



Statistics for Business and Economics

Eighth Edition

Paul Newbold
William L. Carlson
Betty M. Thorne

ALWAYS LEARNING

PEARSON

EIGHTH EDITION

Statistics for Business and Economics

Global Edition

Paul Newbold

University of Nottingham

William L. Carlson

St. Olaf College

Betty M. Thorne

Stetson University

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam CapeTown Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director: Sally Yagan
Editor in Chief: Donna Battista
Senior Acquisitions Editor: Chuck Synovec
Senior Acquisitions Editor, Global Edition: Steven Jackson
Editor, Global Edition: Leandra Paoli
Senior Editorial Project Manager: Mary Kate Murray
Editorial Assistant: Ashlee Bradbury
Director of Marketing: Maggie Moylan
Executive Marketing Manager: Anne Fahlgren
Marketing Manager, International: Dean Erasmus

Senior Managing Editor: Judy Leale
Production Project Manager: Jacqueline A. Martin
Senior Operations Supervisor: Arnold Vila
Operations Specialist: Cathleen Petersen
Art Director: Steve Frim
Cover Designer: Jodi Notowitz
Cover Art: © Zoe - Fotolia.com
Media Project Manager: John Cassar
Associate Media Project Manager: Sarah Peterson
Full-Service Project Management: PreMediaGlobal, Inc.

Pearson Education Limited

Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearson.com/uk

© Pearson Education Limited 2013

The rights of Paul Newbold, William L. Carlson and Betty Thorne to be identified as authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Authorised adaptation from the United States edition, entitled Statistics for Business and Economics, 8th Edition, ISBN: 978-0-13-274565-9 by Paul Newbold, William L. Carlson and Betty Thorne, published by Pearson Education, Inc., © 2013.

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 licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

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

Microsoft® and Windows® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided “as is” without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on the appropriate page within the text.

ISBN 13: 978-0-273-76706-0
ISBN 10: 0-273-76706-2

British Library Cataloguing-in-Publication Data

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

10 9 8 7 6 5 4 3 2 1
16 15 14 13 12

Typeset in Palatino LT Std by PreMediaGlobal, Inc.
Printed and bound by Courier Kendallville in The United States of America

The publisher's policy is to use paper manufactured from sustainable forests.

I dedicate this book to Sgt. Lawrence Martin Carlson, who gave his life in service to his country on November 19, 2006, and to his mother, Charlotte Carlson, to his sister and brother, Andrea and Douglas, to his children, Savannah, and Ezra, and to his nieces, Helana, Anna, Eva Rose, and Emily.

William L. Carlson

I dedicate this book to my husband, Jim, and to our family, Jennie, Ann, Renee, Jon, Chris, Jon, Hannah, Leah, Christina, Jim, Wendy, Marius, Mihaela, Cezara, Anda, and Mara Iulia.

Betty M. Thorne



Dr. Bill Carlson is professor emeritus of economics at St. Olaf College, where he taught for 31 years, serving several times as department chair and in various administrative functions, including director of academic computing. He has also held leave assignments with the U.S. government and the University of Minnesota in addition to lecturing at many different universities. He was elected an honorary member of Phi Beta Kappa. In addition, he spent 10 years in private industry and contract research prior to beginning his career at St. Olaf. His education includes engineering degrees from Michigan Technological University (BS) and from the Illinois Institute of Technology (MS) and a PhD in quantitative management from the Rackham Graduate School at the University of Michigan. Numerous research projects related to management, highway safety, and statistical education have produced more than 50 publications. He received the Metropolitan Insurance Award of Merit for Safety Research. He has previously published two statistics textbooks. An important goal of this book is to help students understand the forest and not be lost in the trees. Hiking the Lake Superior trail in Northern Minnesota helps in developing this goal. Professor Carlson led a number of study-abroad programs, ranging from 1 to 5 months, for study in various countries around the world. He was the executive director of the Cannon Valley Elder Collegium and a regular volunteer for a number of community activities. He is a member of both the Methodist and Lutheran disaster-relief teams and a regular participant in the local Habitat for Humanity building team. He enjoys his grandchildren, wood-working, travel, reading, and being on assignment on the North Shore of Lake Superior.



Dr. Betty M. Thorne, author, researcher, and award-winning teacher, is professor of statistics and director of undergraduate studies in the School of Business Administration at Stetson University in DeLand, Florida. Winner of Stetson University's McEniry Award for Excellence in Teaching, the highest honor given to a Stetson University faculty member, Dr. Thorne is also the recipient of the Outstanding Teacher of the Year Award and Professor of the Year Award in the School of Business Administration at Stetson. Dr. Thorne teaches in Stetson University's undergraduate business program in DeLand, Florida, and also in Stetson's summer program in Innsbruck, Austria; Stetson University's College of Law; Stetson University's Executive MBA program; and Stetson University's Executive Passport program. Dr. Thorne has received various teaching awards in the JD/MBA program at Stetson's College of Law in Gulfport, Florida. She received her BS degree from Geneva College and MA and PhD degrees from Indiana University. She has co-authored statistics textbooks which have been translated into several languages and adopted by universities, nationally and internationally. She serves on key school and university committees. Dr. Thorne, whose research has been published in various refereed journals, is a member of the American Statistical Association, the Decision Science Institute, Beta Alpha Psi, Beta Gamma Sigma, and the Academy of International Business. She and her husband, Jim, have four children. They travel extensively, attend theological conferences and seminars, participate in international organizations dedicated to helping disadvantaged children, and do missionary work in Romania.

| | | |
|-------------------|---|-----|
| | Preface | 13 |
| | Data File Index | 19 |
| CHAPTER 1 | Using Graphs to Describe Data | 21 |
| CHAPTER 2 | Using Numerical Measures to Describe Data | 59 |
| CHAPTER 3 | Elements of Chance: Probability Methods | 93 |
| CHAPTER 4 | Discrete Probability Distributions | 146 |
| CHAPTER 5 | Continuous Probability Distributions | 197 |
| CHAPTER 6 | Distributions of Sample Statistics | 244 |
| CHAPTER 7 | Confidence Interval Estimation: One Population | 284 |
| CHAPTER 8 | Confidence Interval Estimation: Further Topics | 328 |
| CHAPTER 9 | Hypothesis Tests of a Single Population | 346 |
| CHAPTER 10 | Two Population Hypothesis Tests | 385 |
| CHAPTER 11 | Two Variable Regression Analysis | 417 |
| CHAPTER 12 | Multiple Variable Regression Analysis | 473 |
| CHAPTER 13 | Additional Topics in Regression Analysis | 551 |
| CHAPTER 14 | Introduction to Nonparametric Statistics | 602 |
| CHAPTER 15 | Analysis of Variance | 645 |
| CHAPTER 16 | Forecasting with Time-Series Models | 684 |
| CHAPTER 17 | Sampling: Stratified, Cluster, and Other Sampling Methods | 716 |
| | Appendix Tables | 738 |
| | Index | 783 |

This page intentionally left blank

| | |
|-----------------|----|
| Preface | 13 |
| Data File Index | 19 |

CHAPTER 1 Using Graphs to Describe Data 21

| | | |
|------------|---|----|
| 1.1 | Decision Making in an Uncertain Environment | 22 |
| | <i>Random and Systematic Sampling</i> | 22 |
| | <i>Sampling and Nonsampling Errors</i> | 24 |
| 1.2 | Classification of Variables | 25 |
| | <i>Categorical and Numerical Variables</i> | 25 |
| | <i>Measurement Levels</i> | 26 |
| 1.3 | Graphs to Describe Categorical Variables | 28 |
| | <i>Tables and Charts</i> | 28 |
| | <i>Cross Tables</i> | 29 |
| | <i>Pie Charts</i> | 31 |
| | <i>Pareto Diagrams</i> | 32 |
| 1.4 | Graphs to Describe Time-Series Data | 35 |
| 1.5 | Graphs to Describe Numerical Variables | 40 |
| | <i>Frequency Distributions</i> | 40 |
| | <i>Histograms and Ogives</i> | 44 |
| | <i>Shape of a Distribution</i> | 44 |
| | <i>Stem-and-Leaf Displays</i> | 46 |
| | <i>Scatter Plots</i> | 47 |
| 1.6 | Data Presentation Errors | 51 |
| | <i>Misleading Histograms</i> | 51 |
| | <i>Misleading Time-Series Plots</i> | 53 |

CHAPTER 2 Using Numerical Measures to Describe Data 59

| | | |
|------------|---|----|
| 2.1 | Measures of Central Tendency and Location | 59 |
| | <i>Mean, Median, and Mode</i> | 60 |
| | <i>Shape of a Distribution</i> | 62 |
| | <i>Geometric Mean</i> | 63 |
| | <i>Percentiles and Quartiles</i> | 64 |
| 2.2 | Measures of Variability | 68 |
| | <i>Range and Interquartile Range</i> | 69 |
| | <i>Box-and-Whisker Plots</i> | 69 |
| | <i>Variance and Standard Deviation</i> | 71 |
| | <i>Coefficient of Variation</i> | 75 |
| | <i>Chebyshev's Theorem and the Empirical Rule</i> | 75 |
| | <i>z-Score</i> | 77 |
| 2.3 | Weighted Mean and Measures of Grouped Data | 80 |
| 2.4 | Measures of Relationships Between Variables | 84 |
| | <i>Case Study: Mortgage Portfolio</i> | 91 |

CHAPTER 3 Elements of Chance: Probability Methods 93

- 3.1** Random Experiment, Outcomes, and Events 94
- 3.2** Probability and Its Postulates 101
 - Classical Probability* 101
 - Permutations and Combinations* 102
 - Relative Frequency* 106
 - Subjective Probability* 107
- 3.3** Probability Rules 111
 - Conditional Probability* 113
 - Statistical Independence* 116
- 3.4** Bivariate Probabilities 122
 - Odds* 126
 - Overinvolvement Ratios* 126
- 3.5** Bayes' Theorem 132
 - Subjective Probabilities in Management Decision Making* 138

CHAPTER 4 Discrete Probability Distributions 146

- 4.1** Random Variables 147
- 4.2** Probability Distributions for Discrete Random Variables 148
- 4.3** Properties of Discrete Random Variables 152
 - Expected Value of a Discrete Random Variable* 152
 - Variance of a Discrete Random Variable* 153
 - Mean and Variance of Linear Functions of a Random Variable* 155
- 4.4** Binomial Distribution 159
 - Developing the Binomial Distribution* 160
- 4.5** Poisson Distribution 167
 - Poisson Approximation to the Binomial Distribution* 171
 - Comparison of the Poisson and Binomial Distributions* 172
- 4.6** Hypergeometric Distribution 173
- 4.7** Jointly Distributed Discrete Random Variables 176
 - Conditional Mean and Variance* 180
 - Computer Applications* 180
 - Linear Functions of Random Variables* 180
 - Covariance* 181
 - Correlation* 182
 - Portfolio Analysis* 186

CHAPTER 5 Continuous Probability Distributions 197

- 5.1** Continuous Random Variables 198
 - The Uniform Distribution* 201
- 5.2** Expectations for Continuous Random Variables 203
- 5.3** The Normal Distribution 206
 - Normal Probability Plots* 215
- 5.4** Normal Distribution Approximation for Binomial Distribution 219
 - Proportion Random Variable* 223
- 5.5** The Exponential Distribution 225
- 5.6** Jointly Distributed Continuous Random Variables 228
 - Linear Combinations of Random Variables* 232
 - Financial Investment Portfolios* 232
 - Cautions Concerning Finance Models* 236

CHAPTER 6 Distributions of Sample Statistics 244

- 6.1** Sampling from a Population 245
Development of a Sampling Distribution 246
- 6.2** Sampling Distributions of Sample Means 249
Central Limit Theorem 254
Monte Carlo Simulations: Central Limit Theorem 254
Acceptance Intervals 260
- 6.3** Sampling Distributions of Sample Proportions 265
- 6.4** Sampling Distributions of Sample Variances 270

CHAPTER 7 Confidence Interval Estimation: One Population 284

- 7.1** Properties of Point Estimators 285
Unbiased 286
Most Efficient 287
- 7.2** Confidence Interval Estimation for the Mean of a Normal Distribution: Population Variance Known 291
Intervals Based on the Normal Distribution 292
Reducing Margin of Error 295
- 7.3** Confidence Interval Estimation for the Mean of a Normal Distribution: Population Variance Unknown 297
Student's t Distribution 297
Intervals Based on the Student's t Distribution 299
- 7.4** Confidence Interval Estimation for Population Proportion (Large Samples) 303
- 7.5** Confidence Interval Estimation for the Variance of a Normal Distribution 306
- 7.6** Confidence Interval Estimation: Finite Populations 309
Population Mean and Population Total 309
Population Proportion 312
- 7.7** Sample-Size Determination: Large Populations 315
Mean of a Normally Distributed Population, Known Population Variance 315
Population Proportion 317
- 7.8** Sample-Size Determination: Finite Populations 319
Sample Sizes for Simple Random Sampling: Estimation of the Population Mean or Total 320
Sample Sizes for Simple Random Sampling: Estimation of Population Proportion 321

CHAPTER 8 Confidence Interval Estimation: Further Topics 328

- 8.1** Confidence Interval Estimation of the Difference Between Two Normal Population Means: Dependent Samples 329
- 8.2** Confidence Interval Estimation of the Difference Between Two Normal Population Means: Independent Samples 333
Two Means, Independent Samples, and Known Population Variances 333
Two Means, Independent Samples, and Unknown Population Variances Assumed to Be Equal 335
Two Means, Independent Samples, and Unknown Population Variances Not Assumed to Be Equal 337
- 8.3** Confidence Interval Estimation of the Difference Between Two Population Proportions (Large Samples) 340

CHAPTER 9 Hypothesis Tests of a Single Population 346

- 9.1 Concepts of Hypothesis Testing 347
- 9.2 Tests of the Mean of a Normal Distribution: Population Variance Known 352
 - p-Value* 354
 - Two-Sided Alternative Hypothesis* 360
- 9.3 Tests of the Mean of a Normal Distribution: Population Variance Unknown 362
- 9.4 Tests of the Population Proportion (Large Samples) 366
- 9.5 Assessing the Power of a Test 368
 - Tests of the Mean of a Normal Distribution: Population Variance Known* 369
 - Power of Population Proportion Tests (Large Samples)* 371
- 9.6 Tests of the Variance of a Normal Distribution 375

CHAPTER 10 Two Population Hypothesis Tests 385

- 10.1 Tests of the Difference Between Two Normal Population Means: Dependent Samples 387
 - Two Means, Matched Pairs* 387
- 10.2 Tests of the Difference Between Two Normal Population Means: Independent Samples 391
 - Two Means, Independent Samples, Known Population Variances* 391
 - Two Means, Independent Samples, Unknown Population Variances Assumed to Be Equal* 393
 - Two Means, Independent Samples, Unknown Population Variances Not Assumed to Be Equal* 396
- 10.3 Tests of the Difference Between Two Population Proportions (Large Samples) 399
- 10.4 Tests of the Equality of the Variances Between Two Normally Distributed Populations 403
- 10.5 Some Comments on Hypothesis Testing 406

CHAPTER 11 Two Variable Regression Analysis 417

- 11.1 Overview of Linear Models 418
- 11.2 Linear Regression Model 421
- 11.3 Least Squares Coefficient Estimators 427
 - Computer Computation of Regression Coefficients* 429
- 11.4 The Explanatory Power of a Linear Regression Equation 431
 - Coefficient of Determination, R^2* 433
- 11.5 Statistical Inference: Hypothesis Tests and Confidence Intervals 438
 - Hypothesis Test for Population Slope Coefficient Using the F Distribution* 443
- 11.6 Prediction 446
- 11.7 Correlation Analysis 452
 - Hypothesis Test for Correlation* 452
- 11.8 Beta Measure of Financial Risk 456
- 11.9 Graphical Analysis 458

CHAPTER 12 Multiple Variable Regression Analysis 473

- 12.1 The Multiple Regression Model 474
 - Model Specification* 474
 - Model Objectives* 476
 - Model Development* 477
 - Three-Dimensional Graphing* 480

| | | |
|-------------------|--|------------|
| 12.2 | Estimation of Coefficients | 481 |
| | <i>Least Squares Procedure</i> | 482 |
| 12.3 | Explanatory Power of a Multiple Regression Equation | 488 |
| 12.4 | Confidence Intervals and Hypothesis Tests for Individual Regression Coefficients | 493 |
| | <i>Confidence Intervals</i> | 495 |
| | <i>Tests of Hypotheses</i> | 497 |
| 12.5 | Tests on Regression Coefficients | 505 |
| | <i>Tests on All Coefficients</i> | 505 |
| | <i>Test on a Subset of Regression Coefficients</i> | 506 |
| | <i>Comparison of F and t Tests</i> | 508 |
| 12.6 | Prediction | 511 |
| 12.7 | Transformations for Nonlinear Regression Models | 514 |
| | <i>Quadratic Transformations</i> | 515 |
| | <i>Logarithmic Transformations</i> | 517 |
| 12.8 | Dummy Variables for Regression Models | 522 |
| | <i>Differences in Slope</i> | 525 |
| 12.9 | Multiple Regression Analysis Application Procedure | 529 |
| | <i>Model Specification</i> | 529 |
| | <i>Multiple Regression</i> | 531 |
| | <i>Effect of Dropping a Statistically Significant Variable</i> | 532 |
| | <i>Analysis of Residuals</i> | 534 |
| CHAPTER 13 | Additional Topics in Regression Analysis | 551 |
| 13.1 | Model-Building Methodology | 552 |
| | <i>Model Specification</i> | 552 |
| | <i>Coefficient Estimation</i> | 553 |
| | <i>Model Verification</i> | 554 |
| | <i>Model Interpretation and Inference</i> | 554 |
| 13.2 | Dummy Variables and Experimental Design | 554 |
| | <i>Experimental Design Models</i> | 558 |
| | <i>Public Sector Applications</i> | 563 |
| 13.3 | Lagged Values of the Dependent Variable as Regressors | 567 |
| 13.4 | Specification Bias | 571 |
| 13.5 | Multicollinearity | 574 |
| 13.6 | Heteroscedasticity | 577 |
| 13.7 | Autocorrelated Errors | 582 |
| | <i>Estimation of Regressions with Autocorrelated Errors</i> | 586 |
| | <i>Autocorrelated Errors in Models with Lagged Dependent Variables</i> | 590 |
| CHAPTER 14 | Introduction to Nonparametric Statistics | 602 |
| 14.1 | Goodness-of-Fit Tests: Specified Probabilities | 603 |
| 14.2 | Goodness-of-Fit Tests: Population Parameters Unknown | 609 |
| | <i>A Test for the Poisson Distribution</i> | 609 |
| | <i>A Test for the Normal Distribution</i> | 611 |
| 14.3 | Contingency Tables | 614 |
| 14.4 | Nonparametric Tests for Paired or Matched Samples | 619 |
| | <i>Sign Test for Paired or Matched Samples</i> | 619 |
| | <i>Wilcoxon Signed Rank Test for Paired or Matched Samples</i> | 622 |
| | <i>Normal Approximation to the Sign Test</i> | 623 |

| | | |
|-------------------|---|------------|
| | <i>Normal Approximation to the Wilcoxon Signed Rank Test</i> | 624 |
| | <i>Sign Test for a Single Population Median</i> | 626 |
| 14.5 | Nonparametric Tests for Independent Random Samples | 628 |
| | <i>Mann-Whitney U Test</i> | 628 |
| | <i>Wilcoxon Rank Sum Test</i> | 631 |
| 14.6 | Spearman Rank Correlation | 634 |
| 14.7 | A Nonparametric Test for Randomness | 636 |
| | <i>Runs Test: Small Sample Size</i> | 636 |
| | <i>Runs Test: Large Sample Size</i> | 638 |
| CHAPTER 15 | Analysis of Variance | 645 |
| 15.1 | Comparison of Several Population Means | 645 |
| 15.2 | One-Way Analysis of Variance | 647 |
| | <i>Multiple Comparisons Between Subgroup Means</i> | 654 |
| | <i>Population Model for One-Way Analysis of Variance</i> | 655 |
| 15.3 | The Kruskal-Wallis Test | 658 |
| 15.4 | Two-Way Analysis of Variance: One Observation per Cell, Randomized Blocks | 661 |
| 15.5 | Two-Way Analysis of Variance: More Than One Observation per Cell | 670 |
| CHAPTER 16 | Forecasting with Time-Series Models | 684 |
| 16.1 | Components of a Time Series | 685 |
| 16.2 | Moving Averages | 689 |
| | <i>Extraction of the Seasonal Component Through Moving Averages</i> | 692 |
| 16.3 | Exponential Smoothing | 697 |
| | <i>The Holt-Winters Exponential Smoothing Forecasting Model</i> | 700 |
| | <i>Forecasting Seasonal Time Series</i> | 704 |
| 16.4 | Autoregressive Models | 708 |
| 16.5 | Autoregressive Integrated Moving Average Models | 713 |
| CHAPTER 17 | Sampling: Stratified, Cluster, and Other Sampling Methods | 716 |
| 17.1 | Stratified Sampling | 716 |
| | <i>Analysis of Results from Stratified Random Sampling</i> | 718 |
| | <i>Allocation of Sample Effort Among Strata</i> | 723 |
| | <i>Determining Sample Sizes for Stratified Random Sampling with Specified Degree of Precision</i> | 725 |
| 17.2 | Other Sampling Methods | 729 |
| | <i>Cluster Sampling</i> | 729 |
| | <i>Two-Phase Sampling</i> | 732 |
| | <i>Nonprobabilistic Sampling Methods</i> | 734 |
| | APPENDIX TABLES | 738 |
| | INDEX | 783 |

INTENDED AUDIENCE

Statistics for Business and Economics, 8th edition, was written to meet the need for an introductory text that provides a strong introduction to business statistics, develops understanding of concepts, and emphasizes problem solving using realistic examples that emphasize real data sets and computer based analysis. These examples emphasize business and economics examples for the following:

- MBA or undergraduate business programs that teach business statistics
- Graduate and undergraduate economics programs
- Executive MBA programs
- Graduate courses for business statistics

SUBSTANCE

This book was written to provide a strong introductory understanding of applied statistical procedures so that individuals can do solid statistical analysis in many business and economic situations. We have emphasized an understanding of the assumptions that are necessary for professional analysis. In particular we have greatly expanded the number of applications that utilize data from applied policy and research settings. Data and problem scenarios have been obtained from business analysts, major research organizations, and selected extractions from publicly available data sources. With modern computers it is easy to compute, from data, the output needed for many statistical procedures. Thus, it is tempting to merely apply simple “rules” using these outputs—an approach used in many textbooks. Our approach is to combine understanding with many examples and student exercises that show how understanding of methods and their assumptions lead to useful understanding of business and economic problems.

NEW TO THIS EDITION

The eighth edition of this book has been revised and updated to provide students with improved problem contexts for learning how statistical methods can improve their analysis and understanding of business and economics.

The objective of this revision is to provide a strong core textbook with new features and modifications that will provide an improved learning environment for students entering a rapidly changing technical work environment. This edition has been carefully revised to improve the clarity and completeness of explanations. This revision recognizes the globalization of statistical study and in particular the global market for this book.

1. Improvement in clarity and relevance of discussions of the core topics included in the book.
2. Addition of a number of large databases developed by public research agencies, businesses, and databases from the authors’ own works.

3. Inclusion of a number of new exercises that introduce students to specific statistical questions that are part of research projects.
4. Addition of a number of case studies, with both large and small sample sizes. Students are provided the opportunity to extend their statistical understanding to the context of research and analysis conducted by professionals. These studies include data files obtained from on-going research studies, which reduce for the student, the extensive work load of data collection and refinement, thus providing an emphasis on question formulation, analysis, and reporting of results.
5. Careful revision of text and symbolic language to ensure consistent terms and definitions and to remove errors that accumulated from previous revisions and production problems.
6. Major revision of the discussion of Time Series both in terms of describing historical patterns and in the focus on identifying the underlying structure and introductory forecasting methods.
7. Integration of the text material, data sets, and exercises into new on-line applications including MyMathLab Global.
8. Expansion of descriptive statistics to include percentiles, z-scores, and alternative formulae to compute the sample variance and sample standard deviation.
9. Addition of a significant number of new examples based on real world data.
10. Greater emphasis on the assumptions being made when conducting various statistical procedures.
11. Reorganization of sampling concepts.
12. More detailed business-oriented examples and exercises incorporated in the analysis of statistics.
13. Improved chapter introductions that include business examples discussed in the chapter.
14. Good range of difficulty in the section ending exercises that permit the professor to tailor the difficulty level to his or her course.
15. Improved suitability for both introductory and advanced statistics courses and by both undergraduate and graduate students.
16. Decision Theory, which is covered in other business classes such as operations management or strategic management, has been moved to an online location for access by those who are interested (www.pearsonglobaleditions.com/newbold).

This edition devotes considerable effort to providing an understanding of statistical methods and their applications. We have avoided merely providing rules and canned computer routines for analyzing and solving statistical problems. This edition contains a complete discussion of methods and assumptions, including computational details expressed in clear and complete formulas. Through examples and extended chapter applications, we provide guidelines for interpreting results and explain how to determine if additional analysis is required. The development of the many procedures included under statistical inference and regression analysis are built on a strong development of probability and random variables, which are a foundation for the applications presented in this book. The foundation also includes a clear and complete discussion of descriptive statistics and graphical approaches. These provide important tools for exploring and describing data that represent a process being studied.

Probability and random variables are presented with a number of important applications, which are invaluable in management decision making. These include conditional probability and Bayesian applications that clarify decisions and show counterintuitive results in a number of decision situations. Linear combinations of random variables are developed in detail, with a number of applications of importance, including portfolio applications in finance.

The authors strongly believe that students learn best when they work with challenging and relevant applications that apply the concepts presented by dedicated teachers and the textbook. Thus the textbook has always included a number of data

sets obtained from various applications in the public and private sectors. In the eighth edition we have added a number of large data sets obtained from major research projects and other sources. These data sets are used in chapter examples, exercises, and case studies located at the end of analysis chapters. A number of exercises consider individual analyses that are typically part of larger research projects. With this structure, students can deal with important detailed questions and can also work with case studies that require them to identify the detailed questions that are logically part of a larger research project. These large data sets can also be used by the teacher to develop additional research and case study projects that are custom designed for local course environments. The opportunity to custom design new research questions for students is a unique part of this textbook.

One of the large data sets is the HEI Cost Data Variable Subset. This data file was obtained from a major nutrition-research project conducted at the Economic Research Service (ERS) of the U.S. Department of Agriculture. These research projects provide the basis for developing government policy and informing citizens and food producers about ways to improve national nutrition and health. The original data were gathered in the National Health and Nutrition Examination Survey, which included in-depth interview measurements of diet, health, behavior, and economic status for a large probability sample of the U.S. population. Included in the data is the Healthy Eating Index (HEI), a measure of diet quality developed by ERS and computed for each individual in the survey. A number of other major data sets containing nutrition measures by country, automobile fuel consumption, health data, and more are described in detail at the end of the chapters where they are used in exercises and case studies. A complete list of the data files and where they are used is located at the end of this preface. Data files are also shown by chapter at the end of each chapter.

The book provides a complete and in-depth presentation of major applied topics. An initial read of the discussion and application examples enables a student to begin working on simple exercises, followed by challenging exercises that provide the opportunity to learn by doing relevant analysis applications. Chapters also include summary sections, which clearly present the key components of application tools. Many analysts and teachers have used this book as a reference for reviewing specific applications. Once you have used this book to help learn statistical applications, you will also find it to be a useful resource as you use statistical analysis procedures in your future career.

A number of special applications of major procedures are included in various sections. Clearly there are more than can be used in a single course. But careful selection of topics from the various chapters enables the teacher to design a course that provides for the specific needs of students in the local academic program. Special examples that can be left out or included provide a breadth of opportunities. The initial probability chapter, Chapter 3, provides topics such as decision trees, overinvolvement ratios, and expanded coverage of Bayesian applications, any of which might provide important material for local courses. Confidence interval and hypothesis tests include procedures for variances and for categorical and ordinal data. Random-variable chapters include linear combination of correlated random variables with applications to financial portfolios. Regression applications include estimation of beta ratios in finance, dummy variables in experimental design, nonlinear regression, and many more.

As indicated here, the book has the capability of being used in a variety of courses that provide applications for a variety of academic programs. The other benefit to the student is that this textbook can be an ideal resource for the student's future professional career. The design of the book makes it possible for a student to come back to topics after several years and quickly renew his or her understanding. With all the additional special topics, that may not have been included in a first course, the book is a reference for learning important new applications. And the presentation of those new applications follows a presentation style and uses understandings that are familiar. This reduces the time required to master new application topics.

SUPPLEMENT PACKAGE

Student Resources

Online Resources—These resources, which can be downloaded at no cost from www.pearsonglobaleditions.com/newbold, include the following:

- **Data files**—Excel data files that are used throughout the chapters.
- **PHStat2**—The latest version of PHStat2, the Pearson statistical add-in for Windows-based Excel 2003, 2007, and 2010. This version eliminates the use of the Excel Analysis ToolPak add-ins, thereby simplifying installation and setup.
- **Answers to Selected Even-Numbered Exercises**

MyMathLab Global

MyMathLab Global provides students with direct access to the online resources as well as the following exclusive online features and tools:

- **Interactive tutorial exercises**—These are a comprehensive set of exercises written especially for use with this book that are algorithmically generated for unlimited practice and mastery. Most exercises are free-response exercises and provide guided solutions, sample problems, and learning aids for extra help at point of use.
- **Personalized study plan**—This plan indicates which topics have been mastered and creates direct links to tutorial exercises for topics that have not been mastered. MyMathLab Global manages the study plan, updating its content based on the results of future online assessments.
- **Integration with Pearson eTexts**—A resource for iPad users, who can download a free app at www.apple.com/ipad/apps-for-ipad/ and then sign in using their MyMathLab Global account to access a bookshelf of all their Pearson eTexts. The iPad app also allows access to the Do Homework, Take a Test, and Study Plan pages of their MyMathLab Global course.

Instructor Resources

Instructor's Resource Center—Reached through a link at www.pearsonglobaleditions.com/newbold, the Instructor's Resource Center contains the electronic files for the complete Instructor's Solutions Manual, the Test Item File, and PowerPoint lecture presentations:

- **Register, Redeem, Log In**—At www.pearsonglobaleditions.com/newbold, instructors can access a variety of print, media, and presentation resources that are available with this book in downloadable digital format.
- **Need Help?**—Pearson Education's dedicated technical support team is ready to assist instructors with questions about the media supplements that accompany this text. Visit <http://247pearsoned.com> for answers to frequently asked questions and toll-free user-support phone numbers. The supplements are available to adopting instructors. Detailed descriptions are provided at the Instructor's Resource Center.

Instructor Solutions Manual—This manual includes worked-out solutions for end-of-section and end-of-chapter exercises and applications. Electronic solutions are provided at the Instructor's Resource Center in Word format.

PowerPoint Lecture Slides—A set of chapter-by-chapter PowerPoint slides provides an instructor with individual lecture outlines to accompany the text. The slides include many of the figures and tables from the text. Instructors can use these lecture notes as is or can easily modify the notes to reflect specific presentation needs.

Test-Item File—The test-item file contains true/false, multiple-choice, and short-answer questions based on concepts and ideas developed in each chapter of the text.

TestGen Software—Pearson Education's test-generating software is PC compatible and preloaded with all the Test-Item File questions. You can manually or randomly view test

questions and drag and drop them to create a test. You can add or modify test-bank questions as needed.

MyMathLab Global

MyMathLab Global is a powerful online homework, tutorial, and assessment system that accompanies Pearson Education statistics textbooks. With MyMathLab Global, instructors can create, edit, and assign online homework and tests using algorithmically generated exercises correlated at the objective level to the textbook. They can also create and assign their own online exercises and import TestGen tests for added flexibility. All student work is tracked in the online grade book. Students can take chapter tests and receive personalized study plans based on their test results. Each study plan diagnoses weaknesses and links the student directly to tutorial exercises for the objectives he or she needs to study and retest. Students can also access supplemental animations and video clips directly from selected exercises. MyMathLab Global is available to qualified adopters. For more information, visit www.mymathlab.com/global or contact your sales representative.

MyMathLab Global is a text-specific, easily customizable online course that integrates interactive multimedia instruction with textbook content. MyMathLab Global gives you the tools you need to deliver all or a portion of your course online, whether your students are in a lab setting or working from home. The latest version of MyMathLab Global offers a new, intuitive design that features more direct access to MyMathLab Global pages (Gradebook, Homework & Test Manager, Home Page Manager, etc.) and provides enhanced functionality for communicating with students and customizing courses. Other key features include the following:

- **Assessment Manager** An easy-to-use assessment manager lets instructors create online homework, quizzes, and tests that are automatically graded and correlated directly to your textbook. Assignments can be created using a mix of questions from the exercise bank, instructor-created custom exercises, and/or TestGen test items.
- **Grade Book** Designed specifically for mathematics and statistics, the grade book automatically tracks students' results and gives you control over how to calculate final grades. You can also add offline (paper-and-pencil) grades to the grade book.
- **Exercise Builder** You can use the Exercise Builder to create static and algorithmic exercises for your online assignments. A library of sample exercises provides an easy starting point for creating questions, and you can also create questions from scratch.
- **eText Full Integration** Students who have the appropriate mobile devices can use your eText annotations and highlights for each course, and iPad users can download a free app that allows them access to the Do Homework, Take a Test, and Study Plan pages of their course.
- **"Ask the Publisher" Link in "Ask My Instructor" E-mail** You can easily notify the content team of any irregularities with specific questions by using the "Ask the Publisher" functionality in the "Ask My Instructor" e-mails you receive from students.
- **Tracking Time Spent on Media** Because the latest version of MyMathLab Global requires students to explicitly click a "Submit" button after viewing the media for their assignments, you will be able to track how long students are spending on each media file.

ACKNOWLEDGMENTS

We appreciate the following colleagues who provided feedback about the book to guide our thoughts on this revision: Valerie R. Bencivenga, University of Texas at Austin; Burak Dolar, Augustana College; Zhimin Huang, Adelphi University; Stephen Lich-Tyler, University of North Carolina; Tung Liu, Ball State University; Leonard Presby, William Paterson University; Subarna K. Samanta, The College of New Jersey; Shane Sanders, Nicholls State University; Harold Schneider, Rider University; Sean Simpson, Westchester Community College.

The authors thank Dr. Andrea Carlson, Economic Research Service (ERS), U.S. Department of Agriculture, for her assistance in providing several major data files and for guidance in developing appropriate research questions for exercises and case studies. We also thank Paula Dutko and Empharim Leibtag for providing an example of complex statistical analysis in the public sector. We also recognize the excellent work by Annie Puciloski in finding our errors and improving the professional quality of this book.

We extend appreciation to two Stetson alumni, Richard Butcher (RELEVANT Magazine) and Lisbeth Mendez (mortgage company), for providing real data from their companies that we used for new examples, exercises, and case studies.

In addition, we express special thanks for continuing support from our families. Bill Carlson especially acknowledges his best friend and wife, Charlotte, their adult children, Andrea and Doug, and grandchildren, Ezra, Savannah, Helena, Anna, Eva Rose, and Emily. Betty Thorne extends special thanks to her best friend and husband, Jim, and to their family Jennie, Ann, Renee, Jon, Chris, Jon, Hannah, Leah, Christina, Jim, Wendy, Marius, Mihaela, Cezara, Anda, and Mara Iulia. In addition, Betty acknowledges (in memory) the support of her parents, Westley and Jennie Moore.

The authors acknowledge the strong foundation and tradition created by the original author, Paul Newbold. Paul understood the importance of rigorous statistical analysis and its foundations. He realized that there are some complex ideas that need to be developed, and he worked to provide clear explanations of difficult ideas. In addition, he realized that these ideas become useful only when used in realistic problem-solving situations. Thus, many examples and many applied student exercises were included in the early editions. We have worked to continue and expand this tradition in preparing a book that meets the needs of future business leaders in the information age.

Pearson wish to thank and acknowledge the following people for their work on the Global Edition:

Contributors

Tracey Holker, Department of Strategy and Applied Management, Coventry Business School, United Kingdom

Stefania Paladini, Department of Strategy and Applied Management, Coventry Business School, United Kingdom

Xavier Pierron, Department of Strategy and Applied Management, Coventry Business School, United Kingdom

Reviewers

Rosie Ching Ju Mae, School of Economics, Singapore Management University, Singapore

Patrick Kuok-Kun Chu, Department of Accounting and Information Management, FBA, University of Macau, China

Mohamed Madi, Faculty of Business and Economics, United Arab Emirates University, United Arab Emirates

- Acme LLC Earnings per Share**—Exercise 16.9
- Advertising Retail**—Example 13.6, Exercise 13.38
- Advertising Revenue**—Exercise 11.62
- Anscombe**—Exercise 11.68
- Apple Stock Prices**—Exercise 1.70
- Automobile Fuel Consumption**—Chapter 12
Case Study
- Beef Veal Consumption**—Exercises 13.63–13.65
- Benefits Research**—Example 12.60
- Bigfish**—Exercise 9.68
- Births Australia**—Exercise 13.17
- Bishop**—Exercise 1.43
- Boat Production**—Example 12.12
- Bottles**—Exercise 6.82
- Britain Sick Leave**—Exercise 13.56
- Broccoli**—Example 9.4
- Browser Wars**—Example 1.3, Exercises 1.19, 1.25
- Citydatr**—Examples 12.7, 12.8, 12.9, Exercises 1.46, 11.84, 12.31, 12.100, 12.103, 12.111, 13.22, 13.60
- Closing Stock Prices**—Example 14.5
- Completion Times**—Example 1.9, Exercises 1.7, 2.23, 2.34, 2.53, 13.6
- Cotton**—Chapter 12 Case Study
- Crime Study**—Exercise 11.69
- Currency-Exchange Rates**—Example 1.6, Exercise 1.24
- Developing Country**—Exercise 12.82
- Dow Jones**—Exercises 11.23, 11.29, 11.37, 11.51, 11.60
- Earnings per Share**—Exercises 1.29, 16.2, 16.7, 16.14, 16.24, 16.27
- East Anglica Realty Ltd**—Exercise 13.29
- Economic Activity**—Exercises 11.36, 11.52, 11.53, 11.85, 12.81, 12.104, 13.28
- Exchange Rate**—Exercises 1.49, 14.48
- Fargo Electronics Earnings**—Exercise 16.3
- Fargo Electronics Sales**—Exercise 16.4
- Finstad and Lie Study**—Exercise 1.17
- Florin**—Exercises 1.68, 2.25
- Food Nutrition Atlas**—Exercises 9.66, 9.67, 9.72, 9.73, 10.33, 10.34, 10.42, 10.43, 10.46, 11.92–11.96
- Food Prices**—Exercise 16.20
- Gender and Salary**—Examples 12.13, 12.14
- German Import**—Exercises 12.61
- German Income**—Exercises 13.53
- Gilotti's Pizzeria**—Examples 2.8–2.10, Exercise 2.46
- Gold Price**—Exercises 1.27, 16.5, 16.12
- Grade Point Averages**—Examples 1.10, 2.3, Exercises 1.73, 2.9
- Granola**—Exercise 6.84
- Health Care Cost Analysis**—Exercises 13.66–13.68
- HEI Cost Data Variable Subset**—Examples 1.1, 1.2, 2.7, 7.5, Exercises 1.8, 1.18, 7.23, 8.34, 8.35, 9.74–9.78, 10.51–10.58, 11.97–11.101, 12.114–12.117, 14.17, Chapter 13 Case Study
- Hourly Earnings**—Exercises 16.19, 16.31
- Hours**—Example 14.13
- House Selling Price**—Exercises 10.4, 12.110
- Housing Starts**—Exercises 1.28, 16.1, 16.6, 16.13, 16.26
- Improve Your Score**—Example 8.2
- Income**—Example 14.12
- Income Canada**—Exercise 13.16
- Income Clusters**—Example 17.5
- Indonesia Revenue**—Exercise 13.52
- Industrial Production Canada**—Exercise 16.18
- Insurance**—Example 1.4
- Inventory Sales**—Exercises 1.50, 14.49, 16.11
- Japan Imports**—Exercise 13.54
- Macro2009**—Examples 1.5, 1.7, Exercise 1.22,
- Macro2010**—Example 13.8, Exercises 11.86, 12.105, 13.58, 13.61, 13.62, 16.40–16.43
- Market**—Exercise 13.5
- Mendez Mortgage**—Chapter 2 Case Study, Exercises 7.5, 7.35, 7.36
- Metals**—Exercise 13.59
- Money UK**—Exercises 13.14, 13.31, 13.35
- Motors**—Exercises 12.13, 12.14, 12.48, 13.21

New York Stock Exchange Gains and Losses—
Exercises 11.24, 11.30, 11.38, 11.46

Ole—Exercise 10.48

Pension Funds—Exercise 13.15

Power Demand—Exercise 12.12

Private Colleges—Exercises 11.87–11.91, 12.112, 12.113

Production Cost—Example 12.11

Product Sales—Exercises 16.37, 16.39

Profit Margins—Exercise 16.21

Quarterly Earnings—Exercises 16.22, 16.36, 16.38

Quarterly Sales—Exercise 16.23

Rates—Exercise 2.24

RELEVANT Magazine—Examples 1.8, 2.19,
Exercises 1.71, 14.51

Retail Sales—Examples 11.2, 11.3, 13.13

Return on Stock Price, 60 months—Examples 5.17,
11.5, Exercises 5.104, 5.106, 11.63 – 11.67

Returns—Exercise 1.38

Rising Hills—Example 11.1

Salary Study—Exercise 12.107

Salorg—Exercise 12.72

SAT Math—Example 1.14

Savings and Loan—Examples 12.3, 12.10,
Example 13.7

Shares Traded—Example 14.16

Shiller House Price Cost—Example 16.2,
Exercise 12.109

Shopping Times—Example 2.6, Exercises 1.72, 2.54

Snappy Lawn Care—Exercises 1.66, 2.41, 2.45

Staten—Exercise 12.106

Stock Market Index—Exercise 14.50

Stock Price File—Exercises 5.101–5.105

Stordata—Exercise 1.45

Storet—Exercise 10.47

Student Evaluation—Exercise 11.61

Student GPA—Exercises 2.48, 11.81, 12.99, 12.108

Student Pair—Exercises 8.32, 10.5

Student Performance—Exercise 12.71

Study—Exercises 2.10, 7.86

Sugar—Exercise 7.24

Sugar Coated Wheat—Exercises 6.83, 8.14

Sun—Exercises 1.39, 2.11

Teacher Rating—Exercise 12.92

Tennis—Exercise 1.15

Thailand Consumption—Exercises 13.18, 13.36

TOC—Exercise 7.45

Trading Volume—Exercise 16.25

Trucks—Example 7.4

Turkey Feeding—Examples 10.1, 10.4

Vehicle Travel State—Exercises 11.82, 11.83, 12.80,
12.101, 12.102

Water—Exercises 1.37, 2.22, 7.6, 7.103

Weekly Sales—Example 14.17

Using Graphs to Describe Data

CHAPTER OUTLINE

- 1.1 Decision Making in an Uncertain Environment
 - Random and Systematic Sampling
 - Sampling and Nonsampling Errors
- 1.2 Classification of Variables
 - Categorical and Numerical Variables
 - Measurement Levels
- 1.3 Graphs to Describe Categorical Variables
 - Tables and Charts
 - Cross Tables
 - Pie Charts
 - Pareto Diagrams
- 1.4 Graphs to Describe Time-Series Data
- 1.5 Graphs to Describe Numerical Variables
 - Frequency Distributions
 - Histograms and Ogives
 - Shape of a Distribution
 - Stem-and-Leaf Displays
 - Scatter Plots
- 1.6 Data Presentation Errors
 - Misleading Histograms
 - Misleading Time-Series Plots

Introduction

What are the projected sales of a new product? Will the cost of Google shares continue to increase? Who will win the next presidential election? How satisfied were you with your last purchase at Starbucks, Best Buy, or Sports Authority? If you were hired by the National Nutrition Council of the United States, how would you determine if the Council's guidelines on consumption of fruit, vegetables, snack foods, and soft drinks are being met? Do people who are physically active have healthier diets than people who are not physically active? What factors (perhaps disposable income or federal funds) are significant in forecasting the aggregate consumption of durable goods? What effect will a 2% increase in interest rates have on residential investment? Do

credit scores, current balance, or outstanding maintenance balance contribute to an increase in the percentage of a mortgage company's delinquent accounts increasing? Answers to questions such as these come from an understanding of statistics, fluctuations in the market, consumer preferences, trends, and so on.

Statistics are used to predict or forecast sales of a new product, construction costs, customer-satisfaction levels, the weather, election results, university enrollment figures, grade point averages, interest rates, currency-exchange rates, and many other variables that affect our daily lives. We need to absorb and interpret substantial amounts of data. Governments, businesses, and scientific researchers spend billions of dollars collecting data. But once data are collected, what do we do with them? How do data impact decision making?

In our study of *statistics* we learn many tools to help us process, summarize, analyze, and interpret data for the purpose of making better decisions in an uncertain environment. Basically, an understanding of statistics will permit us to make sense of all the data.

In this chapter we introduce tables and graphs that help us gain a better understanding of data and that provide visual support for improved decision making. Reports are enhanced by the inclusion of appropriate tables and graphs, such as frequency distributions, bar charts, pie charts, Pareto diagrams, line charts, histograms, stem-and-leaf displays, or ogives. Visualization of data is important. We should always ask the following questions: What does the graph suggest about the data? What is it that we see?

1.1 DECISION MAKING IN AN UNCERTAIN ENVIRONMENT

Decisions are often made based on limited information. Accountants may need to select a portion of records for auditing purposes. Financial investors need to understand the market's fluctuations, and they need to choose between various portfolio investments. Managers may use surveys to find out if customers are satisfied with their company's products or services. Perhaps a marketing executive wants information concerning customers' taste preferences, their shopping habits, or the demographics of Internet shoppers. An investor does not know with certainty whether financial markets will be buoyant, steady, or depressed. Nevertheless, the investor must decide how to balance a portfolio among stocks, bonds, and money market instruments while future market movements are unknown.

For each of these situations, we must carefully define the problem, determine what data are needed, collect the data, and use statistics to summarize the data and make inferences and decisions based on the data obtained. Statistical thinking is essential from initial problem definition to final decision, which may lead to reduced costs, increased profits, improved processes, and increased customer satisfaction.

Random and Systematic Sampling

Before bringing a new product to market, a manufacturer wants to arrive at some assessment of the likely level of demand and may undertake a market research survey. The manufacturer is, in fact, interested in *all* potential buyers (the population). However, populations are often so large that they are unwieldy to analyze; collecting complete information for a population could be impossible or prohibitively expensive. Even in circumstances where sufficient resources seem to be available, time constraints make the examination of a subset (sample) necessary.

Population and Sample

A **population** is the complete set of all items that interest an investigator. Population size, N , can be very large or even infinite. A **sample** is an observed subset (or portion) of a population with sample size given by n .

Examples of populations include the following:

- All potential buyers of a new product
- All stocks traded on the NYSE Euronext
- All registered voters in a particular city or country
- All accounts receivable for a corporation

Our eventual aim is to make statements based on sample data that have some validity about the population at large. We need a sample, then, that is representative of the population. How can we achieve that? One important principle that we must follow in the sample selection process is randomness.

Random Sampling

Simple random sampling is a procedure used to select a sample of n objects from a population in such a way that each member of the population is chosen strictly by chance, the selection of one member does not influence the selection of any other member, each member of the population is equally likely to be chosen, and every possible sample of a given size, n , has the same chance of selection. This method is so common that the adjective *simple* is generally dropped, and the resulting sample is called a **random sample**.

Another sampling procedure is systematic sampling (stratified sampling and cluster sampling are discussed in Chapter 17).

Systematic Sampling

Suppose that the population list is arranged in some fashion unconnected with the subject of interest. **Systematic sampling** involves the selection of every j th item in the population, where j is the ratio of the population size N to the desired sample size, n ; that is, $j = N/n$. Randomly select a number from 1 to j to obtain the first item to be included in your systematic sample.

Suppose that a sample size of 100 is desired and that the population consists of 5,000 names in alphabetical order. Then $j = 50$. Randomly select a number from 1 to 50. If your number is 20, select it and every 50th number, giving the systematic sample of elements numbered 20, 70, 120, 170, and so forth, until all 100 items are selected. A systematic sample is analyzed in the same fashion as a simple random sample on the grounds that, relative to the subject of inquiry, the population listing is already in random order. The danger is that there could be some subtle, unsuspected link between the ordering of the population and the subject under study. If this were so, bias would be induced if systematic sampling was employed. Systematic samples provide a good representation of the population if there is no cyclical variation in the population.

Sampling and Nonsampling Errors

Suppose that we want to know the average age of registered voters in the United States. Clearly, the population size is so large that we might take only a random sample, perhaps 500 registered voters, and calculate their average age. Because this average is based on sample data, it is called a *statistic*. If we were able to calculate the average age of the entire population, then the resulting average would be called a *parameter*.

Parameter and Statistic

A **parameter** is a numerical measure that describes a specific characteristic of a population. A **statistic** is a numerical measure that describes a specific characteristic of a sample.

Throughout this book we will study ways to make decisions about a population parameter, based on a sample statistic. We must realize that some element of uncertainty will always remain, as we do not know the exact value of the parameter. That is, when a sample is taken from a population, the value of any population parameter will not be able to be known *precisely*. One source of error, called **sampling error**, results from the fact that information is available on only a subset of all the population members. In Chapters 6, 7, and 8 we develop statistical theory that allows us to characterize the nature of the sampling error and to make certain statements about population parameters.

In practical analyses there is the possibility of an error unconnected with the kind of sampling procedure used. Indeed, such errors could just as well arise if a complete census of the population were taken. These are referred to as **nonsampling errors**. Examples of nonsampling errors include the following:

1. **The population actually sampled is not the relevant one.** A celebrated instance of this sort occurred in 1936, when *Literary Digest* magazine confidently predicted that Alfred Landon would win the presidential election over Franklin Roosevelt. However, Roosevelt won by a very comfortable margin. This erroneous forecast resulted from the fact that the members of the *Digest's* sample had been taken from telephone directories and other listings, such as magazine subscription lists and automobile registrations. These sources considerably underrepresented the poor, who were predominantly Democrats. To make an inference about a population (in this case the U.S. electorate), it is important to sample that population and not some subgroup of it, however convenient the latter course might appear to be.
2. **Survey subjects may give inaccurate or dishonest answers.** This could happen because questions are phrased in a manner that is difficult to understand or in a way that appears to make a particular answer seem more palatable or more desirable. Also, many questions that one might want to ask are so sensitive that it would be foolhardy to expect uniformly honest responses. Suppose, for example, that a plant manager wants to assess the annual losses to the company caused by employee thefts. In principle, a random sample of employees could be selected and sample members asked, What have you stolen from this plant in the past 12 months? This is clearly not the most reliable means of obtaining the required information!
3. **There may be no response to survey questions.** Survey subjects may not respond at all, or they may not respond to certain questions. If this is substantial, it can induce additional sampling and nonsampling errors. The sampling error arises because the achieved sample size will be smaller than that intended. Nonsampling error possibly occurs because, in effect, the population being sampled is not the population of interest. The results obtained can be regarded as a random sample *from the population that is willing to respond*. These people may differ in important ways from the larger population. If this is so, a bias will be induced in the resulting estimates.

There is no general procedure for identifying and analyzing nonsampling errors. But nonsampling errors could be important. The investigator must take care in such matters as identifying the relevant population, designing the questionnaire, and dealing with non-response in order to minimize the significance of nonsampling errors. In the remainder of this book it is assumed that such care has been taken, and our discussion centers on the treatment of sampling errors.

To think statistically begins with problem definition: (1) What information is required? (2) What is the relevant population? (3) How should sample members be selected? (4) How should information be obtained from the sample members? Next we will want to know how to use sample information to make decisions about our population of interest. Finally, we will want to know what conclusions can be drawn about the population.

After we identify and define a problem, we collect data produced by various processes according to a design, and then we analyze that data using one or more statistical procedures. From this analysis, we obtain information. Information is, in turn, converted into knowledge, using understanding based on specific experience, theory, literature, and additional statistical procedures. Both descriptive and inferential statistics are used to change data into knowledge that leads to better decision making. To do this, we use descriptive statistics and inferential statistics.

Descriptive and Inferential Statistics

Descriptive statistics focus on graphical and numerical procedures that are used to summarize and process data. **Inferential statistics** focus on using the data to make predictions, forecasts, and estimates to make better decisions.

1.2 CLASSIFICATION OF VARIABLES

A variable is a specific characteristic (such as age or weight) of an individual or object. Variables can be classified in several ways. One method of classification refers to the type and amount of information contained in the data. Data are either categorical or numerical. Another method, introduced in 1946 by American psychologist Stanley Smith Stevens is to classify data by levels of measurement, giving either qualitative or quantitative variables. Correctly classifying data is an important first step to selecting the correct statistical procedures needed to analyze and interpret data.

Categorical and Numerical Variables

Categorical variables produce responses that belong to groups or categories. For example, responses to yes/no questions are categorical. Are you a business major? and Do you own a car? are limited to yes or no answers. A health care insurance company may classify incorrect claims according to the type of errors, such as procedural and diagnostic errors, patient information errors, and contractual errors. Other examples of categorical variables include questions on gender or marital status. Sometimes categorical variables include a range of choices, such as “strongly disagree” to “strongly agree.” For example, consider a faculty-evaluation form where students are to respond to statements such as the following: The instructor in this course was an effective teacher (1: strongly disagree; 2: slightly disagree; 3: neither agree nor disagree; 4: slightly agree; 5: strongly agree).

Numerical variables include both discrete and continuous variables. A **discrete numerical variable** may (but does not necessarily) have a finite number of values. However, the most common type of discrete numerical variable produces a response that comes from a counting process. Examples of discrete numerical variables include the number of students enrolled in a class, the number of university credits earned by a student at the end of a particular semester, and the number of Microsoft stocks in an investor’s portfolio.

A **continuous numerical variable** may take on any value within a given range of real numbers and usually arises from a measurement (not a counting) process. Someone might say that he is 6 feet (or 72 inches) tall, but his height could actually be 72.1 inches, 71.8 inches, or some other similar number, depending on the accuracy of the instrument used to measure height. Other examples of continuous numerical variables include the weight of a cereal box, the time to run a race, the distance between two cities, or the temperature. In each case the value could deviate within a certain amount, depending on the precision of the measurement instrument used. We tend to truncate continuous variables in daily conversation and treat them as though they were the same as discrete variables without even giving it a second thought.

Measurement Levels

We can also describe data as either *qualitative* or *quantitative*. With **qualitative data** there is no measurable meaning to the “difference” in numbers. For example, one basketball player is assigned the number 20 and another player has the number 10. We cannot conclude that the first player plays twice as well as the second player. However, with **quantitative data** there is a measurable meaning to the difference in numbers. When one student scores 90 on an exam and another student scores 45, the difference is measurable and meaningful.

Qualitative data include nominal and ordinal levels of measurement. Quantitative data include interval and ratio levels of measurement.

Nominal and ordinal levels of measurement refer to data obtained from categorical questions. Responses to questions on gender, country of citizenship, political affiliation, and ownership of a mobile phone are nominal. **Nominal data** are considered the lowest or weakest type of data, since numerical identification is chosen strictly for convenience and does not imply ranking of responses.

The values of nominal variables are words that describe the categories or classes of responses. The values of the gender variable are male and female; the values of Do you own a car? are yes and no. We arbitrarily assign a code or number to each response. However, this number has no meaning other than for categorizing. For example, we could code gender responses or yes/no responses as follows:

1 = Male; 2 = Female
1 = Yes; 2 = No

Ordinal data indicate the rank ordering of items, and similar to nominal data the values are words that describe responses. Some examples of ordinal data and possible codes are as follows:

1. Product quality rating (1: poor; 2: average; 3: good)
2. Satisfaction rating with your current Internet provider (1: very dissatisfied; 2: moderately dissatisfied; 3: no opinion; 4: moderately satisfied; 5: very satisfied)
3. Consumer preference among three different types of soft drink (1: most preferred; 2: second choice; 3: third choice)

In these examples the responses are ordinal, or put into a rank order, but there is no measurable meaning to the “difference” between responses. That is, the difference between your first and second choices may not be the same as the difference between your second and third choices.

Interval and ratio levels of measurement refer to data obtained from numerical variables, and meaning is given to the *difference* between measurements. An interval scale indicates rank and distance from an arbitrary zero measured in unit intervals. That is, data are provided relative to an arbitrarily determined benchmark. Temperature is a classic example of this level of measurement, with arbitrarily determined benchmarks generally based on either Fahrenheit or Celsius degrees. Suppose that it is 80°F in Orlando, Florida, and only 20°F in St. Paul, Minnesota. We can conclude that the difference in temperature is 60°, but we cannot say that it is four times as warm in Orlando as it is in St. Paul. The year is another example of an interval level of measurement, with benchmarks based most commonly on the Gregorian calendar.

Ratio data indicate both rank and distance from a natural zero, with ratios of two measures having meaning. A person who weighs 200 pounds is twice the weight of a person who weighs 100 pounds; a person who is 40 years old is twice the age of someone who is 20 years old.

After collecting data, we first need to classify responses as categorical or numerical or by measurement scale. Next, we assign an arbitrary ID or code number to each response. Some graphs are appropriate for categorical variables, and others are used for numerical variables.

Note that data files usually contain “missing values.” For example, respondents to a questionnaire may choose not to answer certain questions about gender, age, income, or some other sensitive topic. Missing values require a special code in the data entry stage. Unless missing values are properly handled, it is possible to obtain erroneous output. Statistical software packages handle missing values in different ways.

EXERCISES



Visit www.mymathlab.com/global or www.pearsonglobal Editions.com/newbold to access the data files.


Basic Exercises


- 1.1 A mortgage company randomly samples accounts of their time-share customers. State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?
 - a. The original purchase price of a customer’s time-share unit
 - b. The state (or country) of residence of a time-share owner
 - c. A time-share owner’s satisfaction level with the maintenance of the unit purchased (1: very dissatisfied to 5: very satisfied)
 - d. The number of times a customer’s payment was late
- 1.2 Visitors to a supermarket in Singapore were asked to complete a customer service survey. Are the answers to the following survey questions categorical or numerical? If an answer is categorical, give the level of measurement. If an answer is numerical, is it discrete or continuous?
 - a. Have you visited this store before?
 - b. How would you rate the level of customer service you received today on a scale from 1 (very poor) to 5 (very good)?
 - c. How much money did you spend in the store today?
- 1.3 A questionnaire was distributed at a large university to find out the level of student satisfaction with various activities and services. For example, concerning parking availability, students were asked to indicate their level of satisfaction on a scale from 1 (very dissatisfied) to 5 (very satisfied). Is a student’s response to this question numerical or categorical? If numerical, is it discrete or continuous? If categorical, give the level of measurement.
- 1.4 Faculty at one university were asked a series of questions in a recent survey. State the type of data for each question.
 - a. Indicate your level of satisfaction with your teaching load (very satisfied, moderately satisfied, neutral, moderately dissatisfied, or very dissatisfied).
 - b. How many of your research articles were published in refereed journals during the last 5 years?

- c. Did you attend the last university faculty meeting?
- d. Do you think that the teaching evaluation process needs to be revised?

- 1.5 A number of questions were posed to a random sample of visitors to a London tourist information center. For each question below, describe the type of data obtained.
 - a. Are you staying overnight in London?
 - b. How many times have you visited London previously?
 - c. Which of the following attractions have you visited?
 - Tower of London
 - Buckingham Palace
 - Big Ben
 - Covent Garden
 - Westminster Abbey
 - d. How likely are you to visit London again in the next 12 months: (1) unlikely, (2) likely, (3) very likely?
- 1.6 Residents in one housing development were asked a series of questions by their homeowners’ association. Identify the type of data for each question.
 - a. Did you play golf during the last month on the development’s new golf course?
 - b. How many times have you eaten at the country club restaurant during the last month?
 - c. Do you own a camper?
 - d. Rate the new security system for the development (very good, good, poor, or very poor).

Application Exercises

- 1.7  The supervisor of a very large plant obtained the times (in seconds) to complete a task for a random sample of employees. This information and other data about the employees are stored in the data file **Completion Times**.
 - a. Give an example of a categorical variable with ordinal responses.
 - b. Give an example of a categorical variable with nominal responses.
 - c. Give an example of a numerical variable.

- 1.8  The U.S. Department of Agriculture (USDA) Center for Nutrition Policy and Promotion (CNPP) developed and administered the Healthy Eating Index–2005 to measure how well the population follows the recommendations of the 2005 Dietary Guidelines for Americans. The data are contained in the data file **HEI Cost Data Variable Subset**.
- Give an example of a categorical variable with ordinal responses.
 - Give an example of a categorical variable with nominal responses.
 - Give an example of a numerical variable with continuous responses.
 - Give an example of a numerical variable with discrete responses.

1.3 GRAPHS TO DESCRIBE CATEGORICAL VARIABLES

We can describe categorical variables using frequency distribution tables and graphs such as bar charts, pie charts, and Pareto diagrams. These graphs are commonly used by managers and marketing researchers to describe data collected from surveys and questionnaires.

Frequency Distribution

A **frequency distribution** is a table used to organize data. The left column (called classes or groups) includes all possible responses on a variable being studied. The right column is a list of the frequencies, or number of observations, for each class. A **relative frequency distribution** is obtained by dividing each frequency by the number of observations and multiplying the resulting proportion by 100%.

Tables and Charts

The classes that we use to construct frequency distribution tables of a categorical variable are simply the possible responses to the categorical variable. Bar charts and pie charts are commonly used to describe categorical data. If our intent is to draw attention to the *frequency* of each category, then we will most likely draw a **bar chart**. In a bar chart the height of a rectangle represents each frequency. There is no need for the bars to touch.

Example 1.1 Healthy Eating Index 2005 (HEI–2005): Activity Level (Frequency Distribution and Bar Chart)

The U.S. Department of Agriculture (USDA) Center for Nutrition Policy and Promotion (CNPP) and the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), conduct surveys to assess the health and nutrition of the U.S. population. The CNPP conducts the Healthy Eating Index (Guenther et al. 2007) and the NCHS conducts the National Health and Nutrition Examination Survey (CDC 2003–2004). The Healthy Eating Index (HEI) monitors the diet quality of the U.S. population, particularly how well it conforms to dietary guidance. The HEI–2005 measures how well the population follows the recommendations of the 2005 *Dietary Guidelines for Americans* (Guenther et al.). In particular it measures, on a 100-point scale, the adequacy of consumption of vegetables, fruits, grains, milk, meat and beans, and liquid oils.

The data file **HEI Cost Data Variable Subset** contains considerable information on randomly selected individuals who participated in two extended interviews and medical examinations. Data for the first interview are identified by daycode = 1; data for the second interview are identified by daycode = 2. Other variables in the data file are described in the data dictionary in the Chapter 10 Appendix.

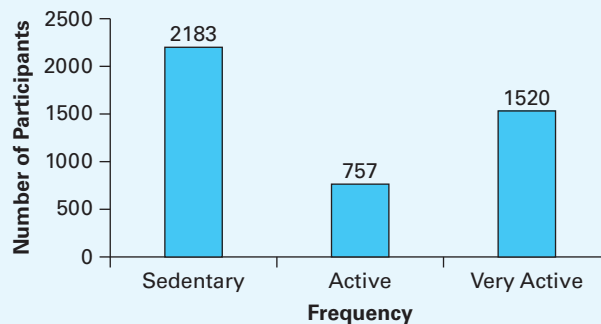
One variable in the HEI–2005 study is a participant’s activity level coded as 1 = sedentary, 2 = active, and 3 = very active. Set up a frequency distribution and relative frequency distribution and construct a simple bar chart of activity level for the HEI–2005 participants during their first interview.

Solution Table 1.1 is a frequency distribution and a relative frequency distribution of the categorical variable “activity level.” Figure 1.1 is a bar chart of this data.

Table 1.1 HEI–2005 Participants’ Activity Level: First Interview

| | <i>PARTICIPANTS</i> | <i>PERCENT</i> |
|-------------|---------------------|----------------|
| Sedentary | 2,183 | 48.9 |
| Active | 757 | 17.0 |
| Very active | 1,520 | 34.1 |
| Total | 4,460 | 100.0 |

Figure 1.1 HEI–2005 Participants’ Activity Level: First Interview (Simple Bar Chart)



Cross Tables

There are situations in which we need to describe relationships between categorical or ordinal variables. Market-research organizations describe attitudes toward products, measured on an ordinal scale, as a function of educational levels, social status measures, geographic areas, and other ordinal or categorical variables. Personnel departments study employee evaluation levels versus job classifications, educational levels, and other employee variables. Production analysts study relationships between departments or production lines and performance measures to determine reasons for product change, reasons for interruption of production, and quality of output. These situations are usually described by cross tables and pictured by component or cluster bar charts. These bar charts are useful extensions of the simple bar chart in Figure 1.1.

Cross Table

A **cross table**, sometimes called a crosstab or a contingency table, lists the number of observations for every combination of values for two categorical or ordinal variables. The combination of all possible intervals for the two variables defines the cells in a table. A cross table with r rows and c columns is referred to as an $r \times c$ cross table.

Example 1.2 illustrates the use of cross tables, component bar charts, and cluster bar charts to describe graphically two categorical variables from the HEI–2005 study.

Example 1.2 HEI-2005: Activity Level and Gender (Component and Cluster Bar Charts)

Consider again the data in Table 1.1. Sometimes a comparison of one variable (activity level) with another variable (such as gender) is of interest. Construct component and cluster bar charts that compare activity level and gender. Use the data coded daycode = 1 in the data file **HEI Cost Data Variable Subset**.

Solution Table 1.2 is a cross table of activity levels (1 = sedentary; 2 = active; and 3 = very active) and gender (0 = male; 1 = female) obtained from the first interview for HEI-2005 participants.

Table 1.2 HEI-2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

| | MALES | FEMALES | TOTAL |
|-------------|------------|------------|--------------|
| Sedentary | 957 | 1,226 | 2,183 |
| Active | 340 | 417 | 757 |
| Very active | <u>842</u> | <u>678</u> | <u>1,520</u> |
| Total | 2,139 | 2,321 | 4,460 |

Figure 1.2 displays this information in a *component* or *stacked bar chart*. Figure 1.3 is a *cluster*, or *side-by-side*, bar chart of the same data.

Figure 1.2 HEI-2005 Participants' Activity Level (First Interview) by Gender (Component Bar Chart)

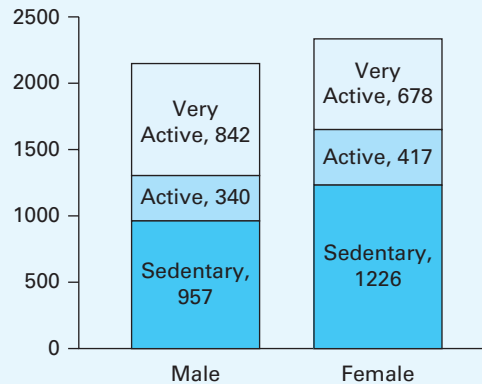


Figure 1.3 HEI-2005 Participants' Activity Level (First Interview) by Gender (Cluster Bar Chart)

