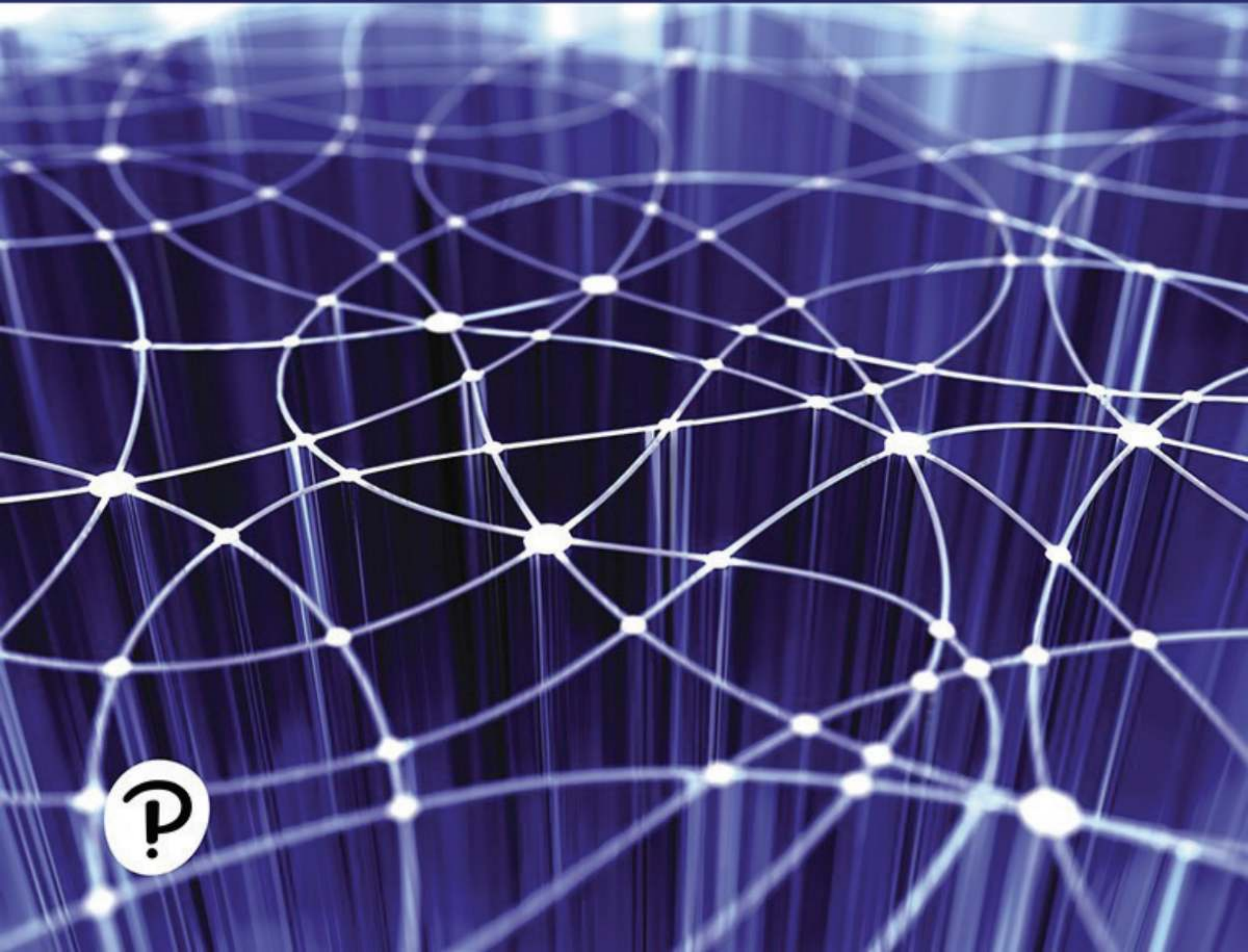
A SECOND COURSE IN STATISTICS

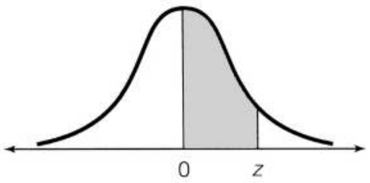
REGRESSION ANALYSIS

Eighth Edition

William Mendenhall
Terry Sincich



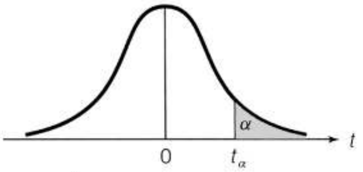
Normal curve areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Abridged from Table 1 of A. Hald, *Statistical Tables and Formulas* (New York: John Wiley & Sons, Inc.), 1952. Reproduced by permission of the publisher.

Critical values for Student's t



ν	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Source: This table is reproduced with the kind permission of the trustees of *Biometrika* from E. S. Pearson and H. O. Hartley (Eds.), *The Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., *Biometrika*, 1996.

A Second Course in Statistics
Regression Analysis

Eighth Edition

William Mendenhall
University of Florida

Terry Sincich
University of South Florida

Pearson

Director, Portfolio Management: Deirdre Lynch
Courseware Portfolio Manager: Patrick Barbera
Courseware Portfolio Management Assistant:
Morgan Danna
Content Producer: Angela Montoya
Managing Producer: Karen Wernholm
Producer: Shana Siegmund
Product Marketing Manager: Yvonne Vannatta
Product Marketing Assistant: Jon Bryant
Field Marketing Manager: Evan St. Cyr
Senior Author Support/Technology
Specialist: Joe Vetere
Manager, Rights and Permissions: Gina Cheselka
Text and Cover Design: Jerilyn Bockorick
Production Coordination, Composition,
and Illustrations: Pearson CSC
Manufacturing Buyer: Carol Melville,
LSC Communications
Cover Image: enot-poloskun/Getty Images

[For instructor editions: This work is solely for the use of instructors and administrators for the purpose of teaching courses and assessing student learning. Unauthorized dissemination, publication or sale of the work, in whole or in part (including posting on the internet) will destroy the integrity of the work and is strictly prohibited.]

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.

Library of Congress Cataloging-in-Publication Data

Names: Mendenhall, William, author. | Sincich, Terry, author.
Title: A second course in statistics : regression analysis / William
Mendenhall, Terry Sincich.
Other titles: Corrected title: Second course in statistics : regression
analysis
Description: Eighth edition. | [Hoboken, New Jersey] : Pearson Education,
Inc., [2020] | Includes bibliographical references and index.
Identifiers: LCCN 2018040176 | ISBN 9780135163795 | ISBN 013516379X
Subjects: LCSH: Commercial statistics. | Statistics. | Regression analysis.
Classification: LCC HF1017 .M46 2020 | DDC 519.5/36--dc23 LC record available at https://lccn.loc.gov/2018040176

Pearson

ISBN 13: 978-0-13-516379-5
ISBN 10: 0-13-516379-X

Contents

6 Variable Screening Methods **337**

Case Study 3 Deregulation of the Intrastate Trucking Industry **355**

7 Some Regression Pitfalls **365**

8 Residual Analysis **393**

Case Study 4 An Analysis of Rain Levels in California **449**

Case Study 5 An Investigation of Factors Affecting the Sale Price of Condominium Units Sold at Public Auction **457**

Online

Preface

Overview

This text is designed for two types of statistics courses. The early chapters, combined with a selection of the case studies, are designed for use in the second half of a two-semester (two-quarter) introductory statistics sequence for undergraduates with statistics or non-statistics majors. Or, the text can be used for a course in applied regression analysis for masters or Ph.D. students in other fields.

At first glance, these two uses for the text may seem inconsistent. How could a text be appropriate for both undergraduate and graduate students? The answer lies in the content. In contrast to a course in statistical theory, the level of mathematical knowledge required for an applied regression analysis course is minimal. Consequently, the difficulty encountered in learning the mechanics is much the same for both undergraduate and graduate students. The challenge is in the application – diagnosing practical problems, deciding on the appropriate linear model for a given situation, and knowing which inferential technique will answer the researcher's practical question. This *takes experience*, and it explains why a student with a non-statistics major can take an undergraduate course in applied regression analysis and still benefit from covering the same ground in a graduate course.

Introductory Statistics Course

It is difficult to identify the amount of material that should be included in the second semester of a two-semester sequence in introductory statistics. Optionally, a few lectures should be devoted to Chapter 1 (A Review of Basic Concepts) to make certain that all students possess a common background knowledge of the basic concepts covered in a first-semester (first-quarter) course. Chapter 2 (Introduction to Regression Analysis), Chapter 3 (Simple Linear Regression), Chapter 4 (Multiple Regression Models), Chapter 5 (Model Building), Chapter 6 (Variable Screening Methods), Chapter 7 (Some Regression Pitfalls), and Chapter 8 (Residual Analysis) provide the core for an applied regression analysis course. These chapters could be supplemented by the addition of Chapter 10 (Time Series Modeling and Forecasting), Chapter 11 (Principles of Experimental Design), and Chapter 12 (The Analysis of Variance for Designed Experiments).

Applied Regression for Graduates

In our opinion, the quality of an applied graduate course is not measured by the number of topics covered or the amount of material memorized by the students. The measure is how well they can apply the techniques covered in the course to the solution of real problems encountered in their field of study. Consequently, we advocate moving on to new topics only after the students have demonstrated ability (through testing) to apply the techniques under discussion. In-class consulting sessions, where a case study is presented and the students have the opportunity to diagnose the problem and recommend an appropriate method of analysis, are very helpful in teaching applied regression analysis. This approach is particularly useful in helping students master the difficult topic of model selection and model building (Chapters 4-8) and relating questions about the model to real-world questions. The seven case studies (which follow relevant chapters) illustrate the type of material that might be useful for this purpose.

A course in applied regression analysis for graduate students would start in the same manner as the undergraduate course, but would move more rapidly over the review material and would more than likely be supplemented by Appendix A (Derivation of the Least Squares Estimates), Appendix B (The Mechanics of a Multiple Regression Analysis), and/or Appendix C (A Procedure for Inverting a Matrix), one of the statistical software Windows tutorials available at the course website (SAS, SPSS, MINITAB, or R), Chapter 9 (Special Topics in Regression), and other chapters selected by the instructor. As in the undergraduate course, we recommend the use of case studies and in-class consulting sessions to help students develop an ability to formulate appropriate statistical models and to interpret the results of their analyses.

Features

1. **Readability.** We have purposely tried to make this a teaching (rather than a reference) text. Concepts are explained in a logical intuitive manner using worked examples.
2. **Emphasis on model building.** The formulation of an appropriate statistical model is fundamental to any regression analysis. This topic is treated in Chapters 4-8 and is emphasized throughout the text.
3. **Emphasis on developing regression skills.** In addition to teaching the basic concepts and methodology of regression analysis, this text stresses its use, as a tool, in solving applied problems. Consequently, a major objective of the text is to develop a skill in applying regression analysis to appropriate real-life situations.
4. **Real data-based examples and exercises.** The text contains many worked examples that illustrate important aspects of model construction, data analysis, and the interpretation of results. Nearly every exercise is based on data and research extracted from a news article, magazine, or journal. Exercises are located at the ends of key sections and at the ends of chapters.
5. **Case studies.** The text contains seven case studies, each of which addresses a real-life research problem. The student can see how regression analysis was used to answer the practical questions posed by the problem, proceeding with the formulation of appropriate statistical models to the analysis and interpretation of sample data.
6. **Data sets.** The online resource provides complete data sets that are associated with the case studies, exercises and examples. These can be used by instructors and students to practice model-building and data analyses.
7. **Extensive use of statistical software.** Tutorials on how to use any of four popular statistical software packages – SAS, SPSS, MINITAB, and R – are provided online. Printouts associated with the respective software packages are presented and discussed throughout the text.
8. **End-of-Chapter Summaries.** Important points are reinforced through flow graphs (which aid in selecting the appropriate statistical method) and boxed notes with key words, formulas, definitions, lists, and key concepts.

New to the 8th Edition

Although the scope and coverage remain the same, the eighth edition contains several substantial changes, additions, and enhancements:

1. **New and Updated Case Studies.** *Case Study 2: Modeling Sale Prices of Residential Properties*, has been updated with current data. A new case study (*Case Study 7: Voice Versus Face Recognition – Does One Follow the Other?*) now follows the chapter on analysis of variance.
2. **Real Data-based Exercises.** Many new and updated exercises, based on contemporary studies and real data in a variety of fields, have been added. Most of these exercises foster and promote critical thinking skills.
3. **Statistical Software Output.** All statistical software printouts shown in the text have been updated to reflect the most recent version of the software: Minitab, SAS, and SPSS.
4. **Updated Statistical Software Tutorials.** They can be found at the following website: www.pearson.com/math-stats-resources. The text's online resource provides updated instructions on how to use the Windows versions of SAS, SPSS, MINITAB, and R. Step-by-step instructions and screen shots for each method presented in the text are shown.
5. **Updated and New Sections in Chapter 9: Special Topics in Regression.** The section on logistic regression (Section 9.6) has been expanded. A new section (Section 9.7) on Poisson regression has been added. And, in addition to ridge regression, Section 9.8 now includes a discussion of Lasso regression.

Numerous less obvious changes in details have been made throughout the text in response to suggestions by current users of the earlier editions.

Supplements

The text is also accompanied by the following supplementary material:

1. **Instructor's solutions manual.** The instructor's exercise solutions manual presents the full solutions to the other half (the even) exercises contained in the text. For adopters, the manual is complimentary from the publisher.
2. **Data Files.** They can be found at the book's resource website: www.pearson.com/math-stats-resources. The text's online resource provides data files for all data sets marked with a data (💿) icon in the text. These include data sets for text examples, exercises, and case studies. The data files are saved in ".csv" format for easy importing into statistical software such as R, as well as in SAS (".sas7bdat"), SPSS (".sav"), and Minitab (".mtw") format.

Acknowledgments

We want to thank the many people who contributed time, advice, and other assistance to this project. We owe particular thanks to the many reviewers who provided suggestions and recommendations at the onset of the project and for the succeeding editions (including the 8th):

Jack Miller (University of Michigan)
Scott Grimshaw (Brigham Young University)

Liam O'Brien (Colby College)
Subarna K Samanta (The College of New Jersey)
Wolde Woubneh (Kean University)
Alan Huebner (University of Notre Dame)
Jen-Wen Lin (University of Toronto)
Karen Keating (Kansas State University)
Seamus Freyne (Mississippi State University)
Martin Tanner (Northwestern University)
Rebecca L. Pierce (Ball State University)
Julius Esunge (University of Mary Washington)
Brant Deppa (Winona State University)
Ross Hosky (Appalachian State University)
David Holmes (College of New Jersey)
Patrick McKnight (George Mason University)
David Kidd (George Mason University)
W.R. Stephenson (Iowa State University)
Lingyun Ma (University of Georgia)
Pinyuen Chen (Syracuse University)
Gokarna Aryal (Purdue University, Calumet)
Monnie McGee (Southern Methodist University)
Ruben Zamar (University of British Columbia)
Tom O'Gorman (Northern Illinois University)
William Bridges, Jr. (Clemson University)
Paul Maiste (Johns Hopkins University)
Mohammed Askalani, Mankato State University (Minnesota)
Ken Boehm, Pacific Telesis (California)
Andrew C. Brod, University of North Carolina at Greensboro
James Daly, California State Polytechnic Institute at San Luis Obispo
Assane Djeto, University of Nevada - Las Vegas
Robert Elrod, Georgia State University
James Ford, University of Delaware
Carol Ghomi, University of Houston
James Holstein, University of Missouri at Columbia
Steve Hora, Texas Technological University
K. G. Janardan, Eastern Michigan University
Thomas Johnson, North Carolina State University
Ann Kittler, Ryerson College (Toronto)
James T. McClave, University of Florida
John Monahan, North Carolina State University
Kris Moore, Baylor University
Farrokh Nasri, Hofstra University
Robert Pavur, University of North Texas

P. V. Rao, University of Florida
Tom Rothrock, Info Tech, Inc.
Ray Twery, University of North Carolina at Charlotte
Joseph Van Matre, University of Alabama at Birmingham
William Weida, United States Air Force Academy
Dean Wichern, Texas A&M University
James Willis, Louisiana State University

We are particularly grateful to Charles Bond, Evan Anderson, Jim McClave, Herman Kelting, Rob Turner, P.J. Taylor and Mike Jacob, who provided data sets and/or background information used in the case studies.

This page intentionally left blank

Chapter 1

A Review of Basic Concepts (Optional)

Contents

Objectives

1. Review some basic concepts of sampling.
2. Review methods for describing both qualitative and quantitative data.
3. Review inferential statistical methods: confidence intervals and hypothesis tests.

Although we assume students have had a prerequisite introductory course in statistics, courses vary somewhat in content and in the manner in which they present statistical concepts. To be certain that we are starting with a common background, we use this chapter to review some basic definitions and concepts. Coverage is optional.

1.1 Statistics and Data

According to *The Random House College Dictionary* (2001 ed.), statistics is "the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data." In short, statistics is the **science of data**—a science that will enable you to be proficient data producers and efficient data users.

> **Definition 1.1** **Statistics** is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data.

Data are obtained by measuring some characteristic or property of the objects (usually people or things) of interest to us. These objects upon which the measurements (or observations) are made are called **experimental units**, and the properties being measured are called **variables** (since, in virtually all studies of interest, the property varies from one observation to another).

Definition 1.2 An **experimental unit** is an object (person or thing) upon which we collect data.

Definition 1.3 A **variable** is a characteristic (property) of the experimental unit with outcomes (data) that vary from one observation to the next.

All data (and consequently, the variables we measure) are either **quantitative** or **qualitative** in nature. Quantitative data are data that can be measured on a naturally occurring numerical scale. In general, qualitative data take values that are nonnumerical; they can only be classified into categories. The statistical tools that we use to analyze data depend on whether the data are quantitative or qualitative. Thus, it is important to be able to distinguish between the two types of data.

Definition 1.4 **Quantitative data** are observations measured on a naturally occurring numerical scale.

Definition 1.5 Nonnumerical data that can only be classified into one of a group of categories are said to be **qualitative data**.

Example 1.1

Chemical and manufacturing plants often discharge toxic waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:

1. River/creek where each fish was captured
2. Number of miles upstream where the fish was captured
3. Species (channel catfish, largemouth bass, or smallmouth buffalofish)
4. Length (centimeters)
5. Weight (grams)
6. DDT concentration (parts per million)

The data are saved in the FISHDDT file. Data for 10 of the 144 captured fish are shown in Table 1.1.

(a) Identify the experimental units.

(b) Classify each of the five variables measured as quantitative or qualitative.

Solution

(a) Because the measurements are made for each fish captured in the Tennessee River and its tributaries, the experimental units are the 144 captured fish.

(b) The variables upstream that capture location, length, weight, and DDT concentration are quantitative because each is measured on a natural numerical scale: upstream in miles from the mouth of the river, length in centimeters, weight in grams, and DDT in parts per million. In contrast, river/creek and species cannot be measured quantitatively; they can only be classified into categories (e.g., channel catfish, largemouth bass, and smallmouth buffalofish for species). Consequently, data on river/creek and species are qualitative. ■

FISHDDT

Table 1.1 Data collected by U.S. Army Corps of Engineers (selected observations)

River/Creek	Upstream	Species	Length	Weight	DDT
FLINT	5	CHANNELCATFISH	42.5	732	10.00
FLINT	5	CHANNELCATFISH	44.0	795	16.00
SPRING	1	CHANNELCATFISH	44.5	1133	2.60
TENNESSEE	275	CHANNELCATFISH	48.0	986	8.40
TENNESSEE	275	CHANNELCATFISH	45.0	1023	15.00
TENNESSEE	280	SMALLMOUTHBUFF	49.0	1763	4.50
TENNESSEE	280	SMALLMOUTHBUFF	46.0	1459	4.20
TENNESSEE	285	LARGEMOUTHBASS	25.0	544	0.11
TENNESSEE	285	LARGEMOUTHBASS	23.0	393	0.22
TENNESSEE	285	LARGEMOUTHBASS	28.0	733	0.80

1.1 Exercises

1.1 College application data. Colleges and universities are requiring an increasing amount of information about applicants before making acceptance and financial aid decisions. Classify each of the following types of data required on a college application as quantitative or qualitative.

(a) High school GPA
(b) Country of citizenship
(c) Applicant's score on the SAT or ACT
(d) Gender of applicant
(e) Parents' income
(f) Age of applicant

1.2 Fuel Economy Guide. The data in the accompanying table were obtained from the *Model Year 2018 Fuel Economy Guide* for new automobiles.

(a) Identify the experimental units.
(b) State whether each of the variables measured is quantitative or qualitative.

1.3 Performance-based logistics. In industry, performance based logistics (PBL) strategies are increasingly popular ways to reduce cost, increase revenue, and attain customer satisfaction. The *Journal of Business Logistics* (Vol. 36, 2015) used the opinions of a sample of 17 upper-level employees of the U.S. Department of Defense and its suppliers to determine the factors that lead to successful PBL projects. The current position (e.g., vice president, manager [mgr.]), type of organization (commercial or government), and years of experience were measured for each employee interviewed. These data are listed on the next page. Identify each variable measured as producing quantitative or qualitative data.

MODEL NAME	MFG	TRANSMISSION TYPE	ENGINE SIZE (LITERS)	NUMBER OF CYLINDERS	EST. CITY MILEAGE (MPG)	EST. HIGHWAY MILEAGE (MPG)
TSX	Acura	Automatic	2.4	4	23	33
Jetta	VW	Manual	1.4	4	28	40
M2	BMW	Manual	3.0	6	18	26
Fusion	Ford	Automatic	2.0	4	43	41
Camry	Toyota	Automatic	2.5	4	51	53
Escalade	Cadillac	Automatic	6.2	8	14	23

Source: Model Year 2018 Fuel Economy Guide, U.S. Dept. of Energy, U.S. Environmental Protection Agency (www.fueleconomy.gov).

PBL

INTERVIEWEE	POSITION	ORGANIZATION	EXPERIENCE (YEARS)
1	Vice president	Commercial	30
2	Post production	Government	15
3	Analyst	Commercial	10
4	Senior mgr.	Government	30
5	Support chief	Government	30
6	Specialist	Government	25
7	Senior analyst	Commercial	9
8	Division chief	Government	6
9	Item mgr.	Government	3
10	Senior mgr.	Government	20
11	MRO mgr.	Government	25
12	Logistics mgr.	Government	30
13	MRO mgr.	Commercial	10
14	MRO mgr.	Commercial	5
15	MRO mgr.	Commercial	10
16	Specialist	Government	20
17	Chief	Government	25

1.4 **Satellite Database.** The Union for Concerned Scientists (UCS) maintains the Satellite Database—a listing of the more than 1,000 operational satellites currently in orbit around Earth. Several of the many variables stored in the database include country of operator/owner, primary use (civil, commercial, government, or military), class of orbit (low Earth, medium Earth, or geosynchronous), longitudinal position (degrees), apogee (i.e., altitude farthest from Earth's center of mass, in kilometers), launch mass (kilograms), usable electric power (watts), and expected lifetime (years). Identify the experimental unit for these data. Which of the variables measured are qualitative? Which are quantitative?

1.5 **Medicinal value of plants.** Sea-buckthorn (*Hippophae*), a plant that typically grows at high altitudes in Europe and Asia, has been found to have medicinal value. The medicinal properties of berries collected from sea-buckthorn were investigated in *Academia Journal of Medicinal Plants* (August 2013). The following variables were measured for each plant sampled. Identify each as producing quantitative or qualitative data.

(a) Species of sea buckthorn (*H. rhamnoides, H. gyantsensis, H. neurocarpa, H. tibetana,* or *H. salicifolia*)

(b) Altitude of collection location (meters)

(c) Total flavonoid content in berries (milligrams per gram)

1.6 **Accounting and Machiavellianism.** *Behavioral Research in Accounting* (January 2008) published a study of Machiavellian traits in accountants. *Machiavellian* describes negative character traits that include manipulation, cunning, duplicity, deception, and bad faith. A questionnaire was administered to a random sample of 700 accounting alumni of a large southwestern university. Several variables were measured, including age, gender, level of education, income, job satisfaction score, and Machiavellian ("Mach") rating score. What type of data (quantitative or qualitative) is produced by each of the variables measured?

1.2 Populations, Samples, and Random Sampling

When you examine a data set in the course of your study, you will be doing so because the data characterize a group of experimental units of interest to you. In statistics, the data set that is collected for all experimental units of interest is called a **population**. This data set, which is typically large, either exists in fact or is part of an ongoing operation and hence is conceptual. Some examples of statistical populations are given in Table 1.2.

Table 1.2 Some typical populations

Variable	Experimental Units	Population Data Set	Type
a. Starting salary of a graduating Ph.D. biologist	All Ph.D. biologists graduating this year	Set of starting salaries of all Ph.D. biologists who graduated this year	Existing
b. Breaking strength of water pipe in Philadelphia	All water pipe sections in Philadelphia	Set of breakage rates for all water pipe sections in Philadelphia	Existing
c. Quality of an item produced on an assembly line	All manufactured items	Set of quality measurements for all items manufactured over the recent past and in the future	Part existing, part conceptual
d. Sanitation inspection level of a cruise ship	All cruise ships	Set of sanitation inspection levels for all cruise ships	Existing

Definition 1.6 A **population data set** is a collection (or set) of data measured on all experimental units of interest to you.

Many populations are too large to measure (because of time and cost); others cannot be measured because they are partly conceptual, such as the set of quality measurements (population c in Table 1.2). Thus, we are often required to select a subset of values from a population and to make **inferences** about the population based on information contained in a **sample**. This is one of the major objectives of modern statistics.

Definition 1.7 A **sample** is a subset of data selected from a population.

Definition 1.8 A **statistical inference** is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

Example 1.2

According to the research firm TVNewser (May 2017), the average age of viewers of CNN news programming is 60 years. Suppose a FOX network executive hypothesizes that the average age of FOX News viewers is greater than 60. To test her hypothesis, she samples 500 FOX News viewers and determines the age of each.

(a) Describe the population.

(b) Describe the variable of interest.

(c) Describe the sample.

(d) Describe the inference.

Solution

(a) The population is the set of units of interest to the FOX executive, which is the set of all FOX News viewers.

(b) The age (in years) of each viewer is the variable of interest.

(c) The sample must be a subset of the population. In this case, it is the 500 FOX News viewers selected by the executive.

(d) The inference of interest involves the *generalization* of the information contained in the sample of 500 viewers to the population of all FOX News viewers. In particular, the executive wants to estimate the average age of the viewers in order to determine whether it is greater than 60 years. She might accomplish this by calculating the average age in the sample and using the sample average to estimate the population average. ■

Whenever we make an inference about a population using sample information, we introduce an element of uncertainty into our inference. Consequently, it is important to report the **reliability** of each inference we make. Typically, this is accomplished by using a probability statement that gives us a high level of confidence that the inference is true. In Example 1.2, we could support the inference about the average age of all FOX News viewers by stating that the population average falls within 2 years of the calculated sample average with "95% confidence." (Throughout the text, we demonstrate how to obtain this measure of reliability—and its meaning—for each inference we make.)

Definition 1.9 A **measure of reliability** is a statement (usually quantified with a probability value) about the degree of uncertainty associated with a statistical inference.

The level of confidence we have in our inference, however, will depend on how **representative** our sample is of the population. Consequently, the sampling procedure plays an important role in statistical inference.

Definition 1.10 A **representative sample** exhibits characteristics typical of those possessed by the population.

The most common type of sampling procedure is one that gives every different sample of fixed size in the population an equal probability (chance) of selection. Such a sample—called a **random sample**—is likely to be representative of the population.

Definition 1.11 A **random sample** of n experimental units is one selected from the population in such a way that every different sample of size n has an equal probability (chance) of selection.

How can a random sample be generated? If the population is not too large, each observation may be recorded on a piece of paper and placed in a suitable container. After the collection of papers is thoroughly mixed, the researcher can remove n pieces of paper from the container; the elements named on these n pieces of paper are the ones to be included in the sample. Lottery officials utilize such a technique in generating the winning numbers for Florida's weekly 6/52 Lotto game. Fifty-two white ping-pong balls (the population), each identified from 1 to 52 in black numerals, are placed into a clear plastic drum and mixed by blowing air into the container. The ping-pong balls bounce at random until a total of six balls "pop" into a tube attached to the drum. The numbers on the six balls (the random sample) are the winning Lotto numbers.

This method of random sampling is fairly easy to implement if the population is relatively small. It is not feasible, however, when the population consists of a large number of observations. Since it is also very difficult to achieve a thorough mixing, the procedure only approximates random sampling. Most scientific studies, however, rely on computer software (with built-in random-number generators) to automatically generate the random sample. Almost all of the popular statistical software packages available (e.g., SAS, SPSS, MINITAB and R) have procedures for generating random samples.

1.2 Exercises

1.7 Guilt in decision making. The effect of guilt emotion on how a decision-maker focuses on the problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 155 volunteer students participated in the experiment, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading/writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers concluded that a higher proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups.

(a) Identify the population, sample, and variables measured for this study.

(b) What inference was made by the researcher?

1.8 Jamming attacks on wireless networks. Terrorists often use wireless networks to communicate. To disrupt these communications, the U.S. military uses jamming attacks on the wireless networks. The *International Journal of Production Economics* (Vol. 172, 2016) described a study of 80 such jamming attacks. The configuration of the wireless network attacked was determined in each case. Configuration consists of network type (WLAN, WSN, or AHN) and number of channels (single- or multi-channel).

(a) Suppose the 80 jamming attacks represent all jamming attacks by the U.S. military over the past several years, and these attacks are the only attacks of interest to the researchers. Do the data associated with these 80 attacks represent a population or a sample? Explain.

(b) The 80 jamming attacks actually represent a sample. Describe the population for which this sample is representative.

1.9 Can money spent on gifts buy love? Is the gift you purchased for that special someone really appreciated? This was the question investigated in the *Journal of Experimental Social Psychology* (Vol. 45, 2009). Researchers examined the link between engagement ring price (dollars) and level of appreciation of the recipient (measured on a 7-point scale, where 1 = "not at all" and 7 = "to a great extent"). Participants for the study had used a popular website for engaged couples. The website's directory was searched for those with "average" American names (e.g., "John Smith," "Sara Jones"). These individuals were then invited to participate in an online survey in exchange for a $10 gift certificate each. Of the respondents, those who paid really high or really low prices for the ring were excluded, leaving a sample size of 33 respondents.

(a) Identify the experimental units for this study.

(b) What are the variables of interest? Are they quantitative or qualitative in nature?

(c) Describe the population of interest.

(d) Do you believe the sample of 33 respondents is representative of the population? Explain.

1.10 Gallup Youth Poll. A Gallup Youth Poll was conducted to determine the topics that teenagers most want to discuss with their parents. The findings show that 46% would like more discussion about the family's financial situation, 37% would like to talk about school, and 30% would like to talk about religion. The survey was based on a national sampling of 505 teenagers, selected at random from all U.S. teenagers.

(a) Describe the sample.

(b) Describe the population from which the sample was selected.

(c) Is the sample representative of the population?
(d) What is the variable of interest?
(e) How is the inference expressed?
(f) Newspaper accounts of most polls usually give a *margin of error* (e.g., plus or minus 3%) for the survey result. What is the purpose of the margin of error and what is its interpretation?

1.11 **STEM experiences for girls.** The National Science Foundation (NSF) promotes girls' participation in informal science, technology, engineering or mathematics (STEM) programs. What has been the impact of these informal STEM experiences? This was the question of interest in the published study *Cascading Influences: Long-Term Impacts of Informal STEM Experiences for Girls* (March 2013). A sample of 159 young women who recently participated in a STEM program were recruited to complete an online survey. Of these, only 27% felt that participation in the STEM program increased their interest in science.

(a) Identify the population of interest to the researchers.
(b) Identify the sample.
(c) Use the information in the study to make an inference about the relevant population.

1.12 **Accounting and Machiavellianism.** Refer to the *Behavioral Research in Accounting* (January 2008) study of Machiavellian traits in accountants, Exercise 1.6 (p. 4). Recall that a questionnaire was administered to a random sample of 700 accounting alumni of a large southwestern university; however, due to nonresponse and incomplete answers, only 198 questionnaires could be analyzed. Based on this information, the researchers concluded that Machiavellian behavior is not required to achieve success in the accounting profession.

(a) What is the population of interest to the researcher?
(b) Identify the sample.
(c) What inference was made by the researcher?
(d) How might the nonresponses impact the inference?

1.3 Describing Qualitative Data

Consider a study of aphasia published in the *Journal of Communication Disorders*. Aphasia is the "impairment or loss of the faculty of using or understanding spoken or written language." Three types of aphasia have been identified by researchers: Broca's, conduction, and anomic. They wanted to determine whether one type of aphasia occurs more often than any other, and, if so, how often. Consequently, they measured aphasia type for a sample of 22 adult aphasiacs. Table 1.3 gives the type of aphasia diagnosed for each aphasiac in the sample.

For this study, the variable of interest, aphasia type, is qualitative in nature. Qualitative data are nonnumerical in nature; thus, the value of a qualitative variable can only be classified into categories called *classes*. The possible aphasia types—Broca's, conduction, and anomic—represent the classes for this qualitative variable. We can summarize such data numerically in two ways: (1) by computing the *class frequency*—the number of observations in the data set that fall into each class; or (2) by computing the *class relative frequency*—the proportion of the total number of observations falling into each class.

Definition 1.12 A **class** is one of the categories into which qualitative data can be classified.

Definition 1.13 The **class frequency** is the number of observations in the data set falling in a particular class.

Definition 1.14 The **class relative frequency** is the class frequency divided by the total number of observations in the data set, i.e.,

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

APHASIA

Table 1.3 Data on 22 adult aphasiacs

Subject	Type of Aphasia
1	Broca's
2	Anomic
3	Anomic
4	Conduction
5	Broca's
6	Conduction
7	Conduction
8	Anomic
9	Conduction
10	Anomic
11	Conduction
12	Broca's
13	Anomic
14	Broca's
15	Anomic
16	Anomic
17	Anomic
18	Conduction
19	Broca's
20	Anomic
21	Conduction
22	Anomic

Source: Reprinted from *Journal of Communication Disorders*, March 1995, Vol. 28, No. 1, E. C. Li, S. E. Williams, and R. D. Volpe, "The effects of topic and listener familiarity of discourse variables in procedural and narrative discourse tasks," p. 44 (Table 1) Copyright © 1995, with permission from Elsevier.

Examining Table 1.3, we observe that 5 aphasiacs in the study were diagnosed as suffering from Broca's aphasia, 7 from conduction aphasia, and 10 from anomic aphasia. These numbers—5, 7, and 10—represent the class frequencies for the three classes and are shown in the summary table, Table 1.4.

Table 1.4 also gives the relative frequency of each of the three aphasia classes. From Definition 1.14, we know that we calculate the relative frequency by dividing the class frequency by the total number of observations in the data set. Thus, the relative frequencies for the three types of aphasia are

$$\begin{aligned} \text{Broca's:} \quad & \frac{5}{22} = .227 \\ \text{Conduction:} \quad & \frac{7}{22} = .318 \\ \text{Anomic:} \quad & \frac{10}{22} = .455 \end{aligned}$$

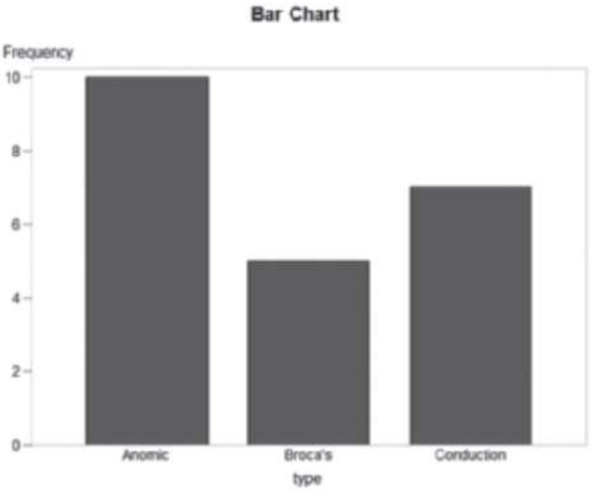
From these relative frequencies we observe that nearly half (45.5%) of the 22 subjects in the study are suffering from anomic aphasia.

Although the summary table in Table 1.4 adequately describes the data in Table 1.3, we often want a graphical presentation as well. Figures 1.1 and 1.2 show two of the most widely used graphical methods for describing qualitative data—bar graphs and pie charts. Figure 1.1 shows the frequencies of aphasia types in a **bar graph** produced with SAS. Note that the height of the rectangle, or "bar," over each class is equal to the class frequency. (Optionally, the bar heights can be proportional to class relative frequencies.)

Table 1.4 Summary table for data on 22 adult aphasiacs

Class	Frequency	Relative Frequency
(Type of Aphasia)	(Number of Subjects)	(Proportion)
Broca's	5	.227
Conduction	7	.318
Anomic	10	.455
Totals	22	1.000

Figure 1.1 SAS bar graph for data on 22 aphasiacs



In contrast, Figure 1.2 shows the relative frequencies of the three types of aphasia in a **pie chart** generated with SPSS. Note that the pie is a circle (spanning 360°) and the size (angle) of the "pie slice" assigned to each class is proportional to the class relative frequency. For example, the slice assigned to anomic aphasia is 45.5% of 360°, or $(.455)(360^\circ) = 163.8^\circ$.

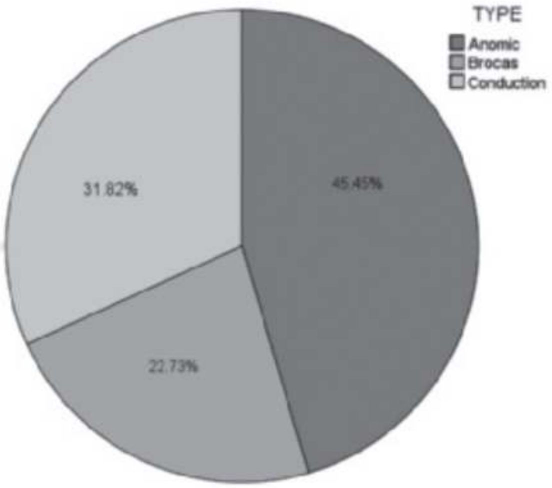


Figure 1.2 SPSS pie chart for data on 22 aphasiacs

1.3 Exercises

1.13 Estimating the rhino population. The International Rhino Federation estimates there are nearly 29,000 rhinoceroses living in the wild in Africa and Asia. A breakdown of the approximate number of rhinos of each species is reported in the accompanying table.

RHINO SPECIES	POPULATION ESTIMATE (ROUNDED)
African Black	5,000
African White	20,000
(Asian) Sumatran	100
(Asian) Javan	60
(Asian) Greater One-Horned	3,500
Total	28,660

Source: International Rhino Federation, 2018 (https://rhinos.org).

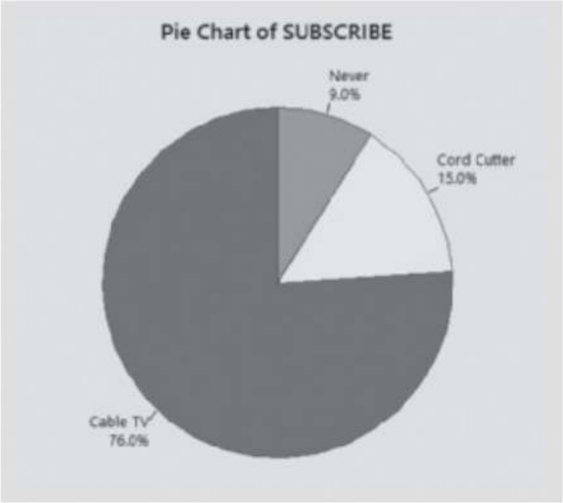
(a) Construct a relative frequency table for the data.

(b) Display the relative frequencies in a bar graph.

(c) What proportion of the 28,660 rhinos are African rhinos? Asian? Construct a pie chart to illustrate these proportions.

CABLETV

1.14 Cable TV subscriptions and "cord cutters." Has the increasing popularity of smartphones and video streaming over the internet impacted cable and satellite TV subscriptions? This was one of the questions of interest in a recent Pew Research Center survey (December 2015). Telephone (both landline and cell phone) interviews were conducted on a representative sample of 2,001 adults living in the United States. For this sample, 1,521 adults reported that they received cable or satellite TV service at home, 180 revealed that they never subscribed to cable/satellite TV service at home, and the remainder (300 adults) indicated that they were "cord cutters," i.e., they canceled the cable/satellite TV service. The results are summarized in the MINITAB pie chart shown.



(a) According to the pie chart, what proportion of the sample have a cable/satellite TV subscription at home? Verify the accuracy of this proportion using the survey results.

(b) Now consider only the 1,821 adults in the sample that have at one time or another subscribed to cable/satellite TV service. Create a graph that compares the proportions of these adults who currently subscribe to cable/satellite TV service with the proportion who are "cord cutters."

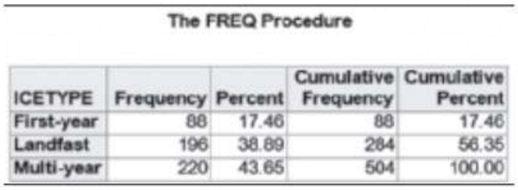
1.15 Motivation and right-oriented bias. Evolutionary theory suggests that motivated decision-makers tend to exhibit a right-oriented bias. (For example, if presented with two equally valued brands of detergent on a supermarket shelf, consumers are more likely to choose the brand on the right.) In *Psychological Science* (November 2011), researchers tested this theory using data on all penalty shots attempted in World Cup soccer matches (a total of 204 penalty shots). The researchers believed that goalkeepers, motivated to make a penalty-shot save but with little time to make a decision, would tend to dive to the right. The results of the study (percentages of dives to the left, middle, or right) are provided in the table. Note that the percentages in each row, corresponding to a certain match situation, add to 100%. Use graphs to illustrate the distribution of dives for the three match situations. What inferences can you draw from the graphs?

MATCH SITUATION	DIVE LEFT	STAY MIDDLE	DIVE RIGHT
Team behind	29%	0%	71%
Tied	48%	3%	49%
Team ahead	51%	1%	48%

Source: Based on M. Roskes et al., "The Right Side? Under Time Pressure, Approach Motivation Leads to Right-Oriented Bias," *Psychological Science*, Vol. 22, No. 11, November 2011 (adapted from Figure 2).

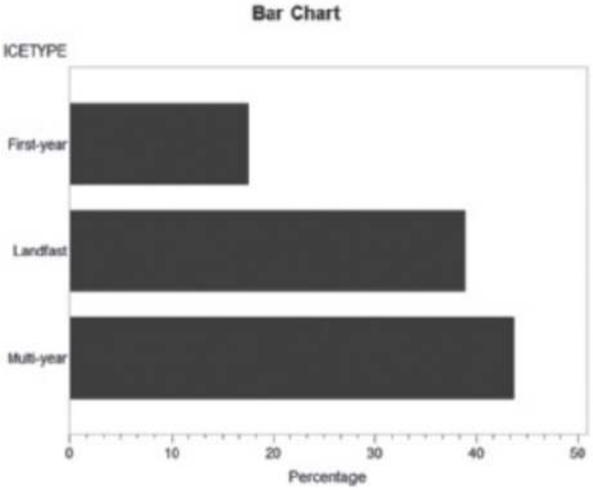
PONDICE

1.16 Characteristics of ice melt ponds. The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice melt ponds in the Canadian arctic. Environmental engineers at the University of Colorado are using these data to study how climate impacts the sea ice. Data for 504 ice melt ponds located in the Barrow Strait in the Canadian arctic are saved in the PONDICE file. One variable of interest is the type of ice observed for each pond. Ice type is classified as first-year ice, multi-year ice, or landfast ice. A SAS summary table and horizontal bar graph that describe the ice types of the 504 melt ponds are shown below.



The FREQ Procedure

ICETYPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
First-year	88	17.46	88	17.46
Landfast	196	38.89	284	56.35
Multi-year	220	43.65	504	100.00



(a) Of the 504 melt ponds, what proportion had landfast ice?

(b) The University of Colorado researchers estimated that about 17% of melt ponds in the Canadian arctic have first-year ice. Do you agree?

(c) Interpret the horizontal bar graph.

1.17 Groundwater contamination in wells. In New Hampshire, about half the counties mandate the use of reformulated gasoline. This has led to an increase in the contamination of groundwater with methyl *tert*-butyl ether (MTBE). *Environmental Science and Technology* (January 2005) reported on the factors related to MTBE contamination in private and public New Hampshire wells. Data were collected for a sample of 223 wells. These data