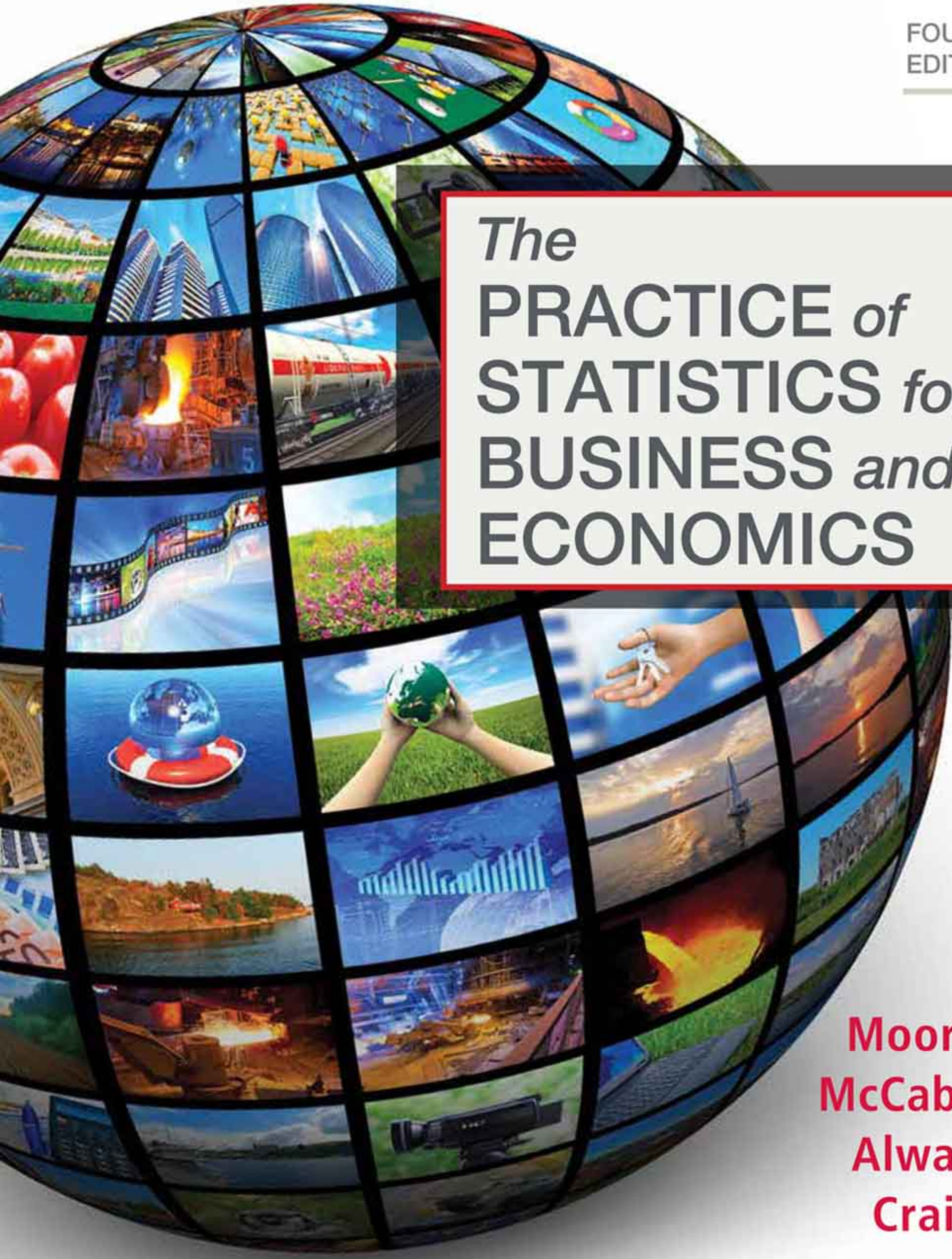
FOURTH EDITION



The PRACTICE *of* STATISTICS *for* BUSINESS *and* ECONOMICS

Moore
McCabe
Alwan
Craig

Table entry for z is the area under the standard Normal curve to the left of z.

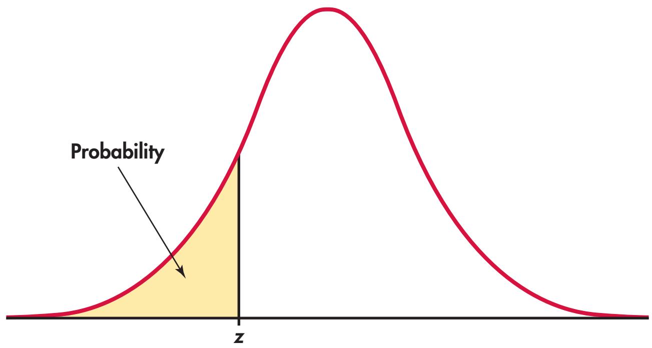


TABLE A Standard Normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table entry for z is the area under the standard Normal curve to the left of z.

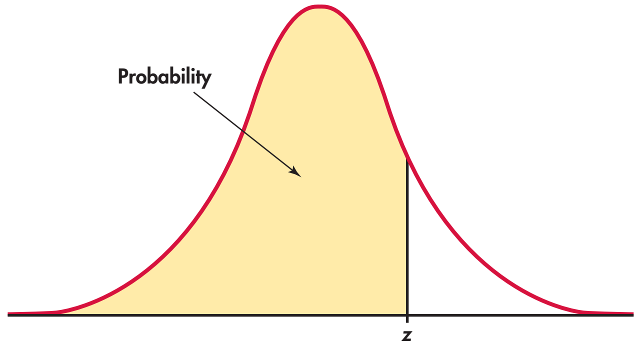


TABLE A Standard Normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The PRACTICE *of* STATISTICS *for* BUSINESS *and* ECONOMICS

this page left intentionally blank

The PRACTICE of STATISTICS for BUSINESS and ECONOMICS

FOURTH EDITION

David S. Moore
Purdue University

George P. McCabe
Purdue University

Layth C. Alwan
University of Wisconsin–Milwaukee

Bruce A. Craig
Purdue University

W. H. Freeman and Company
A Macmillan Education Imprint

Publisher: Terri Ward
Senior Acquisitions Editor: Karen Carson
Marketing Manager: Cara LeClair
Senior Developmental Editor: Katrina Mangold
Media Editor: Catriona Kaplan
Associate Editor: Marie Dripchak
Editorial Assistant: Victoria Garvey
Photo Editor: Cecilia Varas
Photo Researcher: Elyse Rieder
Cover and Text Designer: Blake Logan
Illustrations: MPS Ltd.
Senior Production Supervisor: Susan Wein
Project Management: MPS North America LLC
Composition: MPS Ltd.
Printing and Binding: RR Donnelley
Cover and Title Page Image: Oleksiy Mark/Shutterstock

Library of Congress Control Number: 2015948762

Instructor Complimentary Copy:
ISBN-13: 978-1-4641-3226-1
ISBN-10: 1-4641-3226-7

Hardcover:
ISBN-13: 978-1-4641-2564-5
ISBN-10: 1-4641-2564-3

Loose-leaf:
ISBN-13: 978-1-4641-3227-8
ISBN-10: 1-4641-3227-5

Printed in the United States of America

First printing

W. H. Freeman and Company
One New York Plaza
Suite 4500
New York, NY
10004-1562
www.macmillanhighered.com

BRIEF CONTENTS

The Core book includes Chapters 1–14. Chapters 15–17 are individual optional Companion Chapters and can be found at www.macmillanhighered.com/psbe4e.

this page left intentionally blank

CONTENTS

The following optional Companion Chapters can be found online at **www.macmillanhighered.com/psbe4e**.

TO INSTRUCTORS: *ABOUT THIS BOOK*

Statistics is the science of data. ***The Practice of Statistics for Business and Economics (PSBE)*** is an introduction to statistics for students of business and economics based on this principle. We present methods of basic statistics in a way that emphasizes working with data and mastering statistical reasoning. *PSBE* is elementary in mathematical level but conceptually rich in statistical ideas. After completing a course based on our text, we would like students to be able to think objectively about conclusions drawn from data and use statistical methods in their own work.

In *PSBE* we combine attention to basic statistical concepts with a comprehensive presentation of the elementary statistical methods that students will find useful in their work. We believe that you will enjoy using *PSBE* for several reasons:

1. *PSBE* examines the nature of modern statistical practice at a level suitable for beginners. We focus on the production and analysis of data as well as the traditional topics of probability and inference.
2. *PSBE* has a logical overall progression, so data production and data analysis are a major focus, while inference is treated as a tool that helps us to draw conclusions from data in an appropriate way.
3. *PSBE* presents data analysis as more than a collection of techniques for exploring data. We emphasize systematic ways of thinking about data. Simple principles guide the analysis: always plot your data; look for overall patterns and deviations from them; when looking at the overall pattern of a distribution for one variable, consider shape, center, and spread; for relations between two variables, consider form, direction, and strength; always ask whether a relationship between variables is influenced by other variables lurking in the background. We warn students about pitfalls in clear cautionary discussions.
4. *PSBE* uses real examples and exercises from business and economics to illustrate and enforce key ideas. Students learn the technique of least-squares regression and how to interpret the regression slope. But they also learn the conceptual ties between regression and correlation and the importance of looking for influential observations.
5. *PSBE* is aware of current developments both in statistical science and in teaching statistics. Brief, optional "Beyond the Basics" sections give quick overviews of topics such as density estimation, the bootstrap, scatterplot smoothers, data mining, nonlinear regression, and meta-analysis.

Themes of This Book

Look at your data is a consistent theme in *PSBE*. Rushing to inference—often automated by software—without first exploring the data is the most common source of statistical error that we see in working with users from many fields. A second theme is that *where the data come from matters.* When we do statistical inference, we are acting as if the data come from a properly randomized sample or experimental design. A basic understanding of these designs helps students grasp how inference works. The distinction between observational and experimental data helps students understand the truth of the mantra that "association does not imply causation." Moreover, managers need to understand the use of sample surveys for market research and customer satisfaction and the use of statistically designed experiments for product and process development and improvement.

Another strand that runs through *PSBE* is that data lead to decisions in a specific setting. A calculation or graph or "reject H0" is not the conclusion of an exercise in statistics. We encourage students to state a conclusion in the specific problem context, and we hope that you will require them to do so.

Finally, we think that a first course in any discipline should focus on the essentials. We have not tried to write an encyclopedia, but to equip students to use statistics (and learn more statistics as needed) by presenting the major concepts and most-used tools of the discipline. Longer lists of procedures "covered" tend to reduce student understanding and ability to use any procedures to deal with real problems.

What's New in the Fourth Edition

- **Chapter opener questions** Each chapter begins with a bulleted list of practical business questions that can be addressed by the methods in the chapter.
- **Data** Chapter 1 now begins with a short section giving a basic overview of data.
- **Categorical data** The material on descriptive statistics for categorical data in Chapter 2 as well as inference in Chapter 9 has been expanded to include mosaic plots as a visual tool to understand relationships.
- **Producing data** Chapter 3 now begins with a short section giving a basic overview of data sources.
- **Probability** We have reorganized the sections on probability models, general probability rules, and random variables so that they are now self-contained in one chapter (Chapter 4).
- **Distributions** Our reorganization of probability topics allows for a natural transition to Chapter 5 to be devoted to distributions on counts and proportions. New material has been added on the exploration of real data to check for compatibility with binomial and Poisson assumptions.
- **Inference** We have reorganized the sections on inference and sampling distributions so that they now flow in sequence. Material that previously appeared in Chapter 3 with a focus on proportions, concepts of sampling distributions, and estimation now appears in the last section of Chapter 5 ("Toward Statistical Inference"). This section is immediately followed by Chapter 6, which provides a complete treatment on inference for the mean.
- **Inference for means** Chapter 7 is retitled ("Inference for Means"), and the section on inference for population spread was moved to the one-way analysis of variance chapter (Chapter 14). In addition, Section 7.1 was streamlined by moving the discussion of inference for non-Normal populations to Section 7.3.
- **Sample size determination for means and proportions** Additional material on choosing sample sizes for one and two means or proportions using software is included in Chapters 7 and 8, respectively.
- **Equivalence testing** This topic is now included in Chapter 7, and the power calculations now appear in a separate section in this chapter.

- **Inference for categorical data** Chapter 9 is retitled ("Inference for Categorical Data"), and now includes goodness of fit as well as inference for two-way tables.
- **Quality control** Chapter 12 ("Statistics for Quality: Control and Capability") introduces the new topic of the moving-range chart for the monitoring of individual measurement processes. In addition, the calculations of process capability indices are now presented in manner typically reported in statistical software.
- **Time series** Chapter 13 ("Time Series Forecasting") introduces several new techniques, including the autocorrelation function (ACF) and partial autocorrelation function (PACF). In addition, we have introduced a new section on random walks. We also newly introduce to this chapter the use of moving averages and centered moving averages to estimate seasonal ratios and show how to use these ratios to deseasonalize a time series.
- **Exercises and examples** Approximately 50% of the exercises and examples are new or revised. We have placed additional emphasis on making the business or economics relevance of the exercises clear to the reader.
- **Increased emphasis on software** We have increased our emphasis on graphical displays of data. Software displays have been updated and are given additional prominence.

REMINDER

- **Reminders** At key points in the text, Reminder margin notes direct the reader to the first explanation of a topic, providing page numbers for easy reference.
- **Data file names** Data file names now include a short description of the content as well as the exercise or example number. Marginal icons show data set names for examples and in-text icons show the data set names for exercises.
- **Software basics** These have been expanded to include more software options and moved from the appendices at the end of each chapter to our online resources. These can now be found at **www.macmillanhighered.com/psbe4e**.

Content and Style

PSBE adapts to the business and economics statistics setting the approach to introductory instruction that was inaugurated and proved successful in the best-selling general statistics texts *Introduction to the Practice of Statistics* (eighth edition, Freeman 2014). *PSBE* features use of real data in examples and exercises and emphasizes statistical thinking as well as mastery of techniques. As the continuing revolution in computing automates most tiresome details, an emphasis on statistical concepts and on insight from data becomes both more practical for students and teachers and more important for users who must supply what is not automated.

Chapters 1 and 2 present the methods and unifying ideas of data analysis. Students appreciate the usefulness of data analysis, and realizing they can actually do it relieves a bit of their anxiety about statistics. We hope that they will grow accustomed to examining data and will continue to do so even when formal inference to answer a specific question is the ultimate goal. Note in particular that Chapter 2 gives an extended treatment of correlation and regression as descriptive tools, with attention to issues such as influential observations and the dangers posed by lurking variables. These ideas and tools have wider scope than an emphasis on inference (Chapters 10 and 11) allows. We think that a full discussion of data analysis for both one and several variables before students meet inference in these settings both reflects statistical practice and is pedagogically helpful.

Teachers will notice some nonstandard ideas in these chapters, particularly regarding the Normal distributions—we capitalize “Normal” to avoid suggesting that these distributions are “normal” in the usual sense of the word. We introduce density curves and Normal distributions in Chapter 1 as models for the overall pattern of some sets of data. Only later (Chapter 4) do we see that the same tools can describe probability distributions. Although unusual, this presentation reflects the historical origin of Normal distributions and also helps break up the mass of probability that is so often a barrier that students fail to surmount.

We use the notation $N(\mu, \sigma)$ rather than $N(\mu, \sigma^2)$ for Normal distributions. The traditional notation is, in fact, indefensible other than as inherited tradition. The standard deviation, not the variance, is the natural measure of scale in Normal distributions, visible on the density curve, used in standardization, and so on. We want students to think in terms of mean and standard deviation, so we talk in these terms.

In Chapter 3, we discuss random sampling and randomized comparative experiments. The exposition pays attention to practical difficulties, such as nonresponse in sample surveys, that can greatly reduce the value of data. We think that an understanding of such broader issues is particularly important for managers who must use data but do not themselves produce data. Discussion of statistics in practice alongside more technical material is part of our emphasis on data leading to practical decisions. We include a section on data ethics, a topic of increasing importance for business managers. Chapters 4 and 5 then present probability. We have chosen an unusual approach: Chapter 4 contains only the probability material that is needed to understand statistical inference, and this material is presented quite informally. The sections on probability models, general probability rules, and random variables have been reorganized so that they are now self-contained in this chapter. Chapter 5 now focuses on distributions of counts and proportions with new material on checking binomial and Poisson assumptions. It also concludes with a section titled “Toward Statistical Inference,” which introduces the concepts of parameters and statistics, sampling distributions, and bias and precision. This section provides a nice lead in to Chapter 6, which provides the reasoning of inference.

The remaining chapters present statistical inference, still encouraging students to ask where the data come from and to look at the data rather than quickly choosing a statistical test from an Excel menu. Chapter 6, which describes the reasoning of inference, is the cornerstone. Chapters 7 and 8 discuss one-sample and two-sample procedures for means and proportions, respectively, which almost any first course will cover. We take the opportunity in these core “statistical practice” chapters to discuss practical aspects of inference in the context of specific examples. Chapters 9, 10, and 11 present selected and more advanced topics in inference: two-way tables and simple and multiple regression. Chapters 12, 13, and 14 present additional advanced topics in inference: quality control, time series forecasting, and one-way analysis of variance.

Instructors who wish to customize a single-semester course or to add a second semester will find a wide choice of additional topics in the Companion Chapters that extend *PSBE*. These chapters are:

Chapter 15 Two-Way Analysis of Variance
Chapter 16 Nonparametric Tests
Chapter 17 Logistic Regression

Companion Chapters can be found on the book’s website:
www.macmillanhighered.com/psbe4e.

Accessible Technology

Any mention of the current state of statistical practice reminds us that quick, cheap, and easy computation has changed the field. Procedures such as our recommended two-sample t and logistic regression depend on software. Even the mantra "look at your data" depends—in practice—on software because making multiple plots by hand is too tedious when quick decisions are required. What is more, automating calculations and graphs increases students' ability to complete problems, reduces their frustration, and helps them concentrate on ideas and problem recognition rather than mechanics.

We therefore strongly recommend that a course based on PSBE *be accompanied by software of your choice.* Instructors will find using software easier because all data sets for *PSBE* can be found in several common formats on the website **www.macmillanhighered.com/psbe4e**.

The Microsoft Excel spreadsheet is by far the most common program used for statistical analysis in business. Our displays of output, therefore, emphasize Excel, though output from several other programs also appears. *PSBE* is not tied to specific software. Even so, one of our emphases is that a student who has mastered the basics of, say, regression can interpret and use regression output from almost any software.

We are well aware that Excel lacks many advanced statistical procedures. More seriously, Excel's statistical procedures have been found to be inaccurate, and they lack adequate warnings for users when they encounter data for which they may give incorrect answers. There is good reason for people whose profession requires continual use of statistical analysis to avoid Excel. But there are also good, practical reasons why managers whose life is not statistical prefer a program that they regularly use for other purposes. Excel appears to be adequate for simpler analyses of the kind that occur most often in business applications.

Some statistical work, both in practice and in *PSBE*, can be done with a calculator rather than software. Students should have at least a "two-variable statistics" calculator with functions for correlation and the least-squares regression line as well as for the mean and standard deviation. Graphing calculators offer considerably more capability. Because students have calculators, the text doesn't discuss "computing formulas" for the sample standard deviation or the least-squares regression line.

Technology can be used to assist learning statistics as well as doing statistics. The design of good software for learning is often quite different from that of software for doing. We want to call particular attention to the set of statistical applets available on the *PSBE* website: **www.macmillanhighered.com/psbe4e**. These interactive graphical programs are by far the most effective way to help students grasp the sensitivity of correlation and regression to outliers, the idea of a confidence interval, the way ANOVA responds to both within-group and among-group variation, and many other statistical fundamentals. Exercises using these applets appear throughout the text, marked by a distinctive icon. We urge you to assign some of these, and we suggest that if your classroom is suitably equipped, the applets are very helpful tools for classroom presentation as well.

APPLET

Carefully Structured Pedagogy

Few students find statistics easy. An emphasis on real data and real problems helps maintain motivation, and there is no substitute for clear writing. Beginning with data analysis builds confidence and gives students a chance to become familiar with your chosen software before the statistical content becomes intimidating. We have adopted several structural devices to aid students. Major settings that drive the exposition are presented as cases with more background information than other examples. (But we avoid the temptation to give so much information that the case obscures the statistics.) A distinctive icon ties together examples and exercises based on a case.

CASE

The *exercises* are structured with particular care. Short "Apply Your Knowledge" sections pose straightforward problems immediately after each major new idea. These give students stopping points (in itself a great help to beginners) and also tell them that "you should be able to do these things right now." Most numbered sections in the text end with a substantial set of exercises, and more appear as review exercises at the end of each chapter.

Acknowledgments

We are grateful to the many colleagues and students who have provided helpful comments about *PSBE*, as well as those who have provided feedback about *Introduction to the Practice of Statistics*. They have contributed to improving *PSBE* as well. In particular, we would like to thank the following colleagues who, as reviewers and authors of supplements, offered specific comments on *PSBE*, Fourth Edition:

Ala Abdelbaki, *University of Virginia*
Diane Bean, *Kirkwood Community College*
Tadd Colver, *Purdue University*
Bryan Crissinger, *University of Delaware*
Douglas Antola Crowe, *Bradley University*
John Daniel Draper, *The Ohio State University*
Anne Drougas, *Dominican University*
Gary Evans, *Purdue University*
Homi Fatemi, *Santa Clara University*
Mark A. Gebert, *University of Kentucky*
Kim Gilbert, *University of Georgia*
Matt Gnagey, *Weber State University*
Deborah J. Gougeon, *University of Scranton*
Betsy Greenberg, *University of Texas at Austin*
Susan Herring, *Sonoma State University*
Paul Holmes, *University of Georgia*
Patricia Humphrey, *Georgia Southern University*
Ronald Jorgensen, *Milwaukee School of Engineering*
Leigh Lawton, *University of St. Thomas*
James Manley, *Towson University*
Lee McClain, *Western Washington University*
Glenn Miller, *Pace University*
Carolyn H. Monroe, *Baylor University*
Hayley Nathan, *University of Wisconsin–Milwaukee*
Joseph Nolan, *Northern Kentucky University*
Karah Osterberg, *Northern Illinois University*
Charles J. Parker, *Wayne State College*
Hilde E. Patron Boenheim, *University of West Georgia*
Cathy D. Poliak, *University of Houston*
Michael Racer, *University of Memphis*
Terri Rizzo, *Lakehead University*
Stephen Robertson, *Southern Methodist University*
Deborah Rumsey, *The Ohio State University*
John Samons, *Florida State College at Jacksonville*
Bonnie Schroeder, *The Ohio State University*
Caroline Schruth, *University of Washington*
Carl Schwarz, *Simon Fraser University*
Sarah Sellke, *Purdue University*
Jenny Shook, *Pennsylvania State University*
Jeffrey Sklar, *California Polytechnic State University*
Rafael Solis, *California State University, Fresno*
Weixing Song, *Kansas State University*
Christa Sorola, *Purdue University*
Lynne Stokes, *Southern Methodist University*
Tim Swartz, *Simon Fraser University*
Elizabeth J. Wark, *Worchester State University*
Allen L. Webster, *Bradley University*
Mark Werner, *University of Georgia*
Blake Whitten, *University of Iowa*
Yuehua Wu, *York University*
Yan Yu, *University of Cincinnati*

MEDIA AND SUPPLEMENTS

The following electronic and print supplements are available with *The Practice of Statistics for Business and Economics*, Fourth Edition:

W. H. Freeman's new online homework system, **LaunchPad**, offers our quality content curated and organized for easy assignability in a simple but powerful interface. We have taken what we have learned from thousands of instructors and hundreds of thousands of students to create a new generation of W. H. Freeman/ Macmillan technology.

Curated units. Combining a curated collection of videos, homework sets, tutorials, applets, and e-Book content, LaunchPad's interactive units give instructors a building block to use as-is or as a starting point for customized learning units. A majority of exercises from the text can be assigned as online homework, including an abundance of algorithmic exercises. An entire unit's worth of work can be assigned in seconds, drastically reducing the amount of time it takes for instructors to have their course up and running.

Easily customizable. Instructors can customize the LaunchPad units by adding quizzes and other activities from our vast wealth of resources. They can also add a discussion board, a dropbox, and an RSS feed, with a few clicks. LaunchPad allows instructors to customize the student experience as much or as little as desired.

Useful analytics. The gradebook quickly and easily allows instructors to look up performance metrics for classes, individual students, and individual assignments.

Intuitive interface and design. The student experience is simplified. Students' navigation options and expectations are clearly laid out, ensuring they never get lost in the system.

Assets integrated into LaunchPad include the following:

Interactive e-Book Every LaunchPad e-Book comes with powerful study tools for students, video and multimedia content, and easy customization for instructors. Students can search, highlight, and bookmark, making it easier to study and access key content. And teachers can ensure that their classes get just the book they want to deliver: customizing and rearranging chapters; adding and sharing notes and discussions; and linking to quizzes, activities, and other resources.

LearningCurve provides students and instructors with powerful adaptive quizzing, a game-like format, direct links to the e-Book, and instant feedback. The quizzing system features questions tailored specifically to the text and adapts to students' responses, providing material at different difficulty levels and topics based on student performance.

SolutionMaster offers an easy-to-use Web-based version of the instructor's solutions, allowing instructors to generate a solution file for any set of homework exercises.

Statistical Video Series consists of StatClips, StatClips Examples, and Statistically Speaking "Snapshots." View animated lecture videos, whiteboard lessons, and documentary-style footage that illustrate key statistical concepts and help students visualize statistics in real-world scenarios.

NEW Video Technology Manuals available for TI-83/84 calculators, Minitab, Excel, JMP, SPSS (an IBM Company),* R (with and without Rcmdr), and CrunchIt!® provide 50 to 60 brief videos for using each specific statistical software in conjunction with a variety of topics from the textbook.

NEW StatBoards videos are brief whiteboard videos that illustrate difficult topics through additional examples, written and explained by a select group of statistics educators.

UPDATED StatTutor Tutorials offer multimedia tutorials that explore important concepts and procedures in a presentation that combines video, audio, and interactive features. The newly revised format includes built-in, assignable assessments and a bright new interface.

UPDATED Statistical Applets give students hands-on opportunities to familiarize themselves with important statistical concepts and procedures, in an interactive setting that allows them to manipulate variables and see the results graphically. Icons in the textbook indicate when an applet is available for the material being covered.

CrunchIt!® is W. H. Freeman's Web-based statistical software that allows users to perform all the statistical operations and graphing needed for an introductory business statistics course and more. It saves users time by automatically loading data from *PSBE*, and it provides the flexibility to edit and import additional data.

JMP Student Edition (developed by SAS) is easy to learn and contains all the capabilities required for introductory business statistics. JMP is the leading commercial data analysis software of choice for scientists, engineers, and analysts at companies throughout the world (for Windows and Mac).

Stats@Work Simulations put students in the role of the statistical consultant, helping them better understand statistics interactively within the context of real-life scenarios.

***EESEE* Case Studies** (*Electronic Encyclopedia of Statistical Examples and Exercises*), developed by The Ohio State University Statistics Department, teach students to apply their statistical skills by exploring actual case studies using real data.

Data files are available in JMP, ASCII, Excel, TI, Minitab, SPSS, R, and CSV formats.

Student Solutions Manual provides solutions to the odd-numbered exercises in the text.

Instructor's Guide with Full Solutions includes worked out solutions to all exercises, teaching suggestions, and chapter comments.

*SPSS was acquired by IBM in October 2009.

Test bank offers hundreds of multiple-choice questions and is available in LaunchPad. The test bank is also available at the website **www.macmillanhighered.com/psbe4e** (user registration as an instructor required) for Windows and Mac, where questions can be downloaded, edited, and resequenced to suit each instructor's needs.

Lecture slides offer a detailed lecture presentation of statistical concepts covered in each chapter of *PSBE*.

Additional Resources Available with *PSBE*, 4e

Website www.macmillanhighered.com/psbe4e This open-access website includes statistical applets, data files, and companion Chapters 15, 16, and 17. Instructor access to the website requires user registration as an instructor and features all the open-access student Web materials, plus

- **Image slides** containing all textbook figures and tables
- **Lecture slides**

Special Software Packages Student versions of JMP and Minitab are available for packaging with the text. JMP is available inside LaunchPad at no additional cost. Contact your W. H. Freeman representative for information, or visit **www.macmillanhighered.com**.

i-clicker is a two-way radio-frequency classroom response solution developed by educators for educators. Each step of i-clicker's development has been informed by teaching and learning. To learn more about packaging i-clicker with this textbook, please contact your local sales rep, or visit **www1.iclicker.com**.

TO STUDENTS: *WHAT IS STATISTICS?*

Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data.* We are bombarded by data in our everyday lives. The news mentions movie box-office sales, the latest poll of the president's popularity, and the average high temperature for today's date. Advertisements claim that data show the superiority of the advertiser's product. All sides in public debates about economics, education, and social policy argue from data. A knowledge of statistics helps separate sense from nonsense in this flood of data.

The study and collection of data are also important in the work of many professions, so training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisers, as well as policymakers in government and business, study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal consumer tastes and preferences. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers and, therefore, also make use of the methods of statistics. This means it is extremely likely that your undergraduate research projects will involve, at some level, the use of statistics.

Learning from Data

The goal of statistics is to learn from data. To learn, we often perform calculations or make graphs based on a set of numbers. But to learn from data, we must do more than calculate and plot because data are not just numbers; they are numbers that have some context that helps us learn from them.

Two-thirds of Americans are overweight or obese according to the Center for Disease Control and Prevention (CDC) website (**www.cdc.gov/nchs/nhanes.htm**). What does it mean to be obese or to be overweight? To answer this question, we need to talk about body mass index (BMI). Your weight in kilograms divided by the square of your height in meters is your BMI. A person who is 6 feet tall (1.83 meters) and weighs 180 pounds (81.65 kilograms) will have a BMI of $81.65/(1.83)^2 = 24.4$ kg/m^2. How do we interpret this number? According to the CDC, a person is classified as overweight or obese if their BMI is 25 kg/m^2 or greater and as obese if their BMI is 30 kg/m^2 or more. Therefore, two-thirds of Americans have a BMI of 25 kg/m^2 or more. The person who weighs 180 pounds and is 6 feet tall is not overweight or obese, but if he gains 5 pounds, his BMI would increase to 25.1 and he would be classified as overweight. What does this have to do with business and economics? Obesity in the United States costs about \$147 billion per year in direct medical costs!

When you do statistical problems, even straightforward textbook problems, don't just graph or calculate. Think about the context, and state your conclusions in the specific setting of the problem. As you are learning how to do statistical calculations and graphs, remember that the goal of statistics is not calculation for its own sake, but gaining understanding from numbers. The calculations and graphs can be automated by a calculator or software, but you must supply the understanding. This book presents only the most common specific procedures for statistical analysis. A thorough grasp of the principles of statistics will enable you to quickly learn more advanced methods as needed. On the other hand, a fancy computer analysis carried out without attention to basic principles will often produce elaborate nonsense. As you read, seek to understand the principles as well as the necessary details of methods and recipes.

The Rise of Statistics

Historically, the ideas and methods of statistics developed gradually as society grew interested in collecting and using data for a variety of applications. The earliest origins of statistics lie in the desire of rulers to count the number of inhabitants or measure the value of taxable land in their domains. As the physical sciences developed in the seventeenth and eighteenth centuries, the importance of careful measurements of weights, distances, and other physical quantities grew. Astronomers and surveyors striving for exactness had to deal with variation in their measurements. Many measurements should be better than a single measurement, even though they vary among themselves. How can we best combine many varying observations? Statistical methods that are still important were invented in order to analyze scientific measurements.

By the nineteenth century, the agricultural, life, and behavioral sciences also began to rely on data to answer fundamental questions. How are the heights of parents and children related? Does a new variety of wheat produce higher yields than the old and under what conditions of rainfall and fertilizer? Can a person's mental ability and behavior be measured just as we measure height and reaction time? Effective methods for dealing with such questions developed slowly and with much debate.

As methods for producing and understanding data grew in number and sophistication, the new discipline of statistics took shape in the twentieth century. Ideas and techniques that originated in the collection of government data, in the study of astronomical or biological measurements, and in the attempt to understand heredity or intelligence came together to form a unified "science of data." That science of data—statistics—is the topic of this text.

Business Analytics

The business landscape has become increasingly dominated with the terms of "business analytics," "predictive analytics," "data science," and "big data." These terms refer to the skills, technologies, and practices in the exploration of business performance data. Companies (for-profit and nonprofit) are increasingly making use of data and statistical analysis to discover meaningful patterns to drive decision making in all functional areas including accounting, finance, human resources, marketing, and operations. The demand for business managers with statistical and analytic skills has been growing rapidly and is projected to continue for many years to come. In 2014, LinkedIn reported the skill of "statistical analysis" as the number one hottest skill that resulted in a job hire.[1] In a *New York Times* interview, Google's senior vice president of people operations Laszlo Bock stated, "I took statistics at business school, and it was transformative for my career. Analytical training gives you a skill set that differentiates you from most people in the labor market."[2] Our goal with this text is to provide you with a solid foundation on a variety of statistical methods and the way to think critically about data. These skills will serve you well in a data-driven business world.

The Organization of This Book

The text begins with a discussion of data analysis and data production. The first two chapters deal with statistical methods for organizing and describing data. These chapters progress from simpler to more complex data. Chapter 1 examines data on a single variable, and Chapter 2 is devoted to relationships among two or more variables. You will learn both how to examine data produced by others and how to organize and summarize your own data. These summaries will first be graphical, then numerical, and then, when appropriate, in the form of a mathematical model that gives a compact description of the overall pattern of the data. Chapter 3 outlines arrangements (called designs) for producing data that answer specific questions. The principles presented in this chapter will help you to design proper samples and experiments for your research projects and to evaluate other such investigations in your field of study.

The next part of this book, consisting of Chapters 4 through 8, introduces statistical inference—formal methods for drawing conclusions from properly produced data. Statistical inference uses the language of probability to describe how reliable its conclusions are, so some basic facts about probability are needed to understand inference. Probability is the subject of Chapters 4 and 5. Chapter 6, perhaps the most important chapter in the text, introduces the reasoning of statistical inference. Effective inference is based on good procedures for producing data (Chapter 3), careful examination of the data (Chapters 1 and 2), and an understanding of the nature of statistical inference as discussed in Section 5.3 and Chapter 6. Chapters 7 and 8 describe some of the most common specific methods of inference, for drawing conclusions about means and proportions from one and two samples.

The five shorter chapters in the latter part of this book introduce somewhat more advanced methods of inference, dealing with relations in categorical data, regression and correlation, and analysis of variance. Supplementary chapters, available from the text website, present additional statistical topics.

What Lies Ahead

The Practice of Statistics for Business and Economics is full of data from many different areas of life and study. Many exercises ask you to express briefly some understanding gained from the data. In practice, you would know much more about the background of the data you work with and about the questions you hope the data will answer. No textbook can be fully realistic. But it is important to form the habit of asking "What do the data tell me?" rather than just concentrating on making graphs and doing calculations.

You should have some help in automating many of the graphs and calculations. You should certainly have a calculator with basic statistical functions. Look for keywords such as "two-variable statistics" or "regression" when you shop for a calculator. More advanced (and more expensive) calculators will do much more, including some statistical graphs. You may be asked to use software as well. There are many kinds of statistical software, from spreadsheets to large programs for advanced users of statistics. The kind of computing available to learners varies a great deal from place to place—but the big ideas of statistics don't depend on any particular level of access to computing.

Because graphing and calculating are automated in statistical practice, the most important assets you can gain from the study of statistics are an understanding of the big ideas and the beginnings of good judgment in working with data. Ideas and judgment can't (at least yet) be automated. They guide you in telling the computer what to do and in interpreting its output. This book tries to explain the most important ideas of statistics, not just teach methods. Some examples of big ideas that you will meet are "always plot your data," "randomized comparative experiments," and "statistical significance."

You learn statistics by doing statistical problems. "Practice, practice, practice." Be prepared to work problems. The basic principle of learning is persistence. Being organized and persistent is more helpful in reading this book than knowing lots of math. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. The gain will be worth the pain.

NOTES

1. See **blog.linkedin.com/2014/12/17/the-25-hottest-skills-that-got-people-hired-in-2014/**.
2. See **www.nytimes.com/2014/04/20/opinion/sunday/friedman-how-to-get-a-job-at-google-part-2.html?_r=0**.

INDEX OF CASES

INDEX OF DATA TABLES

BEYOND THE BASICS INDEX

ABOUT THE AUTHORS

David S. Moore is Shanti S. Gupta Distinguished Professor of Statistics, Emeritus, at Purdue University and was 1998 president of the American Statistical Association. He received his A.B. from Princeton and his Ph.D. from Cornell, both in mathematics. He has written many research papers in statistical theory and served on the editorial boards of several major journals. Professor Moore is an elected fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He has served as program director for statistics and probability at the National Science Foundation.

In recent years, Professor Moore has devoted his attention to the teaching of statistics. He was the content developer for the Annenberg/Corporation for Public Broadcasting college-level telecourse *Against All Odds: Inside Statistics* and for the series of video modules *Statistics: Decisions through Data,* intended to aid the teaching of statistics in schools. He is the author of influential articles on statistics education and of several leading texts. Professor Moore has served as president of the International Association for Statistical Education and has received the Mathematical Association of America's national award for distinguished college or university teaching of mathematics.

George P. McCabe is the Associate Dean for Academic Affairs in the College of Science and a Professor of Statistics at Purdue University. In 1966, he received a B.S. degree in mathematics from Providence College and in 1970 a Ph.D. in mathematical statistics from Columbia University. His entire professional career has been spent at Purdue with sabbaticals at Princeton; the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Melbourne, Australia; the University of Berne (Switzerland); the National Institute of Standards and Technology (NIST) in Boulder, Colorado; and the National University of Ireland in Galway. Professor McCabe is an elected fellow of the American Association for the Advancement of Science and of the American Statistical Association; he was 1998 Chair of its section on Statistical Consulting. In 2008–2010, he served on the Institute of Medicine Committee on Nutrition Standards for the National School Lunch and Breakfast Programs. He has served on the editorial boards of several statistics journals. He has consulted with many major corporations and has testified as an expert witness on the use of statistics in several cases.

Professor McCabe's research interests have focused on applications of statistics. Much of his recent work has focused on problems in nutrition, including nutrient requirements, calcium metabolism, and bone health. He is author or coauthor of more than 160 publications in many different journals.

Layth C. Alwan is an Associate Professor of Supply Chain, Operations Management and Business Statistics, Sheldon B. Lubar School of Business, University of Wisconsin–Milwaukee. He received a B.A. in mathematics, a B.S. in statistics, an

M.B.A., and a Ph.D. in business statistics/operations management, all from the University of Chicago, and an M.S. in computer science from DePaul University. Professor Alwan is an author of many research articles related to statistical process control and business forecasting. He has consulted for many leading companies on statistical issues related to quality, forecasting, and operations/supply chain management applications. On the teaching front, he is focused on engaging and motivating business students on how statistical thinking and data analysis methods have practical importance in business. He is the recipient of several teaching awards, including Business School Teacher of the Year and Executive MBA Outstanding Teacher of the Year.

Bruce A. Craig is Professor of Statistics and Director of the Statistical Consulting Service at Purdue University. He received his B.S. in mathematics and economics from Washington University in St. Louis and his Ph.D. in statistics from the University of Wisconsin–Madison. He is an elected fellow of the American Statistical Association and was Chair of its section on Statistical Consulting in 2009. He is also an active member of the Eastern North American Region of the International Biometrics Society and was elected by the voting membership to the Regional Committee between 2003 and 2006. Professor Craig has served on the editorial board of several statistical journals and has been a member of several data and safety monitoring boards, including Purdue's institutional review board.

Professor Craig's research interest focuses on the development of novel statistical methodology to address research questions, primarily in the life sciences. Areas of current interest are diagnostic testing and assessment, protein structure determination, and animal abundance estimation.

RANDY FARIS/CORBIS

CHAPTER 1

Examining Distributions

Introduction

Statistics is the science of learning from data. Data are numerical or qualitative descriptions of the objects that we want to study. In this chapter, we will master the art of examining data.

Data are used to inform decisions in business and economics in many different settings.

- Why has the AC Nielsen company been studying the habits of customers since it was founded in 1923?
- Who uses the databases of information maintained by the Better Business Bureau to make business decisions?
- How can data collected by the U.S. Chamber of Commerce be analyzed to provide summaries used to evaluate business opportunities?

We begin in Section 1.1 with some basic ideas about data. We learn about the different types of data that are collected and how data sets are organized.

Section 1.2 starts our process of learning from data by looking at graphs. These visual displays give us a picture of the overall patterns in a set of data. We have excellent software tools that help us make these graphs. However, it takes a little experience and a lot of judgment to study the graphs carefully and to explain what they tell us about our data.

Section 1.3 continues our process of learning from data by computing numerical summaries. These sets of numbers describe key characteristics of the patterns that we saw in our graphical summaries.

A statistical model is an idealized framework that helps us to understand variables and relationships between variables. In the first three sections, we focus on numerical and graphical ways to describe data. In Section 1.4, the final section of this chapter, we introduce the idea of a density curve as a

CHAPTER OUTLINE

way to describe the distribution of a variable. The most important statistical model is the Normal distribution, which is introduced here. Normal distributions are used to describe many sets of data. They also play a fundamental role in the methods that we use to draw conclusions from many sets of data.

1.1 Data

A statistical analysis starts with a set of data. We construct a set of data by first deciding what *cases* or units we want to study. For each case, we record information about characteristics that we call *variables*.

Cases, Labels, Variables, and Values

Cases are the objects described by a set of data. Cases may be customers, companies, subjects in a study, or other objects.

A **label** is a special variable used in some data sets to distinguish the different cases.

A **variable** is a characteristic of a case.

Different cases can have different **values** for the variables.

THE PHOTO WORKS

EXAMPLE 1.1 Restaurant Discount Coupons

DATA COUPONS

A website offers coupons that can be used to get discounts for various items at local restaurants. Coupons for food are very popular. Figure 1.1 gives information for seven restaurant coupons that were available for a recent weekend. These are the cases. Data for each coupon are listed on a different line, and the first column has the coupons numbered from 1 to 7. The next columns gives the type of restaurant, the name of the restaurant, the item being discounted, the regular price, and the discount price.

FIGURE 1.1 Food discount coupons, Example 1.1.

Excel

	A	B	C	D	E	F
1	ID	Type	Name	Item	RegPrice	DiscPrice
2	1	Italian	Domo's	Pizza	20	10
3	2	Italian	Mama Rita's	Pizza	20	12
4	3	BBQ	Smokey McSween's	Barbecue	30	17
5	4	BBQ	Smokey Grill	Ribs	20	11
6	5	Mexican	Dos Amigos	Tacos	16	8
7	6	Mexican	Holy Guacamole	Steak fajitas	13	8
8	7	Seafood	Sea Grille	Shrimp platter	20	11

Some variables, like the type of restaurant, the name of the restaurant, and the item simply place coupons into categories. The regular price and discount price columns have numerical values for which we can do arithmetic. It makes sense to give an average of the regular prices, but it does not make sense to give an "average" type of restaurant. We can, however, do arithmetic to compare the regular prices classified by type of restaurant.

Categorical and Quantitative Variables

A **categorical variable** places a case into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations, such as adding and averaging, make sense.

EXAMPLE 1.2 Categorical and Quantitative Variables for Coupons

DATA COUPONS

The restaurant discount coupon file has six variables: coupon number, type of restaurant, name of restaurant, item, regular price, and discount price. The two price variables are quantitative variables. Coupon number, type of restaurant, name of restaurant, and item are categorical variables.

An appropriate label for your cases should be chosen carefully. In our food coupon example, a natural choice of a label would be the name of the restaurant. However, if there are two or more coupons available for a particular restaurant, or if a restaurant is a chain with different discounts offered at different locations, then the name of the restaurant would not uniquely label each of the coupons.

APPLY YOUR KNOWLEDGE

DATA COUPONS

1.1 How much is the discount worth? Refer to Example 1.1. Add another column to the spreadsheet that gives the value of the coupon. Explain how you computed the entries in this column. Does the new column contain values for a categorical variable or for a quantitative variable? Explain your answer.

In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Who?** What **cases** do the data describe? **How many** cases appear in the data?

2. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? In what **unit of measurement** is each variable recorded?

3. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about cases other than the ones we actually have data for? Are the variables that are recorded suitable for the intended purpose?

APPLY YOUR KNOWLEDGE

1.2 Read the spreadsheet. Refer to Figure 1.1. Give the regular price and the discount price for the Smokey Grill ribs coupon.

DATA COUPONS

1.3 Who, what, and why for the restaurant discount coupon data. What cases do the data describe? How many cases are there? How many variables are there? What are their definitions and units of measurement? What purpose do the data have?

spreadsheet

The display in Figure 1.1 is from an Excel **spreadsheet.** Spreadsheets are very useful for doing the kind of simple computations that you did in Exercise 1.1. You can type in a formula and have the same computation performed for each row.

Note that the names we have chosen for the variables in our spreadsheet do not have spaces. For example, we could have used the name "Restaurant Name" for the name of the restaurant rather than Name. In some statistical software packages, however, spaces are not allowed in variable names. For this reason, when creating spreadsheets for eventual use with statistical software, it is best to avoid spaces in variable names. Another convention is to use an underscore (_) where you would normally use a space. For our data set, we could have used Regular_Price and Discount_Price for the two price variables.

EXAMPLE 1.3 Accounting Class Data

Suppose that you are a teaching assistant for an accounting class and one of your jobs is to keep track of the grades for students in two sections of the course. The cases are the students in the class. There are weekly homework assignments that are graded, two exams during the semester, and a final exam. Each of these components is given a numerical score, and the components are added to get a total score that can range from 0 to 1000. Cutoffs of 900, 800, 700, etc., are used to assign letter grades of A, B, C, etc.

The spreadsheet for this course will have seven variables:

- an identifier for each student
- the number of points earned for homework
- the number of points earned for the first exam
- the number of points earned for the second exam
- the number of points earned for the final exam
- the total number of points earned
- the letter grade earned.

There are no units of measurement for student identifier and the letter grade. These are categorical variables. The student identifier is a label. The other variables are measured in "points." Because we can do arithmetic with their values, these variables are quantitative variables.

EXAMPLE 1.4 Accounting Class Data for a Different Purpose

Suppose the data for the students in the accounting class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set that includes other variables such as Gender, PrevAcct (whether or not the student has taken an accounting course in high school), and Year (student classification as first, second, third, or fourth year). The label, student identifier, is a categorical variable, variables involving points are quantitative, and the remaining variables are all categorical.

In our examples of accounting class data, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using A = 4, B = 3, C = 2, D = 1, and F = 0. The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average.

Sometimes, experts argue about numerical scales such as this. They ask whether or not the difference between an A and a B is the same as the difference between a D and an F. Similarly, many questionnaires ask people to respond on a 1 to 5 scale, with 1 representing strongly agree, 2 representing agree, etc. Again we could ask whether or not the five possible values for this scale are equally spaced in some sense. From a practical point of view, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data.

APPLY YOUR KNOWLEDGE

1.4 Apartment rentals for students. A data set lists apartments available for students to rent. Information provided includes the monthly rent, whether or not a fitness center is provided, whether or not pets are allowed, the number of bedrooms, and the distance to the campus. Describe the cases in the data set, give the number of variables, and specify whether each variable is categorical or quantitative.

Knowledge of the context of data includes an understanding of the variables that are recorded. Often, the variables in a statistical study are easy to understand: height in centimeters, study time in minutes, and so on. But each area of work also has its own special variables. A marketing research department measures consumer behavior using a scale developed for its customers. A health food store combines various types of data into a single measure that it will use to determine whether or not to put a new store in a particular location. These kinds of variables are measured with special **instruments.** Part of mastering your field of work is learning what variables are important and how they are best measured.

instrument

Be sure that each variable really does measure what you want it to. A poor choice of variables can lead to misleading conclusions. Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple count of occurrences.

rate

EXAMPLE 1.5 Comparing Colleges Based on Graduates

Think about comparing colleges based on the numbers of graduates. This view tells you something about the relative sizes of different colleges. However, if you are interested in how well colleges succeed at graduating students whom they admit, it would be better to use a rate. For example, you can find data on the Internet on the six-year graduation rates of different colleges. These rates are computed by examining the progress of first-year students who enroll in a given year. Suppose that at College A there were 1000 first-year students in a particular year, and 800 graduated within six years. The graduation rate is

$$\frac{800}{1000} = 0.80$$

or 80%. College B has 2000 students who entered in the same year, and 1200 graduated within six years. The graduation rate is

$$\frac{1200}{2000} = 0.60$$

or 60%. How do we compare these two colleges? College B has more graduates, but College A has a better graduation rate.

APPLY YOUR KNOWLEDGE

1.5 Which variable would you choose? Refer to the previous example on colleges and their graduates.

(a) Give a setting where you would prefer to evaluate the colleges based on the numbers of graduates. Give a reason for your choice.
(b) Give a setting where you would prefer to evaluate the colleges based on the graduation rates. Give a reason for your choice.

adjusting one variable to create another

In Example 1.5, when we computed the graduation rate, we used the total number of students to adjust the number of graduates. We constructed a new variable by dividing the number of graduates by the total number of students. Computing a rate is just one of several ways of **adjusting one variable to create another.** In Exercise 1.1 (page 3), you computed the value of the discount by subtracting the discount price from the regular price. We often divide one variable by another to compute a more meaningful variable to study.

Exercise 1.5 illustrates an important point about presenting the results of your statistical calculations. *Always consider how to best communicate your results to a general audience.* For example, the numbers produced by your calculator or by statistical software frequently contain more digits than are needed. Be sure that you do not include extra information generated by software that will distract from a clear explanation of what you have found.

SECTION 1.1 Summary

- A data set contains information on a number of **cases.** Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.
- For each case, the data give values for one or more **variables.** A variable describes some characteristic of a case, such as a person's height, gender, or salary. Variables can have different **values** for different cases.
- A **label** is a special variable used to identify cases in a data set.
- Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each case, such as height in centimeters or annual salary in dollars.
- The **key characteristics** of a data set answer the questions Who?, What?, and Why?
- A **rate** is sometimes a more meaningful measure than a count.

SECTION 1.1 Exercises

For Exercises 1.1 to 1.3, see page 3; for 1.4, see page 5; and for 1.5, see page 6.

1.6 Summer jobs. You are collecting information about summer jobs that are available for college students in your area. Describe a data set that you could use to organize the information that you collect.
(a) What are the cases?
(b) Identify the variables and their possible values.
(c) Classify each variable as categorical or quantitative. Be sure to include at least one of each.
(d) Use a label and explain how you chose it.
(e) Summarize the key characteristics of your data set.

1.7 Employee application data. The personnel department keeps records on all employees in a company. Here is the information kept in one of the data files: employee identification number, last name, first name, middle initial, department, number of years with the company, salary, education (coded as high school, some college, or college degree), and age.
(a) What are the cases for this data set?
(b) Identify each item in the data file as a label, a quantitative variable, or a categorical variable.
(c) Set up a spreadsheet that could be used to record the data. Give appropriate column headings, and include three sample cases.

1.8 Where should you locate your business? You are interested in choosing a new location for your business. Create a list of criteria that you would use to rank cities. Include at least six variables, and give reasons for your choices. Will you use a label? Classify each variable as quantitative or categorical.

1.9 Survey of customers. A survey of customers of a restaurant near your campus wanted opinions regarding the following variables: (a) quality of the restaurant; (b) portion size; (c) overall satisfaction with the restaurant; (d) respondent's age; (e) whether the respondent is a college student; (f) whether the respondent ate there at least once a week. Responses for items (a), (b), and (c) are given a scale of 1 (very dissatisfied) to 5 (very satisfied). Classify each of these variables as categorical or quantitative, and give reasons for your answers.

1.10 Your survey of customers. Refer to the previous exercise. Make up your own customer survey with at least six questions. Include at least two categorical variables and at least two quantitative variables. Tell which variables are categorical and which are quantitative. Give reasons for your answers.

1.11 Study habits of students. You are planning a survey to collect information about the study habits of college students. Describe two categorical variables and two quantitative variables that you might measure for each student. Give the units of measurement for the quantitative variables.

1.12 How would you rate colleges? Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

1.13 Attending college in your state or in another state. The U.S. Census Bureau collects a large amount of information concerning higher education.[1] For example, the bureau provides a table that includes the following variables: state, number of students from the state who attend college, and number of students who attend college in their home state.

(a) What are the cases for this set of data?

(b) Is there a label variable? If yes, what is it?

(c) Identify each variable as categorical or quantitative.

(d) Consider a variable computed as the number of students in each state who attend college in the state divided by the total number of students from the state who attend college. Explain how you would use this variable to describe something about the states.

1.14 Alcohol-impaired driving fatalities. A report on drunk-driving fatalities in the United States gives the number of alcohol-impaired driving fatalities for each state.[2] Discuss at least two different ways that these numbers could be converted to rates. Give the advantages and disadvantages of each.

1.2 Displaying Distributions with Graphs

exploratory data analysis

Statistical tools and ideas help us examine data to describe their main features. This examination is called **exploratory data analysis.** Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We follow these principles in organizing our learning. The rest of this chapter presents methods for describing a single variable. We study relationships among two or more variables in Chapter 2. Within each chapter, we begin with graphical displays, then add numerical summaries for a more complete description.

Categorical variables: Bar graphs and pie charts

distribution of a categorical variable

The values of a categorical variable are labels for the categories, such as "Yes" and "No." The **distribution of a categorical variable** lists the categories and gives either the **count** or the **percent** of cases that fall in each category.

EXAMPLE 1.6 How Do You Do Online Research?

A study of 552 first-year college students asked about their preferences for online resources. One question asked them to pick their favorite.[3] Here are the results:

Resource	Count (n)
Google or Google Scholar	406
Library database or website	75
Wikipedia or online encyclopedia	52
Other	19
Total	552

Resource is the categorical variable in this example, and the values are the names of the online resources.

Note that the last value of the variable resource is "Other," which includes all other online resources that were given as selection options. For data sets that have a large number of values for a categorical variable, we often create a category such as this that includes categories that have relatively small counts or percents. *Careful judgment is needed when doing this.* You don't want to cover up some important piece of information contained in the data by combining data in this way.

EXAMPLE 1.7 Favorites as Percents

When we look at the online resources data set, we see that Google is the clear winner. We see that 406 reported Google or Google Scholar as their favorite. To interpret this number, we need to know that the total number of students polled was 552. When we say that Google is the winner, we can describe this win by saying that 73.6% (406 divided by 552, expressed as a percent) of the students reported Google as their favorite. Here is a table of the preference percents:

Resource	Percent (%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0

The use of graphical methods will allow us to see this information and other characteristics of the data easily. We now examine two types of graphs.

EXAMPLE 1.8 Bar Graph for the Online Resource Preference Data

bar graph

Figure 1.2 displays the online resource preference data using a **bar graph.** The heights of the four bars show the percents of the students who reported each of the resources as their favorite.

The categories in a bar graph can be put in any order. In Figure 1.2, we ordered the resources based on their preference percents. For other data sets, an

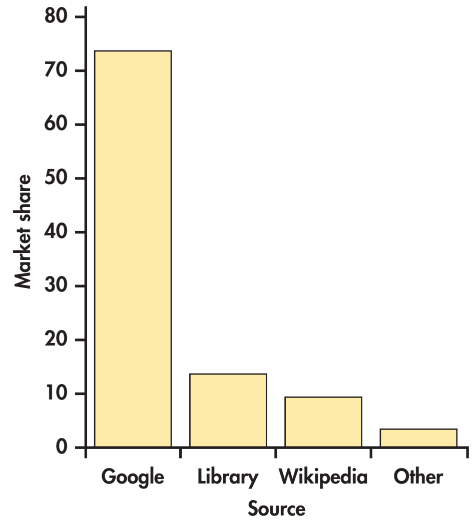


FIGURE 1.2 Bar graph for the online resource preference data, Example 1.8.

alphabetical ordering or some other arrangement might produce a more useful graphical display.

You should always consider the best way to order the values of the categorical variable in a bar graph. Choose an ordering that will be useful to you. If you have difficulty, ask a friend if your choice communicates what you expect.

DATA ONLINE

EXAMPLE 1.9 Pie Chart for the Online Resource Preference Data

pie chart

The **pie chart** in Figure 1.3 helps us see what part of the whole each group forms. Here it is very easy to see that Google is the favorite for about three-quarters of the students.

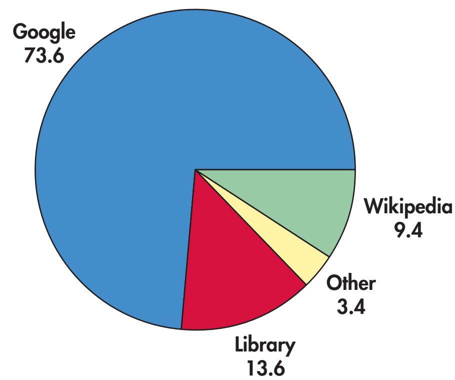


FIGURE 1.3 Pie chart for the online resource preference data in Example 1.9.

APPLY YOUR KNOWLEDGE

1.15 Compare the bar graph with the pie chart. Refer to the bar graph in Figure 1.2 and the pie chart in Figure 1.3 for the online resource preference data. Which graphical display does a better job of describing the data? Give reasons for your answer.

We use graphical displays to help us learn things from data. Here is another example.