that helps with the contextual understanding of the statistical question under consideration or provides additional information.

Graphical Approach

Because many students are visual learners, we have taken extra care to make our **figures** informative. We've annotated many of the figures with labels that clearly identify the noteworthy aspects of the illustration. Further, many figure captions include a question (answered in the Chapter Review) designed to challenge the student to interpret and think about the information being communicated by the graphic. The graphics also feature a pedagogical use of color to help students recognize patterns and distinguish between statistics and parameters. The use of color is explained in the very front of the book for easy reference.

Hands-On Activities and Simulations

Each chapter contains activities that allow students to become familiar with a number of statistical methodologies and tools. The instructor can elect to carry out the activities in class, outside of class, or a combination of both. These hands-on activities and simulations encourage students to learn by doing.

Statistics: In Practice

We realize that there is a difference between proper academic statistics and what is actually done in practice. Data analysis in practice is an art as well as a science. Although statistical theory has foundations based on precise assumptions and

conditions, in practice the real world is not so simple. **In Practice** boxes and text references alert students to the way statisticians actually analyze data in practice. These comments are based on our extensive consulting experience and research and by observing what well-trained statisticians do in practice.

Examples and Exercises

Innovative Example Format

Recognizing that the worked examples are the major vehicle for engaging and teaching students, we have developed a unique structure to help students learn by relying on five components:

- **Picture the Scenario** presents background information so students can visualize the situation. This step places the data to be investigated in context and often provides a link to previous examples.
- **Questions to Explore** reference the information from the scenario and pose questions to help students focus on what is to be learned from the example and what types of questions are useful to ask about the data.
- **Think It Through** is the heart of each example. Here, the questions posed are investigated and answered using appropriate statistical methods. Each solution is clearly matched to the question so students can easily find the response to each Question to Explore.
- **Insight** clarifies the central ideas investigated in the example and places them in a broader context that often states the conclusions in less technical terms. Many of the Insights also provide connections between seemingly disparate topics in the text by referring to concepts learned previously and/or foreshadowing techniques and ideas to come.
- **Try Exercise:** Each example concludes by directing students to an end-of-section exercise that allows immediate practice of the concept or technique within the example.

TRY

Residuals

APP

Concept tags are included with each example so that students can easily identify the concept covered in the example.

App integration allows instructors to focus on the data and the results instead of on the computations. Many of our Examples show screenshots and their output is easily reproduced, live in class or later by students on their own.


Relevant and Engaging Exercises

The text contains a strong emphasis on real data in both the examples and exercises. We have updated the exercise sets in the fifth edition to ensure that students have ample opportunity (with more than 100 exercises in almost every chapter) to practice techniques and apply the concepts. We show how statistics addresses a wide array of applications, including opinion polls, market research, the environment, and health and human behavior. Because we believe that most students benefit more from focusing on the underlying concepts and interpretations of data analyses than from the actual calculations, many exercises can be carried out using one of our web apps.

We have exercises in two places:

- **At the end of each section.** These exercises provide immediate reinforcement and are drawn from concepts within the section.
- **At the end of each chapter.** This more comprehensive set of exercises draws from all concepts across all sections within the chapter. There, you will also find multiple-choice and True/False-type exercises and those covering more advanced material.

Each exercise has a descriptive label. Exercises for which an app can be used are indicated with the **APP** icon. Exercises that use data from the General Social

Survey are indicated by a  icon. Larger data sets used in examples and exercises are referenced in the text, listed in the back endpapers, and made available for **download** on the book's website ArtofStat.com.

Technology Integration

Web Apps

Screenshots of web apps are shown throughout the text. An overview of the apps is available on pages vii–ix. At the end of each chapter, in the Statistical Software section, the functionality of each app used in the chapter is described in detail. A list of apps by chapter appears on the second to last last page of this book.

R

We now provide R code in the end of chapter section on statistical software. Every R function used is explained and illustrated with output. In addition, annotated R scripts can be found on ArtofStat.com under “R Code.” While only base R functions are used in the book, the scripts available online use functions from the tidyverse.

Data Sets

We use a wealth of real data sets throughout the textbook. These data sets are available for download at ArtofStat.com. The same data set is often used in several chapters, helping reinforce the four components of the statistical investigative process and allowing the students to see the big picture of statistical reasoning. Many data sets are also preloaded and accessible from within the apps.

StatCrunch[®]

StatCrunch[®] is powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. The vibrant online community offers more than 40,000 data sets for instructors to use and students to analyze.

- **Collect.** Users can upload their own data to StatCrunch or search a large library of publicly shared data sets, spanning almost any topic of interest. Also, an online survey tool allows users to collect data quickly through web-based surveys.
- **Crunch.** A full range of numerical and graphical methods allows users to analyze and gain insights from any data set. Interactive graphics help users understand statistical concepts and are available for export to enrich reports with visual representations of data.
- **Communicate.** Reporting options help users create a wide variety of visually appealing representations of their data.

Full access to StatCrunch is available with MyLab Statistics, and StatCrunch is available by itself to qualified adopters. For more information, visit www.statcrunch.com or contact your Pearson representative.

An Invitation Rather Than a Conclusion

We hope that students using this textbook will gain a lasting appreciation for the vital role that statistics plays in presenting and analyzing data and informing decisions. Our major goals for this textbook are that students learn how to:

- Recognize that we are surrounded by data and the importance of becoming statistically literate to interpret these data and make informed decisions based on data.
- Become critical readers of studies summarized in mass media and of research papers that quote statistical results.

- Produce data that can provide answers to properly posed questions.
- Appreciate how probability helps us understand randomness in our lives and grasp the crucial concept of a sampling distribution and how it relates to inference methods.
- Choose appropriate descriptive and inferential methods for examining and analyzing data and drawing conclusions.
- Communicate the conclusions of statistical analyses clearly and effectively.
- Understand the limitations of most research, either because it was based on an observational study rather than a randomized experiment or survey or because a certain lurking variable was not measured that could have explained the observed associations.

We are excited about sharing the insights that we have learned from our experience as teachers and from our students through this book. Many students still enter statistics classes on the first day with dread because of its reputation as a dry, sometimes difficult, course. It is our goal to inspire a classroom environment that is filled with creativity, openness, realistic applications, and learning that students find inviting and rewarding. We hope that this textbook will help the instructor and the students experience a rewarding introductory course in statistics.

Get the *most* out of MyLab Statistics



MyLab Statistics Online Course for *Statistics: The Art and Science of Learning from Data* by Agresti, Franklin, and Klingenberg (access code required)

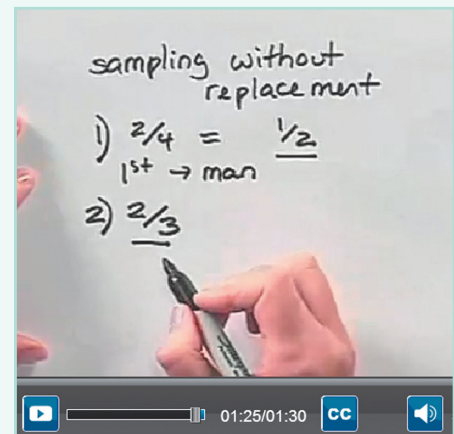
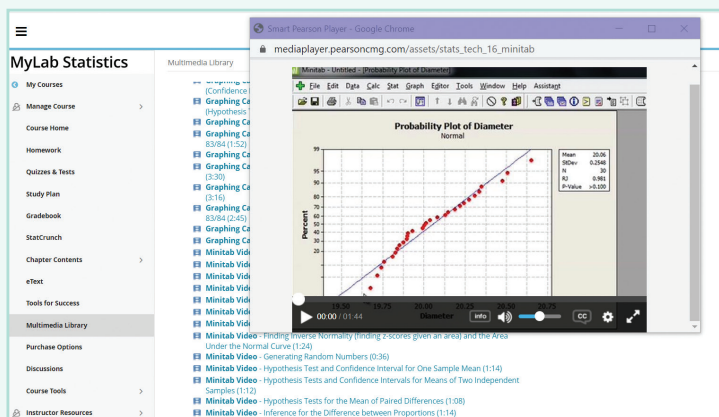
MyLab Statistics is available to accompany Pearson's market leading text offerings. To give students a consistent tone, voice, and teaching method, each text's flavor and approach is tightly integrated throughout the accompanying MyLab Statistics course, making learning the material as seamless as possible.

Technology Tutorials and Study Cards

Technology tutorials provide brief video walkthroughs and step-by-step instructional study cards on common statistical procedures for MINITAB[®], Excel[®], and the TI family of graphing calculators.

Example-Level Resources

Students looking for additional support can use the example-based videos to help solve problems, provide reinforcement on topics and concepts learned in class, and support their learning.



Resources for Success



Instructor Resources

Additional resources can be downloaded from www.pearson.com.

Online Resources Students and instructors have a full library of resources, including apps developed for in-text activities, data sets and instructor-to-instructor videos. Available through MyLab Statistics and at the Pearson Global Editions site, <https://media.pearsoncmg.com/intl/ge/abp/resources/index.html>.

Updated! Instructor to Instructor Videos provide an opportunity for adjuncts, part-timers, TAs, or other instructors who are new to teaching from this text or have limited class prep time to learn about the book's approach and coverage from the authors. The videos focus on those topics that have proven to be most challenging to students and offer suggestions, pointers, and ideas about how to present these topics and concepts effectively based on many years of teaching introductory statistics. The videos also provide insights on how to help students use the textbook in the most effective way to realize success in the course. The videos are available for download from Pearson's online catalog at www.pearson.com and through MyLab Statistics.

Instructor's Solutions Manual, by James Lapp, contains fully worked solutions to every textbook exercise.

PowerPoint Lecture Slides are fully editable and printable slides that follow the textbook. These slides can be used during lectures or posted to a

website in an online course. Fully accessible versions are also available.

TestGen[®] (www.pearson.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions.

The Online Test Bank is a test bank derived from TestGen[®]. It includes multiple choice and short answer questions for each section of the text, along with the answer keys.

Student Resources

Additional resources to help student success.

Online Resources Students and instructors have a full library of resources, including apps developed for in-text activities and data sets (Available through MyLab Statistics and the Pearson Global Editions site, <https://media.pearsoncmg.com/intl/ge/abp/resources/index.html>).

Study Cards for Statistics Software This series of study cards, available for Excel[®], MINITAB[®], JMP[®], SPSS[®], R[®], StatCrunch[®], and the TI family of graphing calculators provides students with easy, step-by-step guides to the most common statistics software. These study cards are available through MyLab Statistics.

Acknowledgments

We appreciate all of the thoughtful contributions and additions to the fourth edition by Michael Posner Villanova University. We are indebted to the following individuals, who provided valuable feedback for the fifth edition:

John Bennett, *Halifax Community College, NC*; Yishi Wang, *University of North Carolina Wilmington, NC*; Scott Crawford, *University of Wyoming, WY*; Amanda Ellis, *University of Kentucky, KY*; Lisa Kay, *Eastern Kentucky University, KY*; Inyang Inyang, *Northern Virginia Community College, VA*; Jason Moliterno, *Sacred Heart University, CT*; Daniel Turek, *Williams College, MA*.

We are also indebted to the many reviewers, class testers, and students who gave us invaluable feedback and advice on how to improve the quality of the book.

ARIZONA Russel Carlson, University of Arizona; Peter Flanagan-Hyde, Phoenix Country Day School ■ **CALIFORNIA** James Curl, Modesto Junior College; Christine Drake, University of California at Davis; Mahtash Esfandiari, UCLA; Brian Karl Finch, San Diego State University; Dawn Holmes, University of California Santa Barbara; Rob Gould, UCLA; Rebecca Head, Bakersfield College; Susan Herring, Sonoma State University; Colleen Kelly, San Diego State University; Marke Mavis, Butte Community College; Elaine McDonald, Sonoma State University; Corey Manchester, San Diego State University; Amy McElroy, San Diego State University; Helen Noble, San Diego State University; Calvin Schmall, Solano Community College; Scott Nickleach, Sonoma State University; Ann Kalinoskii, San Jose University ■ **COLORADO** David Most, Colorado State University ■ **CONNECTICUT** Paul Bugl, University of Hartford; Anne Doyle, University of Connecticut; Pete Johnson, Eastern Connecticut State University; Dan Miller, Central Connecticut State University; Kathleen McLaughlin, University of Connecticut; Nalini Ravishanker, University of Connecticut; John Vangar, Fairfield University; Stephen Sawin, Fairfield University ■ **DISTRICT OF COLUMBIA** Hans Engler, Georgetown University; Mary W. Gray, American University; Monica Jackson, American University ■ **FLORIDA** Nazanin Azarnia, Santa Fe Community College; Brett Holbrook; James Lang, Valencia Community College; Karen Kinard, Tallahassee Community College; Megan Mocko, University of Florida; Maria Ripol, University of Florida; James Smart, Tallahassee Community College; Latricia Williams, St. Petersburg Junior College, Clearwater; Doug Zahn, Florida State University ■ **GEORGIA** Carrie Chmielarski, University of Georgia; Ouida Dillon, Oconee County High School; Kim Gilbert, University of Georgia; Katherine Hawks, Meadowcreek High School; Todd Hendricks, Georgia Perimeter College; Charles LeMarsh, Lakeside High School; Steve Messig, Oconee County High School; Broderick Oluyede, Georgia Southern University; Chandler Pike, University of Georgia; Kim Robinson, Clayton State University; Jill Smith, University of Georgia; John Seppala, Valdosta State University; Joseph Walker, Georgia State University; Evelyn Bailey, Oxford College of Emory University; Michael Roty, Mercer University ■ **HAWAII** Erica Bernstein, University of Hawaii at Hilo ■ **IOWA** John Cryer, University of Iowa; Kathy Rogotzke, North Iowa Community College; R. P. Russo, University of Iowa; William Duckworth, Iowa State University ■ **ILLINOIS** Linda Brant Collins, University of Chicago; Dagmar Budikova, Illinois State University; Ellen Fireman, University of Illinois; Jinadasa Gamage, Illinois State University; Richard Maher, Loyola University Chicago; Cathy Poliak, Northern Illinois University; Daniel Rowe, Heartland Community College ■ **KANSAS** James Higgins, Kansas State University; Michael Mosier, Washburn University; Katherine C. Earles, Wichita State University ■ **KENTUCKY** Lisa Kay, Eastern Kentucky University ■ **MASSACHUSETTS** Richard Cleary, Bentley University; Katherine Halvorsen, Smith College; Xiaoli Meng, Harvard University; Daniel Weiner, Boston University ■ **MICHIGAN** Kirk Anderson, Grand Valley State University; Phyllis Curtiss, Grand Valley State University; Roy Erickson, Michigan State University; Jann-Huei Jinn, Grand Valley State University; Sango Otieno, Grand Valley State University; Alla Sikorskii, Michigan State University; Mark Stevenson, Oakland Community College; Todd Swanson, Hope College; Nathan Tintle, Hope College; Phyllis Curtiss, Grand Valley State; Ping-Shou Zhong, Michigan State University ■ **MINNESOTA** Bob Dobrow, Carleton College; German J. Pliego, University of St. Thomas;

Peihua Qui, University of Minnesota; Engin A. Sungur, University of Minnesota–Morris ■ **MISSOURI** Lynda Hollingsworth, Northwest Missouri State University; Robert Paige, Missouri University of Science and Technology; Larry Ries, University of Missouri–Columbia; Suzanne Tourville, Columbia College ■ **MONTANA** Jeff Banfield, Montana State University ■ **NEW JERSEY** Harold Sackrowitz, Rutgers, The State University of New Jersey; Linda Tappan, Montclair State University ■ **NEW MEXICO** David Daniel, New Mexico State University ■ **NEW YORK** Brooke Fridley, Mohawk Valley Community College; Martin Lindquist, Columbia University; Debby Lurie, St. John’s University; David Mathiason, Rochester Institute of Technology; Steve Stehman, SUNY ESF; Tian Zheng, Columbia University; Bernadette Lanciaux, Rochester Institute of Technology ■ **NEVADA** Alison Davis, University of Nevada–Reno ■ **NORTH CAROLINA** Pamela Arroway, North Carolina State University; E. Jacquelin Dietz, North Carolina State University; Alan Gelfand, Duke University; Gary Kader, Appalachian State University; Scott Richter, UNC Greensboro; Roger Woodard, North Carolina State University ■ **NEBRASKA** Linda Young, University of Nebraska ■ **OHIO** Jim Albert, Bowling Green State University; John Holcomb, Cleveland State University; Jackie Miller, The Ohio State University; Stephan Pelikan, University of Cincinnati; Teri Rysz, University of Cincinnati; Deborah Rumsey, The Ohio State University; Kevin Robinson, University of Akron; Dottie Walton, Cuyahoga Community College - Eastern Campus ■ **OREGON** Michael Marciniak, Portland Community College; Henry Mesa, Portland Community College, Rock Creek; Qi-Man Shao, University of Oregon; Daming Xu, University of Oregon ■ **PENNSYLVANIA** Winston Crawley, Shippensburg University; Douglas Frank, Indiana University of Pennsylvania; Steven Gendler, Clarion University; Bonnie A. Green, East Stroudsburg University; Paul Lupinacci, Villanova University; Deborah Lurie, Saint Joseph’s University; Linda Myers, Harrisburg Area Community College; Tom Short, Villanova University; Kay Somers, Moravian College; Sister Marcella Louise Wallowicz, Holy Family University ■ **SOUTH CAROLINA** Beverly Diamond, College of Charleston; Martin Jones, College of Charleston; Murray Siegel, The South Carolina Governor’s School for Science and Mathematics; Ellen Breazel, Clemson University ■ **SOUTH DAKOTA** Richard Gayle, Black Hills State University; Daluss Siewert, Black Hills State University; Stanley Smith, Black Hills State University ■ **TENNESSEE** Bonnie Daves, Christian Academy of Knoxville; T. Henry Jablonski, Jr., East Tennessee State University; Robert Price, East Tennessee State University; Ginger Rowell, Middle Tennessee State University; Edith Seier, East Tennessee State University; Matthew Jones, Austin Peay State University ■ **TEXAS** Larry Ammann, University of Texas, Dallas; Tom Bratcher, Baylor University; Jianguo Liu, University of North Texas; Mary Parker, Austin Community College; Robert Paige, Texas Tech University; Walter M. Potter, Southwestern University; Therese Shelton, Southwestern University; James Surlles, Texas Tech University; Diane Resnick, University of Houston–Downtown; Rob Eby, Blinn College—Bryan Campus ■ **UTAH** Patti Collings, Brigham Young University; Carolyn Cuff, Westminster College; Lajos Horvath, University of Utah; P. Lynne Nielsen, Brigham Young University ■ **VIRGINIA** David Bauer, Virginia Commonwealth University; Ching-Yuan Chiang, James Madison University; Jonathan Duggins, Virginia Tech; Steven Garren, James Madison University; Hasan Hamdan, James Madison University; Debra Hydorn, Mary Washington College; Nusrat Jahan, James Madison University; D’Arcy Mays, Virginia Commonwealth University; Stephanie Pickle, Virginia Polytechnic Institute and State University ■ **WASHINGTON** Rich Alldredge, Washington State University; Brian T. Gill, Seattle Pacific University; June Morita, University of Washington; Linda Dawson, Washington State University, Tacoma ■ **WISCONSIN** Brooke Fridley, University of Wisconsin–LaCrosse; Loretta Robb Thielman, University of Wisconsin–Stout ■ **WYOMING** Burke Grandjean, University of Wyoming ■ **CANADA** Mike Kowalski, University of Alberta; David Loewen, University of Manitoba

The detailed assessment of the text fell to our accuracy checkers, Nathan Kidwell and Joan Sanuik.

Thank you to James Lapp, who took on the task of revising the solutions manuals to reflect the many changes to the fifth edition.

We would like to thank the Pearson team who has given countless hours in developing this text; without their guidance and assistance, the text would not have come to completion. We thank Amanda Brands, Suzanna Bainbridge, Rachel Reeve, Karen Montgomery, Bob Carroll, Jean Choe, Alicia Wilson, Demetrius

Hall, and Joe Vetere. We also thank Kim Fletcher, Senior Project Manager at Integra-Chicago, for keeping this book on track throughout production.

Alan Agresti would like to thank those who have helped us in some way, often by suggesting data sets or examples. These include Anna Gottard, Wolfgang Jank, René Lee-Pack, Jacalyn Levine, Megan Lewis, Megan Mocko, Dan Nettleton, Yongyi Min, and Euijung Ryu. Many thanks also to Tom Piazza for his help with the General Social Survey. Finally, Alan Agresti would like to thank his wife, Jacki Levine, for her extraordinary support throughout the writing of this book. Besides putting up with the evenings and weekends he was working on this book, she offered numerous helpful suggestions for examples and for improving the writing.

Chris Franklin gives a special thank-you to her husband and sons, Dale, Corey, and Cody Green. They have patiently sacrificed spending many hours with their spouse and mom as she has worked on this book through five editions. A special thank-you also to her parents, Grady and Helen Franklin, and her two brothers, Grady and Mark, who have always been there for their daughter and sister. Chris also appreciates the encouragement and support of her colleagues and her many students who used the book, offering practical suggestions for improvement. Finally, Chris thanks her coauthors, Alan and Bernhard, for the amazing journey of writing a textbook together.

Bernhard Klingenberg wants to thank his statistics teachers in Graz, Austria; Sheffield, UK; and Gainesville, Florida, who showed him all the fascinating aspects of statistics throughout his education. Thanks also to the Department of Mathematics & Statistics at Williams College and the Graduate Program in Data Science at New College of Florida for being such great places to work. Finally, thanks to Chris Franklin and Alan Agresti for a wonderful and inspiring collaboration.

ALAN AGRESTI, *GAINESVILLE, FLORIDA*

CHRISTINE FRANKLIN, *ATHENS, GEORGIA*

BERNHARD KLINGENBERG, *WILLIAMSTOWN, MASSACHUSETTS*

Global Edition Acknowledgments

Pearson would like to thank the following people for contributing to and reviewing the Global edition:

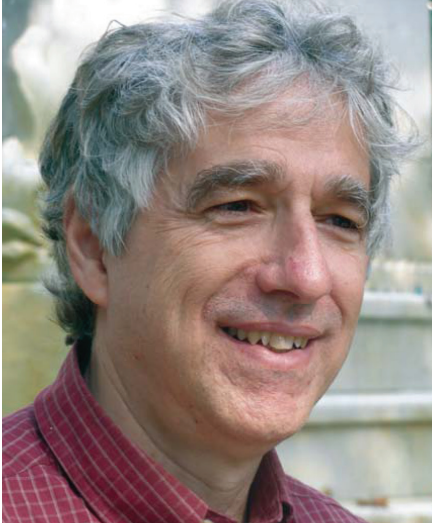
Contributors

■ DELHI Vikas Arora, professional statistician; Abhishek K. Umrawal, Delhi University ■ LEBANON Mohammad Kacim, Holy Spirit University of Kaslik

Reviewers

■ DELHI Amit Kumar Misra, Babasaheb Bhimrao Ambedkar University ■ TURKEY Ümit Işlak, Boğaziçi Üniversitesi; Özlem İlk Dağ, Middle East Technical University

About the Authors



Alan Agresti is Distinguished Professor Emeritus in the Department of Statistics at the University of Florida. He taught statistics there for 38 years, including the development of three courses in statistical methods for social science students and three courses in categorical data analysis. He is author of more than 100 refereed articles and six texts, including *Statistical Methods for the Social Sciences* (Pearson, 5th edition, 2018) and *An Introduction to Categorical Data Analysis* (Wiley, 3rd edition, 2019). He is a Fellow of the American Statistical Association and recipient of an Honorary Doctor of Science from De Montfort University in the UK. He has held visiting positions at Harvard University, Boston University, the London School of Economics, and Imperial College and has taught courses or short courses for universities and companies in about 30 countries worldwide. He has also received teaching awards from the University of Florida and an excellence in writing award from John Wiley & Sons.



Christine Franklin is the K-12 Statistics Ambassador for the American Statistical Association and elected ASA Fellow. She is retired from the University of Georgia as the Lothar Tresp Honoratus Honors Professor and Senior Lecturer Emerita in Statistics. She is the co-author of two textbooks and has published more than 60 journal articles and book chapters. Chris was the lead writer for American Statistical Association Pre-K-12 Guidelines for the Assessment and Instruction in Statistics Education (GAISE) Framework document, co-chair of the updated Pre-K-12 GAISE II, and chair of the ASA Statistical Education of Teachers (SET) report. She is a past Chief Reader for Advance Placement Statistics, a Fulbright scholar to New Zealand (2015), recipient of the United States Conference on Teaching Statistics (USCOTS) Lifetime Achievement Award and the ASA Founders Award, and an elected member of the International Statistical Institute (ISI). Chris loves being with her family, running, hiking, scoring baseball games, and reading mysteries.



Bernhard Klingenberg is Professor of Statistics in the Department of Mathematics & Statistics at Williams College, where he has been teaching introductory and advanced statistics classes since 2004, and in the Graduate Data Science Program at New College of Florida, where he teaches statistics and data visualization for data scientists. Bernhard is passionate about making statistical software accessible to everyone and in addition to co-authoring this book is actively developing the web apps featured in this edition. A native of Austria, Bernhard frequently returns there to hold visiting positions at universities and gives short courses on categorical data analysis in Europe and the United States. He has published several peer-reviewed articles in statistical journals and consults regularly with academia and industry. Bernhard enjoys photography (some of his pictures appear in this book), scuba diving, hiking state parks, and spending time with his wife and four children.

Gathering and Exploring Data



Chapter 1

Statistics: The Art and Science of Learning From Data

Chapter 2

Exploring Data With Graphs and Numerical Summaries

Chapter 3

Exploring Relationships Between Two Variables

Chapter 4

Gathering Data

- 1.1 Using Data to Answer Statistical Questions
- 1.2 Sample Versus Population
- 1.3 Organizing Data, Statistical Software, and the New Field of Data Science



Statistics: The Art and Science of Learning From Data

Example 1

How Statistics Helps Us Learn About the World

Picture the Scenario

In this book, you will explore a wide variety of everyday scenarios. For example, you will evaluate media reports about opinion surveys, medical research studies, the state of the economy, and environmental issues. You'll face financial decisions, such as choosing between an investment with a sure return and one that could make you more money but could possibly cost you your entire investment. You'll learn how to analyze the available information to answer necessary questions in such scenarios. One purpose of this book is to show you why an understanding of statistics is essential for making good decisions in an uncertain world.

Questions to Explore

This book will show you how to collect appropriate information and how to apply statistical methods so you can better evaluate that information

and answer the questions posed. Here are some examples of questions we'll investigate in coming chapters:

- How can you evaluate evidence about global warming?
- Are cell phones dangerous to your health?
- What's the chance your tax return will be audited?
- How likely are you to win the lottery?
- Is there bias against women in appointing managers?
- What "hot streaks" should you expect in basketball?
- How successful is the flu vaccine in preventing the flu?
- How can you predict the selling price of a house?

Thinking Ahead

Each chapter uses questions like these to introduce a topic and then introduces tools for making sense of the available information. We'll see that **statistics** is the art and science of designing studies and analyzing the information that those studies produce.

In the business world, managers use statistics to analyze results of marketing studies about new products, to help predict sales, and to measure employee performance. In public policy discussions, statistics is used to support or discredit proposed measures, such as gun control or limits on carbon emissions. Medical studies use statistics to evaluate whether new ways to treat a disease are better than existing ones. In fact, most professional occupations today rely heavily on statistical methods. In a competitive job market, understanding how to reason statistically provides an important advantage. New jobs, such as data scientist, are created to process the amount of information generated in a digital and connected world and have statistical thinking as one of their core components.

But it's important to understand statistics even if you will never use it in your job. Understanding statistics can help you make better choices. Why? Because every day you are bombarded with statistical information from news reports, advertisements, political campaigns, surveys, or even the health app on your smartphone. How do you know how to process and evaluate all the information? An understanding of the statistical reasoning—and, in some cases, statistical misconceptions—underlying these pronouncements will help. For instance, this book will enable you to evaluate claims about medical research studies more effectively so that you know when you should be skeptical. For example, does taking an aspirin daily truly lessen the chance of having a heart attack?

We realize that you are probably not reading this book in the hope of becoming a statistician. (That's too bad, because there's a severe shortage of statisticians and data scientists—more jobs than trained people. And with the ever-increasing ways in which statistics is being applied, it's an exciting time to work in these fields.) You may even suffer from math phobia. Please be assured that to learn the main concepts of statistics, logical thinking and perseverance are more important than high-powered math skills. Don't be frustrated if learning comes slowly and you need to read about a topic a few times before it starts to make sense. Just as you would not expect to sit through a single foreign language class session and be able to speak that language fluently, the same is true with the language of statistics. It takes time and practice. But we promise that your hard work will be rewarded. Once you have completed even part of this book, you will understand much better how to make sense of statistical information and, hence, the world around you.

1.1 Using Data to Answer Statistical Questions

Does a low-carbohydrate diet result in significant weight loss? Are people more likely to stop at a Starbucks if they've seen a recent Starbucks TV commercial? Information gathering is at the heart of investigating answers to such questions. The information we gather with experiments and surveys is collectively called **data**.

For instance, consider an experiment designed to evaluate the effectiveness of a low-carbohydrate diet. The data might consist of the following measurements for the people participating in the study: weight at the beginning of the study, weight at the end of the study, number of calories of food eaten per day, carbohydrate intake per day, body mass index (BMI) at the start of the study, and gender. A marketing survey about the effectiveness of a TV ad for Starbucks could collect data on the percentage of people who went to a Starbucks since the ad aired and analyze how it compares for those who saw the ad and those who did not see it.

Today, massive amounts of data are collected automatically, for instance, through smartphone apps. These often personal data are used to answer statistical questions about consumer behavior, such as who is more likely to buy which product or which type of applicant is least likely to pay back a consumer loan.

Defining Statistics

You already have a sense of what the word *statistics* means. You hear statistics quoted about sports events (number of points scored by each player on a basketball team), statistics about the economy (median income, unemployment rate), and statistics about opinions, beliefs, and behaviors (percentage of students who indulge in binge drinking). In this sense, a statistic is merely a number calculated from data. But statistics as a field is a way of thinking about data and quantifying uncertainty, not a maze of numbers and messy formulas.

Statistics

Statistics is the art and science of collecting, presenting, and analyzing data to answer an investigative question. Its ultimate goal is translating data into knowledge and an understanding of the world around us. In short, *statistics is the art and science of learning from data.*

The statistical investigative process involves four components: (1) formulate a statistical question, (2) collect data, (3) analyze data, and (4) interpret and communicate results. The following scenarios ask questions that we'll learn how to answer using such a process.

Scenario 1: Predicting an Election Using an Exit Poll In elections, television networks often declare the winner well before all the votes have been counted. They do this using exit polling, interviewing voters after they leave the voting booth. Using an exit poll, a network can often predict the winner after learning how several thousand people voted, out of possibly millions of voters.

The 2010 California gubernatorial race pitted Democratic candidate Jerry Brown against Republican candidate Meg Whitman. A TV exit poll used to project the outcome reported that 53.1% of a sample of 3,889 voters said they had voted for Jerry Brown. Was this sufficient evidence to project Brown as the winner, even though information was available from such a small portion of the more than 9.5 million voters in California? We'll learn how to answer that question in this book.

Scenario 2: Making Conclusions in Medical Research Studies Statistical reasoning is at the foundation of the analyses conducted in most medical research studies. Let's consider three examples of how statistics can be relevant.

Heart disease is the most common cause of death in industrialized nations. In the United States and Canada, nearly 30% of deaths yearly are due to heart disease, mainly heart attacks. Does regular aspirin intake reduce deaths from heart attacks? Harvard Medical School conducted a landmark study to investigate. The people participating in the study regularly took either an aspirin or a placebo (a pill with no active ingredient). Of those who took aspirin, 0.9% had heart attacks during the study. Of those who took the placebo, 1.7% had heart attacks, nearly twice as many.

Can you conclude that it's beneficial for people to take aspirin regularly? Or, could the observed difference be explained by how it was decided which people would receive aspirin and which would receive the placebo? For instance, might those who took aspirin have had better results merely because they were healthier, on average, than those who took the placebo? Or, did those taking aspirin have a better diet or exercise more regularly, on average?

For years there has been controversy about whether regular intake of large doses of vitamin C is beneficial. Some studies have suggested that it is. But some scientists have criticized those studies' designs, claiming that the subsequent statistical analysis was meaningless. How do we know when we can trust the statistical results in a medical study that is reported in the media?

Suppose you wanted to investigate whether, as some have suggested, heavy use of cell phones makes you more likely to get brain cancer. You could pick half the students from your school and tell them to use a cell phone each day for the next 50 years and tell the other half never to use a cell phone. Fifty years from now you could see whether more users than nonusers of cell phones got brain cancer. Obviously it would be impractical and unethical to carry out such a study. And who wants to wait 50 years to get the answer? Years ago, a British statistician figured out how to study whether a particular type of behavior has an effect on cancer, using already available data. He did this to answer a then controversial question: Does smoking cause lung cancer? How did he do this?

This book will show you how to answer questions like these. You'll learn when you can trust the results from studies reported in the media and when you should be skeptical.

Scenario 3: Using a Survey to Investigate People's Beliefs How similar are your opinions and lifestyle to those of others? It's easy to find out. Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of a few thousand adult Americans provides data about the opinions and behaviors of the American public. You can 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?" "How much TV do you watch per day?" and "How many sexual partners have you had in the past year?" Similar surveys occur in other countries, such as the Eurobarometer survey within the European Union. We'll use data from such surveys to illustrate the proper application of statistical methods.

Reasons for Using Statistical Methods

The scenarios just presented illustrate the three main components for answering a statistical investigative question:

- **Design:** Stating the goal and/or statistical question of interest and planning how to obtain data that will address them
- **Description:** Summarizing and analyzing the data that are obtained
- **Inference:** Making decisions and predictions based on the data for answering the statistical question

Design refers to planning how to obtain data that will efficiently shed light on the statistical question of interest. How could you conduct an experiment to determine reliably whether regular large doses of vitamin C are beneficial? In marketing, how do you select the people to survey so you'll get data that provide good predictions about future sales? Are there already available (public) data that you can use for your analysis, and what is the quality of the data?

Description refers to exploring and summarizing patterns in the data. Files of raw data are often huge. For example, over time the GSS has collected data about hundreds of characteristics on many thousands of people. Such raw data are not easy to assess—we simply get bogged down in numbers. It is more informative to use a few numbers or a graph to summarize the data, such as an average amount of TV watched or a graph displaying how the number of hours of TV watched per day relates to the number of hours per week exercising.

Inference refers to making decisions or predictions based on the data. Usually the decision or prediction applies to a larger group of people, not merely those in the study. For instance, in the exit poll described in Scenario 1, of 3,889 voters sampled, 53.1% said they voted for Jerry Brown. Using these data, we can predict (infer) that a majority of the 9.5 million voters voted for him. Stating the percentages for the sample of 3,889 voters is *description*, whereas predicting the outcome for all 9.5 million voters is *inference*.

In Words

The verb **infer** means to arrive at a decision or prediction by reasoning from known evidence. **Statistical inference** does this using data as evidence.

Statistical description and inference are complementary ways of analyzing data. Statistical description provides useful summaries and helps you find patterns in the data; inference helps you make predictions and decide whether observed patterns are meaningful. You can use both to investigate questions that are important to society. For instance, “Has there been global warming over the past decade?” “Is having the death penalty available for punishment associated with a reduction in violent crime?” “Does student performance in school depend on the amount of money spent per student, the size of the classes, or the teachers’ salaries?”

Long before we analyze data, we need to give careful thought to posing the questions to be answered by that analysis. The nature of these questions has an impact on all stages—design, description, and inference. For example, in an exit poll, do we just want to predict which candidate won, or do we want to investigate *why* by analyzing how voters’ opinions about certain issues related to how they voted? We’ll learn how questions such as these and the ones posed in the previous paragraph can be phrased in terms of statistical summaries (such as percentages and means) so that we can use data to investigate their answers.

Finally, a topic that we have not mentioned yet but that is fundamental for statistical reasoning is **probability**, which is a framework for quantifying how likely various possible outcomes are. We’ll study probability because it will help us answer questions such as, “If Brown were actually going to lose the election (that is, if he were supported by less than half of all voters), what’s the chance that an exit poll of 3,889 voters would show support by 53.1% of the voters?” If the chance were extremely small, we’d feel comfortable making the inference that his reelection was supported by the majority of all 9.5 million voters.

▶ Activity 1

Using Data Available Online to Answer Statistical Questions

Did you ever wonder about how much time people spend watching TV? We could use the survey question “On the average day, about how many hours do you personally watch television?” from the General Social Survey (GSS) to investigate this statistical question. Data from the GSS are accessible online via <http://sda.berkeley.edu>. Click on Archive (the second box in the top row) and then scroll down and click on the link labeled “General Social Survey (GSS) Cumulative Datafile 1972–2018.” You can also access this site directly and more easily via the link provided on the book’s website www.ArtofStat.com. (You may want to bookmark this link, as many activities in this book use data from the GSS.)

As shown in the screenshot,

- Enter TVHOURS in the field labeled Row. TVHOURS is the name the GSS uses for identifying the responses to the TV question.
- Enter YEAR(2018) in the Selection Filter field. This will select only the data for the 2018 GSS, not data from earlier (or later) iterations of the GSS, when this question was also included.
- Change the Weight to No Weight. This will specify that we see the actual frequency for each of the possible responses to the survey question and not some numbers adjusting for the way the survey was taken.

The screenshot shows the SDA Frequencies/Crosstabulation Program interface. At the top, there are tabs for Tables, Means, Correl. matrix, Comp. correl., and Regression. Below these are buttons for Logit/Probit and List values. The main area is titled "SDA Frequencies/Crosstabulation Program" with a help link. The form contains the following fields:

- Row: TVHOURS (Required)
- Column: (empty)
- Control: (empty)
- Selection Filter(s): YEAR(2018)
- Weight: No Weight

Below the form are three expandable sections: Output Options, Chart Options, and Decimal Options. At the bottom are two buttons: Run the Table and Clear Fields.

Press the Run the Table button on the bottom, and a new browser window will open that shows the frequencies (e.g., 145 and 349) and, in bold, the percentages (e.g., 9.3 and 22.4) of people who made each of the possible responses. (See the screenshot on the next page, which displays the top and bottom of the table.)

A **statistical analysis question** we can ask is, “What is the most common response?” From the bottom of the table we see that a total of 1,555 persons answered the question, of which 376 or, as shown, 24.2% ($= 100 \times (376/1,555)$) watched two

Frequency Distribution		
Cells contain: -Column percent -N of cases		Distribution
	0	9.3 145
	1	22.4 349
	2	24.2 376
	3	15.4 240
	4	10.9 169
	5	6.4 100

	20	.3 4
	24	.3 4
	COL TOTAL	100.0 1,555

hours of TV. This is the most common response. What is the second most common response? What percent of respondents watched zero hours of TV?

Another survey question asked by the GSS is, “Taken all together, would you say that you are very happy, pretty happy, or not too happy?” The GSS name for this question

is HAPPY. Return to the initial table but now enter HAPPY instead of TVHOURS as the row variable. Keep YEAR(2018) for the selection filter (if you leave this field blank, you get the cumulative data for all years for which this question was included in the GSS) and No Weight for the Weight box. Press Run the Table to replicate the screenshot shown below, which tells us that 29.9% ($= 100 \times (701/2,344)$) of the 2,344 respondents to this question in 2018 said that they are very happy.

Frequency Distribution		
Cells contain: -Column percent -N of cases		Distribution
HAPPY	1: VERY HAPPY	29.9 701
	2: PRETTY HAPPY	55.8 1,307
	3: NOT TOO HAPPY	14.3 336
	COL TOTAL	100.0 2,344

You can use the GSS to investigate statistical questions such as what kinds of people are more likely to be very happy. Those who have high incomes (GSS variable INCOME16) or are in good health (GSS variable HEALTH)? Are those who watch less TV more likely to be happy? We’ll explore how to investigate and answer these statistical questions using the statistical problem solving process throughout this book.

Try Exercises 1.3 and 1.5 ◀

1.1 Practicing the Basics

- 1.1 Aspirin, the wonder drug** An analysis by Professor Peter M Rothwell and his colleagues (Nuffield Department of Clinical Neuroscience, University of Oxford, UK) published in 2012 in the medical journal *The Lancet* (<http://www.thelancet.com>) assessed the effects of daily aspirin intake on cancer mortality. They looked at individual patient data from 51 randomized trials (77,000 participants) of daily intake of aspirin versus no aspirin or other anti-platelet agents. According to the authors, aspirin reduced the incidence of cancer, with maximum benefit seen when the scheduled duration of trial treatment was five years or more and resulted in a relative reduction in cancer deaths of about 15% (562 cancer deaths in the aspirin group versus 664 cancer deaths in the control group). Specify the aspect of this study that pertains to (a) design, (b) description, and (c) inference.
- 1.2 Poverty and age** The Current Population Survey (CPS) is a survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics. It provides a comprehensive body of data on the labor force, unemployment, wealth, poverty, and so on. The data can be found online at www.census.gov/cps/. The 2014 CPS ASEC (Annual Social and Economic Supplement) had redesigned questions for income that were implemented to a sample of approximately 30,000 addresses that were eligible to receive these. The

report indicated that 21.1% of children under 18 years, 13.5% of people between 18 to 64 years, and 10.0% of people 65 years and older were below the poverty line. Based on these results, the report concluded that the percentage of all people between the ages of 18 and 64 in poverty lies between 13.2% and 13.8%. Specify the aspect of this study that pertains to (a) description and (b) inference.

- 1.3 GSS and heaven** Go to the General Social Survey website, <http://sda.berkeley.edu>, press Archive, scroll down a bit and then select the “General Social Survey (GSS) Cumulative Datafile 1972–2018”, or whichever appears as the top (most recent) one. Alternatively, go there directly via the link provided on the book’s website, www.ArtofStat.com (see Activity 1). Enter the variable HEAVEN as the row variable; leave all other fields blank, but be certain to select No Weight for the Weight field. HEAVEN refers to the GSS question, “Do you believe in heaven?”
- Overall, how many people gave a valid response to the question whether they believe in heaven?
 - What percent of respondents said yes, definitely; yes, probably; no, probably not; and no, definitely not? (You can compute these percentages from the numbers

given in each category and the total shown in the last row of the frequency table, or just read them off from the percentages given in bold.)

- c. The questions in parts a and b refer to the cumulative data over all the years the GSS included this question (which was in 1991, 1998, 2008, and 2018). To get the percentages for each year separately, enter HEAVEN in the row variable and YEAR in the column variable. Press Run the Table (keeping Weight as No Weight). What is the proportion of respondents that answered yes, definitely in 2018? How does it compare to the proportion 10 years earlier?

1.4 GSS and heaven and hell Refer to the previous exercise, but now type HELL into the row variable box to obtain data on the question, “Do you believe in hell?” Keep all other fields blank and select No Weight.

GSS

- a. What percent of respondents said yes, definitely; yes, probably; no, probably not; and no, definitely not for the belief in hell question?

- b. To get data for this question for each year separately, type YEAR into the column variable field. Find the percentage of respondents who said yes, definitely to the hell question in 2018. How does this compare to the proportion of respondents in 2018 who believe in heaven?

1.5

TRY

GSS

GSS and work hours Go to the General Social Survey website, <http://sda.berkeley.edu>, press Archive, scroll down a bit and then select the “General Social Survey (GSS) Cumulative Datafile 1972–2018”, or whichever appears as the top (most recent) one. Alternatively, go there directly via the link provided on the book’s website, www.ArtofStat.com (see Activity 1). Enter the variable USUALHRS as the row variable, type YEAR(2016) in the selection filter, and choose No Weight for the Weight field. USUALHRS refers to the question, “How many hours per week do you usually work?”

- a. Overall, how many people responded to this question?
b. What was the most common response, and what percentage of people selected it?

1.2 Sample Versus Population

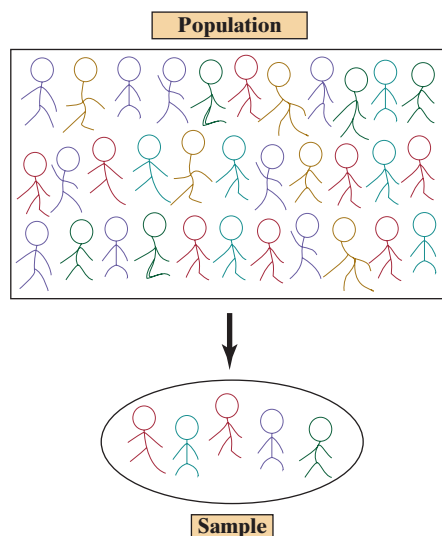
We’ve seen that statistics consists of methods for **designing** investigative studies, **describing** (summarizing) data obtained for those studies, and making **inferences** (decisions and predictions) based on those data to answer a statistical question of interest.

We Observe Samples but Are Interested in Populations

The entities that we measure in a study are called the **subjects**. Usually subjects are people, such as the individuals interviewed in a General Social Survey (GSS). But they need not be. For instance, subjects could be schools, countries, or days. We might measure characteristics such as the following:

- For each school: the per-student expenditure, the average class size, the average score of students on an achievement test
- For each country: the percentage of residents living in poverty, the birth rate, the percentage unemployed, the percentage who are computer literate
- For each day in an Internet café: the amount spent on coffee, the amount spent on food, the amount spent on Internet access

The **population** is the set of all the subjects of interest. In practice, we usually have data for only *some* of the subjects who belong to that population. These subjects are called a **sample**.



Population and Sample

The **population** is the total set of subjects in which we are interested. A **sample** is the subset of the population for whom we have (or plan to have) data, often randomly selected.

In the 2018 GSS, the sample consisted of the 2,348 people who participated in this survey. The population was the set of all adult Americans at that time—about 240 million people.