



SECOND EDITION

Visualizing **HEALTH CARE STATISTICS**

A Data-Mining Approach

Zada T. Wicker
J. Burton Browning

SECOND EDITION

Visualizing **HEALTH CARE STATISTICS**

A Data-Mining Approach

Zada T. Wicker, MBA, RHIT,
CCS, CCS-P

Professor Health Information Technology
Sullivan University

Dr. J. Burton Browning, EdD

Retired Professor
Author and Entrepreneur



JONES & BARTLETT
LEARNING



World Headquarters
Jones & Bartlett Learning
5 Wall Street
Burlington, MA 01803
978-443-5000
info@jblearning.com
www.jblearning.com

Jones & Bartlett Learning books and products are available through most bookstores and online booksellers. To contact Jones & Bartlett Learning directly, call 800-832-0034, fax 978-443-8000, or visit our website, www.jblearning.com.

Substantial discounts on bulk quantities of Jones & Bartlett Learning publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones & Bartlett Learning via the above contact information or send an email to specialsales@jblearning.com.

Copyright © 2021 by Jones & Bartlett Learning, LLC, an Ascend Learning Company

All rights reserved. No part of the material protected by this copyright may be reproduced or utilized in any form, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the copyright owner.

The content, statements, views, and opinions herein are the sole expression of the respective authors and not that of Jones & Bartlett Learning, LLC. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not constitute or imply its endorsement or recommendation by Jones & Bartlett Learning, LLC and such reference shall not be used for advertising or product endorsement purposes. All trademarks displayed are the trademarks of the parties noted herein. *Visualizing Health Care Statistics: A Data-Mining Approach, Second Edition* is an independent publication and has not been authorized, sponsored, or otherwise approved by the owners of the trademarks or service marks referenced in this product.

There may be images in this book that feature models; these models do not necessarily endorse, represent, or participate in the activities represented in the images. Any screenshots in this product are for educational and instructive purposes only. Any individuals and scenarios featured in the case studies throughout this product may be real or fictitious, but are used for instructional purposes only.

This publication is designed to provide accurate and authoritative information in regard to the Subject Matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the service of a competent professional person should be sought.

21529-8

Production Credits

VP, Product Management: Amanda Martin
Director of Product Management: Laura Pagluica
Product Manager: Sophie Fleck Teague
Content Strategist: Tess Sackmann
Manager, Project Management: Lori Mortimer
Project Specialist: Kathryn Leeber
Digital Project Specialist: Angela Dooley
Senior Marketing Manager: Susanne Walker
Production Services Manager: Colleen Lamy

VP, Manufacturing and Inventory Control: Therese Connell
Composition: Exela Technologies
Cover Design: Michael O'Donnell
Text Design: Kristin E. Parker
Senior Media Development Editor: Troy Liston
Rights Specialist: Maria Leon Maimone
Cover Image (Title Page, Chapter Opener):
© Mad Dog/Shutterstock
Printing and Binding: McNaughton & Gunn

Library of Congress Cataloging-in-Publication Data

Names: Browning, J. Burton, author. | Wicker, Zada T., author.
Title: Visualizing health care statistics : a data-mining approach / J. Burton Browning, Zada T. Wicker.
Description: Second edition. | Burlington, MA : Jones & Bartlett Learning, [2021] | Includes bibliographical references and index.
Identifiers: LCCN 2020007462 | ISBN 9781284197525 (paperback)
Subjects: MESH: Data Mining | Data Interpretation, Statistical | Data Collection | Health Information Management
Classification: LCC R859.7.D35 | NLM W 26.55.I4 | DDC 610.285—dc23
LC record available at <https://lcn.loc.gov/2020007462>

6048

Printed in the United States of America
24 23 22 21 20 10 9 8 7 6 5 4 3 2 1

We wish to thank our parents: Coleen Grower, who passed in April 2019, and Betty Browning. You both taught us to give back to the world and leave things better than we found them. As such, we knew there was a need for an engaging and real-world healthcare statistics textbook to help our students, and this work is the result of that multi-year effort. The effort involved took time away from you both, but we tried to always be available when needed. Your enduring support and encouragement have given us the strength to not only finish the original edition, but revise a second edition.

-J. Burton Browning

As a sincere thought from a daughter-in-law: Betty, thank you for accepting me into the family and being there when I needed you. As a final word to my mother Coleen, thank you for showing me the world and helping me to become the person I am today. As a single mother, you are still my inspiration and have given me the strength and ability I now am trying to impart to others.

-Zada T. Wicker

Brief Contents

Foreword	ix
Preface	x
Reviewers	xi
Acknowledgments	xiii

CHAPTER 1	Introduction to Healthcare Statistics	1
CHAPTER 2	Central Tendency, Variance, and Variability	14
CHAPTER 3	Patient Data	29
CHAPTER 4	Occupancy and Utilization Data	55
CHAPTER 5	Morbidity and Mortality Data	79
CHAPTER 6	Autopsy Data	108
CHAPTER 7	Infection, Consultation, and Other Data	130
CHAPTER 8	Health Information Management Statistics	152
CHAPTER 9	Research Methodology and Ethics	170
CHAPTER 10	Data Collection and Reporting Methods	189
CHAPTER 11	The Future of Healthcare Statistics	210
APPENDIX A	Common Statistical Abbreviations Used in This Text	229
APPENDIX B	Resources for Further Information	230
APPENDIX C	Historical Abuse of Human Research Subjects	231
APPENDIX D	Formulas for Healthcare Statistics	234
APPENDIX E	CAHIIM Competency Exercises	237

Glossary	240
Index	249

Contents

Foreword	ix
Preface	x
Reviewers	xi
Acknowledgments	xiii

CHAPTER 1 Introduction to Healthcare Statistics 1

Introduction	2
History and Rationale of Healthcare Statistics	2
Definition of Statistics	3
The Use of Statistics in Health Care	4
Key Producers and Users of Healthcare Statistics	4
Data Mining	5
Definition	5
History	6
Current Applications	6
Basic Statistical Concepts	7
Dataset	7
Variables	8
Data Distribution	8
Types of Data	8
Types of Data Mining Models	9
Predictive Models	9
Descriptive Models	9
Decision Models	9
Obtaining Data	10
Global Perspective	11
Chapter Summary	11
Apply Your Knowledge	12
References	12
Web Links	12

CHAPTER 2 Central Tendency, Variance, and Variability 14

Introduction	15
Measures of Central Tendency	15

Mean	15
Median	17
Mode	17
Frequency Distribution	18
Variance and Measures of Dispersion or Variability	19
Min and Max	19
Range	20
Outlier Data	21
Interquartile Range	21
Standard Deviation	22
Variance	22
Data Harvesting	25
Global Perspective	26
Chapter Summary	26
Apply Your Knowledge	26
References	28
Web Links	28

CHAPTER 3 Patient Data 29

Introduction	30
Census Data and Their Importance	31
Calculation and Reporting of Patient Census Data	33
Inpatient Service Days	33
Average Daily Census	34
Data Visualization	35
Visually Examine Data With Sparklines and Microcharts	38
Newborn Services	41
Open-Source Software	41
Freeware and Shareware	42
Types of Databases	44
Flat-File Database	44
Relational Database	45
Data Formats	45
R-Project	46
Data Stored in R	48
Global Perspective	52

Chapter Summary	53
Apply Your Knowledge	53
References	54
Web Links	54

CHAPTER 4 Occupancy and Utilization Data55

Introduction	56
Bed Count Computation	56
Definition of Inpatient Bed Count	56
Importance of Inpatient Bed Count	57
Bed Occupancy Ratios	57
Certificate of Need	58
Calculating Newborn Bassinet Occupancy Ratio	58
Bed Turnover Rate	59
Two Formulas for Bed Turnover	59
Length of Stay	59
Discharge Days and How to Calculate Them	60
Total Length of Stay	61
Average Length of Stay	61
Median and Standard Deviation for Length of Stay	66
Median Length of Stay	66
Why Might You Use Median Instead of Mean?	66
Standard Deviation of Length of Stay	67
Visually Representing Data	68
PowerPoint Presentation	69
Data Mining: Association Rules With R-Project	70
Global Perspective	74
Chapter Summary	77
Apply Your Knowledge	77
References	78
Web Links	78

CHAPTER 5 Morbidity and Mortality Data79

Introduction	80
Morbidity Rates	81
Incidence	81
Prevalence	82
Mortality Rates	83
Gross Mortality Rate	83
Fatality Rate	84
Net Mortality Rate	86

Postoperative Mortality Rate	87
Maternal Mortality Rate	89
Maternal Mortality Rate as the Number of Deaths per 100,000 or 10,000 Births	90
Anesthesia Mortality Rate	91
Newborn Mortality Rate	92
Fetal Mortality Rate	92
Mortality Rates for Cancer	94
Mortality-Adjusted Rates	94
Interpreting Mortality Rate Results	95
Conducting Formal Research	95
Research Design	95
The Hypothesis and Null Hypothesis Statements	96
Statistical Measures	96
P Values and Significance	96
Type I Errors	97
Type II Errors	97
Either End of the Curve: Tails	97
The Normal Distribution of Data	98
Parametric and Nonparametric Tests	98
z Score	98
t Test	100
Infant Mortality by Race and County	104
Global Perspective	104
Chapter Summary	104
Apply Your Knowledge	104
References	107
Web Links	107

CHAPTER 6 Autopsy Data108

Introduction	109
What Is an Autopsy and Why Is It Important?	109
Centers for Disease Control and Prevention Data	112
Types of Autopsy Data	112
Autopsy Rate	112
Net Autopsy Rates	113
Inpatient Hospital Autopsies	114
Adjusted Hospital Autopsy Rate	114
Autopsy Rate for Newborns	115
Fetal Autopsy Rate	116
Statistical Measures	116
F Test: Comparison of Two Variances	116
Analysis of Variance	119
One-Way Analysis of Variance	119
Two-way Analysis of Variance	119

Global Perspective	126
Chapter Summary	126
Apply Your Knowledge	126
References	129
Web Links	129

CHAPTER 7 Infection, Consultation, and Other Data . . . 130

Introduction	131
Infection Rates	131
Infection Control Committee	131
Nosocomial Infection Rate	132
Specific Infection Rate	133
Postoperative Infection Rate	134
Complication Rates	136
Cesarean Section Rate	137
Consultation Rates	138
General Occurrence Rates	139
Statistical Measures and Tools	140
R Commander Graphical User Interface	140
Post-Hoc Analysis of Variance	144
Tukey Honest Significant Difference	144
HAI Reporting and Tracking	145
Text Mining and Visualization Using R-Project and Wordle	146
Global Perspective	149
Chapter Summary	149
Apply Your Knowledge	149
References	150
Web Links	151

CHAPTER 8 Health Information Management Statistics 152

Introduction	153
Functions of Health Information Management	153
Requirements to Work in Health Information Management	153
Labor Cost and Compensation	154
Transcription Cost and New Technology	155
Other Costs Associated With Health Information Management	157
Productivity	159
Healthcare Facility Staffing	159
Measuring Utilization	161

Types of Financial Reports	161
Readmission Rate Reports	161
Discharge Reports	162
Two Types of Budgets: Operational and Capital	162
Operational Budget	162
Capital Budget	163
Data Mining With Naive Bayes and R-Project	166
Global Perspective	168
Chapter Summary	168
Apply Your Knowledge	168
References	169
Web Links	169

CHAPTER 9 Research Methodology and Ethics 170

Introduction	171
Types of Research	172
Research Process	172
Step 1: Identify the Problem	172
Step 2: Research the Problem	172
Step 3: Develop Research Questions	172
Step 4: Determine the Type of Data Needed, Sample Size, and Methods of Analysis	172
Step 5: Collect Data	173
Step 6: Analyze the Data	173
Step 7: Draw a Conclusion	173
Step 8: Report the Results and Implications for Further Research	173
Research Ethics and the Abuse of Human Subjects	173
Late 1700s: Edward Jenner and the First Smallpox Vaccine	173
1850s: J. Marion Sims, the Father of Gynecology	174
1900: Walter Reed and Yellow Fever Transmission	174
1932 to 1972: Tuskegee Study of Untreated Syphilis	175
1964: The Declaration of Helsinki	175
1979: The Belmont Report	175
The Institutional Review Board	176
Review Process	176
Exemption and Types of Review	176
Informed Consent	177
Membership	177
Data Dictionary	177

Statistical Measures and Tools	178
Common Nonparametric Statistics	178
Regression Analysis: Simple Linear Regression	181
Global Perspective	186
Chapter Summary	186
Apply Your Knowledge	187
References	187
Web Links	188

CHAPTER 10 Data Collection and Reporting Methods 189

Introduction	190
Descriptive Research and Information Collection	191
Sampling	191
Nonprobability Sampling	191
Probability Sampling	192
Random Numbers and Random Sampling	193
Quality Question Design for Data Collection	196
Bloom's Taxonomy and Question Design	196
Guidelines for Question Writing	196
Types of Questions	197
Types of Studies	199
Longitudinal Study	199
Case Study	200
Documentary Study	200
Data Collection Methods	200
Survey	200
Interview	201
Questionnaire	201
Observation	201
Appraisal	201
Survey Tools	201
Pivot Tables	202
Multiple Regression Analysis	205
Global Perspective	208
Chapter Summary	208
Apply Your Knowledge	208
References	209
Web Links	209

CHAPTER 11 The Future of Healthcare Statistics 210

Introduction	211
The Future of Healthcare Statistics	211
Radio Frequency Identification	211
Automatic Medication Dispenser	212
Health Information Exchange	212
Efforts to Decrease Healthcare Costs	213
Time Series Analysis of Data	213
Forecasting Future Data	215
Project Management General Concepts	217
Analysis of Covariance	220
Coronavirus—The Next Pandemic	224
Ebola	225
Global Perspective	225
Chapter Summary	225
Apply Your Knowledge	226
References	227
Web Links	228

APPENDIX A Common Statistical Abbreviations Used in This Text 229

APPENDIX B Resources for Further Information 230

APPENDIX C Historical Abuse of Human Research Subjects 231

APPENDIX D Formulas for Healthcare Statistics 234

APPENDIX E CAHIIM Competency Exercises 237

Glossary 240

Index 249

Foreword

As a career public health specialist working in a community college setting in a non-public health position, I am often asked to speak to students about public health-related topics. Moreover, as someone who specialized in the translation of data at a state public health level, I am in a position to understand the value of data and how it should be taught in an introductory course.

With this textbook, *Visualizing Health Care Statistics: A Data Mining Approach*, authors J. Burton Browning and Zada Wicker have captured complex topics in a way that makes understanding them much easier and more accessible to the student. Examples include their explanations of biostatistics data points and their relevance to the workplace and in general society across a range of issues impacting health data such as morbidity and mortality. Hopefully, students will see the benefit of these statistical concepts and basic

calculations to the health field in general, but will also understand the relevance of this kind of information to themselves personally. Each of us represents a part to a whole in a data point somewhere.

For some, this textbook can help build foundational skills to launch a career pathway in the future working with health statistics in a meaningful way. Browning has nurtured this project with Ms. Wicker as a coauthor to capitalize on her many years of experience in health information technology. Their combined 35-plus years of experience translate into a project that will help students cross the bridge from learning to understanding.

Pamela Federline, MPH
Director, Office of Planning, Research, and Effectiveness
College of the Albemarle Elizabeth City, NC

Preface

Too often, texts on statistics are filled with examples that are designed to teach concepts in a vacuum and not related to real data. More often than that, the critical step of applying statistical knowledge to real-world situations or global situations is missing. Use of massive amounts of data in a data mining context is so new that most texts ignore big data altogether. It is in that light that the authors have tried to create an introductory text that is real-world, helps students develop data-gathering skills, and all the while prepares them for a national Registered Health Information Technician (RHIT) exam, which has required healthcare statistical knowledge.

The authors have designed this textbook as an introductory healthcare statistics text for students in healthcare-related curricula, most specifically the health information technology (HIT) field. It is with the RHIT exam in mind that this text has been written. Yet it is of great value to any student in healthcare informatics or related fields. As such it is most appropriate for second-year students, although certainly other levels could benefit.

A key feature of this text is that it covers basic statistical methods to give students a solid grounding in the theoretical framework of statistical measures. Moreover, it provides all key statistical measures required on the RHIT exam and introduces students to the very important topic of data mining within a framework of harvesting real-world healthcare data via a global perspective. Each chapter reinforces the concept of data visualization, a key aspect of reporting data to end users in a meaningful way.

The chapters are sequenced such that each chapter builds on knowledge from the previous chapter. Thus, it is important to present chapters as sequenced in the text. In most instances, either commonly available commercial software (such as Microsoft Excel or a web browser) or open-source free software is used to minimize costs to both students and schools and to provide students with cutting-edge software.

Key pedagogical features of the text include the following:

- Chapter-opening quote
- Chapter outline
- Learning outcomes
- Chapter key terms list and end-of-book glossary
- Chapter introduction
- Data mining basic concepts
- Hands-on Statistics: A step-by-step numbered list of procedures for performing a statistics-related calculation or task either by hand or using a common computer application, such as Excel
- How Does Your Hospital Rate?: A progressive case study that starts near the beginning of each chapter, resumes in the middle of the chapter, and concludes at the end of the chapter. Each box includes critical-thinking questions
- Global Perspective: An exercise that allows the student to practice data mining from international data sources
- Apply Your Knowledge: End-of-chapter statistics calculation and research exercises
- Software information to highlight integrated software
- Calculation by hand exercises to highlight manual calculation opportunities
- Did You Know?: Sidebar boxes containing interesting information related to statistics and other topics
- Chapter summary
- Screenshots of integrated computer applications

To aid both instructors and students, the authors have provided an Instructor's Guide with ideas on how to present each chapter. PowerPoint presentations are available for each chapter as well. These ancillary resources are available via the Jones & Bartlett Learning website.

Reviewers

Karen Bakuzonis, PhD, MSHA, RHIA

Program Chair
Ashford University
San Diego, California

Lynda Carlson, PhD, MS, MPH

Professor and Program Director
Borough of Manhattan Community College
New York, New York

Cheryl Christopher, MPA, RHIA

Assistant Professor
Borough of Manhattan Community College
New York, New York

Charlotte E. Creason, RHIA

Adjunct Faculty
Tyler Junior College
Tyler, Texas

Kelly Davis, EdD, MSN, RN, CNE

RN to BSN Program Coordinator
Delaware Technical Community College
Georgetown, Delaware

Barbara Dimanlig, RHIA, CHPS

Health Information Technology
City College of San Francisco
San Francisco, California

Sandra Hertkorn

Health Information
Bryan College
Gold River, California

Deb Honstad, MA, RHIA

Program Director and Associate Professor
San Juan College
Farmington, New Mexico

Margo Imel, MBA/TM, RHIT

Instructor
Araphoe Community College
Littleton, Colorado

Mariann Davidson Jeffrey, RHIT

Health Careers
Central Arizona College
Apache Junction, Arizona

Christine Kowalski, EdD, RHIA, CP-EHR

Course Mentor
Western Governors University
Salt Lake City, Utah

Meg Kurek, MSIS, PMP

Workforce Development
Community College of Allegheny County
Monroeville, Pennsylvania

Maribeth S. Lane, RHIA

Program Director
Northwest Iowa Community College
Sheldon, Iowa

Cynthia Lundquist, BHA

Kaplan College
Stockton, California

**Barbara Marchelletta,
CMA (AAMA), RHIT, CPC, CPT**

Program Director
Beal College
Bangor, Maine

Vince Ochotorena, MBA, MAM, BS

Instructor
Anthem College
Phoenix, Arizona

Terri L. Randolph, MBA/HCM

Stratford University
Falls Church, Virginia

Patricia L. Shaw, EdD, RHIA, FAHIMA

Chair and Program Director
Weber State University
Ogden, Utah

Jody Smith, PhD, RHIA, FAHIMA

Professor Emeritus
St. Louis University
St. Louis, Missouri

Karen J. Smith, PhD, RHIA, FAHIMA

Health Informatics and Information
Management
Saint Louis University
St. Louis, Missouri

Jasper Xu, PhD, MHA, CHFP

Assistant Professor
University of North Florida
Jacksonville, Florida

Heather Watson, RHIA

Program Director
Davidson County Community College
Thomasville, North Carolina

Mary Worsley, MH, RHIA, CCS

Associate Professor
Miami Dade College
Miami, Florida

Acknowledgments

It is our pleasure to give a special thank you to the following individuals for their input on the second edition: Donna Estes, RHIT, HIMT Program Director at Georgia Northwestern Technical College in Rock Spring, Georgia, CPHQ-Retired, MPM, and Katrina Putman, RHIT.

Thank you for the kind comments and notes from the *First Edition* reviewers and the *Second Edition* reviewers. With reviewers' comments, we were able to make positive changes to an already popular text. Also, a special thank you to Tess Sackmann from Jones & Bartlett Learning, who helped orchestrate this text.

CHAPTER 1

Introduction to Healthcare Statistics

Far and away the best prize that life has to offer is the chance to work hard at work worth doing.

—Theodore Roosevelt

CHAPTER OUTLINE

Introduction	Types of Data
History and Rationale of Healthcare Statistics	Types of Data Mining Models
Definition of Statistics	Predictive Models
The Use of Statistics in Health Care	Descriptive Models
Key Producers and Users of Healthcare Statistics	Decision Models
Data Mining	Obtaining Data
Definition	Global Perspective
History	Chapter Summary
Current Applications	Apply Your Knowledge
Basic Statistical Concepts	References
Dataset	Web Links
Variables	
Data Distribution	

LEARNING OUTCOMES

After completing this chapter, you should be able to do the following:

1. Define and describe the history of statistics.
2. Describe how statistics are used in health care.
3. Identify common users of healthcare statistics.
4. Define data mining.
5. Describe the history of data mining and how it is used in health care.
6. Explain what a dataset is and how it is used.
7. Identify the four types of data.
8. Discuss five basic elements of data mining.

KEY TERMS

Aspect	Flowchart	Primary data
Data	Independent variable	Ratio data
Data mining	Interval data	Sample
Data-driven	Machine learning	Secondary data
Dataset	Nominal data	Statistics
Decision model	Ordinal data	Telehealth
Dependent variable	Parameters	Variable
Descriptive model	Population	Viewpoint
Evidence-based medicine (EBM)	Predictive model	

How Does Your Hospital Rate?

Paul, a 54-year-old man who has been diagnosed with congestive heart failure, is facing heart valve replacement surgery. He and his family would like the best surgeon and hospital available in his area for this major surgery. They are considering three different hospitals. One site for statistical information that Paul and his family might review would be the Health section of the *US News and World Report* website (<https://health.usnews.com>), specifically the information relating to cardiology. This website ranks the specialist, survival, patient safety, patient volume, and nursing staff.

Consider the following:

1. What statistics should Paul be considering as he makes this decision?
2. How would the hospitals in your area rate?
3. Which one would you pick, and why?
4. Why is it important for a hospital to keep healthcare statistics?

Introduction

Healthcare statistics allow a hospital to assess, improve, and communicate its quality of patient care, develop better policies and procedures for infection control, and achieve many other goals we will be discussing throughout the following chapters. In this chapter, we will introduce to you the history and definition of healthcare statistics and their importance to a hospital. We will also discuss organizations that keep statistics, not just here in the United States, but around the globe, such as the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and other groups. Keeping statistical data allows us to watch for trends in health care and make proactive rather than reactive choices.

Because the word data will be used so often throughout this text and can be used so many different ways in everyday life, it is worth taking a moment to define this term for our purposes. **Data** (singular, *datum*) are units of information, such as measurements, that can

be collected and interpreted. They are the commodity in which we deal as healthcare statisticians.

Data mining is another important concept that we will discuss. We will explain its use in the healthcare arena and show you how you can use data mining and benchmarks with your local hospital. We will also cover how to use Excel and R-Project, both powerful data analysis tools, to examine statistical methods.

Let's begin our exploration of healthcare statistics in the United States and around the globe.

History and Rationale of Healthcare Statistics

The history and study of statistics is as much an examination of historical events as it is a study of the probability, logic, and mathematics behind statistical analysis. As the famous mathematics educator Freudenthal and others have noted, learning the history of a topic often aids in the overall understanding

of the focus of a student's study (Leen, 1994). It is in this light that this text will discuss statistics in general and statistical analysis in health care.

There are differing views as to the first uses of statistical analysis, and in fact even the word *statistics*, but certainly most would agree that a large majority of the first uses were descriptive in nature, before the term statistics was formally used.

An early example can be found in Mackenzie's book of ancient stories from India. He notes that Rituparna estimated the number of leaves and berries on two branches of a fruit tree and estimated probabilities of dice rolls. With regard to the fruit tree, he estimated the number of leaves and berries on the basis of a twig, which he multiplied by the estimated number of twigs on the two branches. After a night of counting, he found that his estimate was very close to the real number. Most likely the use of two branches provided a way to take the count from each and determine an average, to be used in estimation for the entire tree (Mackenzie, 1913).

In 431 BCE, the author of *The History of the Peloponnesian War*, Thucydides, notes that the origins of probability can be found in the Athenians' evaluation of the height of the wall of Platae. This estimation was done by determining an average size for a brick, counting the number of bricks in an area, and multiplying by the area they were trying to attack to determine the height they would need for ladders to scale the walls. To determine the height, they multiplied the mode (most frequently occurring value of several sampled bricks) by the count of the number of bricks (Thucydides, n.d). As you will later learn, *mode* is the most frequently occurring value in a set of values. Since bricks were not necessarily uniform in size, determining the mode of a standard brick would be important if you were estimating the size of a wall of bricks you needed to scale.

Early accounts of the use of the undefined term *statistics* vary but include the ninth century work "A Manuscript on Deciphering Cryptographic Messages" written by Al-Kindi. It included a thorough account of how to use statistics and frequency analysis to decipher secret encrypted messages. As a formal term to describe the subject of this text, a version of the word first appears as *statistik* by German author Gottfried Achenwall in 1749 in describing data about the state or arithmetic of the state; however, earlier civilizations, such as the Romans, collected state demographic information and other related data earlier than 1749, though not necessarily using the term *statistic* (Johnson & Kotz, 2012). Unrelated to state data, other early recordings of statistical data

concerned sailing, temperature, astrology, and related data that were used for predictions. If you are familiar with the *Farmer's Almanac* or similar texts, then you are aware of the direct impact that predictive statistics can have on people and the state.

Today, statistics are used to express everything from temperature and demographic data averages to mortality rates for cardiac surgery in health care and pattern analysis in massive datasets. In a data-driven society, almost every aspect of life in some way has statistical factors associated with it, from your interest rate at the bank to the target heart rate you are trying to meet for a fitness plan. In short, we use statistics to make a great many important decisions, including those relating to finances, work, and, most specifically for the purposes of this text, health care.

DID YOU KNOW?

Florence Nightingale was a member of the Royal Statistical Society. As one of the first women to collect statistics on health policy, she led the way for other female statisticians to work in the field. She was also credited with using graphs to present her findings to Queen Victoria in an effort to reform the sanitary conditions in military hospitals, so she was also an early proponent of data visualization methods (Lancaster, 2013).

Now you know!

Definition of Statistics

Before we discuss in detail how statistics are used in health care, identify some of the users of statistics, and examine some of the statistical methods, we should formally define the term statistics as it will apply to our uses. According to Batten (1986), "Statistics is a series of methods to collect, analyze, and interpret masses of numerical data." However, statistics are not typically just an end unto themselves but rather are used to solve real-world problems. Thus, a practical definition of **statistics** might be the collection of data for the purpose of making predictions (inferences) or considerations to answer a question. Along with this definition are many strategies, rules, and procedures that define and formalize our collection and use of data and resultant findings. We will focus on health-care statistics, but the underlying methods and processes are applicable to other areas, such as research statistics, which is integral to the field of health care.

The Use of Statistics in Health Care

Statistics are frequently used by many healthcare organizations, including hospitals and insurance companies. Such organizations use statistics to aid in making beneficial business decisions based on data they have collected over time. Hospitals collect and summarize data to improve quality of care, analyze cost of patient care, measure utilization of services provided to patients, examine target marketing decisions, and improve potential offerings of services to patients.

For example, a healthcare professional might examine average length of stay at several area hospitals. On finding that one hospital had a considerably longer average length of stay for patients, the hospital administrators might look for some underlying cause. By finding the issue or issues and resolving them, the hospital's quality of service (QOS) would be enhanced.

Hospitals who are accredited are required to retain data from certain areas to maintain their accreditation standards. The main hospital accrediting body is The Joint Commission. As an example of what is required, a hospital might have to keep fetal monitor strips or complete patient records for 7 years on-site, with older data being held off-site in some archived form for long-term storage (perhaps converted to a different medium, such as microfiche, tape, or optical storage).

The federal government is pushing legislation to move forward with the electronic health record. The *Federal Register* contains all the rules and regulations regarding the implementation of the electronic health record. Included in this legislation are antikickback rules for physicians who refer a patient to a lab or other type of facility in which they have ownership. Rules and regulations are also set forth by the Department of Health and Human Services (DHHS) and the Centers for Medicare and Medicaid Services (CMS).

Third-party payers may require hospitals to collect and maintain performance data. Data can be collected and abstracted for these purposes using many different tools and methods. Of critical importance is how the data are analyzed and used. There are some very specific statistical methods that are used to analyze these data for reporting to external and internal consumers of the data, such as third-party payers and marketing data consumers, respectively.

Key Producers and Users of Healthcare Statistics

The Bureau of Labor Statistics, the National Center for Health Statistics (NCHS), the CDC, and CMS are just some of the agencies who produce and maintain healthcare statistics. Vital statistics are also kept in each state and are another source of statistical information, along with census information. These organizations provide and use statistics to improve health care, summarize findings, and examine trends in the United States and around the globe.

The Bureau of Labor Statistics is a unit of the US Department of Labor and serves as the principal agent for the US Federal Statistical System. The primary mission of the Bureau of Labor Statistics is to collect and analyze essential statistical data for use by the public and the US Congress. The most common statistics that are kept by the Bureau of Labor Statistics relate to prices, employment, unemployment, compensation for injuries, and work injuries. For example, according to the US Department of Labor agency, Occupational Safety and Health Administration (OSHA), in 2018 there were 1008 fatalities to workers who performed construction.

The NCHS is the principal agent that delivers statistical information and directs policies and actions that will improve the health of the public. The NCHS is housed within the CDC. The CDC and the NCHS compile statistics for all types of disease for the United States and worldwide. For example, in the United States, rates of the sexually transmitted disease gonorrhea increased by 8% from the year 2010 to 2011, totaling 321,849. By 2017, the number of new gonorrhea cases had climbed all the way to 555,608 (CDC, 2018).

The primary responsibility of CMS is to administer Medicare and Medicaid, the Children's Health Insurance Program, the Health Insurance Portability and Accountability Act (HIPAA), and Clinical Laboratory Improvement Amendments. CMS keeps statistics for each state as to how many people are on Part A or Part B or both for Medicare. As of July 2016, 56.5 million were on only Part A of Medicare, and 52.1 million were on Part B (National Committee to Preserve Social Security and Medicare, 2020). Keeping this type of statistic is vital to health care because people are living longer and will eventually require more healthcare services.

In addition to the organizations already discussed, users of statistical data include the following:

- Federal government agencies gather information that references public health issues such as HIV/AIDS, cancer, births, and deaths.

- Accreditation agencies use statistics to show the most common diagnoses and procedures and the amount of resources used to treat those patients.
- Managed care organizations use statistics to review costs for the level of care that is being provided to their patients.
- Healthcare researchers use the data from health law and regulations, physician practices, and **telehealth** (services that use electronic information and telecommunications technologies to support long-distance clinical health care). Technologies can include videoconferencing, the internet, store-and-forward imaging, streaming media, terrestrial communication, and wireless communication. There are many other types of information used for research.
- Mental health facilities and drug and alcohol facilities use this information to measure the success of the services being provided and success rates of patients.

Note that these parties are external to the health-care provider (e.g., the hospital). In fact, the hospital would be a primary consumer of this valuable data. Activities such as QOS, as mentioned earlier, are but one reason for this information. Another is governmental oversight via the certificate of need (CON), which is a statement issued by a government agency for projected construction or modification of a health-care facility. The facility must meet the requirements statistically to meet the CON criteria. It ensures that the new facility will be needed at the time of completion for those additional services. Basically, you might consider that a CON is an assurance that facilities are not built or expanded beyond the requirements of the community.

Hospitals and health information management organizations use statistical information as well. Both of these organizations typically use similar types of information in their statistical analysis. Let's examine the departments that would perform and use the various data and findings.

- Healthcare administrators use health statistics to make data-driven decisions. Think of **data-driven** as making decisions based on statistical information instead of by guessing. For example, if a hospital had data for 3 years on nursing needs of the emergency department on New Year's, you might be able to expand staffing on that day, based on previous years' data. Or if data showed that the number of patients in labor and delivery increased by 10% each year over a span of 4 years, the facility might consider adding more beds and staff to that area.
- Healthcare department managers use statistics to set goals for the department, such as annual budgets.
- Cancer registries use statistical information regarding the different types of cancer, stages, and treatment. They also maintain survival rates of cancer patients. Cancer registries can receive accreditation by the American College of Surgeons (ACS). The cancer registry must meet the standards set by the ACS to be accredited.
- Nursing facilities use statistical information to review the different types of payers of insurance their patients have.
- Home healthcare organizations keep statistical data to track patients and their outcomes. The information includes the following: the number of visits, dressing changes, oxygen machine use, and many other services that home health organizations make. Further data include how many patients are taking their medication as directed, how many are improving, and how many are being readmitted.
- Hospice provides services either in the home or in a healthcare facility. The services they provide are linked to the patient's diagnosis.

Data Mining

Now that we have learned about some consumers of statistical data, we should examine sources for these data and how they are collected.

Definition

The process of extracting information from a large set of data is known as **data mining**. Tan, Steinbach, and Kumar (2003) note that data mining is a "confluence of many disciplines" and show an overlap of statistics, data mining, and artificial intelligence, machine learning, and pattern recognition. As a process, data mining involves the steps of defining, finding, and extracting data or knowledge that is buried in large sets of raw data. In this process, "data is retrieved, consolidated, managed and prepared for analysis" (Valova & Noirhomme, 2009). Following data mining, the resultant data are analyzed and then organized into a usable format.

History

To better understand this concept, let's cover some of the history of data mining while setting some boundaries for how we will obtain data, review some processes and strategies to make sense of mined data, and integrate these data and methods into software analysis tools.

Data mining to make healthcare medical treatment decisions is not new, although the use of formal computerized data mining tools is. In fact, data mining and **evidence-based medicine (EBM)**—medical practice that is based on the best available current methods of diagnosis and treatment, as revealed in research—have existed since the time of Hippocrates (460 BCE) in ancient Greece. Other notable “data miners” in history include Aulus Cornelius Celsus of ancient Rome, and John Snow, the father of modern epidemiology. Celsus wrote that wound cleansing and hygiene were important in health care, though these practices did not take hold until the late 1800s. John Snow tracked via maps the source of cholera in 1854. Thus, we see that statistics has a role in not only medical administration, but treatment as well.

In the 1960s, when the computer age was just starting to become commercialized and heavily adopted by the business community, data were collected on magnetic tapes, punch cards, and disks—media that offered considerably less storage capacity than today's options. Data mining actually took a major step in revolution in the 1980s when relational databases and structured query languages (SQLs) were developed. By using the data sorting and retrieval power afforded by structured query language, hospitals were better able to analyze and make sense of large sets of data.

Data warehousing, the centralized storage of large sets of data, was introduced in the 1990s. It supported online analytic processing and multidimensional databases, which helped it to grow rapidly. In today's world of big business, hospitals and other large corporations use collected information to make large-scale assessments, such as predicted growth over the next 5 years or total revenue over the last 3 years.

Three different areas have provided the growth to make data mining what it is today: statistics, artificial intelligence, and machine learning. Statistics has enabled organizations like the CDC to provide better health care and services to patients, enabling its top 10 achievements in the 20th century: improved vaccinations, improved motor vehicle safety, safer and healthier foods, better control of infectious diseases, improved workplace safety, reduced deaths from heart disease and stroke, better family planning, increased

awareness of tobacco as a health hazard, fluoridation in drinking water, and healthier mothers and babies. Statistics play a vital role in keeping the general population healthy. With the advent of computers and the internet, we can now harvest and refine medical decisions to an even finer degree, to the benefit of patients and medical establishments.

Current Applications

Today, data mining or predictive analysis in health care is a growing field. In fact, it is being used more and more to not only predict trends and analyze findings, but as an important tool for medical establishments to improve patient care, improve service offerings, and decrease losses, not the least of which is lost revenue in patient billing. Data mining has long been used in other fields but is still an emerging area within the healthcare industry. Canlas notes that data mining tools are being used not only for e-business and marketing but also by healthcare providers for “analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims” (Canlas, 2009). Moreover, the recent changes in health care, billing, and administration in the United States offer great opportunities for data mining.

Another current trend in data mining is **machine learning**. In addition to predictive analysis for the future, this process can be used to analyze past historical data. Therefore, data mining is both a process and an analysis method. It is a process, as there are procedures for collecting and preparing the raw data. The appropriate analysis method is chosen based on the type of data needed and the desired results. You will examine data mining in more detail in this chapter as a formal (and modern) method. Lastly you will examine some statistical terms and processes and examine how to handle some common statistical formulas using computer applications.

How Does Your Hospital Rate?

Now that Paul has found a website, he can compare facilities. The hospital Paul should review is Hospital A because it has a score of 100 out of 100, but what if that hospital is located in Ohio and Paul lives in Maryland, which is quite a distance to travel? In that case, Paul should look at Hospital B, which is nearby and has a score of 76.6 out of 100.

After choosing to review Hospital B, he finds that the reputation with the specialist is only 23.3%, survival rate is better than expected, the safety rating for patients is moderate, there is a high volume of cardiac patients, it has the highest rating for magnet nursing recognition, and, most important, it is rated seven out of seven for advanced technologies and key patient services, including an advanced trauma center and intensivist staff, meaning they have a staff physician in the intensive care unit at all times. This information has hopefully answered some of Paul's questions about the physicians and quality of care for cardiac patients.

Consider the following:

1. What other information should Paul consider before making his decision?
2. Which facility would you choose if you were Paul?

Basic Statistical Concepts

Although a detailed discussion of statistics goes beyond the scope of this text, the next sections will introduce some of the major statistical concepts relevant to healthcare professionals. Let's start with some basic items applicable to most statistical measures.

Dataset

A **dataset** is the data collected on a subject under examination. When your dataset includes information on every member of the group being investigated, you are examining a **population**. Or you may have a **sample** of data, which is a subset of data that statistically represents the entire population, ideally. A capital letter N is used to represent a population size and a lowercase n refers to a sample size. These concepts are important and ones we will refer to throughout the text, so it is important to note them.

As an example of the previous definitions, consider the Nielsen media rating group. Suppose they survey a sample of 1 out of 1,000 households, asking them which television shows they like or dislike. Considering how many households there are in the United States, even using this seemingly small sample size will yield thousands of survey results. Similarly, in the healthcare setting, collecting data from 50 randomly

selected hospitals from across the country might be a fair and manageable representation of the greater population, which consists of every hospital in the United States. Imagine trying to survey all the hospitals in the United States! Obviously, it would be far easier to sort through and make inferences or summaries about the data when working with a sample rather than an entire population.

Typically, data mining involves very large samples or sets of data, or "big data," as you may hear it referred to. There are significant qualitative differences between the conclusions you can draw when examining a population versus a sample. When examining an entire population, you are not dealing with estimates or probability of an outcome but actual facts about an outcome. For example, consider a case in which out of 100 people surveyed, only 10 had a terminal disease. You know in this case that 10 of the 100 have a terminal disease—not just 10% but 10 real people. In this case, you are dealing with a population and not statistical or inference data. You have all the data and know all of the variables, so there is no need to estimate or infer. On the other hand, if the 100 people surveyed are only a random sample of a larger population rather than the entire population itself, you could generalize that the 10% who reported having a terminal disease represent the entire population. This, however, would be an inference only, not the statement of a fact.

Moving along, **parameters** are descriptive measures of a population and are sometimes referred to as fixed references (Batten, 1986). The use of parameters is different from that of statistical data, in which you are making assumptions about a larger population based on a random sampling of the population. An example of a parameter could be that we have 10,000 people under study in a population. We know we are examining 10,000 people. For example, imagine that you have 10,000 people to make a generalization about. In contrast, using statistical data methods (strictly speaking), you would randomly gather information from, say, 500 of the 10,000 and assume that the other 9,500 would answer the same way to the questions asked. As you can infer, having all the data is typically better than relying only on a sample of data.

However, in statistics, you typically do not have all the data, so you must collect data from many similar sources, which hopefully can lead to a valid finding for the unique situation at hand. In other words, we have data on the basis of which we can generalize findings that *should* be applicable to other groups.

For example, of 10,000 people surveyed, we found that 99% believed that hospital stays following major surgery, as covered by third-party payers, were not long enough. That being determined, we could infer that most others in the population would be closely aligned to the 99% mark, given that we could survey them as well. This of course would still be an assumption as to the outcome, but because the percentage of people who responded in this manner is so high (99%), it is probably true. For example, the data from five hospitals in our state of North Carolina seem to show that if patients are given information on a certain lifestyle change (such as hand washing) on their first visit, their rate of secondary infection is lowered by over 50%.

Variables

A **variable** is a characteristic or property of something that may take on different values. For example, the number of patients admitted to a hospital between the hours of 12 am (midnight) and 7 am might be tabulated and examined over a week's time frame. Each day's data would be a variable. An **independent variable** (also known as an experimental or predictor variable) is a factor we can measure, manipulate, or control for to produce a change in another variable, which is known as the **dependent variable** or outcome variable. For example, imagine that you would like to determine whether using twice the amount of a certain drug for a disease will help patients recover more quickly. In this example, the amount of the drug administered is the independent variable, and the recovery rate of the patients is the dependent variable.

Data Distribution

Data distribution refers to the characteristic pattern that data assume when represented in graphical form. A normal distribution of data is one that is characterized by data that average around a central value with no real tendency to skew left or right. In this case, 50% of data falls to the left of the peak of the curve, and 50% falls to the right, forming a symmetric curve that possesses a single peak when graphed. Because the curve is shaped like a bell, you will hear it referred to as a "bell curve" (**Figure 1.1**). This data distribution pattern is one of the most important statistical concepts to understand.

However, if the distribution is not normal, data could be spread out to the left or to the right, or it could be randomly distributed.

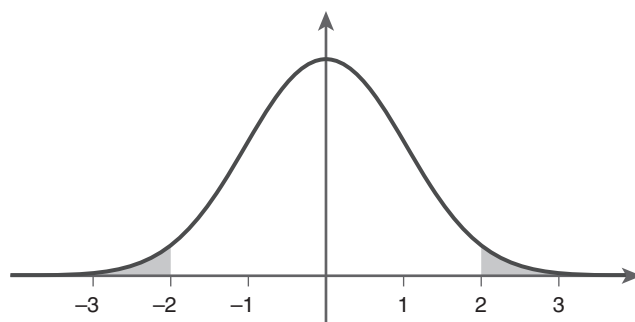


Figure 1.1 Bell curve.

Types of Data

The four types of data you might encounter are nominal (categorical), ordinal, interval, and ratio.

1. **Nominal data:** This type of data indicates categories and cannot be ordered. Examples include the following: techniques (technique A, technique B), gender (male or female), or occupation (e.g., students, professional programmers).
2. **Ordinal data:** This type of data can be ordered (e.g., in terms of size), but the difference between any of the two values may not always be equal. Examples include responses to a Likert-scale question, such as the following: *use it every day*, *use it once a week*, *use it once a month*, *use it once a year*, and *have never used it*. Likert-scale questions will be discussed in more detail in a later chapter.
3. **Interval data:** This type of data can be ordered, and the difference between any two consecutive values is the same, but there is no absolute zero, which allows you to have meaningful negative values. The most famous example of this type of data is temperature in degrees Celsius ($^{\circ}\text{C}$) or Fahrenheit ($^{\circ}\text{F}$). The zero values (i.e., 0°C and 0°F) are artificially defined, and negative temperature values are possible. But the difference between any two consecutive values at any point on the scale (i.e., between 0°C and 1°C and between 100°C and 101°C) is the same.

Note that when working with data from Likert-scale questions, if you can assume that the differences between any two options are equal, you can treat them as interval data. For instance, if your options are *strongly agree*, *agree*, *neutral*, *disagree*, and *strongly disagree*, you may be able to treat them as interval data, which is handy for computer tabulation. You simply assign a numeric value for each selection (e.g., *strongly agree* = 1,

agree = 2) then count the occurrences of each from the dataset, such as “out of 50 people sampled, 5 chose that they strongly agreed with the first question, 15 reported that they only agreed,” and so on.

4. **Ratio data:** This type of data can be ordered, the distance between any two consecutive values is the same (interval data), and there is an absolute zero. This means that a meaningful negative value of interval data does not exist (in statistics). Consider as examples the weight of a llama, the height of a tree, the length of something, or a recorded time or speed. A count could be considered as a ratio, as well.

Different statistical methods are designed only for certain types of data. As you work through this text, you should pay particular attention to this fact, as well as the size of the sample you are working with. Both choice of method and sample size are critical to choosing an appropriate statistical measure.

Types of Data Mining Models

Because this text offers the opportunity to mine real data for use in statistics, it will be helpful to discuss some key data mining models in detail here, including predictive, descriptive, and decision models. Subsequent chapters will give you hands-on experience with data mining methods.

Predictive Models

A **predictive model** explicitly predicts future behavior based on past trends. In each case, a model is created or chosen to try to best predict the probability of an outcome. When a predictive model is used in data mining, its main purpose is to forecast probabilities and trends in the future. Data are collected for pertinent predictors. The model is formulated, and finally it is validated and revised as new data become available.

For example, predictive models can be used by insurance companies and government programs such as Medicaid to assist in the prediction of future medical needs. A predictive model can also identify those who may be at high risk for developing a chronic disease or having poor health outcomes. However, initial medical information (data) on these patients is needed to make such predictions. Without their personal health information, identification of their

risks would be significantly delayed. This could potentially lead to unfavorable outcomes or a delayed improvement in the patient services offered.

Descriptive Models

A **descriptive model** describes patterns in existing data, including the main features or a summary of the data under examination by the researcher. It provides a hypothetical basis for the system under study. Descriptive models can recognize relationships while being used with other models to make predictions. In this approach, maps, charts, and graphics portray and promote understanding of real-world, complex, and sometimes redundant systems or services. Think of it as a way to describe visually the data in question. There are two general dimensions that are used for this description of mined data, viewpoints and aspects.

A **viewpoint** is a specific context or approach that you intentionally adopt when examining the data that allows you to focus on relevant details in the study and ignore irrelevant data.

An **aspect** is a specific category of data that is used in conjunction with your viewpoint. You might have one or more aspects associated with each viewpoint. Think of a viewpoint as a book chapter, with aspects being headings within the chapter. For example, a viewpoint might be data on patients being treated for lung cancer, such as survival rates and length of survival times. After patients who are cured are removed from consideration, aspects would include data on patients who survived 1 to 6 months, 7 to 12 months, and 1 to 2 years.

Decision Models

A **decision model**, also known as a business rule or business logic, is a logic system to determine desired actions for the business based on thresholds, conditions, or events. When using decision models, all elements in a relationship are examined to forecast or predict results. Units are arranged into groups according to the relationships between the units. All information would be organized into a logic **flowchart**, which is a graphical way to depict a series of actions, given a certain series of events. For example, a “Get up and go to school” flowchart would include events in the order in which they should occur, such as “wake up and turn off the alarm,” “take a shower,” “dry off,” “eat breakfast,” etc. Note the importance of the sequence of events in a flowchart. If you took a shower after you dried off, you might be a bit wet at the breakfast table!

As another example, consider diagramming a process to detail patient admittance to a medical facility. Each step would be graphically presented, with details and substeps, as a part of the flowchart. The first step might be to determine whether an arriving patient is an emergency case. This would be a logical true/false decision, with one of two paths being taken, depending on the case. The flowchart then would present the alternate steps for true and for false responses in a graphical fashion.

With decision models, the relationships among all the known data of a situation and the result of the logical decision process can be used to forecast future patterns and thus better prepare the organization with regard to planning. Thus, this model can help optimize and streamline the business and help the business to provide better end service and maximize cost savings. In later chapters we will examine some specific formulas and tools used for flowcharting and decision logic, such as for forecasting. Although there are many ways to examine data, this text will generally be limited to models most appropriate for the healthcare facilities.

Obtaining Data

If you are conducting an exhaustive search for information from many sources, use of data mining and statistical inference might be the best way to proceed. Regarding how to refer to your data, if a dataset was collected by other people, it is considered a **secondary data** source; if you collected the dataset yourself, it is a **primary data** source.

Related to obtaining data, there are five basic steps in data mining:

1. Extract and transform the data and load the data into the data warehouse system.
2. Store and manage data using a relational database system.
3. Provide data access to users.
4. Use application software to analyze the data.
5. Present the gathered information in a meaningful manner.

Keep in mind that if you are examining all of the data available, say from a local source such as the hospital where you work, then you are examining a population (capital *N*) and are using local data, otherwise known as a primary data source. Regardless of where you get your data, you should cite your sources, along with what measures and tools you used. Your

results might be suspect if you do not validate where your information came from, so give as many details as you can.

Some may assume that secondary data are inferior to primary data, but make no mistake, using mined secondary data in health care will improve patient care and increase benefits in the future. As paper medical records are replaced with electronic ones, mined secondary data will increase significantly, allowing for quality improvement in patient care and increases in cost savings, patient satisfaction, and revenue. Thus, a leading goal for healthcare facilities at this time is increased use of mined secondary data internally in their facilities, especially on identified areas that are in need of quality improvement.

How Does Your Hospital Rate?

Paul and all Americans want high-quality health care by the best professionals. However, we certainly cannot afford to travel to other states to receive treatment from the top-rated hospitals unless we are rich. When shopping for a medical care facility, consider that you might want the following: timely service, a safe environment, current medical technologies and treatments, and reliable patient-centered care. As Paul found out, the Hospital B has a high score for the use of technology. However, their moderate safety score concerns him; no matter how great the care may be in other areas, receiving a staph infection while in the hospital would hinder his journey back to full health.

To facilitate the sort of rating process that Paul undertook, the US federal government, via Medicare.gov, has created Hospital Compare, a website that allows people to compare area healthcare facilities based on zip code or by facility name.

Consider the following:

1. Use the following website: <https://www.medicare.gov/hospitalcompare/search.html>. How does your hospital rate for cardiac procedures?
2. Would you have a cardiac procedure performed in your area facility, or would you travel to another facility?
3. What do you think Paul should do?

Global Perspective

Taking our healthcare statistics globally allows us to compare diseases of other countries to those in the United States. In this Global Perspective section, we will examine rheumatic heart disease. This disease is caused by rheumatic fever, which affects the mitral valve in the heart. If a child has strep throat or scarlet fever that is not treated properly with antibiotics, he or she will have a higher risk of acquiring rheumatic heart disease. The mitral valve between the left atrium and left ventricle is affected by this condition. This condition affects children mainly between the ages of 5 and 15 years. Surgery to repair the mitral valve is one way to treat this condition.

The data below were retrieved from WHO. **Table 1.1** shows data for rheumatic heart disease in the United States and India, comparing males and females between the ages of 0 and 14 years in 2008. The population of males followed in India was 195,436, and the population of females was 179,323. The population of males in the United States was 32,646, and the population of females was 31,056.

On an interesting note, further research could be done to understand why India has a 2.3% rate of rheumatic heart disease in males and 3.0% rate in females. Questions that could be asked might include the following:

- Why is the female rate of rheumatic heart disease higher than that of males?
- Do children in India receive the antibiotic at all?
- Are incorrect dosages of the medication given?
- Are children not able to get the surgery necessary to repair the valve?
- Why do children in India acquire strep throat or rheumatic fever?

Table 1.1 Incidence of Rheumatic Heart Disease in India and the United States in 2008

Country	Sex	Population with Rheumatic Heart Disease
India	Male	2.3%
	Female	3.0%
United States	Male	0%
	Female	0%

Many other questions could be asked, and further research could be done with this information to follow up with types of treatment given to children in India compared with those in the United States.

This comparison of a disease's prevalence in the United States versus India is just a small sample of some of the interesting global trends and statistics that we will cover throughout subsequent chapters.

Hands-on Statistics 1.1: Examine Other Data from WHO

1. Using the same data source (WHO) as referenced in the Global Perspective section, compare India with a different country and discuss the results of your comparison. See http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html.
2. What are some strengths and weaknesses of this data source?

Chapter Summary

This chapter established a foundation for learning more about healthcare statistics and examining sources of data. It also examined some parts of the process involved in data analysis. The study of statistics is complex, involving extensive mathematical algorithms and research considerations, but rest assured that subsequent chapters will further explain and provide real-world examples of how to use the statistical methods involved in healthcare statistics. When you finish this text, you will be confident and well versed in the subject and will be able to put your knowledge to immediate practical use.

Also covered in this chapter is an introduction to data analysis and presentation of your findings. Hopefully, this chapter has whet your appetite for these subjects. In coming chapters, you will learn more topics related to statistics, data harvesting, data analysis, and presentation of your findings, all emphasizing real-world data. This chapter is only a small step toward an exciting journey to find out how powerful statistics can be.

Apply Your Knowledge

1. Consider users of healthcare statistics other than those mentioned in the chapter. List at least three users of healthcare statistics, and describe the statistics they keep.
2. Other than Florence Nightingale, who were some of the first users of statistics, and what statistics did they keep?
3. Name at least two countries that use telehealth.
4. Write a pro and con for some health issue, such as “smoking increases your risk of lung disease” or “running on pavement is bad for your knees.” Be creative!
5. Discussion: When researching *data mining*, you may come across the terms *data dredging* or *data snooping*. What do they mean? What implications should a researcher know about these terms? How might they be avoided?
6. Go to the following website: http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html. Review data for males and females in India and in the United States, comparing the percentages between the different age groups of those with rheumatic heart disease: 15 to 50 years and 60+ years. Write a brief explanation of your findings.
7. Go to <http://www.cdc.gov/DataStatistics/>. Click on a topic of your choice and compare health statistics for your state with those of another state. Did your state have a higher statistic than your comparison?

References

- Batten, J. (1986). *Research in education* (rev. ed.). Greenville, NC: Morgan Printers.
- Canlas, R., Jr. (2009). *Data mining in healthcare: Current applications and issues* (Master's thesis). Adelaide, South Australia: Carnegie Mellon University in Australia.
- Centers for Disease Control and Prevention. (2018). Table 10. Selected nationally notifiable disease rates and number of new cases: United States, selected years 1950–2017. Retrieved from <https://www.cdc.gov/nchs/data/hus/2018/010.pdf>
- Johnson, N., & Kotz, S. (2012). *Leading personalities in statistical sciences: From the seventeenth century to the present*. Wiley Online Library. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9781118150719.ch2/summary>.
- Lancaster, L. (2013). Celebrating statisticians: Florence Nightingale. JMP Blog. Retrieved from <https://community.jmp.com/t5/JMP-Blog/Celebrating-statisticians-Florence-Nightingale/ba-p/30247>
- Leen, S. (1994). The legacy of Hans Freudenthal. *Educational Studies in Mathematics*, 25(1-2), 164.
- Mackenzie, D. (1913). Indian myth and legend. Sacred Texts. Retrieved from <http://www.sacred-texts.com/hin/iml/index.htm>
- National Committee to Preserve Social Security and Medicare. (n.d.). Medicare. Retrieved February 21, 2020, from <https://www.ncpssm.org/our-issues/medicare/medicare-fast-facts/>
- Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education.
- Thucydides. (n.d.). *The history of the Peloponnesian War* (R. Crawley, Trans.). Project Gutenberg. Retrieved from <http://www.gutenberg.org/files/7142/7142-h/7142-h.htm>
- Valova, I., & Noirhomme, M. (2009). Comparative analysis of advanced technologies for processing of large data sets. *Information Technologies and Control*, 1-13.

Web Links

- Using Excel to do Basic Statistical Analysis: <https://www.statisticshowto.datasciencecentral.com/mode/#excel>
- Excel Tutorials for Statistical Data Analysis: http://www.statstutorials.com/EXCEL/EXCEL_TTEST2.html
- Introductory Statistics: Concepts, Models, and Applications: <http://www.psychstat.missouristate.edu/introbook/sbk25m.htm>
- MS Excel: How to Use the QUARTILE Function (WS): <http://www.techonthenet.com/excel/formulas/quartile.php>
- Excel and Quartiles: <http://www.meadinkent.co.uk/excel-quartiles.htm>
- Drawing a Normal Curve: http://www.tushar-mehta.com/excel/charts/normal_distribution/
- Sexually Transmissible Infections: <http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Main+Features10Jun+2012>
- Sexually Transmitted Diseases (STDs): Data and Statistics: <http://www.cdc.gov/std/stats11/trends-2011.pdf>
- Medicare Enrollment—Aged Beneficiaries: As of July 2010: <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareEnrpts/Downloads/10Aged.pdf>

Has Statistics Made Us Healthier? The Role of Statistics in Public Health: <http://www.statisticsviews.com/details/feature/5025891/Has-statistics-made-us-healthier-The-role-of-statistics-in-public-health.html>

Descriptive Statistics: <http://www.businessdictionary.com/definition/descriptive-statistics.html>

Discrete/Continuous: <http://www.chegg.com/homework-help/definitions/discrete-continuous-31>

Hospital Compare: <https://www.medicare.gov/hospitalcompare/search.html>

CHAPTER 2

Central Tendency, Variance, and Variability

Facts are stubborn, but statistics are more pliable.

—Mark Twain

CHAPTER OUTLINE

Introduction	Interquartile Range
Measures of Central Tendency	Standard Deviation
Mean	Variance
Median	Data Harvesting
Mode	Global Perspective
Frequency Distribution	Chapter Summary
Variance and Measures of Dispersion or Variability	Apply Your Knowledge
Min and Max	References
Range	Web Links
Outlier Data	

LEARNING OUTCOMES

After completing this chapter, you should be able to do the following:

1. Discuss the uses of mean, median, and mode in a hospital setting.
2. Discuss the uses of measures of central tendency.
3. Demonstrate how to find mean, median, mode, and standard deviation.
4. Describe the difference between continuous data and discrete data.
5. Describe the term *skewed*.

KEY TERMS

Continuous data	Max	Outlier data
Descriptive statistics	Mean	Range
Discrete data	Median	Skewed
Frequency distribution	Min	Standard deviation
Interquartile range	Mode	Variance

How Does Your Hospital Rate?

Dr. Barker, chief of the medical staff at his hospital, is reviewing cases of heart failure and the length of stay, readmission rates, and average age of patients with this condition. The purpose of this review is to improve patient care, reduce readmission rates, and potentially reduce length of stay in the hospital. Heart failure is the heart's inability to pump enough blood and oxygen to support the other organs in the body. Heart failure expenditures average about \$32 billion each year in the United States. This cost includes healthcare services, medication, and missed days of work. Dr. Barker went to the website of the Centers for Disease Control and Prevention (CDC) to find data on his county (CDC Interactive Heart Disease site: <https://nccd.cdc.gov/DHDSPAtlas/>).

Consider the following:

1. If you were in Dr. Barker's position, what specific goals might you have to improve patient care in this situation?
2. What might a high readmission rate indicate, and what are measures that could be taken to reduce this number?
3. Why would Dr. Barker want to reduce length of stay among heart failure patients? How does length of stay relate to the quality of patient care?

Introduction

The Mark Twain quote that introduces this chapter is quite telling in that although the data you have may be reliable and accurate, the interpretation of the data can be questionable, as this and other chapters will reveal. In this chapter, you will learn more about key statistical methods and provide simple and real-world examples of each. These examples will involve computing your results with a Microsoft Excel spreadsheet or other software tools. First you will learn about measures of central tendency, including mode, mean, median, variables, and frequency distribution. These measures are part of **descriptive statistics** and are concerned with summarizing and interpreting some of the properties of sets of data, but they do not suggest necessarily the properties of the entire population from which the sample was taken (which would require determining whether your collected data were in fact representative of the entire population).

Following the discussion of measures of central tendency, you will learn about measures of difference among the numbers in a dataset. Standard deviation, standard error, range, and variance are covered in this section and are known as measures of dispersion or variability.

Measures of Central Tendency

Statistical methods used to determine the shape of a distribution of data are mean, median, and mode. These methods are known as measures of central tendency,

or measures of central location and summary statistics. Each is a valid measure of central tendency but is used under different conditions. You will recall from the first chapter how mode was used to estimate the height of a brick wall, which is only one example of this measure.

If all the data are perfectly normal, the mean, median, and mode will be identical in reliability and will represent summary values accurately in a given dataset. Follow along and try the examples to see how these methods work.

Mean

Mean is a measure of central tendency that can be determined by mathematically calculating the average of observations (e.g., data elements) in a frequency distribution. Mean is the most common measure of central tendency. The mean can be used with discrete data (e.g., "choose 1, 2, or 3") but is most commonly used with continuous data (e.g., weight of a patient). In this case, the mean is just a model of the dataset.

One of the most important factors relating to the mean is that it minimizes errors in predicting any one value in the dataset. In fact, it produces the fewest errors compared with median and mode. Another important characteristic of this measure is that it includes all values in a dataset. Remember that a population is all of your data, whereas a sample is just a subset, preferably a random but representative one.

However, using the mean has one disadvantage: it is susceptible to the influence of outlier data, or data elements that are very different or far away from most

of the other data elements. Outlier data, discussed later in the chapter, may also be referred to as *flier*, *maverick*, *aberrant*, or *straggler*, although *outlier* is most commonly used. Outlier data may be detected by statistical tests (such as a Dixon or Grubbs test) or by a graphical display of the data.

Moreover, the more **skewed** (or varied) the data become, the less effective the mean is in locating a central tendency and typical value. Skewed simply means that the dataset contains both very small and very large numerical values. Mean is best used with data that are not skewed, such as in the set 1, 3, 2, 1, 3.

There are actually several types of calculation for mean. We have already covered the standard mean, also known as the arithmetic mean. Another is the sample mean, also known as the average. By using the sample mean, outlier data can be used. Skewed data will make no difference in the sample mean. Sample mean is an estimate of the population, and the dataset may be quite varied, such as 8, 34, 56, 25, 41, 2, 17, 25.

Other means are the harmonic mean, for rate or speed measures, and the geometric mean, for when the ranges you are comparing are different and you need to equalize them. Next we will compute a sample mean.

Hands-on Statistics 2.1: Use Excel to Find Mean

To compute the mean, add all items together and divide by the number of elements in the dataset. To perform this calculation in Excel, examine **Figure 2.1** and follow these directions:

- 1. Open a new, blank spreadsheet in Excel.
- 2. Key in the following data about inpatient hospital days, entering one numeral per cell in consecutive cells of the same row, beginning in row 2, column C (C2), and ending in cell G2:

1, 3, 2, 1, 3

- 3. Use the average function to find the mean of the dataset. To do so, click in cell H2 and then select the average function from the Formulas tab, or type the following formula in cell H2:

=AVERAGE(C2:G2)

- 4. Click out of cell H2, and the mean should appear in this cell: 2.

fx =AVERAGE(C2:G2)						
C	D	E	F	G	H	
1	3	2	1	3	2	

Figure 2.1 Using the average function in Excel.

DID YOU KNOW?

One in 33 babies are born with some type of birth defect. Birth defects are one of the leading causes of infant deaths and account for more than 20% of infant deaths. Statistics show that 1 in 691 infants born will have Down syndrome, meaning there are 6,037 infants diagnosed with Down syndrome each year. Anencephaly (congenital absence of all or a major part of the brain) occurs in 1 in 4,859 newborns, and 859 cases are reported per year (CDC, 2018b). What can a pregnant woman do to aid in the prevention of these types of birth defects? She should take folic acid every day, avoid alcohol and smoking, prevent infections, discuss current medications with her physician, maintain a healthy weight, and maintain regular office visits.

Now you know!

Median

Median is a measure of central tendency that reflects the midpoint of a frequency distribution when observations are arranged in order from the lowest to the highest. Median is used more often when data are skewed because it can retain its position and is less affected by skewed values. As stated by Agresti and Finlay (1997), “It is a measure of central tendency that better describes a typical value when the sample distribution of measurements is highly skewed.” The more skewed the data, the larger the difference between the median and the mean, which is the rationale for using the median in such cases. Median is best used with ordinal and interval/ratio (skewed) types of variables.

For example, consider the following dataset: 72, 2, 60, 44, 38, 51, 83, 25, 16, 69, 45. To calculate the median of this group of numbers, first put the data in order from the lowest to the highest, as follows: 2, 16, 25, 38, 44, 45, 51, 60, 69, 72, 83. In an odd set of numbers, the median will be the number in the very center of the series when arranged in order; that is, there will be exactly as many numbers before the median as after it. Our sample is an odd set of numbers, meaning our median is the number in the center, 45.

For an even set, you find the middle two numbers, add them together, and divide by 2. Your result will be the median value, which could well be a decimal value. You can round the resultant value depending on the type of number you need to obtain (i.e., whole number versus a fractional value).

Mode

Mode is a measure of central tendency that consists of the most frequent observations in a frequency distribution. Although mode is also regularly used in dataset analysis, Batten notes that mode “is considered to be less important than either the mean or the median” (Batten, 1986). The mode need not be unique, as it is possible to have two or more values that are distributed equally. You may have two numbers, 3 and 45, for example, that occur several times in a dataset along with other values. However, mode does not work well with continuous data and should not be used in that case, unless you are working with ratings that are assigned a numeric value, such as with Likert response variables assigned a numeric value (e.g., strongly disagree = 1, disagree = 2). Continuous data are basically any type of data that have infinite values with connected data points, which can result in an unlimited selection of data. For example, a measure of height could be recorded as a round 72 inches, or it could have infinite values—72.2, 72.24, and so on. It is generally accepted in the reporting of this example that you would say, “The average height was . . .” instead of reporting the most frequently occurring height; though both would work, the average better describes height data. Also, mode does not offer an effective measure of central tendency if the most common data are located far away from the rest of the data. Mode is best used when using a nominal type of variable.

Hands-on Statistics 2.2: Use Excel to Find Median

To find the median using Excel, see **Figure 2.2** and follow these directions:

1. Begin with a blank spreadsheet open in Excel.
2. Key in the following data, entering one numeral per cell in consecutive cells of the same row, beginning in cell A2 and ending in cell K2:

2, 3, 4, 5, 5, 6, 9, 77, 5, 4, 3

3. Use the average function to find the median of the dataset. To do so, click in cell L2 and then select the median function from the Formulas tab, or type the following formula in cell L2:

=MEDIAN(A2:K2)

4. Click out of cell L2, and the median should appear in this cell: 5.

	L2 fx =MEDIAN(A2:K2)											
	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	2	3	4	5	5	6	9	77	5	4	3	5
3												

Figure 2.2 Median using Excel.

Hands-on Statistics 2.3: Use Excel to Find Mode

To find the mode using Excel, see **Figure 2.3** and follow these directions:

- 1. Begin with a blank spreadsheet open in Excel.
- 2. Key in the following data, entering one numeral per cell in consecutive cells of the same row, beginning in cell G2 and ending in cell M2:

4, 4, 5, 6, 4, 2, 3

- 3. Use the mode function to find the median of the dataset. To do so, click in cell F2 and then select the median function from the Formulas tab, or type the following formula in cell F2:

=MODE(G2:M2)

- 4. Click out of cell F2, and the median should appear in this cell: 4.

This is a simple example, and you might be thinking, “I could have figured that out in my head!” and you would be correct. With larger datasets, however, it becomes much more difficult to figure out the answer without using a spreadsheet. Consider that even with a dataset as small as 35 items, it becomes very difficult to “eyeball” the results to find the mode.

Font		Alignment		Number						
fx =MODE(G2:M2)										
	D	E	F	G	H	I	J	K	L	M
			Mode	v1	v2	v3	v4	v5	v6	v7
			4	4	4	5	6	4	2	3

Figure 2.3 Mode using Excel.

At this point, two main types of data should be considered, **discrete data** and **continuous data**. Discrete data are defined and finite. For example, a survey that asks you to choose 1, 2, or 3 is discrete, as there is no 1.2 or 2.5 data point. Use of continuous data is well illustrated in a measurement of height. There certainly could be fractional points of measurement, and not only a set number defined values. Then you will use Excel to find the mode of a dataset.

Frequency Distribution

Frequency distribution describes how often different values are found in a set. Simply put, this function counts the frequency of values and tallies them for you, even in a chart or histogram, should you choose it.

Hands-on Statistics 2.4: Use Excel to Find Frequency Distribution

For this exercise and many that follow, note that you generally will need to install the free data analysis toolpack into Excel, unless it has been installed for you.

To find the frequency distribution using Excel, follow these directions:

- 1. Open the spreadsheet titled “CH02_Frequency_Distribution.xlsx” located in Chapter 2 of the eBook.
- 2. Highlight the two columns of data.
- 3. Select the *Data* menu, then *Data Analysis*.
- 4. Select *Histogram* and click *OK*.
- 5. To the right of *Input Range*, click the red arrow, then highlight the dataset named “data.”
- 6. To the right of *Bin Range*, click the red arrow, then highlight the possible values in the set (the heading in the spreadsheet is named “bin values”).
- 7. Click the down red arrow to maximize the control box again.
- 8. Check *Chart Output* if you want a chart, then click *OK*.
- 9. A new sheet will be inserted in your workbook, with the Bin, or value, and the frequency of occurrence to the right.

How Does Your Hospital Rate?

Dr. Barker continues to review the CDC website (<http://nccd.cdc.gov/DHDSPAtlas/>), which provides him with in-depth information on the statistics of heart failure. His results span the years 2008 through 2010. In his comparison, he looks at race, demographics, and age to see if there is a correlation between them. Dr. Barker finds that in Brunswick County, North Carolina, 78.4% of black male Medicare beneficiaries hospitalized for heart failure were discharged home following treatment, compared with 71.1% of white males in the same circumstances. He also finds that heart disease occurred in 1528 per 100,000 black males in Brunswick County, compared with a national average of 1695.6 per 100,000 among this demographic; in contrast, it occurred in 1413.6 per 100,000 white males in Brunswick County, compared with a national average of 1469.3 per 100,000 in this demographic. These statistics for Brunswick County are very close to the national average.

Consider the following:

1. Visit the website <http://nccd.cdc.gov/DHDSPAtlas/>. Look up and compare the percentages of black and white male Medicare recipients discharged home after a hospitalization for heart failure. How do these rates compare?
2. How do these rates compare with the same demographic groups for the state as a whole? For the nation?
3. What might we learn about discharges following hospitalization for heart failure among whites and blacks? What additional information should we gather to shed light on this issue?

Variance and Measures of Dispersion or Variability

These measures include minimum (min) and maximum (max) values, range, outlier data, interquartile range, standard deviation, and variance. We will first examine min and max.

Min and Max

Min is simply the minimum or smallest value in a dataset, and **max** is the maximum or largest value in a dataset. Excel can return these values to you automatically, which can be very useful for large datasets.

We will use both of these functions in combination to work with range, or the distance between two possible values. This combination will be very handy for many situations when dealing with large datasets.

Hands-on Statistics 2.5: Use Excel to Find Min and Max

For a given set of values, the formula `=MAX(start:end)`, where *start* and *end* are the starting and ending cells, gives you the maximum value from a set of values. The min function, `=MIN(start:end)`, gives you the minimum, using the same logic. To find the min and max using Excel, see **Figure 2.4** and follow these directions:

1. Begin with a blank spreadsheet open in Excel.
2. Key in the following data, entering one numeral per cell in consecutive cells of the same row, beginning in cell B1 and ending in cell K1:

3, 1, 66, 4, 7, 9, 1, 2, 8, 7

f_x =MAX(B2:K2)										
B	C	D	E	F	G	H	I	J	K	L
3	1	66	4	7	9	1	2	8	7	1
3	1	66	4	7	9	1	2	8	7	66

Figure 2.4 Finding the min and max using Excel.

(continues)

- 3. Either highlight the cells indicated in the previous step and drag them down to the row below to autofill it or reenter these numbers manually in cells B2 to K2.
- 4. Use the min function to find the min value of the dataset. Type the following formula in cell L1:

=MIN(B1:K1)
- 5. Click out of cell L1, and the min should appear in this cell: 1.
- 6. Use the max function to find the max value of the dataset. Type the following formula in cell L2:

=MAX(B2:K2)
- 7. Click out of cell L2, and the max should appear in this cell: 66.

Range

Range is the difference between the maximum value in a dataset and the minimum value. When you compute the range, most of the data are ignored because you are using only the largest and smallest extremes. Your range statistic provides information

on the statistical dispersion of a data sample, or the start and end points.

To see how range works, consider the following dataset: 53, 64, 78, 98, 58, 61, 83, 89. Noting that the highest value is 98 and the lowest value is 53, you can find the range by subtracting the minimum number from the maximum number: $98 - 53 = 45$.

Hands-on Statistics 2.6: Use Excel to Find Range

To find the range using Excel, you must first find the max and min values for the dataset and then use the subtract function to calculate the difference between them. See **Figure 2.5** and follow these directions:

- 1. Begin with a blank spreadsheet open in Excel.
- 2. Key in the following data, entering one numeral per cell in consecutive cells of the same row, beginning in cell A2 and ending in cell K2:

2, 3, 4, 5, 5, 6, 9, 77, 5, 4, 3
- 3. Use the max function to find the max value of the dataset. Type the following formula in cell L2:

=MAX(A2:K2)
- 4. Click out of cell L2, and the max should appear in this cell: 77.
- 5. Use the min function to find the min value of the dataset. Type the following formula in cell L3:

=MIN(A2:K2)
- 6. Click out of cell L3, and the min should appear in this cell: 2.
- 7. Use the subtract function to find the range of the dataset. Type the following formula in cell L5:

=L2-L3
- 8. Click out of cell L5, and the range should appear in this cell: 75.

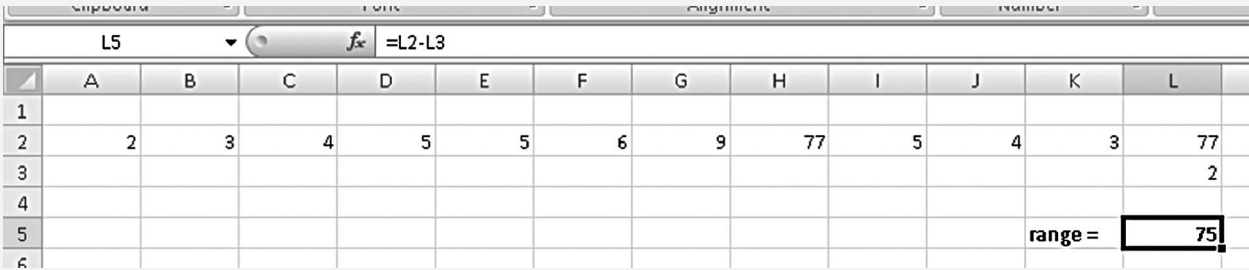


Figure 2.5 Calculation of range in Excel by subtracting min from max.

Outlier Data

Outlier data are elements of a dataset that lie an abnormal distance from other values in the dataset or sample. You, the researcher, must decide what is and is not an outlier. In the preceding example, the 77 would be considered outlier data. If you have decided to eliminate outlier data, you would remove them from the dataset but might make note of them.

Interquartile Range

The **interquartile range** is a measure of the dispersion within a dataset. It is the difference between the third quartile and the first quartile. Quartiles are the three points that divide a dataset into four equal groups, each group making up a quarter of the data. The interquartile range, therefore, is the breadth of the

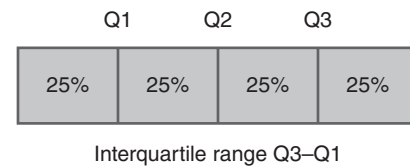


Figure 2.6 Quartiles.

interval that encompasses 50% of the sample. Usually it is smaller than the range and is less affected by outlier data. The interquartile range also tells you the size of the box in a box plot chart, which we will address in later chapters. The quartile function in Excel can be used to find the interquartile range, which gives a measure of the spread of the distribution, ignoring outlier data (**Figure 2.6**).

Next let's examine how to summarize just how different sets of numbers are.

Hands-on Statistics 2.7: Use Excel to Find Interquartile Range

The function in Excel for quartile is as follows:

=QUARTILE({array data}, *nth quartile*)

where *array data* is a sequence of like data (e.g., 1, 4, 5, 9) and the *nth quartile* is the quartile value you wish to return. The quartiles are represented as follows: 0 = the smallest value in the dataset; 1 = the first quartile (Q1), or 25%; 2 = the second quartile (Q2), or 50%; 3 = the third quartile (Q3), or 75%; and 4 = the largest value.

For example, based on the formula of Q3 – Q1, we can get the interquartile range, or one-half the total range. See **Figures 2.7** and **2.8**, and follow these directions:

1. Begin with a blank spreadsheet open in Excel.
2. Use the quartile function to find the Q3 value of the dataset. Type the following formula in cell B1:

=QUARTILE({1,2,3,4,5,6,7,8,9,10},3)

Font		Alignment	
fx		=QUARTILE({1,2,3,4,5,6,7,8,9,10},3)	
B	C	D	E
7.75		3.25	4.5

Figure 2.7 Quartile function in Excel for Q3 in cell B1, with interquartile range in cell F1.

Font		Alignment	
fx		=QUARTILE({1,2,3,4,5,6,7,8,9,10},1)	
B	C	D	E
7.75		3.25	4.5

Figure 2.8 Quartile function in Excel for Q1 in cell D1, with interquartile range in cell F1.

(continues)

- 3. Click out of cell B1, and the Q3 value should appear in this cell: 7.75.
- 4. Use the quartile function to find the Q1 value of the dataset. Type the following formula in cell D1:
$$=QUARTILE(\{1,2,3,4,5,6,7,8,9,10\},1)$$
- 5. Click out of cell D1, and the Q1 value should appear in this cell: 3.25.
- 6. Use the subtract function to find the interquartile range of the dataset. Type the following formula in cell F1:
$$=B1-D1$$
- 7. Click out of cell F1, and the interquartile range should appear in this cell: 4.5.

Thus, in this example we learn that the difference between the 75th and the 25th percentiles is 4.5, which represents the size of the spread of the middle 50% of the data in the distribution.

Standard Deviation

Standard deviation is a measure of variability that describes the deviation from the mean of a frequency distribution of data. The standard deviation is symbolized by *sd* or *s*. Note that the more the data are spread, the greater the standard deviation. For example, if you have test scores that range from 48% to 98%, the standard deviation will be higher than test scores that range from 95% to 98% because in the latter case the data are not as varied as in the first range.

Hands-on Statistics 2.8: Use Excel to Find Standard Deviation

- The formula in Excel for finding the standard deviation is `=STDEV(start:end)`, where *start* and *end* are the starting and ending cells of the range. To find the standard deviation using Excel, see **Figure 2.9** and follow these directions:
- 1. Begin with a blank spreadsheet open in Excel.
 - 2. Type “Scores” in cell B2 and then key in the following data, entering one numeral per cell in consecutive cells of the same column, beginning in cell B3 and ending in cell B12:
$$1, 2, 3, 1, 2, 4, 1, 1, 3, 2$$
 - 3. Use the standard deviation function to find the standard deviation of the dataset. Type the following formula in cell B13:

$$=STDEV(B3:B12)$$

B13		fx		=STDEV(B3:B12)	
A	B	C	D	E	F
	Scores				
	1				
	2				
	3				
	1				
	2				
	4				
	1				
	1				
	3				
	2				
	1.054093				

Figure 2.9 Standard deviation function in Excel.

- 4. Click out of cell B13, and the standard deviation value should appear in this cell: 1.054093.

Variance

Variance is a measure of the spread of observations in a distribution of data. It is equal to the square of the standard deviation, which we just examined. When stating the variance of a dataset, you are giving a measure of how closely aligned your expected value is to the distribution. It is basically a spread of the distribution and its average value.

When you have a large variance, individual values of random variables will tend to be farther from the mean. The smaller the variance, the closer the individual values and random variables tend to be to the

Hands-on Statistics 2.9: Use Excel to Find Variance

There are several variance functions in Excel, and we will use the simplest right now, the VAR function. However, note that there are others, and you will see them via the autocomplete feature of Excel when you go to insert a function.

To find the variance using Excel, see **Figure 2.10** and follow these directions:

1. Begin with a blank spreadsheet open in Excel.
2. Key in the following data, entering one numeral per cell in consecutive cells of the same row, beginning in cell B1 and ending in cell J1:

3, 4, 3, 2, 6, 8, 9, 1, 3

3. Use the VAR function to find the variance of the dataset. Type the following formula in cell K1:

=VAR(B1:J1)

4. Click out of cell K1, and the variance value should appear in this cell: 7.5.

	B	C	D	E	F	G	H	I	J	K	L
1	3	4	3	2	6	8	9	1	3	7.5	
2											
3											

Figure 2.10 Variance function in Excel.

mean. You should also note that variance and standard deviation of random variables are always non-negative.

A sample variance is the sum of the squared deviations from their average divided by one less than the number of observations in the given sample, written as follows:

$$\text{Variance} = ((a - \text{mean})^2 + (b - \text{mean})^2 + (c - \text{mean})^2) / \text{total number of values} - 1$$

In this formula, a , b , and c are all values in the dataset, and mean is the average of all the values in the dataset.

Let's apply this formula to the dataset 10, 20, 30. The average of this dataset is $10 + 20 + 30 = 60 / 3 = 20$. The equation for variance is:

$$\begin{aligned} & ((10 - 20)^2 + (20 - 20)^2 + (30 - 20)^2) / 3 - 1 \\ & = (-10^2 + 0^2 + 10^2) / 2 \\ & = (100 + 0 + 100) / 2 \\ & = 200 / 2 = 100 \end{aligned}$$

The final computation of variance for this dataset is 100.

This type of statistical information has many purposes and can be used in clinical trials of new prescription drugs or patient treatments, for example.

DID YOU KNOW?

According to the *Washington Post*, the cost of an emergency department visit can vary in cost depending on what kind of medical insurance you have and where you seek treatment (Kliff, 2013). A trip to the emergency department can cost you 40% more than the average cost of rent. For example, for sprains and strains, the median cost was \$1,051.00 (982.00–1,110), the mean was \$1,498.00 (1,304.00–1,692.00), and the interquartile range was \$1,018. Charges for sprains and strains ranged from \$4.00 to \$24,110. Kidney stones ranked the highest in interquartile range of the top 10 diagnoses treated in the emergency department, with a median of \$3,437.00 (2,917.00–3,877.00), mean of \$4,247.00 (3,642.00–4,852.00), and interquartile range of \$3,742.00. The cost of kidney stones ranged from \$128.00 to \$39,408. How do the charges for these conditions rank at your local hospital?

Now you know!

Hands-on Statistics 2.10: Examine Typical Hospital Data

Now we will use the knowledge just covered to examine a 6-hour time frame from a hypothetical hospital emergency department. Using Excel, see **Figure 2.11** and follow these directions:

1. Open the file “CH02_Hospital_Time_Frame.xls” located in Chapter 2 of the eBook. The data in that file should appear as in Figure 2.11.
2. From left to right, observe the data in the following columns. The *Medical Record Number* is just a number representing a patient. *Sex* is noted as 1 for male and 2 for female. *Age* is as you would assume. *Diagnosis Code* is an alphanumeric value for what the patient is diagnosed as having—for example, a cold. *Procedure Code* is a code for a specific procedure (such as an X-ray scan) that was performed on the patient. Lastly, *Ins. Code* (Insurance) is a numeric value, with 1 being Blue Cross, 2 being VA, 3 being Kaiser, etc. Now, considering the aforementioned items, let’s examine some different functions and insert them into the spreadsheet to make sense of the raw data.
3. Use the MODE function to find the most frequently occurring sex and insurance payment type. Type the following in cell F51:

=MODE(F4:F50)

Then type the following in cell B51:

=MODE(B4:B50)

As you click out of these cells, the mode values will appear in them.

4. Use the standard deviation function on the age. Type the following formula in cell C51:

=STDEV(C4:C50)

As you click out of this cell, the standard deviation value should appear in it.

Next, find the range value for age by following these steps:

1. Use the max function to find the max value of the dataset. Type the following formula in cell C52:

=MAX(C4:C50)

Click out of cell C52, and the max should appear in this cell.

2. Use the min function to find the min value of the dataset. Type the following formula in cell C53:

=MIN(C4:C50)

Click out of cell C53, and the min should appear in this cell.

3. Use the subtract function to find the range of the dataset. Type the following formula in cell C54:

=C52–C53

Click out of cell C54, and the range should appear in this cell.

	A	B	C	D	E	F	G	H
1	Patient admittance data for 6 hr time frame for Cape of Good Hope Hospital							
2								
3	Medical Record Number	Sex	Age	Diagnosis Code	Procedure Code	Ins. Code		
4	112211	1	22	359.21		2		
5	112201	1	63	346.91	3.31	3		
6	112233	2	47	873.51	86.59	3		
7	112275	1	58	305.01		5		
8	112281	2	30	305.01		4		
9	112287	1	21	873.53	86.59	2		
10	112292	2	34	303.01		2		
11	112298	2	59	780.31		1		

Figure 2.11 Sample data from Cape of Good Hope Hospital.

For some items, such as Medical Record Number, there is no analysis to be performed, perhaps other than a count of the number of patients, performed as follows:

4. Use the Count function to find the number of medical records listed. Type the following formula in cell A51:

=COUNT(A4:A50)

Click out of cell A51, and the number of medical records should appear in this cell.

Data Harvesting

In the previous example, our data were already in an Excel spreadsheet. However, that will not always be the case. Raw data often come in different file

types, such as a Word document, Excel spreadsheet, or image file such as JPEG. Raw data formats that are common include comma-separated values (CSV), in which each cell or data item is separated in the file with a comma, and tab-delimited values, in which

Hands-on Statistics 2.11: Harvest Real-World Data

In this example, we will examine hospital data using the Washington State Department of Health's Comprehensive Hospital Abstract Reporting System (CHARS). Follow these steps:

1. Visit the CHARS primary webpage at <http://www.doh.wa.gov/DataandStatisticalReports/HealthcareinWashington/HospitalandPatientData/HospitalDischargeDataCHARS.aspx>. Alternatively, you can enter the following phrase in a search engine: "Comprehensive Hospital Abstract Reporting System (CHARS) from the Washington State Department of Health." Scroll down the page and click on *Chars Reports*. You will have the option of downloading reports as either PDF (Adobe Acrobat) or Excel files.
2. Scroll down to the bottom of the page to the section titled *2015 Full Year Standard Reports Discharges: Inpatient / Observation*. Or use a newer year if desired.
3. In the subsection titled *Excel files*, click on the link *Hospital Census and Charges*. Save this file to an appropriate folder on your computer and then open it.
4. Find the min and max for discharges (total) for all hospitals. To do so, you will have to write a function that examines only certain cells in the spreadsheet. To get started for min, type "=MIN(c9,c12.....)" in an open cell at the bottom of the document, adding the cell numbers for each bold total for each hospital. Use the same approach to find max.

How Does Your Hospital Rate?

Dr. Barker also considers how Brunswick County compares with the rest of the counties in the state by using the County Health Rankings website. Dr. Barker finds that Brunswick County ranked 54th out of 100 in mortality, 23rd in morbidity, 47th in health behaviors, and 42th in social and economic factors. He finds that preventable hospital stays decreased from the year 2003 to 2010. He finally concludes that social and economic factors and health behaviors had the most influence as well as genetics for those with heart failure. Some of the questions that could be asked are: Are the patients taking their medications correctly, or at all? How are their eating habits?

Consider the following:

1. Visit the County Health Rankings website (<http://www.countyhealthrankings.org/app/>), and look up your own county. How does your county rate compare with the state average in mortality, morbidity, health behaviors, and social and economic factors?
2. In which area does your county receive its highest ranking? In which area does it receive its lowest ranking?
3. Which health behaviors in your county appear to be contributing to poor health?

each cell is separated by a uniform amount of space, such as a tab. There are other formats, including some that are specifically formatted for an Apple Mac or MS-DOS. Just choose the type that is most compatible with the system you are using.

That was a simple example of harvesting data, but taking a quick look at the spreadsheet, you will probably notice many other items worth examining. In later chapters we will dig deeper into the data. For now, though, let's step back and examine what the research process looks like and a few key considerations.

Global Perspective

Consider the following US healthcare statistics:

- The average life expectancy in the United States is 78.6 years according to the CDC for 2017 (CDC, 2018a).
- The infant mortality rate is 5.8 deaths per 1,000 live births (CDC, 2018a).
- In 2007, there were 96 preventable deaths per 100,000 people.
- In 2007, there were 2.4 physicians per 1,000 consumers.
- In 2017, 18% of the gross domestic product was related to health expenditures; this is more than double the average spent among other developed countries (Committee for a Responsible Federal Budget, 2018).

So how do we as a nation compare with other countries? The United States, as of 2020, spends more than any country on health care—in fact, disproportionately more when viewed graphically. An internet search for “How does health spending in the United States compare to other countries?” reveals very interesting data. When measured as a percentage of gross domestic product, the United

States spends more than any other nation; yet US citizens have comparably poor health outcomes. In comparison, Japan spends much less on health care but has the world's lowest infant mortality rate (2.17 deaths per 1,000 live births) and a much higher life expectancy (82 years). Even though healthcare systems vary substantially from one country to another, the goal of these comparisons is to improve health care across the globe. That the United States has the most expensive healthcare system in the world there is no doubt.

Chapter Summary

This chapter covered the basic statistical measures of central tendency—mean, median, and mode—and explained their importance in calculating hospital statistics. It also discussed variability in data, or ways to examine how spread out items in a dataset are from the mean. In doing so, it applied two statistical standards: standard deviation and variance. Standard deviation, the square root of the variance, returned a measure of how far each value was from the computed mean of the dataset. Variance returned a similar value but also returned a value of zero if all values in the set were the same, a low positive value if they were similar, and a high value if they were dispersed and quite separated from the mean. Discrete data, such as a number of patients, were contrasted with continuous data, such as patient weight. The chapter also compared nominal (named) data, such as types of cars or gender, to ordinal (ordered) data, such as a scale of 1, 2, 3, etc. Next, among other key hospital statistics, the concept of outlier or skewed data—that is, values quite different from the majority of other values in the dataset—was examined. Last, but not least, the chapter introduced descriptive statistics, descriptive models, data harvesting, and many other meaningful tools used in healthcare statistics.

Apply Your Knowledge

1. Discussion: Describe in your own words how using the different types of statistical measures we have covered in this chapter would be of benefit to a physician's office or hospital to improve patient care and reimbursement methodologies.
2. Fifty children were diagnosed with leukemia during the past year. The weight of each child was recorded (in pounds) at the time

of diagnosis. Listed are the weights from low (lightest) to high (heaviest): 20, 22, 23, 26, 27, 28, 28, 29, 30, 31, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 42, 43, 43, 44, 45, 47, 48, 48, 49, 49, 51, 52, 52, 52, 54, 55, 56, 58, 58, 58, 60, 61, 62, 63, 64, 65, 67, 68, 70. Use an Excel spreadsheet to complete the following:

- a. Calculate the mean, median, and mode.

- b. Calculate the variance and standard deviation computed from a frequency distribution with a class interval of 1.
 - c. Calculate the weights that are one standard deviation above and one standard deviation below the mean.
 - d. Calculate the weights that are two standard deviations above and two standard deviations below the mean.
3. Compute the mean for the following values *by hand*. Do not use Excel.

5, 6, 7, 1, 2, 5, 6, 7, 1, 9, 33, 2, 9

4. Find the variance (by hand, not using Excel) for the following values. Show your work.

300, 400, 150, 510, 430, 611

5. Use the Excel STDEV function to find the standard deviation for the following heights (cm) of 12 students in a class:

170, 160, 165, 161, 163, 164, 170,
164, 163, 170, 166, 165

6. Use data about the ages of 80 nursing home residents, provided in **Table 2.1**, to complete the following:
 - a. Calculate the mean, median, mode, variance, and standard deviation for the ages listed in the table.
 - b. Prepare a frequency table in Excel that includes the ages listed in the table.
7. The lengths of hospital stay for patients with diverticulitis who had a partial bowel resection performed are recorded as follows. Calculate the mean, median, and mode using Excel.

5, 7, 9, 4, 6, 5, 7, 3, 6, 6, 4, 5, 5, 7, 7, 3

Table 2.1

65	68	71	73	75	76	78	80
65	68	71	74	75	76	78	80
65	69	72	74	75	76	78	82
66	69	72	74	75	77	80	82
66	69	72	75	77	77	79	80
67	70	72	75	77	78	80	83
67	70	73	75	77	79	80	83
68	70	73	75	79	80	80	84
68	71	73	76	79	80	80	84
69	71	73	76	79	80	80	84

8. For this question, use data below in **Table 2.2**, in a spreadsheet. A given medical clinic takes in the specified amounts of money per day in the form of cash co-payments. Find the range and mode for this one day. Use the SUM function to find the day's total cash intake.

Table 2.2

Patient Number	Payments by Patient
1	25
2	10
3	250
4	75
5	35
6	25
7	150
8	100
9	25
10	0
11	25
12	25
13	25
14	250
15	100
16	150
17	75
18	50
19	35
20	50
21	75
22	15
23	35
24	25
25	50
26	150
27	250
28	25
29	25
30	25
31	25

References

- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. (3rd ed.). Upper Saddle, NJ: Prentice-Hall.
- Batten, J. (1986). *Research in education* (rev. ed.). Greenville, NC: Morgan Printers.
- Centers for Disease Control and Prevention. (2018a). Mortality in the United States, 2017. Retrieved June 2018, from <https://www.cdc.gov/nchs/products/databriefs/db328.htm>
- Centers for Disease Control and Prevention. (2018b). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. Retrieved June 2018, from <https://www.cdc.gov/ncbddd/birthdefects/features/birthdefects-keyfindings.html>
- Committee for a Responsible Federal Budget. (2018, May 6). American health care: Health spending and the federal budget. Retrieved August 2019, from <https://www.crfb.org/papers/american-health-care-health-spending-and-federal-budget>
- Kliff, S. (2013, March 2). An average ER visit costs more than an average month's rent. *Washington Post*. Retrieved May 2014, from <http://www.washingtonpost.com/blogs/wonkblog/wp/2013/03/02/an-average-er-visit-costs-more-than-an-average-months-rent/>

Web Links

- Discrete/Continuous: <http://www.chegg.com/homework-help/definitions/discrete-continuous-31>
- Descriptive Statistics: <http://www.businessdictionary.com/definition/descriptive-statistics.html>
- MS Excel: How to Use the QUARTILE Function (WS): <http://www.techonthenet.com/excel/formulas/quartile.php>
- Excel and Quartiles: <http://www.meadinkent.co.uk/excel-quartiles.htm>
- Drawing a Normal Curve: http://www.tushar-mehta.com/excel/charts/normal_distribution/
- Interactive Atlas of Heart Disease and Stroke: <http://nccd.cdc.gov/DHDSPAtlas/reports.aspx?state=NC&themeId=13#report>
- Data and Statistics on Birth Defects: <http://www.cdc.gov/ncbddd/birthdefects/data.html>
- An Average ER Visit Costs More Than an Average Month's Rent: <http://www.washingtonpost.com/blogs/wonkblog/wp/2013/03/02/an-average-er-visit-costs-more-than-an-average-months-rent/>
- Life Expectancy: <http://www.cdc.gov/nchs/fastats/life-expectancy.htm#>
- Relative to the Size of Its Wealth, the US Spends a Disproportionate Amount of Health Care: <https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-start>

CHAPTER 3

Patient Data

He who builds on the people builds on mud.

—Niccolò Machiavelli

CHAPTER OUTLINE

Introduction	Types of Databases
Census Data and Their Importance	Flat-File Database
Calculation and Reporting of Patient	Relational Database
Census Data	Data Formats
Inpatient Service Days	R-Project
Average Daily Census	Data Stored in R
Data Visualization	Global Perspective
Visually Examine Data With Sparklines	Chapter Summary
and Microcharts	Apply Your Knowledge
Newborn Services	References
Open-Source Software	Web Links
Freeware and Shareware	

LEARNING OUTCOMES

After completing this chapter, you should be able to do the following:

1. Describe census data and their use in health care.
2. Describe the admissions, discharge, and transfer report and explain how to calculate it.
3. Define inpatient service days, describe this statistic's importance, and explain how to calculate it.
4. Define average daily census and explain how to calculate it.
5. Create sparklines and microcharts to present data.
6. Create a line chart to present data.
7. Define and explain the differences between open-source software and freeware/shareware.
8. Describe types of databases.
9. Import raw data from healthcare websites into Excel.
10. Describe data formats.
11. Use R-Project for basic statistical analysis.

KEY TERMS

Admissions, discharge, and transfer (ADT)	GNU's not Unix (GNU)	Relational database
Average daily census	Infographics	Scalar
Binary code	Inpatient service days	Script-mode interface
Census data	Intensive care unit (ICU)	Source code
Code	Language	Sparklines
Comma-separated values (CSV)	Matrix	Sunset
Commercial code	Microchart	Tab-delimited values
Flat-file database	Open-source software	Vector
General Public License (GPL)	Patient care unit (PCU)	
	Recapitulation algorithm	

How Does Your Hospital Rate?

Veronica is a nurse at Hospital A, a 25-bed licensed facility located on the southeast coastal region in North Carolina (Figure 3.1). She has been assigned the task of calculating the hospital's rates of admissions, discharges, and transfers.

Consider the following:

- 1. What uses might hospital administrators have for this information?
- 2. What do these measures indicate about the operation of this hospital?
- 3. Where might Veronica find this information?



Figure 3.1 Doshier Memorial Hospital, Southport, North Carolina.
Photo by J. Burton Browning.

Introduction

The **admissions, discharges, and transfer (ADT)** list aids hospitals in calculating how many patients were admitted, discharged, and transferred each month and provides yearly totals for each patient care unit in the facility. This information is useful for forecasting staffing issues, implementing new services,

and reviewing the age and sex of the population. Average length of stay is another important statistic that may be viewed. The average length of stay can reveal trends of patients with certain diagnoses and procedures. For example, imagine that eight patients come in with pneumonia; three of these patients go home in 3 days, and the other five stay 6 days. The medical record could be reviewed to determine why those five

patients had a longer length of stay. The transfer part of the ADT indicates how many patients were transferred out and why. Based on this information, facility administrators should consider questions such as the following: Do we provide the services that patients needed? Were we lacking specialists for certain procedures?

Accumulating, analyzing, and reporting facility data not only help in many ways with daily operation of the hospital but also provide benchmarking and forecasting information important to hospital administration. In this chapter you will learn how to calculate these data, ways to report them, and other information related to patient data. Specifically, you will review census data, ADT reports, inpatient service days, and calculations for these vital hospital data. Next you will learn how to present statistical data using tools such as sparklines and microcharts. Data formats and data mining will be examined, as well, building on topics from the previous chapters. You will also gather statistical data from the World Health Organization (WHO), compare data on diseases, and present the data visually.

Census Data and Their Importance

A census is a regular, ongoing count of the people who make up a population. Besides the US national census that is conducted every 10 years, many smaller censuses are routinely conducted by organizations throughout the country. In the hospital setting, **census data** are data that describe patients who are currently in the hospital during a certain time frame. Each **patient care unit (PCU)** reports the number of patients admitted that day, discharged, or transferred in or out, perhaps to another hospital or unit. Note that these data include only *inpatients*, who are the patients who have been admitted to the healthcare facility for an overnight or longer stay.

The statistical data provided from the hospital census is important, as it affects patient care and operational efficiency of the hospital and helps determine financial performance. These data are used by hospitals to make decisions regarding staffing, budgeting, and planning for the future. Census data can also be used to keep track of patients' age, ethnicity, and management of chronic conditions. All of these data are invaluable in longitudinal studies the facility conducts. For example, if over a 10-year time frame you determined that the patient rate for retirees was increasing at a 5% rate every year, you might

determine that specialized services this patient group requires should be expanded, so you would plan for future growth expenditures in this area. As another example, if you determined that there was no steady increase in child care services, then you would not want to direct financial expansion monies to increasing pediatric services. Most importantly to consumers, however, is that this information can be used to improve patient care.

Haley and Bregman in 1981 noted that staffing decisions, based on census data, had an effect on the spread of infectious diseases among neonates. When the infant-to-nurse ratio was lower, there was a lower rate of infection among patients. This observation, made over 30 years ago, holds true today. Georgetown University School of Nursing and Health Studies in 2008 provided research results on lower staffing levels and the increased risk to patients of contracting nosocomial infections—that is, infections originating in the healthcare facility (Cronin, Leo, & McCleary, 2008). More recently, a culmination of findings from many research studies has reaffirmed this finding (Mitchell, Gardner, Stone, Hall, & Pogorzelska-Maziarz, 2018). The conclusion was that staffing shortages increased the risk of HAIs.

According to the Centers for Disease Control and Prevention (CDC), 90,000 of the 1.7 million people who acquired a nosocomial infection in 2002 would die (Klevens et al., 2007). It is probably no surprise that nosocomial infections not only affect the patient, but the whole healthcare system. Nosocomial infections increase medical costs of patients each year by \$4.5 billion to \$5.7 billion. The Institute of Medicine reported that 98,000 deaths occurred unnecessarily, in part related to nursing shortages (Kohn, Corrigan, & Donaldson, 2000). The study noted that when the nurse-to-patient ratio declined, the amount of time spent with the patient decreased, leading to an increase in nosocomial infections among patients. Without accurate census data, these important findings may never have been appreciated.

Regarding the census of patients, to conduct a census in the hospital, each inpatient care area counts the number of patients who are in that area or unit each day. Examples of inpatient care areas include oncology, pediatrics, and surgical services, with the number and types of units in a given hospital depending on the size of the facility. Note that the timing of when a census is taken is important. Most hospitals will take the census count at midnight because most patients are asleep then. If it is done at any another time, patients could be undergoing surgery, having x-ray scans taken, or otherwise occupied. The time

the census is taken needs to always be consistent, though, regardless of when it is done.

Before computer use became commonplace, census reporting and analysis were done by hand. However, as hospital size and admissions increased, computer use became a necessity. As a result, it is now much quicker and easier to view report data and find errors. With the computer it is easy to see when patients have been transferred to another location, the date of transfer, the date of admission, and the date of discharge, allowing the census taker to efficiently follow the patient during his or her admission to the facility. It may seem strange that computers have not always been used for such tasks, but in some facilities, their integration into healthcare management did not begin until the 1980s. Moreover, other emerging technologies, such as radio frequency identification (RFID), use of barcodes, wireless technologies, and integrated systems, are also helping improve patient care in many ways.

Ashar and Ferriter noted in 2007 that “two general categories are being established for using RFID technology in health care settings.” One category is based on retail use of RFID for inventory control of drugs and devices. The second category relating to RFID use involves the capture of streaming data related to patient point of care. Likewise, using barcodes to track patients reduces errors and ensures that correct patient information has been entered in the system. This system can also be used to allocate beds and bed transfers in real time, making it easier to move patients within the hospital. Wireless technologies are increasing the use of mobile devices in the delivery of health care.

Mobile apps are now available that will check various vital signs, including electrocardiographic signals and blood glucose levels. Some other advancements in wireless technologies allow the devices to be extremely compact and portable, reaching millions of people in third-world countries as well as many rural parts of the United States, affording them healthcare benefits associated with RFID technology. Such devices include a stethoscope that transmits heart sounds to a physician allowing him or her to make either a diagnosis or order tests without being present. With the growing use of this technology, in the future we may not have to visit a doctor’s office in person as often as before such technologies were invented.

Many integrated systems assist in the administration and effective delivery of patient care. Current systems such as those from Novant, Athena Health, and NextGen, and versions of the Veterans Information Systems and Technology Architecture (VistA) used by the US Department of Veterans Affairs, are integrated throughout the facility and can keep track of all needed information from the time of patient admission to discharge in a database known as the master patient index. This index contains data on bed tracking, transfers, discharges, and many other features. When all of the information has been put into the computer system, an ADT list is generated, showing which patients are still categorized as inpatients in the facility and which have been either discharged or transferred (**Figure 3.2**). Certainly all of these technologies, including the use of security controls, will only grow in use by medical facilities.

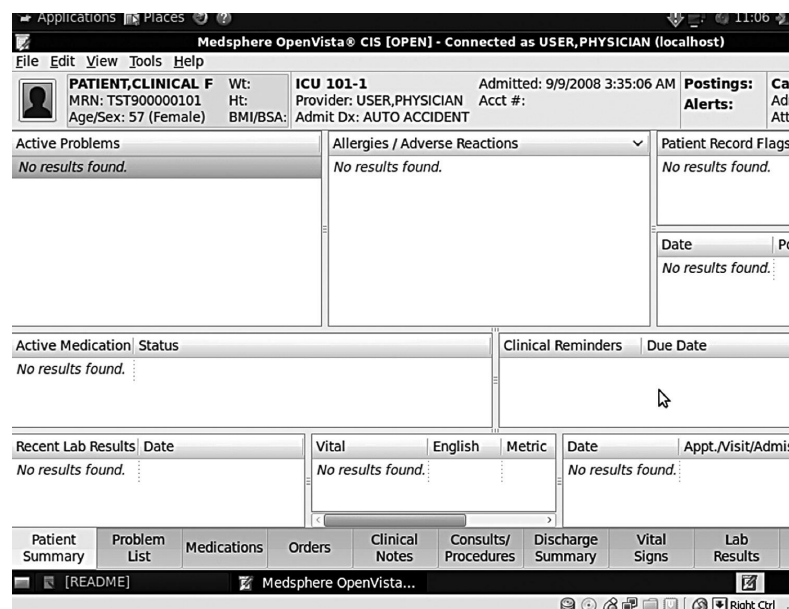


Figure 3.2 The Veterans Information Systems and Technology Architecture (VistA) used by the US Department of Veterans Affairs.

Although these integrated software systems, which are commonly in use at all medical facilities, have effectively automated the ADT list, it is still important for a professional in this field to be able to compute it by hand. In fact, it is critical to understand the statistical processes behind what the automated systems produce because it is the only way we can check behind the system, perhaps detect gross anomalies, and really understand what is taking place.

Therefore, in this chapter and throughout the text, you will learn how to calculate healthcare statistics by hand using formulas and software tools to assist in visualizing the results of your findings.

Calculation and Reporting of Patient Census Data

Two types of patient data that you will need to learn to calculate and report are inpatient service days and the average daily census. Note that although outpatient data are typically not gathered during a census, this information must still be captured so that it can be included in the census report. For instance, consider that you have a patient who was admitted at 9:30 am on June 5 and discharged at 8:30 pm on June 5, the only outpatient in your unit on this day. This patient would not be included in a census taken at 12 midnight on June 5, as he would already be gone. However, he would still need to be counted in the census report. Thus, if you count 45 patients in your care unit during the census, then you would just add the one outpatient to this number, making the total number 46.

Inpatient Service Days

Inpatient service days are a way to measure services that a patient receives within a 24-hour time frame. Other terms for inpatient service day are *patient day*, *census*, *occupancy day*, or *inpatient day*. The *total inpatient service days* refers to all of the inpatient service days for a given period. Keep in mind that physicians and the hospital must ensure that the patient meets severity of illness and medical necessity

criteria before admitting the patient. As mentioned, data regarding any outpatients must be added to the inpatient service days and counted on the ADT report. Keep the following points in mind when calculating inpatient service days:

- The reporting period for inpatient service days begins at 12:01 am and ends at midnight.
- A leave of absence occurs when a patient who has been admitted but not yet discharged is not at the facility at the census-taking hour. Any absence of less than 1 day does not constitute a leave of absence. Normally, leaves of absence are not an everyday occurrence due to shorter length of stays in the healthcare facility. Leaves of absence are more common in long-term care and rehabilitation clinics and for those who are developmentally disabled. Note that patients are now allowed to leave without a physician's order stating the patient has permission to leave or return. If the administrators of a facility do not recognize a leave of absence, they may choose to discharge the patient and then readmit the patient after a few days.
- The day of discharge is not counted for inpatient service days, but the admission date is counted.
- Service days are not reported as fractions or divided for a unit of service.
- If a patient is admitted at 3:00 pm and dies at 7:00 pm, that patient will not be counted in the inpatient service days.
- Administrators of facilities will choose how they want to count their inpatient service days, which could be monthly, quarterly, semiannually, or annually.
- Some facilities may choose to exclude obstetrical and intensive care units due to the moderation of services administered.

One important use of inpatient service days is to measure how well the facility is performing year-to-date and do a comparison with the previous year's performance (**Figure 3.3**). This measure also assists the hospitals and physicians in determining the quality of medical care they are providing each patient and also delivering information about the overall health of patients.

Day	12:01 a.m census		ADM		trf in	Total	DIS	DIS	trf out	11:59 p.m. census		a/d	A/C	serv days	
	A/C	Nb	A/C	b						A/C	Nb				A/C
7-Dec	50	5	4	2	2	56	7	3	1	2	51	6	3	54	6

Figure 3.3 Example of hospital census. *Abbreviations:* A/C, adults and children; ADM, admissions; DIS, discharges; Nb, newborns; trf, transfers.

Hands-on Statistics 3.1: Calculating and Reporting Inpatient Service

Now we will begin our calculations for inpatient census:

1. The starting point, which was the last census, is 50 adults and children (A/C) and five 5 newborns (Nb).
2. Next, the current census is as follows: 50 A/C + 4 admissions + 2 transfers in = 56 A/C.
3. For Nb, the current census would be: 5 Nb + 2 births = 7 Nb.
4. Next, subtract from the census of A/C discharges and transfers out: 56 – 3 discharges (A/C) – 2 transfers out = 51 A/C.
5. Lastly, subtract from the census of Nb discharges: 7 – 1 discharge (Nb) = 6 Nb.

Let's review some topics for clarification:

- The terms *transfer in* and *transfer out* refer to changes within the hospital only. For example, a patient may be admitted to the surgical floor but the next day transferred to the **intensive care unit (ICU)**, the area of the hospital reserved for patients with severe illnesses or injuries that require constant monitoring. This simply means that the patient was transferred off of the surgical floor and into the ICU. These transfers will be included in the ADT report. Alternatively, the physician may say the patient was “transferred,” meaning the patient went to a rehabilitation facility. In this case, the patient is actually discharged from the facility but would still be counted in the ADT report. If the transfers are not counted, the data will be out of balance and a particular care unit will fail to report the transfers in, transfers out, or discharges correctly. Transfers in and out of the PCU will not always be equal, but they will be equal with the overall recapitulation. A **recapitulation algorithm** is one that verifies the data either monthly or yearly, which means it provides a summary of the data collected for a given period.
- For census purposes, newborns are counted separately. However, this is an administrative decision by medical staff or others who would be using the statistical data. Any birth in the facility is considered a newborn admission.
- Another question that might be asked regarding the inpatient census is, how many are in the house? The remainder of patients that are still admitted in the hospital after midnight or as of 11:59 pm are considered “in house.”

Average Daily Census

Average daily census is the average number of patients in a facility on a given day. It provides the administration with information on a specific unit: Are additional beds or services needed? Should new services be added for patient care? It informs even large-scale questions like, is construction of a particular care unit required? Staffing and budgeting of supplies and equipment are affected by average daily census results.

For each PCU in the healthcare facility, the following formula is used to calculate the average daily census:

$$\frac{\text{Total inpatient service days for a period (excluding newborns)}}{\text{Total number of days in the period}}$$

To calculate the census for a month, you need to know the number of days in each of the 12 months (and don't forget the leap year). Remember that adults and children are counted separately from newborns unless your facility directs you otherwise. For example, consider the following: Ocean View Hospital had a total of 5,321 inpatient service days for the month of January. Divide 5,321 by 31 days in January: $5,321/31 = 171.6$. Then round the answer up: 172.

Hands-on Statistics 3.2: Oceanside Hospital Case Study of Average Daily Census

Try your hand at calculating the average daily census in the following exercise:

1. Oceanside Hospital has a 20-bed ICU and has 635 inpatient service days for the month of October.
2. Divide 635 inpatient service days by 31 days in the month of October to get 20.4, which you will round to 20.
3. The census shows that the ICU is filled to capacity each day of the month since you only have 20 beds.
4. Administration will use this information to determine whether to add additional beds in that unit.

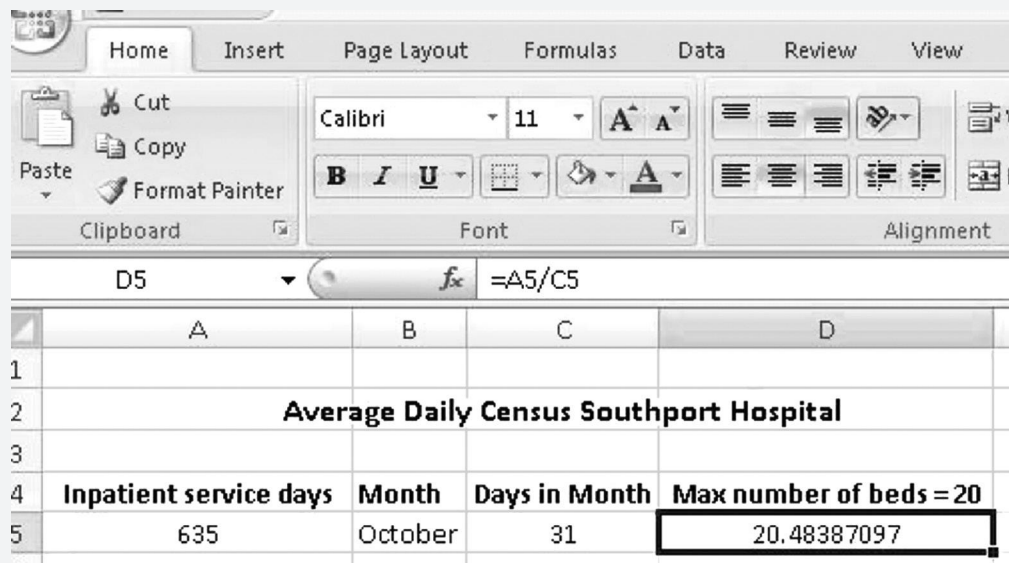
That was fun to work out by hand, but almost all computations are now done on a computer. So, now we will work out an average daily census with a spreadsheet application.

Hands-on Statistics 3.3: Southport Hospital Case Study of Average Daily Census Using Excel

To calculate the average daily census for Southport Hospital using Excel, examine **Figure 3.4** and follow these directions:

1. Open a new, blank spreadsheet in Excel.
2. In a cell in row 2 cell, type "**Average Daily Census Southport Hospital**" as a title in bold and 12-point font.
3. Type "Inpatient service days" in cell A4, "Month" in cell B4, "Days in Month" in cell C4, and "Max number of beds = 20" in cell D4, as in Figure 3.4.
4. Type "635" under *Inpatient service days* in cell A5, "October" under *Month* in cell B5, "31" under *Days in Month* in cell C5, and the formula " $=A5/C5$ " under *Max number of beds = 20* in cell D5.
5. Right click on D5, select *Format cells*, then *Number*, and then change the decimal places to zero (0), as you will not have a fractional bed available for use.
6. When you click out of cell D5, the average daily census should appear: 20.

Save this document, as you will add to these data in the next exercise.



	A	B	C	D
1				
2	Average Daily Census Southport Hospital			
3				
4	Inpatient service days	Month	Days in Month	Max number of beds = 20
5	635	October	31	20.48387097

Figure 3.4 Excel spreadsheet of average daily census.

Data Visualization

Data visualization lets us make better sense visually of complex data. We live in a society in which images are of paramount importance, so it stands to reason we should try to present our data graphically where it is appropriate to do so. David McCandless gives an eloquent description of this in his TED Talks video *The Beauty of Data Visualization*. The old proverb "A picture is worth a thousand words" certainly applies to data visualization; however, in a clinical or scholarly setting, where data are being summarized and reported, both words and graphics work in harmony to represent culminated data accurately to consumers.

To properly choose the right data visualization tool, you must first understand the audience you will be presenting to, as real-world data can be complex and overwhelming for the average person. There are three steps for providing great data visualization: First, know your target audience. Second, make a clear framework for the information you are displaying. Third, make sure it tells a story. Some of the more common forms of data visualization are line graphs, flowcharts, diagrams, pie charts, bar graphs, infographics, maps, cluster charts, and word clouds. These types of data visualization are used in large corporations and all other fields of business, including health care.

Hands-on Statistics 3.4: Chart Average Daily Census Data Using Excel

In this example, we will change a previous example of Southport Hospital to have 12 months of data. As each month has a differing number of days, we will adjust for that. Examine **Figure 3.5** and follow these directions:

	A	B	C	D
1				
2	Average Daily Census Southport Hospital			
3				
4	Inpatient service days	Month	Days in Month	Max number of beds = 20
5				
6	600	January	31	19
7	615	February	28	22
8	635	March	31	20
9	611	April	30	20
10	590	May	31	19
11	600	June	30	20
12	601	July	31	19
13	612	August	31	20
14	613	September	30	20
15	635	October	31	20
16	623	November	30	21
17	601	December	31	19

Figure 3.5 Average daily census: Step 1.

1. With the Excel spreadsheet you created in the previous exercise open, edit it and add the data shown in Figure 3.5 so that you have 12 full months of data.
2. Next, highlight cells B6 through B17, then hold the CTRL key down and also highlight cells D6 through D17. Select Insert, Bar Chart, and 2D Bar. You should now have a chart showing the average daily census for each month. Note **Figure 3.6**.

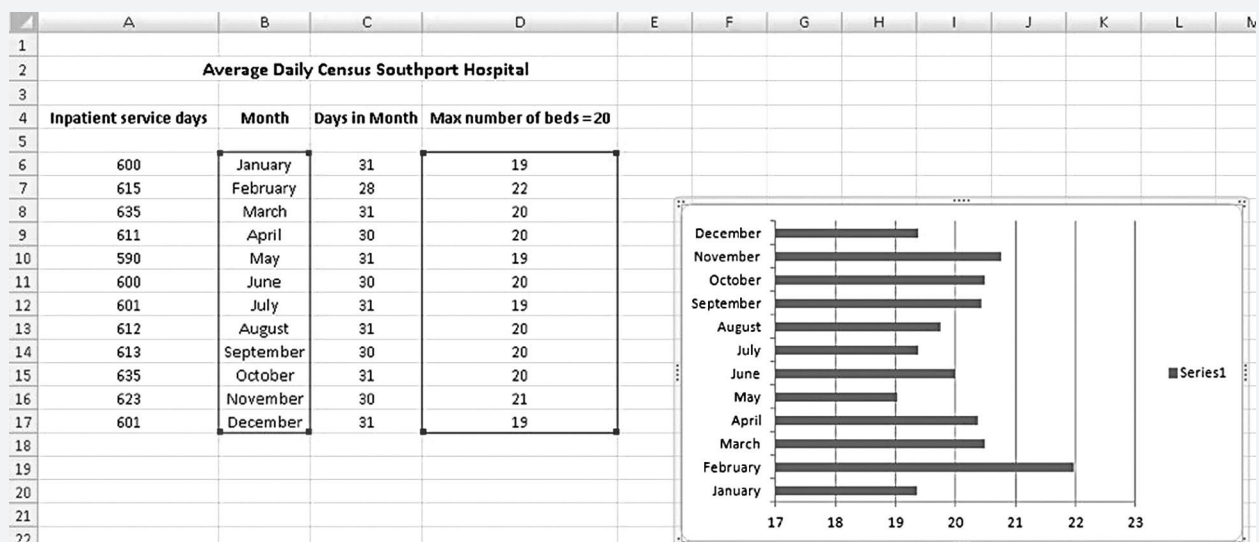


Figure 3.6 Average daily census: Step 2.

? DID YOU KNOW?

Since 45 BCE, and Julius Caesar's Julian Calendar, the number of days in each month has stayed the same. Many school children learn the days of each month with the following rhyme: 30 days hath September, April, June, and November, all the rest have 31, excepting February alone, and that has 28 days clear, with 29 in each leap year.

Now you know!

You now can visualize the usage by month for a calendar year. It might be handy to be able to see whether there are any trends that can be noticed over several years' time frame. In the next Hands-on Statistics exercise, let's assume you have computed the census by month for 3 years. You will learn how to use a line chart to examine whether there are any patterns that can be derived from the data that might not be as noticeable by just looking at the raw numbers.

Hands-on Statistics 3.5: Three-Year Census Data Represented by a Line Chart Using Excel

1. Create in Excel a spreadsheet like the one shown in **Figure 3.7**.

Inpatient Service Days			
Oceanside Hospital	85 Bed Hospital		
	Year 2010	Year 2011	Year 2012
Jan.	1823	1871	1856
Feb.	1856	1858	1839
March	1844	1851	1801
April	1855	1822	1818
May	1821	1830	1825
June	1833	1834	1844
July	1849	1829	1836
August	1837	1811	1827
September	1858	1826	1843
October	1844	1802	1826
November	1816	1799	1800
December	1841	1812	1822

Figure 3.7 Creating a line chart in Excel: Oceanside Hospital raw data.

2. Next, highlight all data, as shown in **Figure 3.8**.
3. Lastly, select *Insert, Line*, and then *Line with Markers*, as shown in **Figure 3.9**.
4. Note that there are definite recurring patterns over 3 years in September through December, as shown in **Figure 3.10**.

(continues)

Inpatient Service Days			
Oceanside Hospital	85 Bed Hospital		
	Year 2010	Year 2011	Year 2012
Jan.	1823	1871	1856
Feb.	1856	1858	1839
March	1844	1851	1801
April	1855	1822	1818
May	1821	1830	1825
June	1833	1834	1844
July	1849	1829	1836
August	1837	1811	1827
September	1858	1826	1843
October	1844	1802	1826
November	1816	1799	1800
December	1841	1812	1822

Figure 3.8 Creating a line chart in Excel: Highlighting data.

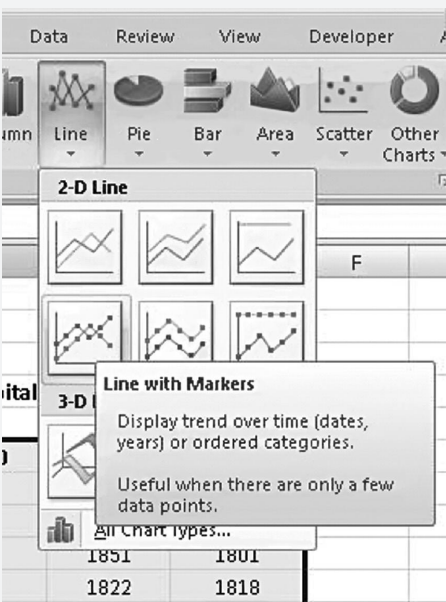


Figure 3.9 Creating a line chart in Excel: Choosing line with markers.

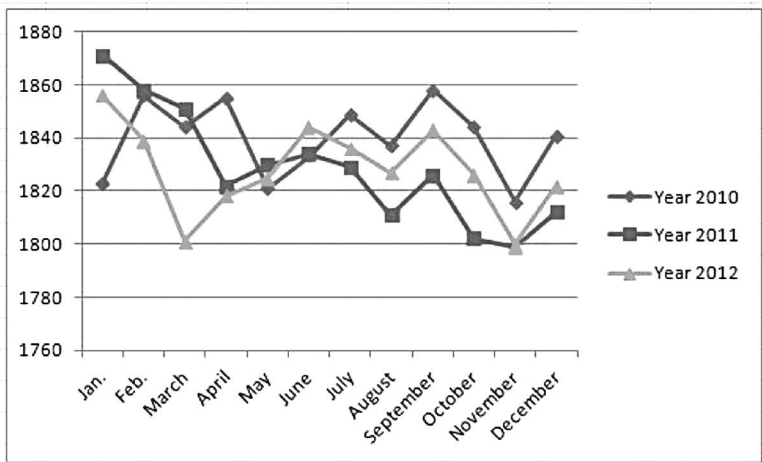


Figure 3.10 Creating a line chart in Excel: Line chart with markers.

Visually Examine Data With Sparklines and Microcharts

We have examined census data for each month as a total for each month. However, what if you wanted to examine census data for every day total in the month? Perhaps you would notice some trends, such as usage increasing typically on Saturdays or Mondays? That might be useful information to justify adding another staff member on those days. Remember, administration as a general rule likes

decisions that are driven by data, and this would be one way you could make data-driven staffing decisions.

Sparklines and microcharts are examples of **infographics**, or graphical representations or charts of data or knowledge, in this case of cells with numeric data; think of them as an electroencephalographic readout of your data. Edward Tufte invented **sparklines** as a way to display data in small graphics. He referred to them as “intense, simple, word-sized graphics,” which is a great way to think about them.

If you have examined any popular stock market site, you might also notice the use of this tiny but powerful infographics tool.

Microcharts, which are known by several different terms, and which are the subject of a patent

dispute involving Microsoft over ownership, are similar graphic representations of data. Other terms you may hear for this tool include *in-cell micro charts* and *microcharting*. Try the next Hands-on exercise to see how easy they both are to use.

Hands-on Statistics 3.6: Average Daily Census for Each Day in a Month With Sparklines and Microcharts Using Excel

Practice creating microcharts and a sparkline for a month's worth of average daily census data. For microcharting, we use the rept function. The formula for rept is as follows:

=REPT("character",# times to repeat)

In the following example, a repeating character will represent the count for the census for each day, noted in a cell next to the numeric value visually represented by the character.

Before beginning the exercise, download and install the proper version (based on your version of Office) of Sparklines from the following website: <http://sparklines-excel.blogspot.com/>. The software is free. Enable macros when asked. Examine **Figures 3.11** through **3.13** and follow these directions:

1. Open a new, blank spreadsheet in Excel.
2. Type in the title "January 2010 Daily Census Report" in cell A4, "Day" in cell B5, and "Census" in cell C5, as in Figure 3.11.

Clipboard Font Alignment Number			
D6 =REPT("*",C6)			
A	B	C	D
1			
2			
3			
4	January 2010 Daily Census Report		
	Day	Census	
6	1	55	*****
7	2	62	*****
8	3	60	*****
9	4	66	*****
10	5	60	*****
11	6	56	*****
12	7	51	*****
13	8	56	*****
14	9	63	*****
15	10	55	*****
16	11	61	*****
17	12	56	*****
18	13	55	*****
19	14	63	*****
20	15	57	*****
21	16	53	*****
22	17	62	*****
23	18	61	*****
24	19	64	*****
25	20	62	*****
26	21	57	*****
27	22	60	*****
28	23	56	*****
29	24	61	*****
30	25	62	*****
31	26	60	*****
32	27	58	*****
33	28	62	*****
34	29	59	*****
35	30	57	*****
36	31	53	*****

Figure 3.11 Microcharts.

(continues)

- 3. Type in the dates for January 2010 under “Day” in column B and the census number for each date in column C, as indicated in Figure 3.11.
- 4. Type the formula “=REPT(“*”, C6)” in cell D6.
- 5. To duplicate this formula for the remaining days, highlight cell D6 and drag its contents down to cell D36.

Now that you have a spreadsheet like what you see in Figure 3.11, add a Sparkline for the month.

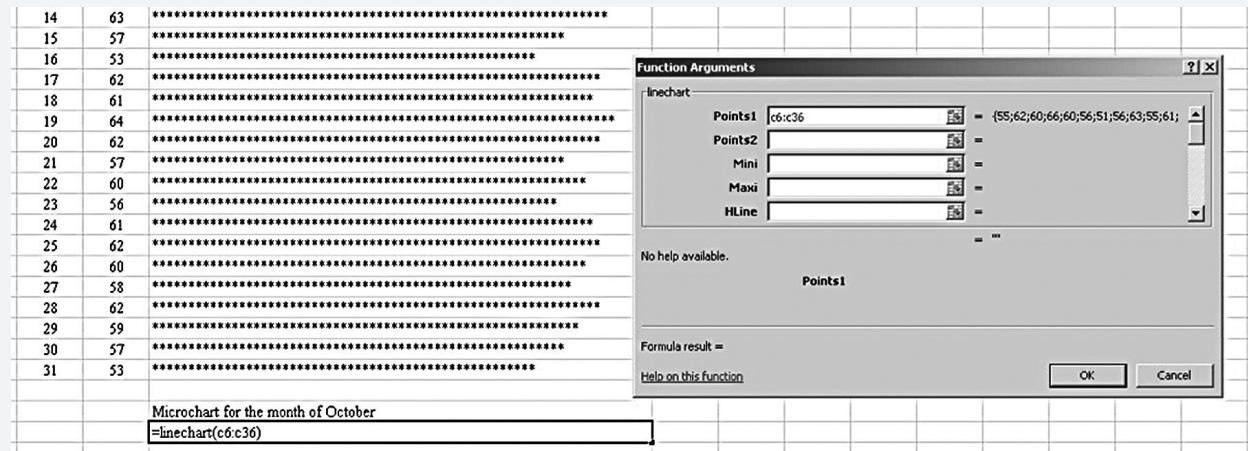


Figure 3.12 Sparkline: Selecting range.

- 6. Type “Sparkline for the month of January” in cell D38.
- 7. Click in cell D39, and then go to the *Add-Ins* menu; select *Sparklines* and then *Line chart*. Note that Microsoft often moves items around in different versions of their software, so depending on which version of Excel you have, you may have a different location for certain features.
- 8. Key in “c6:c36” in the *Points1* field in the pop-up box that appears, as that is your data range to chart, as shown in Figure 3.12.
- 9. Click *OK* for this output and you are done! [See Figure 3.13.]

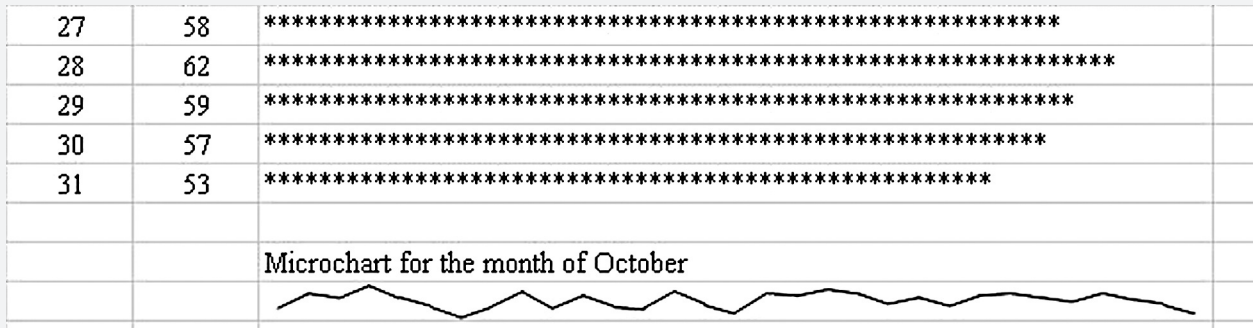


Figure 3.13 Sparkline: Output.

Now you have a microchart for each census day count, as well as a sparkline for the whole month of January.