Essentials of biostatistics in public health free pdf download version 1

Biostatistics is the application of statistical reasoning to the life sciences, and it is the key to unlocking the data gathered by researchers and the evidence presented in the scientific literature. In this course, we'll focus on the use of statistical measurement methods within the world of public health research. Along the way, you'll be introduced to a variety of methods and measures, and you'll practice interpreting data and performing calculations on real data from published studies. Topics include summary measures, visual displays, continuous data, sample size, the normal distribution, binary data, the element of time, and the Kaplan-Meir curve. YUMPU automatically turns print PDFs into web optimized ePapers that Google loves. Copy Download Essentials of Biostatistics in Public Health By Lisa M. Sullivan Extended embed settings VDOC.PUB Authors: Lisa M. Sullivan PDF Download Embed This document was uploaded by our user. The uploader already confirmed that they had the permission to publish it. If you are author/publisher or own the copyright of this documents, please report to us by using this DMCA report form. Report DMCA Essentials of Biostatistics in Public Health Third Edition Lisa M. Sullivan, PhD Professor of Biostatistics Associate Dean for Education Boston University School of Public Health Boston, Massachusetts World Headquarters Jones & Bartlett Learning 5 Wall Street Burlington, MA 01803 978-443-5000 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, p rofessional 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 Copyright © 2018 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. Essentials of Biostatistics in Public Health, Third 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. The author, editor, and publisher have made every effort to provide accurate information. However, they are not responsible for errors, omissions, or for any outcomes related to the use of the contents of this book and take no responsibility for the use of the products and procedures described. Treatments and side effects described in this book may not be applicable to all people; likewise, some people may require a dose or experience a side effect that is not described herein. Drugs and medical devices are discussed that may have limited availability controlled by the Food and Drug Administration (FDA) for use only in a research study or clinical trial. Research, clinical practice, and government regulations often change the accepted standard in this field. When consideration is being given to use of any drug in the clinical setting, the health care provider or reader is responsible for determining FDA status of the drug, reading the package insert, and reviewing prescribing information for the most up-to-date recommendations on dose, precautions, and contraindications, and determining the appropriate usage for the product. This is especially important in the case of drugs that are new or seldom used. 18513-3 Production Credits VP, Executive Publisher: David D. Cella Publisher: Michael Brown Associate Editor: Lindsey M. Sousa Senior Production Editor: Amanda Clerkin Senior Marketing Manager: Sophie Fleck Teague Manufacturing and Inventory Control Supervisor: Amy Bacus Library of Congress Cataloging-in-Publication Data Names: Sullivan, Lisa M. (Lisa Marie), 1961-author. Title: Essentials of biostatistics in public health / Lisa Sullivan. Description: Third edition. | Burlington, Massachusetts : Jones & Bartlett Learning, [2018] | Includes bibliographical references and index. Identifiers: LCCN 2016048158 | ISBN 9781284108194 Subjects: | MESH: Biostatistics—methods | Biometry—methods | Public Health—methods Classification: LCC RA409 | NLM WA 950 | DDC 610.72—dc23 LC record available at 6048 Printed in the United States of America 21 20 19 18 17 10 9 8 7 6 5 4 3 2 1 Composition: Integra Software Services Pvt. Ltd. Cover Design: Scott Moden Rights & Media Specialist: Wes DeShano Media Development Editor: Shannon Sheehan Cover Image: © Sergey Nivens/Shutterstock Printing and Binding: Edwards Brothers Malloy Cover Printing: Edwards Brothers Malloy Contents Acknowledgments vii Preface ix Prologue xi About the Author xiii Chapter 1 Introduction 1 Chapter 2 Study Designs 7 1.1 What Is Biostatistics? 1.2 What Are the Issues? 1.3 Summary 2 2 4 2.1 Vocabulary 2.2 Observational Study Designs 2.3 Randomized Study Designs 2.4 The Framingham Heart Study 2.5 More on Clinical Trials 2.6 Sample Size Implications 2.7 Summary 2.8 Practice Problems 8 9 14 17 18 20 20 20 Chapter 3 Quantifying the Extent of Disease 23 3.1 Prevalence 3.2 Incidence 3.3 Relationships Between Prevalence and Incidence 3.4 Comparing the Extent of Disease Between Groups 3.5 Summary 3.6 Practice Problems 24 24 27 27 30 30 iv Contents Chapter 4 Summarizing Data Collected in the Sample 35 Chapter 5 The Role of Probability 67 4.1 4.2 4.3 4.4 4.5 Dichotomous Variables Ordinal and Categorical Variables Continuous Variables Summary Practice Problems 5.1 Sampling 5.2 Basic Concepts 5.3 Conditional Probability 5.4 Independence 5.5 Bayes' Theorem 5.6 Probability Models 5.7 Summary 5.8 Practice Problems Chapter 6 Confidence Interval Estimates 6.1 Introduction to Estimation 6.2 Confidence Intervals for One Sample, Continuous Outcome 6.3 Confidence Intervals for One Sample, Dichotomous Outcome 6.4 Confidence Intervals for Two Independent Samples, Continuous Outcome 6.5 Confidence Intervals for Matched Samples, Continuous Outcome 6.6 Confidence Intervals for Two Independent Samples, Dichotomous Outcome 6.7 Summary 6.8 Practice Problems Chapter 7 Hypothesis Testing Procedures 7.1 Introduction to Hypothesis Testing 7.2 Tests with One Sample, Continuous Outcome 7.3 Tests with One Sample, Dichotomous Outcome 7.4 Tests with One Sample, Categorical and Ordinal Outcomes 7.5 Tests with Two Independent Samples, Continuous Outcome 7.6 Tests with Matched Samples, Continuous Outcome 7.7 Tests with Two Independent Samples, Dichotomous Outcome 7.8 Tests with More Than Two Independent Samples, Continuous Outcome 7.9 Tests for Two or More Independent Samples, Categorical and Ordinal Outcomes 7.10 Summary 7.11 Practice Problems 37 41 50 61 63 68 70 71 74 74 75 96 97 101 102 104 105 107 110 114 119 120 123 124 131 133 135 139 142 145 149 157 163 164 Contents Chapter 8 Power and Sample Size Determination 171 Chapter 9 Multivariable Methods 193 Chapter 10 Nonparametric Tests 223 Chapter 11 Survival Analysis 249 Chapter 12 Data Visualization 279 8.1 Issues in Estimating Sample Size for Confidence Intervals Estimates 8.2 Issues in Estimating Sample Size for Hypothesis Testing 8.3 Summary 8.4 Practice Problems 9.1 Confounding and Effect Modification 9.2 The Cochran–Mantel–Haenszel Method 9.3 Introduction to Correlation and Regression Analysis 9.4 Multiple Linear Regression Analysis 9.5 Multiple Logistic Regression Analysis 9.6 Summary 9.7 Practice Problems 10.1 Introduction to Nonparametric Testing 10.2 Tests with Two Independent Samples 10.3 Tests with Matched Samples 10.4 Tests with More Than Two Independent Samples 10.5 Summary 10.6 Practice Problems 11.1 Introduction to Survival Data 11.2 Estimating the Survival Function 11.3 Comparing Survival Curves 11.4 Cox Proportional Hazards Regression Analysis 11.5 Extensions 11.6 Summary 11.7 Practice Problems 12.1 Design Principles 12.2 When and How to Use Text, Tables, and Figures 12.3 Presenting Data and Statistical Results in Tables 12.4 Presenting Data and Statistical Results in Figures 12.5 Summary 12.6 Practice Problems 172 179 188 188 195 201 203 212 216 219 219 226 227 232 241 245 246 252 253 261 267 271 272 273 280 281 285 295 328 328 Appendix333 Glossary351 Index361 v Acknowledgments I am very grateful to Dr. Richard Riegelman for his unending energy and enthusiasm for this project. I thank my family for their encouragement; Casey, Ryan, Aislinn, Abbi, and Gavin for their inspiration; most especially, Kimberly Dukes and Kevin Green for their unending love and support. This textbook is dedicated to the memory of my cousin Catherine Render. Catherine was a remarkable woman. She was a loving daughter, sister, wife, mother, cousin, and friend who excelled in every role. She lost a long battle with breast cancer in January 2006. She fought through many setbacks, always pushing forward optimistically, taking care of everyone around her, never asking why. She was and always will be an inspiration to me and to so many others who were fortunate enough to know her. Biostatisticians play an extremely important role in addressing important medical and public health problems. Unfortunately, there are many more problems than solutions. We must never lose sight of the fact that our work is important in improving health and well-being. We need qualified biostatisticians to work in research teams to address problems like breast cancer, cardiovascular disease, diabetes, and so many others. Preface Essentials of Biostatistics in Public Health, Third Edition provides a fundamental and engaging background for students learning to apply and appropriately interpret biostatistical applications in the field of public health. The examples are real, important, and represent timely public health problems. The author aims to make the material relevant, practical, and engaging for students. Throughout the textbook, the author uses data from the Framingham Heart Study and from observational studies and clinical trials in a variety of major areas. The author presents example applications involving important risk factors—such as blood pressure, cholesterol, smoking, and diabetes and their relationships to incident cardiovascular and cerebrovascular disease—throughout. Clinical trials investigating new drugs to lower cholesterol, to reduce pain, and to promote healing following surgery are also considered. The author presents examples with relatively few subjects to illustrate computations while minimizing the actual computation time, as a particular focus is mastery of "by-hand" computations. All of the techniques are then applied to and illustrated on real data from the Framingham Heart Study and large observational studies and clinical trials. For each topic, the author discusses methodology—including assumptions, statistical computations, and the appropriate interpretation of results. Key formulas are summarized at the end of each chapter. Prologue Understanding how to present and interpret data is the foundation for evidence-based public health. It is essential for public health practitioners, future clinicians, and health researchers to know how to use data and how to avoid being deceived by data. In Essentials of Biostatistics in Public Health, Lisa Sullivan ably guides students through this maze. To do so, she uses an abundance of real and relevant examples drawn from her own experience working on the Framingham Heart Study and clinical trials. Essentials of Biostatistics in Public Health takes an intuitive, step-by-step, hands-on approach in walking students though statistical principles. It emphasizes understanding which questions to ask and knowing how to interpret statistical results appropriately. The third edition of Essentials of Biostatistics in Public Health builds upon the success of the previous editions in presenting state-of-the-art biostatistical methods that are widely used in public health and clinical research. A new chapter on data visualization provides important insights about how to produce and interpret data presented as tables, figures, and newer forms of data visualization. Dr. Sullivan provides information on both good and bad data presentations, and teaches the reader how to recognize the difference. Her recommendations are based on sound biostatical principles presented in a way that can be appreciated without extensive statistical background. The Third Edition also features a new series of integrative exercises based on data collected in the Framingham Heart Study. The data set includes real data on more than 4,000 participants and gives students an opportunity to "get their hands dirty" by using real data. In addition, the Third Edition includes a set of key questions for each chapter to engage students—that is, it adopts an inquiry-based approach to teaching and learning. Dr. Sullivan provides links to recent "in the news" articles to encourage students to "dig in" and to critically think about how to draw conclusions from data. The strategies used in Essentials of Biostatistics in Public Health represent a tried-and-true, classroom-tested approach. Lisa Sullivan has more than 2 decades of experience teaching biostatistics to both undergraduates and graduate students. As Assistant Dean for Undergraduate Programs in Public Health at Boston University, she has developed and taught undergraduate courses in biostatistics. She has also served as the chair of the Department of Biostatistics. Today she is the Associate Dean for Education at Boston University School of Public Health. Her background speaks to her unique ability to combine the skills of biostatistics with the skills of education. Dr. Sullivan has won numerous teaching awards for her skills and commitment to education in biostatistics, including the Association of Schools of Public Health Award for Teaching Excellence. She possesses a unique combination of sophisticated biostatistics expertise and a clear and engaging writing style—one that can draw students in and help them understand even the most difficult topic. Even a quick glance through Essentials of Biostatistics in Public Health will convince you of her skills in communication and education. xii Prologue I am delighted that Dr. Sullivan has included her book and workbook in our Essential Public Health series. There is no better book to recommend for the anxious student first confronting the field of biostatistics. Students will find the book and workbook engaging and relevant. Just take a look and see for yourself. Richard Riegelman, MD, MPH, PhD Editor, Essential Public Health series About the Author Lisa M. Sullivan has a PhD in statistics and is Professor and former Chair of the Department of Biostatistics at the Boston University School of Public Health. She is also Associate Dean for Education. She teaches biostatistics for MPH students and lectures in biostatistical methods for clinical researchers. From 2003 to 2015, Lisa was the principal investigator of the National Heart, Lung, and Blood Institute's Summer Institute for Training in Biostatistics, which was designed to promote interest in the field of biostatistics and to expose students to the many exciting career opportunities available to them. Lisa is the recipient of numerous teaching awards, including the Norman A. Scotch Award and the prestigious Metcalf Award, both for excellence in teaching at Boston University. In 2008 she won the Association of Schools of Public Health /Pfizer Excellence in Teaching Award. In 2011, she won the American Statistical Association's Section on Teaching Statistics in the Health Sciences Outstanding Teaching Award. In 2013, she won the Mostellar Statistician of the Year Award, presented by the Boston Chapter of the American Statistical Association. Also in 2013, she won the Massachusetts ACE National Network of Women Leaders Leadership Award. Lisa is also a biostatistician on the Framingham Heart Study, working primarily on developing and disseminating cardiovascular risk functions. She is active in several large-scale epidemiological studies for adverse pregnancy outcomes. Her work has resulted in more than 200 peer-reviewed publications. © Sergey Nivens/Shutterstock 1 chapter Introduction L earning O bjectives By the end of this chapter, the reader will be able to •• •• •• •• •• Define biostatistical applications and their objectives Explain the limitations of biostatistical analysis Compare and contrast a population and a sample Explain the importance of random sampling Develop research questions and select appropriate outcome variables to address important public health problems •• Identify the general principles and explain the role and importance of biostatistical analysis in medical, public health, and biological research Biostatistics is central to public health education and practice; it includes a set of principles and techniques that allows us to draw meaningful conclusions from information or data. Implementing and understanding biostatistical applications is a combination of art and science. Appropriately understanding statistics is important both professionally and personally, as we are faced with statistics every day. For example, cardiovascular disease is the number one killer of men and women in the United States. The American Heart Association reports that more than 2600 Americans die every day of cardiovascular disease, which is approximately one American every 34 seconds. There are over 70 million adults in the United States living with cardiovascular disease, and the annual rates of development are estimated at 7 cases per 1000 in men aged 35–44 years and 68 cases per 1000 in men aged 85–94 years.1 The rates in women are generally delayed about 10 years as compared to men.2 Researchers have identified a number of risk factors for cardiovascular disease including blood pressure, cholesterol, diabetes, smoking, and weight. Smoking and weight (specifically, overweight and -besity) are considered the most and s econd-most, respeco tively, preventable causes of cardiovascular disease death in the United States.3,4 Family history, nutrition, and physical activity are also important risk factors for cardiovascular d isease.5 The previous example describes cardiovascular disease, but similar statistics are available for many other diseases including cancer, diabetes, asthma, and arthritis. Much of what we know about cardiovascular and many other diseases comes from newspapers, news reports, or the Internet. Reporters d escribe or write about research studies on a daily basis. Nightly newscasts almost always contain a report of at least one research study. The results from some studies seem quite obvious, such as the positive effects of exercise on health, whereas other studies describe breakthrough medications that cure disease or prolong a healthy life. Newsworthy topics can include conflicting or contradictory results in medical r esearch. One study might report that a new medical therapy is effective, whereas another study might suggest this new therapy is ineffectual; other studies may show vitamin supplements thought to be effective as being ineffective or even harmful. One study might demonstrate the effectiveness of a drug, and years later it is determined to be harmful due to some serious side effect. To understand and interpret these results requires knowledge of statistical principles and statistical thinking. How are these studies conducted in the first place? For example, how is the extent of disease in a group or region quantified? How is the rate of development of new disease estimated? How are risk factors or characteristics that might be related to development or progression of disease identified? How is the effectiveness of a new drug determined? What could explain contradictory results? These questions are the essence of biostatistics. 2 CHAPTER 1 Introduction 1.1 WHAT IS BIOSTATISTICS? Biostatistics is defined as the application of statistical principles in medicine, public health, or biology. Statistical principles are based in applied mathematics and include tools and techniques for collecting information or data and then summarizing, analyzing, and interpreting those results. These principles extend to making inferences and drawing conclusions that appropriately take uncertainty into account. Biostatistical techniques can be used to address each of the aforementioned questions. In applied biostatistics, the objective is usually to make an inference about a specific population. By definition, this population is the collection of all individuals about whom we would like to make a statement. The population of interest might be all adults living in the United States or all adults living in the city of Boston. The definition of the population depends on the investigator's study question, which is the objective of the analysis. Suppose the population of interest is all adults living in the United States and we want to estimate the proportion of all adults with cardiovascular disease. To answer this question completely, we would examine every adult in the United States and assess whether they have cardiovascular disease. This would be an impossible task! A better and more realistic option would be to use a statistical analysis to estimate the d esired proportion. In biostatistics, we study samples or subsets of the population of interest. In this example, we select a sample of adults living in the United States and assess whether each has cardiovascular disease or not. If the sample is representative of the population, then the proportion of adults in the sample with cardiovascular disease should be a good estimate of the proportion of adults in the population with cardiovascular disease. In biostatistics, we analyze samples and then make inferences about the population based on the analysis of the sample. This inference is quite a leap, especially if the population is large (e.g., the United States population of 300 million) and the sample is relatively small (for example, 5000 people). When we listen to news reports or read about studies, we often think about how results might apply to us personally. The vast majority of us have never been involved in a research study. We often wonder if we should believe results of research studies when we, or anyone we know, never participated in those studies. 1.2 WHAT ARE THE ISSUES? Appropriately conducting and interpreting biostatistical applications require attention to a number of important issues. These include, but are not limited to, the following: • Clearly defining the objective or research question • Choosing an appropriate study design (i.e., the way in which data are collected) • Selecting a representative sample, and ensuring that the sample is of sufficient size • Carefully collecting and analyzing the data • Producing appropriate summary measures or statistics • Generating appropriate measures of effect or association • Quantifying uncertainty • Appropriately accounting for relationships among characteristics • Limiting inferences to the appropriate population In this book, each of the preceding points is addressed in turn. We describe how to collect and summarize data and how to make appropriate inferences. To achieve these, we use biostatistical principles that are grounded in mathematical and probability theory. A major goal is to understand and interpret a biostatistical analysis. Let us now revisit our original questions and think about some of the issues previously identified. How Is the Extent of Disease in a Group or Region Quantified? Ideally, a sample of individuals in the group or region of interest is selected. That sample should be sufficiently large so that the results of the analysis of the sample are adequately precise. (We discuss techniques to determine the appropriate sample size for analysis in Chapter 8.) In general, a larger sample for analysis is preferable; however, we never want to sample more participants than are needed, for both financial and ethical reasons. The sample should also be representative of the population. For example, if the population is 60% women, ideally we would like the sample to be approximately 60% women. Once the sample is selected, each participant is assessed with regard to disease status. The proportion of the sample with disease is computed by taking the ratio of the number with disease to the total sample size. This proportion is an estimate of the proportion of the population with disease. Suppose the sample proportion is computed as 0.17 (i.e., 17% of those sampled have the disease). We estimate the proportion of the population with disease to be approximately 0.17 (or 17%). Because this is an e stimate based on one sample, we must account for uncertainty, and this is reflected in what is called a margin of error. This might result in our estimating the proportion of the population with disease to be anywhere from 0.13 to 0.21 (or 13% to 21%). This study would likely be conducted at a single point in time; this type of study is commonly referred to as a cross- sectional study. Our estimate of the extent of disease refers only to the period under study. It would be inappropriate to make inferences about the extent of disease at future points based on this study. If we had selected adults living in Boston as our population, it would also be inappropriate to infer that the extent What Are the Issues? of disease in other cities or in other parts of Massachusetts would be the same as that observed in a sample of Bostonians. The task of estimating the extent of disease in a region or group seems straightforward on the surface. However, there are many issues that complicate things. For example, where do we get a list of the population, how do we decide who is in the sample, how do we ensure that specific groups are represented (e.g., women) in the sample, and how do we find the people we identify for the sample and convince them to participate? All of these questions must be addressed correctly to yield valid data and correct inferences. How Is the Rate of Development of a New Disease Estimated? To estimate the rate of development of a new disease—say, cardiovascular disease—we need a specific sampling strategy. For this analysis, we would sample only persons free of cardiovascular disease and follow them prospectively (going forward) in time to assess the development of the disease. A key issue in these types of studies is the follow-up period; the investigator must decide whether to follow participants for either 1, 5, or 10 years, or some other period, for the development of the disease. If it is of interest to estimate the development of disease over 10 years, it requires following each participant in the sample over 10 years to determine their disease status. The ratio of the number of new cases of disease to the total sample size reflects the proportion or c umulative incidence of new disease over the predetermined follow-up period. Suppose we follow each of the participants in our sample for 5 years and find that 2.4% develop disease. Again, it is generally of interest to provide a range of plausible values for the proportion of new cases of disease; this is achieved by incorporating a margin of error to reflect the precision in our estimate. Incorporating the margin of error might result in an estimate of the cumulative incidence of disease anywhere from 1.2% to 3.6% over 5 years. Epidemiology is a field of study focused on the study of health and illness in human populations, patterns of health or disease, and the factors that influence these patterns. The study described here is an example of an epidemiological study. Readers interested in learning more about epidemiology should see Magnus.6 How Are Risk Factors or Characteristics That Might Be Related to the Development or Progression of Disease Identified? Suppose we hypothesize that a particular risk factor or exposure is related to the development of a disease. There are several different study designs or ways in which we might 3 collect information to assess the relationship between a potential risk factor and disease onset. The most appropriate study design depends, among other things, on the distribution of both the risk factor and the outcome in the population of interest (e.g., how many participants are likely to have a particular risk factor or not). (We discuss different study designs in Chapter 2 and which design is optimal in a specific situation.) Regardless of the specific design used, both the risk factor and the outcome must be measured on each member of the sample. If we are interested in the relationship between the risk factor and the development of disease, we would again involve participants free of disease at the study's start and follow all participants for the development of disease. To assess whether there is a relationship between a risk factor and the outcome, we estimate the proportion (or percentage) of participants with the risk factor who go on to develop disease and compare that to the proportion (or percentage) of participants who do not have the risk factor and go on to develop disease. There are several ways to make this comparison; it can be based on a difference in proportions or a ratio of proportions. (The details of these comparisons are discussed extensively in Chapter 6 and Chapter 7.) Suppose that among those with the risk factor, 12% develop disease during the follow-up period, and among those free of the risk factor, 6% develop disease. The ratio of the proportions is called a relative risk and here it is equal to 0.12 / 0.06 = 2.0. The interpretation is that twice as many people with the risk f actor develop disease as compared to people without the risk factor. The issue then is to determine whether this estimate, observed in one study sample, reflects an increased risk in the population. Accounting for uncertainty might result in an estimate of the relative risk anywhere from 1.1 to 3.2 times higher for persons with the risk factor. Because the range contains risk values greater than 1, the data reflect an increased risk (because a value of 1 suggests no increased risk). Another issue in assessing the relationship between a particular risk factor and disease status involves understanding complex relationships among risk factors. Persons with the risk factor might be different from persons free of the risk factor; for example, they may be older and more likely to have other risk factors. There are methods that can be used to assess the association between the hypothesized risk factor and disease status while taking into account the impact of the other risk factors. These techniques involve statistical modeling. We discuss how these models are developed and, more importantly, how results are interpreted in Chapter 9. How Is the Effectiveness of a New Drug Determined? The ideal study design from a statistical point of view is the randomized controlled trial or the clinical trial. (The term 4 CHAPTER 1 Introduction c linical means that the study involves people.) For example, suppose we want to assess the effectiveness of a new drug designed to lower cholesterol. Most clinical trials involve specific inclusion and exclusion criteria. For example, we might want to include only persons with total cholesterol levels e xceeding 200 or 220, because the new medication would likely have the best chance to show an effect in persons with elevated cholesterol levels. We might also exclude persons with a history of cardiovascular disease. Once the inclusion and exclusion criteria are determined, we recruit participants. Each participant is randomly assigned to receive either the new experimental drug or a control drug. The randomization component is the key feature in these studies. Randomization theoretically promotes balance between the comparison groups. The control drug could be a placebo (an inert substance) or a cholesterol-lowering medication that is considered the current standard of care. The choice of the appropriate comparator depends on the nature of the disease. For example, with a life-threatening disease, it would be unethical to withhold treatment; thus a placebo comparator would never be appropriate. In this example, a placebo might be appropriate as long as participants' cholesterol levels were not so high as to necessitate treatment. When participants are enrolled and randomized to receive either the experimental treatment or the comparator, they are not told to which treatment they are assigned. This is called blinding or masking. Participants are then instructed on proper dosing and after a predetermined time, cholesterol levels are measured and compared between groups. (Again, there are several ways to make the comparison and we will discuss different options in Chapter 6 and Chapter 7.) Because participants are randomly assigned to treatment groups, the groups should be comparable on all characteristics except the treatment received. If we find that the cholesterol levels are different between groups, the difference can likely be attributed to treatment. Again, we must interpret the observed difference after accounting for chance or uncertainty. If we observe a large difference in cholesterol levels between participants receiving the experimental drug and the comparator, we can infer that the experimental drug is effective. However, inferences about the effect of the drug are only able to be generalized to the population from which participants are drawn—specifically, to the population defined by the inclusion and exclusion criteria. Clinical trials must be carefully designed and analyzed. There exist a number of issues that are specific to clinical trials, and we discuss these in detail in Chapter 2. Clinical trials are discussed extensively in the news, particularly recently. They are heavily regulated in the United States by the Food and Drug Administration (FDA).7 Recent news reports discuss studies involving drugs that were granted approval for specific indications and later removed from the market due to safety concerns. We review these studies and assess how they were conducted and, more important, why they are being reevaluated. For evaluating drugs, randomized controlled trials are considered the gold standard. Still, they can lead to controversy. Studies other than clinical trials are less ideal and are often more controversial. What Could Explain Contradictory Results Between Different Studies of the Same Disease? All statistical studies are based on analyzing a sample from the population of interest. Sometimes, studies are not designed appropriately and results may therefore be questionable. Sometimes, too few participants are enrolled, which could lead to imprecise and even inaccurate results. There are also instances where studies are designed appropriately, yet two different replications produce different results. Throughout this book, we will discuss how and when this might occur. 1.3 SUMMARY In this book, we investigate in detail each of the issues raised in this chapter. Understanding biostatistical principles is critical to public health education. Our approach will be through active learning: examples are taken from the Framingham Heart Study and from clinical trials, and used throughout the book to illustrate concepts. Example applications involving important risk factors such as blood pressure, cholesterol, smoking, and diabetes and their relationships to incident c ardiovascular and cerebrovascular disease are discussed. Examples with relatively few subjects help to illustrate computations while minimizing the actual computation time; a particular focus is mastery of "by-hand" computations. All of the techniques are then applied to real data from the Framingham study and from clinical trials. For each topic, we discuss methodology— including assumptions, statistical formulas, and the appropriate interpretation of results. Key formulas are summarized at the end of each chapter. Examples are selected to represent important and timely public health problems. REFERENCES 1. American Heart Association. Available at .org. 2. Sytkowski, P.A., D'Agostino, R.B., Belanger, A., and Kannel, W.B. "Sex and time trends in cardiovascular disease incidence and mortality: The Framingham Heart Study, 1950–1989." American Journal of Epidemiology 1996; 143(4): 338–350. References 3. W ilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., and Kannel, W.B. "Prediction of coronary heart disease using risk factor categories." Circulation 1998; 97: 1837– 1847. 4. The Expert Panel. "Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults: summary of the second report of the NCEP expert panel (Adult Treatment Panel II)." Journal of the American Medical Association 1993; 269: 3015–3023. 5 5. Kaikkonen, K.S., Kortelainen, M.L., Linna, E., and Huikuri, H.V. "Family history and the risk of sudden cardiac death as a manifestation of an acute coronary event." Circulation 2006; 114(4): 1462– 1467. 6. Magnus, M. Essentials of Infectious Disease Epidemiology. Sudbury, MA: Jones and Bartlett, 2007. 7. United States Food and Drug Administration. Available at http:// www.fda.gov. © Sergey Nivens/Shutterstock © holbox/Shutterstock 2 chapter Study Designs W hen and W hy Key Questions •• How do you know when the results of a study are credible? •• What makes a good study? How do you decide which kind of study is best? •• How do you know when something is a cause of something else? Dig In Choose any one of the studies mentioned previously and consider the following. •• What was the research question that investigators were asking? •• What was the outcome and how was it measured? What was the exposure or risk factor that they were trying to link to the outcome, and how was it measured? •• Is it appropriate to infer causality based on this study? In the News The following are recent headlines of reports summarizing investigations conducted on important public health issues. "Could going to college or being married give you brain cancer?" Sharon Begle of STAT News reports on a study of more than 4 million residents of Sweden and finds that people with 3 years of college or more had a 20% higher risk of developing glioma (a brain cancer) than those with elementary school-level education, as did married men compared to their unmarried counterparts.1 "Does Monsanto's Roundup herbicide cause cancer or not? The controversy explained." Sarah Zhang of Wired comments on conflicting reports from the United Nations and the World Health Organization about glyphosate (a weed killer) and whether it is carcinogenic.2 "Eating pasta does not cause obesity, Italian study finds." Tara John of Time reports that a new study on more than 20,000 Italians found that pasta consumption is not associated with obesity, but rather with a reduction in body mass index. The author does note that the study was partially funded by a pasta company, Barilla, and the Italian government.3 Begle S. STAT News. June 20, 2016. Available at .statnews.com/2016/06/20/brain-cancer-college-marriage/. 2 Zhang S. Wired. May 17, 2016. Available at .com/2016/05/monsantos-roundup-herbicide-cause-cancer-notcontroversy-explained/. 3 John T. Time. July 16, 2016. Available at .com/4393040/pasta-fat-obesity-body-mass-index-good/. 1 L earning O bjectives By the end of this chapter, the reader will be able to •• List and define the components of a good study design •• Compare and contrast observational and experimental study designs •• Summarize the advantages and disadvantages of alternative study designs •• Describe the key features of a randomized controlled trial •• Identify the study designs used in public health and medical studies Once a study objective or research question has been refined— which is no easy task, as it usually involves extensive discussion among investigators, a review of the literature, and an assessment of ethical and practical issues—the next step is to choose the study design to most effectively and efficiently answer the question. The study design is the methodology that is used to collect the information to address the research question. In Chapter 1, we raised a number of questions that might be of interest, including: How is the extent of a disease in a group or region quantified? How is the rate of development of a new disease estimated? How are risk factors or 8 CHAPTER 2 Study Designs c haracteristics that might be related to the development or progression of a disease identified? How is the effectiveness of a new drug determined? To answer each of these questions, a specific study design must be selected. In this chapter, we review a number of popular study d esigns. This review is not meant to be exhaustive but instead illustrative of some of the more popular designs for public health applications. The studies we present can probably be best organized into two broad types: observational and randomized studies. In observational studies, we generally observe a phenomenon, whereas in randomized studies, we intervene and measure a response. Observational studies are sometimes called descriptive or associational studies, nonrandomized, or historical studies. In some cases, observational studies are used to alert the medical community to a specific issue, whereas in other instances, observational studies are used to generate hypotheses. We later elaborate on other instances where observational studies are used to assess specific associations. Randomized studies are sometimes called analytic or experimental studies. They are used to test specific hypotheses or to evaluate the effect of an intervention (e.g., a behavioral or pharmacologic intervention). Another way to describe or distinguish study types is on the basis of the time sequence involved in data collection. Some studies are designed to collect information at a point in time, others to collect information on participants over time, and o thers to evaluate data that have already been collected. In biostatistical and epidemiological research studies, we are often interested in the association between a particular exposure or risk factor (e.g., alcohol use, smoking) and an outcome (e.g., cardiovascular disease, lung cancer). In the following sections, we discuss several observational study designs and several randomized study designs. We describe each design, detail its advantages and disadvantages, and distinguish designs by the time sequence involved. We then describe in some detail the Framingham Heart Study, which is an observational study and one of the world's most important studies of risk factors for cardiovascular disease.1 We then provide more detail on clinical trials, which are often considered the gold standard in terms of study design. At the end of this chapter, we summarize the issues in selecting the a ppropriate study design. Before describing the specific design types, we present some key vocabulary terms that are relevant to study design. 2.1 VOCABULARY •B ias—A systematic error that introduces uncertainty in estimates of effect or association •B lind/double blind—The state whereby a participant is unaware of his or her treatment status (e.g., experimental drug or placebo). A study is said to be double blind when both the participant and the outcome assessor are unaware of the treatment status (masking is used as an equivalent term to blinding). •C linical trial—A specific type of study involving human participants and randomization to the comparison groups •C ohort—A group of participants who usually share some common characteristics and who are monitored or followed over time •C oncurrent—At the same time; optimally, comparison treatments are evaluated concurrently or in parallel •C onfounding—Complex relationships among variables that can distort relationships between the risk factors and the outcome • Cross-sectional—At a single point in time • I ncidence (of disease)—The number of new cases (of disease) over a period of time • I ntention-to-treat—An analytic strategy whereby participants are analyzed in the treatment group they were assigned regardless of whether they followed the study procedures completely (e.g., regardless of whether they took all of the assigned medication) •M atching—A process of organizing comparison groups by similar characteristics •P er protocol—An analytic strategy whereby only participants who adhered to the study protocol (i.e., the specific procedures or treatments given to them) are analyzed (in other words, an analysis of only those assigned to a particular group who followed all procedures for that group) •P lacebo—An inert substance designed to look, feel, and taste like the active or experimental treatment (e.g., saline solution would be a suitable placebo for a clear, tasteless liquid medication) •P revalence (of disease)—The proportion of individuals with the condition (disease) at a single point in time •P rognostic factor—A characteristic that is strongly associated with an outcome (e.g., disease) such that it could be used to reasonably predict whether a person is likely to develop a disease or not •P rospective—A study in which information is collected looking forward in time •P rotocol—A step-by-step plan for a study that details every aspect of the study design and data collection plan • Quasi-experimental design—A design in which subjects are not randomly assigned to treatments •R andomization—A process by which participants are assigned to receive different treatments (this is usually based on a probability scheme) Observational Study Designs •R etrospective—A study in which information is collected looking backward in time • Stratification—A process whereby participants are partitioned or separated into mutually exclusive or nonoverlapping groups 2.2 OBSERVATIONAL STUDY DESIGNS There are a number of observational study designs. We describe some of the more popular designs, from the simplest to the more complex. 2.2.1 The Case Report/Case Series A case report is a very detailed report of the specific features of a particular participant or case. A case series is a systematic review of the interesting and common features of a small collection, or series, of cases. These types of studies are important in the medical field as they have historically served to identify new diseases. The case series does not include a control or comparison group (e.g., a series of disease-free p articipants). These studies are relatively easy to conduct but can be criticized as they are unplanned, uncontrolled, and not designed to answer a specific research question. They are often used to generate specific hypotheses, which are then tested with other, larger studies. An example of an important case s eries was one published in 1981 by Gottlieb et al., who reported on five young homosexual men who sought medical care with a rare form of pneumonia and other unusual infections.2 The initial report was followed by more series with similar presentations, and in 1982 the condition being described was termed Acquired Immune Deficiency Syndrome (AIDS). participants) to allow for estimates of risk factors, diseases, practices, or opinions in different subgroups of interest. However, a major limitation in cross-sectional surveys is the fact that both the exposure or development of a risk factor (e.g., hypertension) and the outcome have occurred. Because the study is conducted at a point in time (see Figure 2–1), it is not possible to assess temporal relationships, specifically whether the exposure or risk factor occurred prior to the outcome of interest. Another issue is related to non-response. While a large sample may be targeted, in some situations only a small fraction of participants approached agree to participate and complete the survey. Depending on the features of the participants and non-participants, non-response can introduce bias or limit generalizability. In Figure 2–1, approximately one-third of the participants have the risk factor and two-thirds do not. Among those with the risk factor, almost half have the disease, as compared to a much smaller fraction of those without the risk factor. Is there an association between the risk factor and the disease? 2.2.3 The Cohort Study A cohort study involves a group of individuals who usually meet a set of inclusion criteria at the start of the study. The cohort is followed and associations are made between a risk factor and a disease. For example, if we are studying risk factors for cardiovascular disease, we ideally enroll a cohort of individuals free of cardiovascular disease at the start of the study. In a prospective cohort study, participants are enrolled and followed going forward in time (see Figure 2–2). In some 2.2.2 The Cross-Sectional Survey A cross-sectional survey is a study conducted at a single point in time. The cross-sectional survey is an appropriate design when the research question is focused on the prevalence of a disease, a present practice, or an opinion. The study is non-randomized and involves a group of participants who are identified at a point in time, and information is collected at that point in time. Cross-sectional surveys are useful for -estimating the prevalence of specific risk factors or prevalence of disease at a point in time. In some instances, it is of interest to make comparisons between groups of participants (e.g., between men and women, between participants under age 40 and those 40 and older). However, inferences from the cross-sectional survey are limited to the time at which data are collected and do not generalize to future time points. Cross-sectional surveys can be easy to conduct, are usually ethical, and are often large in size (i.e., involve many 9 FIGURE 2–1 The Cross-Sectional Survey Risk Factor Present Risk Factor Absent Disease Disease 10 CHAPTER 2 Study Designs FIGURE 2–2 The Prospective Cohort Study Disease Risk Factor Present No Disease Disease Cohort Risk Factor Absent Study Start situations, the cohort is drawn from the general population, whereas in other situations a cohort is assembled. For example, when studying the association between a relatively common risk factor and an outcome, a cohort drawn from the general population will likely include sufficient numbers of individuals who have and do not have the risk factor of interest. When studying the association between a rare risk factor and an outcome, special attention must be paid to constructing the cohort. In this situation, investigators might want to enrich the cohort to include participants with the risk factor (sometimes called a special exposure cohort). In addition, an appropriate comparison cohort would be included. The comparison cohort would include participants free of the risk factor but similar to the exposed cohort in other important characteristics. In a retrospective cohort study, the exposure or risk factor status of the participants is ascertained retrospectively, or looking back in time (see Figure 2–3 and the time of study start). For No Disease Time e xample, s uppose we wish to assess the association between multivitamin use and neural tube defects in newborns. We enroll a cohort of women who deliver live-born infants and ask each to report on their use of multivitamins before becoming pregnant. On the basis of these reports, we have an exposed and unexposed cohort. We then assess the outcome of pregnancy for each woman. Retrospective cohort studies are often based on data gathered from medical records where risk factors and outcomes have occurred and been documented. A study is mounted and records are reviewed to assess risk factor and outcome status, both of which have already occurred. The prospective cohort study is the more common cohort study design. Cohort studies have a major a dvantage in that they allow investigators to assess temporal relationships. It is also possible to estimate the incidence of a disease (i.e., the rate at which participants who are free of a disease develop that disease). We can also compare incidence rates Observational Study Designs 11 FIGURE 2–3 The Retrospective Cohort Study Disease Risk Factor Present No Disease Disease Cohort Risk Factor Absent No Disease Study Start between groups. For example, we might compare the incidence of cardiovascular disease between participants who smoke and participants who do not smoke as a means of quantifying the association between smoking and cardiovascular disease. Cohort studies can be difficult if the outcome or disease under study is rare or if there is a long latency period (i.e., it takes a long time for the disease to develop or be realized). When the disease is rare, the cohort must be sufficiently large so that adequate numbers of events (cases of disease) are observed. By "adequate numbers," we mean specifically that there are sufficient numbers of events to produce stable, precise inferences employing meaningful statistical analyses. When the disease under study has a long latency period, the study must be long enough in duration so that sufficient numbers of events are observed. However, this can introduce another difficulty, namely loss of participant follow-up over a longer study period. Cohort studies can also be complicated by confounding. Confounding is a distortion of the effect of an exposure or risk factor on the outcome by other characteristics. For e xample, suppose we wish to assess the association Time etween smoking and cardiovascular disease. We may find b that smokers in our cohort are much more likely to develop cardiovascular disease. However, it may also be the case that the smokers are less likely to exercise, have higher cholesterol levels, and so on. These complex relationships among the variables must be reconciled by statistical analyses. In Chapter 9, we describe in detail the methods used to handle confounding. 2.2.4 The Case-Control Study The case-control study is a study often used in epidemiologic research where again the question of interest is whether there is an association between a particular risk factor or exposure and an outcome. Case-control studies are particularly useful when the outcome of interest is rare. As noted previously, cohort studies are not efficient when the outcome of interest is rare as they require large numbers of participants to be enrolled in the study to realize a sufficient number of outcome events. In a case-control study, participants are identified on the basis of their outcome status. Specifically, we select a set of cases, or persons with the outcome of interest. We then select 12 CHAPTER 2 Study Designs FIGURE 2–4 The Case-Control Study Risk Factor Present Disease (Cases) Risk Factor Absent Risk Factor Present No Disease (Controls) Risk Factor Absent Study Start Time Direction of Inquiry a set of controls, who are persons similar to the cases except for the fact that they are free of the outcome of interest. We then assess e xposure or risk factor status retrospectively (see Figure 2–4). We hypothesize that the exposure or risk factor is related to the disease and evaluate this by comparing the cases and controls with respect to the proportions that are exposed; that is, we draw inferences about the relationship between exposure or risk factor status and disease. There are a number of important issues that must be addressed in designing casecontrol studies. We detail some of the most important ones. First, cases must be selected very carefully. An explicit definition is needed to identify cases so that the cases are as homogeneous as possible. The explicit definition of a case must be established before any participants are selected or data collected. Diagnostic tests to confirm disease status should be included whenever possible to minimize the possibility of incorrect classification. Controls must also be selected carefully. The controls should be comparable to the cases in all respects except for the fact that they do not have the disease of interest. In fact, the controls should represent non-diseased participants who would have been included as cases if they had the disease. The same diagnostic tests used to confirm disease status in the cases should be applied to the controls to confirm non-disease status. Usually, there are many more controls available for inclusion in a study than cases, so it is often possible to select several controls for each case, thereby increasing the sample size for analysis. Investigators have shown that taking more than four controls for each case does not substantially improve the precision of the analysis.3 (This result will be discussed in subsequent chapters.) In many instances, two controls per case are selected, which is denoted as a 2:1 ("two to one") control to case ratio. Observational Study Designs The next issue is to assess exposure or risk factor status, and this is done retrospectively. Because the exposure or risk factor might have occurred long ago, studies that can establish risk factor status based on documentation or records are preferred over those that rely on a participant's memory of past events. Sometimes, such data are not documented, so participants are queried with regard to risk factor status. This must be done in a careful and consistent manner for all participants, regardless of their outcome status—assessment of exposure or risk factor status must be performed according to the same procedures or protocol for cases and controls. In addition, the individual collecting e xposure data should not be aware of the participant's outcome status (i.e., they should be blind to whether the participant is a case or a control). Case-control studies have several positive features. They are cost- and time-efficient for studying rare diseases. With case-control studies, an investigator can ensure that a sufficient number of cases are included. Case-control studies are also efficient when studying diseases with long latency periods. Because the study starts after the disease has been diagnosed, investigators are not waiting for the disease to occur during the study period. Case-control studies are also useful when there are several p otentially harmful exposures under consideration; data can be collected on each exposure and evaluated. The challenges of the case-control study center mainly around bias. We discuss several of the more common sources of bias here; there are still other sources of bias to consider. Misclassification bias can be an issue in case-control studies and refers to the incorrect classification of o utcome status (case or control) or the incorrect classification of exposure status. If misclassification occurs at random—meaning there is a similar extent of misclassification in both groups—then the association between the exposure and the outcome can be dampened (underestimated). If misclassification is not random—for example, if more cases are incorrectly classified as having the exposure or risk factor—then the association can be exaggerated (overestimated). Another source of bias is called selection bias, and it can result in a distortion of the association (over- or underestimation of the true association) between exposure and outcome status resulting from the selection of cases and controls. Specifically, the relationship between exposure status and disease may be different in those individuals who chose to participate in the study as compared to those who did not. Yet another source of bias is called recall bias, and again, it can result in a distortion of the association between exposure and outcome. It occurs when cases or controls differentially recall exposure status. It is possible 13 that persons with a disease (cases) might be more likely to recall prior exposures than persons free of the disease. The latter might not recall the same information as readily. With case-control studies, it is also not always possible to establish a temporal relationship between exposure and outcome. For example, in the present example both the exposure and outcome are measured at the time of data collection. Finally, because of the way we select participants (on the basis of their outcome status) in case-control studies, we cannot estimate incidence (i.e., the rate at which a disease develops). 2.2.5 The Nested Case-Control Study The nested case-control study is a specific type of case-control study that is usually designed from a cohort study. For example, suppose a cohort study involving 1000 participants is run to assess the relationship between smoking and cardiovascular disease. In the study, suppose that 20 participants d evelop myocardial infarction (MI, i.e., heart attack), and we are interested in assessing whether there is a relationship - between body mass index (measured as the ratio of weight in kilograms to height in meters squared) and MI. With so few participants suffering this very specific outcome, it would be difficult analytically to assess the relationship between body mass index and MI because there are a number of confounding factors that would need to be taken into account. This process generally requires large samples (specifics are discussed in Chapter 9). A nested case-control study could be designed to select suitable controls for the 20 cases that are similar to the cases except that they are free of MI. To facilitate the analysis, we would carefully select the controls and might match the controls to cases on gender, age, and other risk factors known to affect MI, such as blood pressure and cholesterol. Matching is one way of handling confounding. The analysis would then focus specifically on the association between body mass index and MI. Nested case-control studies are also used to assess new biomarkers (measures of biological processes) or to evaluate expensive tests or technologies. For example, suppose a large cohort study is run to assess risk factors for spontaneous preterm delivery. As part of the study, pregnant women provide demographic, medical, and behavioral information through self-administered questionnaires. In addition, each woman submits a blood sample at approximately 13 weeks gestation, and the samples are frozen and stored. Each woman is followed in the study through pregnancy outcome and is classified as having a spontaneous preterm delivery or not (e.g., induced preterm delivery, term delivery, etc.). A new test is developed to measure a hormone in the mother's blood that is hypothesized to be related to spontaneous preterm delivery. A nested case-control study is designed in which women who deliver 14 CHAPTER 2 Study Designs prematurely and spontaneously (cases) are matched to women who do not (controls) on the basis of maternal age, race/ethnicity, and prior history of premature delivery. The hormone is measured in each case and control using the new test applied to the stored (unfrozen) serum samples. The analysis is focused on the association between hormone levels and spontaneous preterm delivery. In this situation the nested case-control study is an efficient way to evaluate whether the risk factor (i.e., hormone) is related to the outcome (i.e., spontaneous preterm delivery). The new test is applied to only those women who are selected into the nested case-control study and not to every woman enrolled in the cohort, thereby reducing cost. 2.3 RANDOMIZED STUDY DESIGNS Cohort and case-control studies often address the question: Is there an association between a risk factor or exposure and an outcome (e.g., a disease)? Each of these observational study designs has its advantages and disadvantages. In the cohort studies, we compare incidence between the exposed and unexposed groups, whereas in the case-control study we compare exposure between those with and without a disease. These are different comparisons, but in both scenarios, we make inferences about associations. (In Chapter 6 and Chapter 7, we detail the statistical methods used to estimate associations and to make statistical inferences.) As we described, o bservational studies can be subject to bias and confounding. In contrast, randomized studies are considered to be the gold standard of study designs as they minimize bias and confounding. The key feature of randomized studies is the random assignment of participants to the comparison groups. In theory, randomizing makes the groups comparable in all respects except the way the participants are treated (e.g., treated with an experimental medication or a placebo, treated with a behavioral intervention or not). We describe two popular randomized designs in detail. 2.3.1 The Randomized Controlled Trial (RCT) or Clinical Trial The randomized controlled trial (RCT) is a design with a key and distinguishing feature—the randomization of participants to one of several comparison treatments or groups. In pharmaceutical trials, there are often two comparison groups; one group gets an experimental drug and the other a control drug. If ethically feasible, the control might be a placebo. If a placebo is not ethically feasible (e.g., it is ethically inappropriate to use a placebo because participants need medication), then a medication currently available and considered the standard of care is an appropriate comparator. This is called an active-controlled trial as opposed to a placebo-controlled trial. In clinical trials, data are collected prospectively (see Figure 2–5). The idea of randomization is to balance the groups in terms of known and unknown prognostic factors (i.e., characteristics that might affect the outcome), which minimizes confounding. Because of the randomization feature, the comparison groups— in theory—differ only in the treatment received. One group receives the experimental treatment and the other does not. With randomized studies, we can make much stronger inferences than we can with observational studies. Specifically, with clinical trials, inferences are made with regard to the effect of treatments on outcomes, whereas with observational studies, inferences are limited to associations between risk factors and outcomes. It is important in clinical trials that the comparison treatments are evaluated concurrently. In the study depicted in Figure 2–5, the treatments are administered at the same point in time, generating parallel comparison groups. Consider a clinical trial evaluating an experimental treatment for allergies. If the experimental treatment is given during the spring and the control is administered during the winter, we might see very different results simply because allergies are highly dependent on the season or the time of year. It is

also important in clinical trials to include multiple study centers, often referred to as multicenter trials. The reason for including multiple centers is to promote generalizability. If a clinical trial is conducted in a single center and the experimental treatment is shown to be effective, there may be a question as to whether the same benefit would be seen in other centers. In multicenter trials, the homogeneity of the effect across centers can be analyzed directly. Ideally, clinical trials should be double blind. Specifically, neither the investigator nor the participant should be aware of the treatment assignment. However, sometimes it is impossible or unethical to blind the participants. For example, consider a trial comparing a medical and a surgical procedure. In this situation, the participant would definitely know whether they underwent a surgical procedure. In some very rare situations, sham surgeries are performed, but these are highly unusual, as participant safety is always of the utmost concern. It is critical that the outcome assessor is blind to the treatment a ssignment. There are many ways to randomize participants in clinical trials. Simple randomization involves essentially flipping a coin and assigning each participant to either the experimental or the control treatment on the basis of the coin toss. In multicenter trials, separate randomization schedules are usually developed for each center. This ensures a balance in the t reatments within each center and does not allow for the possibility that all patients in one center get the same treatment. Sometimes it is important to minimize imbalance between groups with respect to other characteristics. For example, suppose we want to be sure we have participants of similar ages in each of the comparison groups. We could develop separate or stratified randomization schedules for participants less than 40 years of age and participants 40 years of age and older within each center. There are many ways to perform the randomization and the appropriate procedure depends on many factors, including the relationship between important prognostic factors and the outcome, the number of centers involved, and so on. The major advantage of the clinical trial is that it is the cleanest design from an analytic point of view. Randomization minimizes bias and confounding so, theoretically, any benefit (or harm) that is observed can be attributed to the treatment. However, clinical trials are often expensive and very time-consuming. Clinical trials designed around outcomes that are relatively rare require large numbers of participants to demonstrate a significant effect. This increases the time and cost of conducting the trial. There are often a number of challenges in clinical trials that must be faced. First, clinical trials can be ethically challenging. Choosing the appropriate control group requires careful assessment of ethical issues. For example, in cancer trials it would never be possible to use a placebo comparator, as this would put participants at unnecessary risk. Next, clinical trials can be difficult to set up. Recruitment of centers and participants can be difficult. For example, participants might not be willing to participate in a trial because they cannot accept the possibility of being randomly assigned to the control group. Careful monitoring of participants is also a crucial aspect of clinical trials. For example, investigators must be sure that participants are taking the assigned drug as planned and are not taking other medications that might interfere with the study medications (called concomitant medications). Most clinical trials require 16 CHAPTER 2 Study Designs frequent follow-up with participants—for example, every 2 weeks for 12 weeks. Investigators must work to minimize loss to follow-up to ensure that important study data are collected at every time point during the study. Subject retention and adherence to the study protocol are essential for the success of a clinical trial. In some clinical trials, there are very strict inclusion and exclusion criteria. For example, suppose we are evaluating a new medication hypothesized to lower cholesterol. To allow the medication its best chance to demonstrate benefit, we might include only participants with very high total cholesterol levels. This means that inferences about the effect of the medication would then be limited to the population from which the participants were drawn. Clinical trials are sometimes criticized for being too narrow or restrictive. In designing trials, investigators must weigh the impact of the inclusion and exclusion criteria on the observed effects and on their generalizability. Designing clinical trials can be very complex. There are a number of issues that need careful attention, including refining the study objective so that it is clear, concise, and answerable; determining the appropriate participants for the study (detailing inclusion and exclusion criteria explicitly); determining the appropriate outcome variable; deciding on the appropriate control group; developing and implementing a strict monitoring plan; determining the number of participants to enroll; and detailing the randomization plan. While achieving these goals is challenging, a successful randomized clinical trial is considered the best means of establishing the effectiveness of a medical procedure or treatment. 2.3.2 The Crossover Trial The crossover trial is a clinical trial where each participant is assigned to two or more treatments sequentially. When there are two treatments (e.g., an experimental and a control), each participant receives both treatments. For example, half of the participants are randomly assigned to receive the experimental treatment first and then the control; the other half receive the control first and then the experimental treatment. Outcomes are assessed following the administration of each treatment in each participant (see Figure 2–6). Participants receive the FIGURE 2–6 The Crossover Trial Eligible Participants Control Control Experimental Treatment Experimental Treatment R* Study Start *R: Randomization to Initial Treatment Period 1 Wash-out Period Period 2 Time The Framingham Heart Study r andomly assigned treatment in Period 1. The outcome of interest is then recorded for the Period 1 treatment. In most crossover trials, there is then what is called a wash-out period where no treatments are given. The wash-out period is included so that any therapeutic effects of the first treatment are removed prior to the administration of the second treatment in Period 2. In a trial with an experimental and a control treatment, participants who received the control treatment during Period 1 receive the experimental treatment in Period 2 and vice versa. There are several ways in which participants can be assigned to treatments in a crossover trial. The two most popular schemes are called random and fixed assignment. In the random assignment scheme (already mentioned), participants are randomly assigned to the experimental treatment or the control in Period 1. Participants are then assigned the other treatment in Period 2. In a fixed assignment strategy, all participants are assigned the same treatment sequence. For example, everyone gets the experimental treatment first, followed by the control treatment or vice versa. There is an issue with the fixed scheme in that investigators must assume that the outcome observed on the second treatment (and subsequent treatments, if there are more than two) would be equivalent to the outcome that would be observed if that treatment were assigned first (i.e., that there are no carry-over effects). Randomly varying the order in which the treatments are given allows the investigators to assess whether there is any order effect. The major advantage to the crossover trial is that each participant acts as his or her own control; therefore, we do not need to worry about the issue of treatment groups being comparable with respect to baseline characteristics. In this study design, fewer participants are required to demonstrate an effect. A disadvantage is that there may be carry-over effects such that the outcome assessed following the second treatment is affected by the first treatment. Investigators must be careful to include a wash-out period that is sufficiently long to minimize carry-over effects. A participant in Period 2 may not be at the same baseline as they were in Period 1, thus destroying the advantage of the crossover. In this situation, the only useful data may be from Period 1. The wash-out period must be short enough so that participants remain committed to completing the trial. Because participants in a crossover trial receive each treatment, loss to follow-up or dropout is critical because losing one participant means losing outcome data on both treatments. Crossover trials are best suited for short-term treatments of chronic, relatively stable conditions. A crossover trial would not be efficient for diseases that have acute flare-ups because 17 these could influence the outcomes that are observed yet have nothing to do with treatment. Crossover trials are also not suitable for studies with death or another serious condition considered as the outcome. Similar to the clinical trial described previously, adherence or compliance to the study protocol and study medication in the crossover trial is critical. Participants are more likely to skip medication or drop out of a trial if the treatment is unpleasant or if the protocol is long or difficult to follow. Every effort must be made on the part of the investigators to maximize adherence and to minimize loss to follow-up. 2.4 THE FRAMINGHAM HEART STUDY We now describe one of the world's most well-known studies of risk factors for cardiovascular disease. The Framingham Heart Study started in 1948 with the enrollment of a cohort of just over 5000 individuals free of cardiovascular disease who were living in the town of Framingham, Massachusetts.1 The Framingham Heart Study is a longitudinal cohort study that involves repeated assessments of the participants approximately every 2 years. The study celebrated its fiftieth anniversary in 1998 and it still continues today. The original c ohort has been assessed over 30 times. At each assessment, complete physical examinations are conducted (e.g., vital signs, blood pressure, medication history), blood samples are taken to measure lipid levels and novel risk factors, and participants also have echocardiograms in addition to other assessments of cardiovascular functioning. In the early 1970s, approximately 5000 offspring of the original cohort and their spouses were e nrolled into what is called the Framingham Offspring cohort (the second generation of the original cohort). These participants have been followed approximately every 4 years and have been assessed some eight times. In the early 2000s, a third generation of over 4000 participants was enrolled and are being followed approximately every 4 years. Over the past 50 years, hundreds of papers have been published from the Framingham Heart Study identifying important risk factors for cardiovascular disease, such as smoking, blood pressure, cholesterol, physical inactivity, and diabetes. The Framingham Heart Study also identified risk factors for stroke, heart failure, and peripheral artery disease. Researchers have identified psychosocial risk factors for heart disease, and now, with three generations of participants in the Framingham Heart Study, investigators are assessing genetic risk factors for obesity, d iabetes, and cardiovascular disease. More details on the Framingham Heart Study, its design, investigators, research milestones, and publications can be found at http:// www.nhlbi.nih.gov/about/framingham and at http://www.bu.edu/alumni/bostonia/2005/summer/pdfs/heart.pdf. 18 2.5 CHAPTER 2 Study Designs MORE ON CLINICAL TRIALS Clinical trials are extremely important, particularly in medical research. In Section 2.3, we outlined clinical trials from a design standpoint, but there are many more aspects of clinical trials that should be mentioned. First, clinical trials must be conducted at the correct time in the course of history. For example, suppose we ask the research question: "Is the polio vaccine necessary today?" To test this hypothesis, a clinical trial could be initiated in which some children receive the vaccine while others do not. The trial would not be feasible today because it would be unethical to withhold the vaccine from some children. No one would risk the consequences of the disease to study whether the vaccine is necessary. As noted previously, the design of a clinical trial is extremely important to ensure the generalizability and validity of the results. Well-designed clinical trials are very easy to analyze, whereas poorly designed trials are extremely difficult, sometimes impossible, to analyze. The issues that must be considered in designing clinical trials are outlined here. Some have been previously identified but are worth repeating. The number of treatments involved. If there are two treatments involved, statistical analyses are straightforward because only one comparison is necessary. If more than two treatments are involved, then more complicated statistical analyses are required and the issue of multiple comparisons must be addressed (these issues are discussed in Chapter 7 and Chapter 9). The number of treatments involved in a clinical trial should always be based on clinical criteria and not be reduced to simplify statistical analysis. The control treatment. The control treatment may be a treatment that is currently in use and considered the standard of care, or the control treatment may be a placebo. If a standard treatment exists, it should be used as the control because it would be unethical not to offer patients a placebo when a conventional treatment is available. (While clinical trials are considered the gold standard design to evaluate the effectiveness of an experimental treatment, there are instances where a control group is not available. Techniques to evaluate effectiveness in the absence of a control group are described in D'Agostino and Kwan.4) Outcome measures. The outcome or outcomes of interest must be clearly identified in the design phase of the clinical trial. The primary outcome is the one specified in the planned analysis and is used to determine the sample size required for the trial (this is discussed in detail in Chapter 8). The primary outcome is usually more objective than subjective in nature. It is appropriate to specify secondary outcomes, and results based on secondary outcomes should be reported as such. Analyses of secondary outcomes can provide important information and, in some cases, enough evidence for a follow-up trial. Blinding. Blinding refers to the fact that patients are not aware of which treatment (experimental or control) they are receiving in a clinical trial. A single blind trial is one in which the investigator knows which treatment a patient is receiving but the patient does not. Double blinding refers to the situation in which both the patient and the investigator are not aware of which treatment is assigned. In many clinical trials, only the statistician knows which treatment is assigned to each patient. Single-center versus multicenter trials. Some clinical trials are conducted at a single site or clinical center, whereas others are conducted—usually simultaneously—at several centers. There are advantages to including several centers, such as increased generalizability and an increased number of available patients. There are also disadvantages to including multiple centers, such as needing more resources to manage the trial and the introduction of center-specific characteristics (e.g., expertise of personnel, availability or condition of medical equipment, specific characteristics of participants) that could affect the observed outcomes. Randomization. Randomization is a critical component of clinical trials. There are a number of randomization strategies that might be implemented in a given trial. The exact strategy depends on the specific details of the study protocol. Sample size. The number of patients required in a clinical trial depends on the variation in the primary outcome and the expected difference in outcomes between the treated and control patients. Population and sampling. The study population should be explicitly defined by the study investigators (patient inclusion and exclusion criteria). A strategy for patient recruitment must be carefully determined and a system for checking inclusion and exclusion criteria for each potential enrollee must be developed and followed. Ethics. Ethical issues often drive the design and conduct of clinical trials. There are some ethical issues that are common to all clinical trials, such as the safety of the treatments involved. There are other issues that relate only to certain trials. Most institutions have institutional review boards (IRBs) that are responsible for approving research study protocols. Research protocols are evaluated on the basis of scientific accuracy and with respect to potential risks to participants. All participants in clinical trials must provide informed consent, usually on consent forms approved by the appropriate IRB. More on Clinical Trials Protocols. Each clinical trial should have a protocol, which is a manual of operations or procedures in which every aspect of the trial is clearly defined. The protocol details all aspects of subject e nrollment, treatment assignment, data collection, monitoring, data management, and statistical analysis. The protocol ensures consistency in the conduct of the trial and is particularly important when a trial is conducted at several clinical centers (i.e., in a multicenter trial). Monitoring. Monitoring is a critical aspect of all clinical trials. Specifically, participants are monitored with regard to their adherence to all aspects of the study protocol (e.g., attending all scheduled visits, completing study assessments, taking the prescribed medications or treatments). Participants are also carefully monitored for any side effects or adverse events. Protocol violations (e.g., missing scheduled visits) are summarized at the completion of a trial, as are the frequencies of adverse events and side effects. Data management. Data management is a critical part of any study and is particularly important in clinical trials. Data management includes tracking subjects (ensuring that subjects complete each aspect of the trial on time), data entry, quality control (examining data for out-of-range values or inconsistencies), data cleaning, and constructing analytic databases. In most studies, a data manager is assigned to supervise all aspects of data management. The statistical analysis in a well-designed clinical trial is straightforward. Assuming there are two treatments involved (an experimental treatment and a control), there are essentially three phases of analysis: • Baseline comparisons, in which the participants assigned to the experimental treatment group are compared to the patients assigned to the control group with respect to relevant characteristics measured at baseline. These analyses are used to check that the randomization is successful in generating balanced groups. • Crude analysis, in which outcomes are compared between patients assigned to the experimental and control treatments. In the case of a continuous outcome (e.g., weight), the difference in means is estimated; in the case of a dichotomous outcome (e.g., development of disease or not), relative risks are estimated; and in the case of time-to-event data (e.g., time to a heart attack), survival curves are estimated. (The specifics of these analyses are discussed in detail in Chapters 6, 7, 10, and 11.) • Adjusted analyses are then performed, similar to the crude analysis, which incorporate important covariates (i.e., variables that are associated with the outcome) and confounding variables. (The specifics of statistical adjustment are discussed in detail in Chapters 9 and 11.) 19 There are several important sample considerations in statistical analysis of clinical trials. The first is the Intent to Treat (ITT) analysis sample. It includes all patients who were randomized. The second is the Per Protocol analysis sample, and it includes only patients who completed the treatment and followed all components of the protocol. The third is the Safety analysis sample, and it includes all patients who took at least one dose of the assigned treatment even if they did not complete the treatment protocol. All aspects of the design, conduct, and analysis of a clinical trial should be carefully documented. Complete and accurate records of the clinical trial are essential for applications to the Food and Drug Administration (FDA).5 Clinical trials are focused on safety and efficacy. Safety is assessed by the nature and extent of adverse events and side effects. Adverse events may or may not be due to the drug being evaluated. In most clinical trials, clinicians indicate whether the adverse event is likely due to the drug or not. Efficacy is assessed by improvements in symptoms or other aspects of the indication or disease that the drug is designed to address. There are several important stages in clinical trials. Preclinical studies are studies of safety and efficacy in animals. Clinical studies are studies of safety and efficacy in humans. There are three phases of clinical studies, described here. Phase I: First Time in Humans Study. The main objectives in a Phase I study are to assess the toxicology and safety of the proposed treatment in humans and to assess the pharmacokinetics (how fast the drug is absorbed in, flows through, and is secreted from the body) of the proposed treatment. Phase I studies are not generally focused on efficacy (how well the treatment works); instead, safety is the focus. Phase I studies usually involve 10 to 15 patients, and many Phase I studies are performed in healthy, normal volunteers to assess side effects and adverse events. In Phase I studies, one goal is to determine the maximum tolerated dose (MTD) of the proposed drug in humans. Usually, investigators start with very low doses and work up to higher doses. Investigations usually start with three patients, and three patients are added for each elevated dose. Data are collected at each stage to assess safety, and some Phase I studies are placebo-controlled. Usually, two or three separate Phase I studies are conducted. Phase II: Feasibility or Dose-Finding Study. The focus of a Phase II study is still on safety, but of primary interest are side effects and adverse events (which may or may not be directly related to the drug). Another objective in the Phase II study is efficacy, but the efficacy of the drug is based on descriptive analyses in the Phase II study. In some cases, investigators do not know which specific aspects of the indication or disease the drug may affect or which outcome measure best captures 20 CHAPTER 2 Study Designs this effect. Usually, investigators measure an array of outcomes to determine the best outcome for the next phase. In Phase II studies, investigators determine the optimal dosage of the drug with respect to efficacy (e.g., lower doses might be just as effective as the MTD). Phase II studies usually involve 50 to 100 patients who have the indication or disease of interest. Phase II studies are usually placebo-controlled or compared to a standard, currently available treatment. Subjects are randomized and studies are generally double blind. If a Phase II study indicates that the drug is safe but not effective, investigation cycles back to Phase I. Most Phase III studies proceed to Phase III based on observed safety and efficacy. Phase III: Confirmatory Clinical Trial. The focus of the Phase III trial is efficacy, although data are also collected to monitor safety. Phase III trials are designed and executed to confirm the effect of the experimental treatment. Phase III trials usually involve two treatment groups, an experimental treatment at the determined optimal dose and a placebo or standard of care. Some Phase III trials involve three groups: placebo, standard of care, and experimental treatment. Sample sizes can range from 200 to 500 patients, depending on what is determined to be a clinically significant effect. (The exact number is determined by specific calculations that are described in Chapter 8.) At least two successful clinical trials performed by independent investigators at different clinical centers are required in Phase III studies to assess whether the effect of the treatment can be replicated by independent investigators in at least two different sets of participants. More details on the design and analysis of clinical trials can be found in Chow and Liu.6 Investigators need positive results (statistically proven efficacy) in at least two separate trials to submit an FDA application for drug approval. The FDA also requires clinical significance in two trials, with clinical significance specified by clinical investigators in the design phase when the number of subjects is determined (see Chapter 8). The FDA New Drug Application (NDA) contains a summary of results of Phase I, Phase II, and Phase III studies. The FDA reviews an NDA within 6 months to 1 year after submission and grants approval or not. If a drug is approved, the sponsor may conduct Phase IV trials, also called post-marketing trials, that can be retrospective (e.g., based on medical record review) or prospective (e.g., a clinical trial involving many patients to study rare adverse events). These studies are often undertaken to understand the long-term effects (efficacy and safety) of the drug. 2.6 SAMPLE SIZE IMPLICATIONS Biostatisticians have a crucial role in the design of observational studies and clinical trials. At the design stage, one of the most important issues is the number of participants to involve in the study. In Chapter 8, we provide formulas to compute the sample sizes needed to appropriately answer research questions. The sample size needed depends on the study design, the anticipated association between the risk factor and outcome or the effect of the drug, and also on the statistical analysis that will be used to answer the study questions. The sample size should not be too small such that we cannot answer whether the association of the effect of the drug under investigation is not possible, because in this instance, both participants and the investigators have wasted time and money. Alternatively, a sample size should not be too large because again time and money would be wasted but, in addition, participants may be placed at unnecessary risk. Both scenarios are unacceptable from an ethical standpoint, and therefore careful attention must be paid when determining the appropriate sample size for any study or trial. 2.7 SUMMARY To determine which study design is most efficient for a specific application, investigators must have a specific, clearly defined research question. It is also important to understand current knowledge or research on the topic under investigation. The most efficient d esign depends on the expected association or effect, the prevalence or incidence of outcomes, the prevalence of risk factors or exposures, and the expected duration of the study. Also important are practical issues, costs, and—most importantly—ethical issues. Choosing the appropriate study design to address a research question is critical. Whenever possible, prior to mounting a planned study, investigators should try to run a pilot or feasibility study, which is a smaller-scale version of the planned study, as a means to identify potential problems and issues. Whereas pilot studies can be time-consuming and costly, they are usually more than worthwhile. 2.8 PRACTICE PROBLEMS 1. An investigator wants to assess whether smoking is a risk factor for pancreatic cancer. Electronic medical records at a local hospital will be used to identify 50 patients with pancreatic cancer. One hundred patients who are similar but free of pancreatic cancer will also be selected. Each participant's medical record will be analyzed for smoking history. Identify the type of study proposed and indicate its specific strengths and weaknesses. 2. What is the most likely source of bias in the study described in Problem 1? Practice Problems 21 3. An investigator wants to assess whether the use of a specific medication given to infants born prematurely is associated with developmental delay. Fifty infants who were given the medication and 50 comparison infants who were also born prematurely but not given the medication will be selected for the analysis. Each infant will undergo extensive testing at age 2 for various aspects of development. Identify the type of study proposed and indicate its specific strengths and weaknesses. 4. Is bias or confounding a potential problem in the study described in Problem 3? If so, describe how? 5. A study is planned to assess the effect of a new surgical intervention for gallbladder disease. One hundred patients with gallbladder disease will be randomly assigned to receive either the new surgical intervention or the standard surgical intervention. The efficacy of the new surgical intervention will be measured by the time a patient takes to return to normal activities, recorded in days. Identify the type of study proposed and indicate its specific strengths and weaknesses. 6. An investigator wants to assess the association between caffeine consumption and impaired glucose tolerance, a precursor to diabetes. A study is planned to include 70 participants. Each participant will be surveyed with regard to their daily caffeine consumption. In addition, each participant will submit a blood sample that will be used to measure his or her glucose level. Identify the type of study proposed and indicate its specific strengths and weaknesses. 7. Could the study described in Problem 6 be designed as a randomized clinical trial? If so, briefly outline the study design; if not, describe the barriers. 8. A study is planned to compare two weight-loss programs in patients who are obese. The first program is based on a diet that limits caloric intake to 1200 calories per day. The second program is based on specific food combinations. The study will involve 20 participants and each participant will follow each program. The programs will be assigned in random order (i.e., some participants will first follow the restricted-calorie diet and then follow the food-combination diet, whereas others will first follow the food-combination diet and then follow the restricted-calorie diet). The number of pounds lost will be compared between diets. Identify the type of study proposed and indicate its specific strengths and weaknesses. 9. An orthopedic surgeon observes that many of his patients coming in for total knee replacement surgery played organized sports before the age of 10. He plans to collect more extensive data on participation in organized sports from four patients undergoing knee replacement surgery and to report the findings. Identify the type of study proposed and indicate its specific strengths and weaknesses. 10. Suggest an alternative design to address the hypothesis in Problem 9. What are the major issues in addressing this hypothesis? 11. In 1990, 2000 women working in a factory were recruited into a study. Half of the women worked in manufacturing and half in administrative offices. The incidence of bone cancer through 1970 among the 1000 women working in manufacturing was compared with that of the 1000 women working in administrative offices. Thirty of the women in manufacturing developed bone cancer as compared to 9 of the women in administrative offices. This study is an example of a: a. randomized controlled trial b. case-control study c. cohort study d. crossover trial 12. An investigator reviewed the medical records of 200 children seen for care at Boston Medical Center in the past year who were between the ages of 8 and 12 years old, and identified 40 who had asthma. He also identified 40 children of the same ages who were free of asthma. Each child and his or her family were interviewed to assess whether there might be an association between certain environmental factors, such as exposure to secondhand smoke, and asthma. This study is an example of a: a. randomized controlled trial b. case-control study c. cohort study d. crossover trial 13. A study is designed to evaluate the impact of a daily multivitamin on students' academic performance. One hundred study students are randomly assigned to receive either the multivitamin or a placebo and are instructed to take the assigned drug daily for 20 days. On day 20, each student takes a standardized exam and the mean exam scores are compared between groups. This study is an example of a: a. randomized controlled trial b. case-control study c. cohort study d. crossover trial 22 CHAPTER 2 Study Designs 14. A study is performed to assess whether there is an association between exposure to second-hand cigarette smoke in infancy and delayed development. Fifty children with delayed development and 50 children with normal development are selected for investigation. Parents are asked whether their children were exposed to second-hand cigarette smoke in infancy or not. This study is an example of a: a. prospective cohort study b. retrospective cohort study c. case-control study d. clinical trial 15. A study is planned to investigate risk factors for sudden cardiac death. A cohort of men and women between the ages of 35 and 70 is enrolled and followed for up to 20 years. As part of the study, p articipants provide data on demographic and behavioral characteristics; they also undergo testing for cardiac function and provide blood samples to assess lipid profiles and other biomarkers. A new measure of inflammation is hypothesized to be related to sudden cardiac death. What study design is most appropriate to assess the association between the new biomarker and sudden cardiac death? a. Prospective cohort study b. Retrospective cohort study c. case-control study d. clinical trial REFERENCES 1. D'Agostino, R.B. and Kannel, W.B. "Epidemiological background and design: The Framingham Study." Proceedings of the American Statistical Association, Sesquicentennial Invited Paper Sessions, 1989: 707–719. 2. Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A., and Saxon, A. "Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: Evidence of a new acquired cellular immunodeficiency." New England Journal of Medicine 1981; 305(24): 1425–1431. 3. Schlesselman, J.J. Case-Control Studies: Design, Conduct, Analysis. New York: Oxford University Press, 1982. 4. D'Agostino, R.B. and Kwan, H. "Measuring effectiveness: What to expect without a randomized control group." Medical Care 1995; 33(4 Suppl.): AS95–105. 5. United States Food and Drug Administration. Available at http://www.fda.gov. 6. Chow, S.C. and Liu, J.P. Design and Analysis of Clinical Trials: Concepts and Methodologies. New York: John Wiley & Sons, 1998. © Sergey Nivens/Shutterstock 3 chapter Quantifying the Extent of Disease W hen and W hy Key Questions • • How do you know when a disease outbreak is occurring? • • How do you judge whether one group is more at risk for disease than another? In the News Zika is a growing public health issue, made even bigger with the 2016 Summer Olympics taking place in Brazil and the world trying to understand the risks associated with Zika virus infection. Vaccinations are rethinking their travel destinations and are advised to consult resources such as the CDC Traveler's Health website for updates on the spread of Zika around the world.1 The link between Zika virus and microcephaly is under intense study, and new findings are being reported regularly.2 Some cities and towns are taking preventive action against mosquitos. For example, in New York, city and state governments are reportedly investing more than $21 million over a 3-year period as part of their Zika program plan.3 Public health agencies around the globe regularly update their statistics on Zika infection. For example, the World Health Organization puts out weekly situation reports that detail the latest data and statistics from around the world.4 Centers for Disease Control and Prevention. Travelers' health. Available at . 2 Johansson, M.A., Mier-y-Teran-Romero, L., Reefhuis, J., Gilboa, S.M., and Hills, S.L. New England Journal of Medicine 2016; 375: 1. 3 WNYC News. Available at @mosquito-vigilance-zika/?hootPostID=0d3dda8cc9db99e9c80e eaa269e00495. Accessed July 8, 2016. 4 World Health Organization. Zika virus situation reports. Available at . | Dig In • • Which approach would you use to learn more about Zika infection in your local community? How do you actually test for Zika virus infection? • • How might investigators take what we currently know about the global spread of Zika and apply it to estimating risk for Zika infection? • • What would you recommend to a family member or friend planning a trip to Central America? What if that person was thinking about starting a family? Do you have any recommendations for your local community to minimize the risk of Zika infection? L earning O bjectives By the end of this chapter, the reader will be able to • • Define and differentiate prevalence versus incidence • • Select, compute, and interpret the appropriate measure to compare the extent of disease between groups • • Compare and contrast relative risks, risk differences, and odds ratios • • Compute and interpret relative risks, risk differences, and odds ratios In Chapter 2, we presented several different study designs that are popular in public health research. In subsequent chapters, we discuss statistical procedures to analyze the data collected under different study designs. In statistical analyses, we first describe information we collect in our study sample and then estimate or make generalizations about the population based on data observed in the sample. The first step is called descriptive statistics and the second is called inferential statistics. Our goal is to present techniques to describe samples and procedures for generating inferences that appropriately account 24 CHAPTER 3 Quantifying the E xtent of Disease for uncertainty in our estimates. Remember that we analyze only a fraction or subset, called a sample, of the entire population, and based on that sample we make inferences about the larger population. Before we get to those procedures, we focus on some important measures for quantifying disease. Two quantities that are often used in epidemiological and biostatistical analysis are prevalence and incidence. We describe each in turn and then discuss measures that are used to compare groups in terms of prevalence and incidence of disease. 3.1 PREVALENCE Prevalence refers to the proportion of participants with a risk factor or disease at a particular point in time. Consider the prospective cohort study we described in Chapter 2, where a cohort of participants is enrolled at a specific time. We call the initial point or starting point of the study the baseline time point. Suppose in our cohort study each individual undergoes a complete physical examination at baseline. At the baseline examination, we determine—among other things—whether each participant has a history of (meaning a previous diagnosis of) cardiovascular disease (CVD). An estimate of the prevalence of CVD is computed by taking the ratio of the number of existing cases of CVD to the total number of participants examined. This is called the point prevalence (PP) of CVD as it refers to the extent of disease at a specific point in time (i.e., at baseline in our example). Number of persons with disease Point prevalence = }}}}} Number of persons examined at baseline Example 3.1. Table 3–1 The fifth examination of the offspring in the Framingham Heart Study was conducted between 1991 and 1995. A total of n = 3799 participants participated in the fifth examination. Table 3–1 shows the numbers of men and women with diagnosed CVD at the fifth examination. TABLE 3–1 Men and Women With Diagnosed CVD TABLE 3–2 Smoking and Diagnosed CVD Free of CVD History of CVD Total 2757 298 3055 663 81 744 3420 379 3799 Nonsmoker Current smoker Total The point prevalence of CVD among all participants attending the fifth examination of the Framingham Offspring Study is 379 / 3799 = 0.0998, or 9.98%. The point prevalence of CVD among men is 244 / 1792 = 0.1362, or 13.62%, and the point prevalence of CVD among women is 135 / 2007 = 0.0673, or 6.73%. TABLE 3–1 Men and Women With Diagnosed CVD TABLE 3–2 Smoking and Diagnosed CVD Free of CVD History of CVD Total Table 3–2 contains data on prevalent CVD among participants who were and were not currently smoking cigarettes at the time of the fifth examination of the Framingham Offspring Study. Almost 20% (744 / 3799) of the participants attending the fifth examination of the Framingham Offspring Study reported that they were current smokers at the time of the exam. The point prevalence of CVD among nonsmokers is 298 / 3055 = 0.0975, or 9.75%, and the point prevalence of CVD among current smokers is 81 / 744 = 0.1089, or 10.89%. 3.2 INCIDENCE In epidemiological studies, we are often more concerned with estimating the likelihood of developing disease rather than the proportion of people who have disease at a point in time. The latter reflects prevalence, whereas incidence reflects the likelihood of developing a disease among a group of participants free of the disease who are considered at risk of developing the disease over a specified observation period. Consider the study described previously, and suppose we remove participants with a history of CVD from our fixed cohort at baseline so that only participants free of CVD are included (i.e., those who are truly "at risk" of developing a disease). We follow these participants prospectively for 10 years and record, for each individual, whether or not they develop CVD during this follow-up period. If we are able to follow each individual for 10 years and can ascertain whether or not each develops CVD, then we can directly compute the likelihood or risk of -developing CVD over 10 years. Specifically, we take the ratio of the number of new cases of CVD to the total number of participants free of disease at the outset. This is referred to as cumulative incidence (CI): Free of CVD History of CVD Total Cumulative incidence = Men 1548 244 1792 Women 1872 135 2007 Number of persons who develop a disease during a specified period }}}}}}} Number of persons at risk (at baseline) Total 3420 379 3799 Cumulative incidence reflects the proportion of participants who become diseased during a specified observation Incidence eriod. The total number of persons at risk is the same as the p total number of persons included at baseline who are diseasefree. The computation of cumulative incidence assumes that all of these individuals are followed for the entire observation p eriod. This may be possible in some applications— for example, during an acute disease outbreak with a short follow-up or observation period. However, in longer studies it can be difficult to follow every individual for the development of d isease because some individuals may relocate, may not respond to investigators, or may die during the study followup period. In this e xample, the cohort is older (as is the case in most studies of cardiovascular disease, as well as in studies of any other diseases that occur more frequently in older persons) and the follow-up period is long (10 years), making it difficult to follow every individual. The issues that arise and the methods to handle incomplete follow-up are described later. 3.2.1 Problems Estimating the Cumulative Incidence There are a number of problems that can arise that make estimating the cumulative incidence of disease difficult. Because studies of incidence are by definition longitudinal (e.g., 5 or 10 years of follow-up), some study participants may be lost over the course of the follow-up period. Some participants might choose to drop out of the study, others might relocate, and others may die during the follow-up period. Different study designs could also allow for participants to enter at different times (i.e., all participants are not enrolled at baseline, but instead there is a rolling or prolonged enrollment period). For these and other reasons, participants are often followed for different lengths of time. We could restrict attention to only those p articipants who complete the entire follow-up; however, this would result in ignoring valuable information. A better approach involves accounting for the varying follow-up times as described here. 3.2.2 Person-Time Data Again, in epidemiological studies we are generally interested in estimating the probability of developing disease (incidence) over a particular time period (e.g., 10 years). The cumulative incidence assumes that the total population at risk is followed for the entire observation period and that the disease status is ascertained for each member of the population. For the reasons stated previously, it is not always possible to follow each individual for the entire observation period. Making use of the varying amounts of time that different participants contribute to the study results in changing the unit of analysis from the person or study participant to one of person-time, which is explicitly defined. The time unit might be months or years (e.g., person-months or person-years). For example, suppose that an individual enters a study in 1990 and 25 is followed until 2000, at which point they are determined to be disease-free. A second individual enters the same study in 1995 and develops the disease under study in 2000. The first individual contributes 10 years of disease-free follow-up time, whereas the second individual contributes five years of disease-free follow-up time and then contracts the disease. We would want to use all of this information to estimate the incidence of the disease. Together, these two participants contribute 15 years of disease-free time. 3.2.3 Incidence Rate The incidence rate uses all available information and is computed by taking the ratio of the number of new cases to the total follow-up time (i.e., the sum of all disease-free person-time). Rates are estimates that attempt to deal with the problem of varying follow-up times and reflect the likelihood or risk of an i ndividual changing disease status (e.g., developing disease) in a specified unit of time. The denominator is the sum of all of the disease-free follow-up time, specifically time during which participants are considered at risk for developing the disease. Rates are based on a specific time period (e.g., 5 years, 10 years) and are usually expressed as an integer value per a multiple of participants over a specified time (e.g., the incidence of disease is 12 per 1000 person-years). The incidence rate (IR), also called the incidence density (ID), is computed by taking the ratio of the number of new cases of disease to the total number of person-time units available. These person-time units may be person-years (e.g., one individual may contribute 10 years of follow-up, whereas another may contribute 5 years of follow-up) or personmonths (e.g., 360 months, 60 months). The denominator is the sum of all of the participants' time at risk (i.e., disease-free time). The IR or ID is reported as a rate relative to a specific time interval (e.g., 5 per 1000 person-years). The incidence rate is given as follows: Incidence rate = IR = Number of persons who develop disease during a specified period }}}}}}} Sum of the lengths of time during which persons are disease-free For presentation purposes, the incidence rate is usually multiplied by some multiple of 10 (e.g., 100, 1000, 10,000) to produce an integer value (see Example 3.2). Example 3.2. Consider again the fifth examination of the offspring in the Framingham Heart Study. As described in Example 3.1, a total of n = 3799 participants attended the fifth examination, and 379 had a history of CVD. This leaves a total of n = 3420 participants free of CVD at the fifth examination. Suppose we follow each participant for the development 26 CHAPTER 3 Quantifying the E xtent of Disease TABLE 3–3 Men and Women Who Develop CVD No CVD Develop CVD Total Men 1358 190 1548 Women 1753 119 1872 Total 3111 309 3420 TABLE 3–4 Total Disease-Free Time in Men and Women Develop CVD Total Follow-Up Time (years) IR Men 190 9984 0.01903 Women 119 12,153 0.00979 Total 309 22,137 0.01396 of CVD over the next 10 years. Table 3–3 shows the numbers of men and women who develop CVD over a 10-year followup period. Because each participant is not followed for the full 10-year period (the mean follow-up time is 7 years), we cannot correctly estimate cumulative incidence using the previous data. The estimate of the cumulative incidence assumes that each of the 3111 persons free of CVD is followed for 10 years. Because this is not the case, our estimate of cumulative incidence is incorrect; instead, we must sum all of the available follow-up time and estimate an incidence rate. Table 3–4 displays the total disease-free follow-up times for men and women along with the incidence rates. The incidence rates can be reported as 190 per 10,000 p erson-years for men and 98 per 10,000 person-years for women; equivalent to this is 19 per 1000 men per 10 years and 9.8 per 1000 women per 10 years. The denominator of the incidence rate accumulates disease-free time over the entire observation period, and the unit of analysis is person-time (e.g., person-years in Example 3.2). In comparison, the denominator of the cumulative incidence is measured at the beginning of the study (baseline) and the unit of analysis is the person. It is worth noting that rates have dimension (number of new cases per person-time units) and are often confused with proportions (or probabilities) or percentages, which are dimensionless and range from 0 to 1, 0% to 100%. Example 3.3 illustrates the difference between the prevalence, cumulative incidence, and incidence rate. It is important to note that the time component is an integral part of the denominator of the incidence rate, whereas with the cumulative incidence the time component is only part of the interpretation. Example 3.3. Consider a prospective cohort study including six participants. Each participant is enrolled at baseline, and the goal of the study is to follow each participant for 10 years. Over the course of the follow-up period, some participants develop CVD, some drop out of the study, and some die. Figure 3–1 displays the f ollow-up experiences for each participant. In this example, Participant 1 develops CVD 6 years into the study, Participant 2 dies 9 years into the study but is free of CVD, Participant 3 survives the complete follow-up period disease-free, Participant 4 d evelops CVD 2 years into the study and dies after 8 years, Participant 5 drops out of the study after 7d isease-free years, and Participant 6 develops CVD 5 years into the study. Using the data in Example 3.3, we now compute prevalence, cumulative incidence, and incidence rate. Prevalence of CVD at baseline = 0 /6 = 0, or 0% Prevalence of CVD at 5 years = 2 / 6 = 0.333, or 33% Prevalence of CVD at 10 years = 2 / 3 = 0.666, or 67% (Note that we can only assess disease status at 10 years in Participants 1, 3, and 6.) Cumulative incidence of CVD at 5 years = 2 / 6 = 0.333, or 33% The cumulative incidence of CVD at 10 years cannot be estimated because we do not have complete follow-up on Participants 2, 4, or 5. To make use of all available information, we compute the incidence rate. The incidence rate of CVD = 3 / (6 + 9 + 10 + 2 + 7 + 5) = 3 / 39 = 0.0769 We can report this as an incidence rate of CVD of 7.7 per 100 person-years. The incidence rate of death per person-year = 2 / (10 + 9 + 10 + 8 + 7 + 10) = 2 / 54 = 0.037 We can report this as an incidence rate of death of 3.7 per 100 person-years. Notice that the prevalence and cumulative incidence are shown as percentages (these can also be shown as proportions or probabilities), whereas the incidence rates are reported as the number of events per person-time. Comparing the Extent of Disease Between Groups 27 FIGURE 3–1 Course of Follow-up for 6 Participants Participant 1 CVD 2 DTH 3 4 DTH CVD 5 (Drop Out) 6 CVD 0 1 2 3 4 5 6 7 8 9 10 Years 3.3 RELATIONSHIPS BETWEEN PREVALENCE AND INCIDENCE The prevalence (proportion of the population with disease at a point in time) of a disease depends on the incidence (risk of developing disease within a specified time) of the disease as well as the duration of the disease. If the incidence is high but the duration is short, the prevalence (at a point in time) will be low by comparison. In contrast, if the incidence is low but the duration of the disease is long, the prevalence will be high. When the prevalence of disease is low (less than 1%), the prevalence is approximately equal to the product of the incidence and the mean duration of the disease (assuming that there have not been major changes in the course of the disease or its treatment).1 Hypertension is an example of a disease with a relatively high prevalence and low incidence (due to its long duration). Influenza is an example of a condition with low prevalence and high incidence (due to its short duration). The incidence rate is interpreted as the instantaneous potential to change the disease status (i.e., from non-diseased to diseased, also called the hazard) per unit time. An assumption that is implicit in the computation of the IR or ID is that the risk of disease is constant over time. This is not a valid assumption for many diseases because the risk of disease can vary over time and among certain subgroups (e.g., persons of different ages have different risks of disease). It is also very important that only persons who are at risk of developing the disease of interest are included in the denominator of the e stimate of incidence. For example, if a disease is known to affect only people over 65 years of age, then including diseasefree follow-up time measured on study participants less than 65 years of age will underestimate the true incidence. Persontimes measured in these participants should not be included in the denominator as these participants are not truly at risk of developing the disease prior to age 65. 3.4 COMPARING THE EXTENT OF DISEASE BETWEEN GROUPS It is often of interest to compare groups with respect to extent of disease or their likelihood of developing disease. These groups might be defined by an exposure to a potentially harmful agent (e.g., exposed or not), by a particular sociodemographic characteristic (e.g., men or women), or by a particular risk factor (e.g., current smoker or not). Popular comparative measures are generally categorized as difference measures or ratios. Difference measures are used to make absolute comparisons, whereas ratio measures are used to make relative comparisons. Differences or ratios can be constructed to compare prevalence measures or incidence measures between comparison groups. 3.4.1 Difference Measures: Risk Difference and Population Attributable Risk The risk difference (RD), also called excess risk, measures the absolute effect of the exposure or the absolute effect of the risk factor of interest on prevalence or incidence. The risk 28 CHAPTER 3 Quantifying the E xtent of Disease CIoverall – CIunexposed PAR = }}} CIoverall difference is defined as the difference in point prevalence, cumulative incidence, or incidence rates between groups, and is given by the following: IRoverall – IRunexposed PAR = }}} IRoverall Risk difference = RD = PPexposed 2 PPunexposed RD = CIexposed 2 CIunexposed RD = IRexposed 2 IRunexposed In Example 3.1, we computed the point prevalence of CVD in smokers and nonsmokers. Exposed persons are those who reported smoking at the fifth examination of the Framingham Offspring Study. Using data from Example 3.1, the risk (prevalence) difference in CVD for smokers as compared to nonsmokers is computed by subtracting the point prevalence for nonsmokers from the point prevalence for smokers. The risk difference is 0.1089 – 0.0975 = 0.0114, and this indicates that the absolute risk (prevalence) of CVD is 0.0114 higher in smokers as compared to nonsmokers. The risk difference can also be computed by taking the difference in cumulative incidences or incidence rates between comparison groups, and the risk difference represents the excess risk associated with exposure to the risk factor. In Example 3.2 we estimated the incidence rates of CVD in men and women. Here the comparison groups are based on sex as opposed to exposure to a risk factor or not. Thus, we can compute the risk difference by either subtracting the incidence rate in men from the incidence rate in women or vice versa: the approach affects the interpretation. The incidence-rate difference between men and women, using data in Example 3.2, is 190/10,000 person-years in men – 98/10,000 person-years in women = 92/10,000 person-years. This indicates that there are 92 excess CVD events per 10,000 person-years in men as compared to women. The range of possible values for the risk difference in point prevalence or cumulative incidence is –1 to 1. The range of possible values for the risk difference in incidence rates is – ∞ to ∞ events per person-time. The risk difference is positive when the risk for those exposed is greater than that for those unexposed. The risk difference is negative when the risk for those exposed is less than the risk for those unexposed. If exposure to the risk factor is unrelated to the risk of disease, then the risk difference is 0. A value of 0 is the null or no-difference value of the risk difference. The population attributable risk (PAR) is another difference measure that quantifies the association between a risk factor and the prevalence or incidence of disease. The population attributable risk is computed as f ollows: PPoverall – PPunexposed Population attributable risk = PAR = }}} PPoverall The population attributable risk is computed by first assessing the difference in overall risk (exposed and unexposed persons combined) and the risk of those unexposed. This difference is then divided by the overall risk and is usually presented as a percentage. Using data presented in Example 3.1, comparing prevalence of CVD in smokers and nonsmokers, the point prevalence of CVD for all participants attending the fifth examination of the Framingham Offspring Study is 379 / 3799 = 0.0998. The population attributable risk is computed as (0.0998 – 0.0975) / 0.0998 = 0.023 or 2.3% and suggests that 2.3% of the prevalent cases of CVD are attributable to smoking and could be eliminated if the exposure to smoking were eliminated. The population attributable risk is usually expressed as a percentage and ranges from 0% to 100%. The magnitude of the population a ttributable risk is interpreted as the percentage of risk (prevalence or incidence) associated with, or attributable to, the risk factor. If exposure to the risk factor is unrelated to the risk of disease, then the p opulation attributable risk is 0% (i.e., none of the risk is a ssociated with e xposure to the risk factor). The population a ttributable risk assumes a causal relationship between the risk factor and disease and is also interpreted as the percentage of risk (prevalence or incidence) that could be eliminated if the exposure or risk factor were r emoved. 3.4.2 Ratios: Relative Risk, Odds Ratio, and Rate Ratio The relative risk (RR), also called the risk ratio, is a useful measure to compare the prevalence or incidence of d isease between two groups. It is computed by taking the ratio of the respective prevalences or cumulative incidences. Generally, the reference group (e.g., unexposed persons, persons without the risk factor, or persons assigned to the control group in a clinical trial setting) is considered in the denominator: Relative risk = RR = PPexposed PPunexposed RR = CIexposed CIunexposed The ratio of incidence rates between two groups is called the rate ratio or the incidence density ratio.2 Using data presented in Example 3.2, the rate ratio of incident CVD in 29 Comparing the Extent of Disease Between Groups men as compared to women is (190/10,000 person-years)/ (98/10,000 person-years) = 1.94. Thus, the incidence of CVD is 1.94 times higher per person-year in men as compared to women. The relative risk is often felt to be a better measure of the strength of the effect than the risk difference (or attributable risk) because it is relative to a baseline (or comparative) level. Using data presented in Example 3.1, the relative risk of CVD for smokers as compared to nonsmokers is 0.1089 / 0.0975 = 1.12; that is, the prevalence of CVD among smokers is 1.12 times that among nonsmokers. The range of the relative risk is 0 to `. If exposure to the risk factor is unrelated to the risk of disease, then the relative risk and the rate ratio will be 1. A value of 1 is considered the null or no-effect value of the relative risk or the rate ratio. Under some study designs (e.g., the case-control study - described in Chapter 2), it is not possible to compute a relative risk. Instead, an odds ratio is computed as a measure of effect. Suppose that in a case-control study we want to assess the r elationship between exposure to a particular risk factor and disease status. Recall that in a case-control study, we select participants on the basis of their outcome—some have the condition of interest (cases) and some do not (controls). Example 3.4. Table 3–5 shows the relationship between prevalent hypertension and prevalent cardiovascular disease at the fifth examination of the offspring in the Framingham Heart Study. The proportion of persons with hypertension who have CVD is 181 / 840 = 0.215. The proportion of persons free of hypertension but who have CVD is 188 / 2942 = 0.064. Odds are different from probabilities in that odds are computed as the ratio of the number of events to the number of nonevents, whereas a proportion is the ratio of the number of events to the total sample size. The odds that a person with hypertension has CVD are 181 / 659 = 0.275. The odds that a person free of hypertension has CVD are 188 / 2754 = 0.068. The relative risk of CVD for persons with as compared to without hypertension is 0.215 / 0.064 = 3.36, or persons with hypertension are 3.36 times more likely to have prevalent CVD than persons without hypertension. The odds ratio is computed in a similar way but is based on the ratio of odds. The odds ratio is 0.275 / 0.068 = 4.04 and is interpreted as: People with hypertension have 4.04 times the odds of CVD compared to people free of hypertension. Perhaps the most important characteristic of an odds ratio is its invariance property. Using the data in Table 3–5, the odds that a person with CVD has hypertension are 181 / 188 = 0.963. The odds that a person free of CVD has hypertension are 659/ 2754 = 0.239. The odds ratio for hypertension is therefore 0.963 / 0.239 = 4.04. People with CVD have 4.04 times the odds of hypertension compared to people free of CVD. This property does not hold for a relative risk. For example, the proportion of persons with CVD who have hypertension is 181 / 369 = 0.491. The proportion of p ersons free of CVD who have hypertension is 659 / 3413 = 0.193. The r elative risk for hypertension is 0.491 / 0.193 = 2.54. The invariance property of the odds ratio makes it an ideal measure of association for a case-control study. For example, suppose we conduct a case-control study to assess the association between cigarette smoking and a rare form of cancer (e.g., a cancer that is thought to occur in less than 1% of the general population). The cases are individuals with the rare form of cancer and the controls are similar to the cases but free of the rare cancer. Suppose we ask each participant whether he or she formerly or is currently smoking cigarettes or not. For this study, we consider former and current smokers as smokers. The data are shown in Table 3–6. Using these data, we cannot calculate the incidence of cancer in the total sample or the incidence of cancer in smokers or in nonsmokers because of the way in which we collected the data. In this sample, 40 / 69 = 0.58 or 58% of the smokers have cancer and 10 / 31 = 0.32 or 32% of the nonsmokers have cancer—yet this is a rare cancer. These e stimates do not reflect reality because the sample was TABLE 3–6 Smoking and Cancer TABLE 3–5 Prevalent Hypertension and Prevalent CVD No CVD No hypertension Hypertension Total CVD Cancer (Case) No Cancer (Control) Total Smoker 40 29 69 Nonsmoker 10 21 31 Total 50 50 100 Total 2754 188 2942 659 181 840 3413 369 3782 30 CHAPTER 3 Quantifying the E xtent of Disease s pecifically designed to include an equal number of cases and controls. Had we sampled individuals at random from the general population (using a cohort study design), we might have needed to sample more than 10,000 individuals to realize a sufficient number of cases for analysis. With this case-control study, we can estimate an association between smoking and cancer using the odds ratio. The odds of cancer in smokers are 40 / 29 = 1.379 and the odds of cancer in nonsmokers are 10 / 21 = 0.476. The odds ratio is 1.379 / 0.476 = 2.90, suggesting that smokers have 2.9 times the odds of cancer compared to nonsmokers. Note that this is equal to the odds ratio of smoking in cancer cases versus controls, i.e., (40 / 10) / (29 / 21) = 4 / 1.38 = 2.90. The fact that we can estimate an odds ratio in a case- control study is a useful and important property. The odds ratio estimated in a study using a prospective sampling scheme (i.e., sampling representative groups of smokers and nonsmokers and monitoring for cancer incidence) is equivalent to the odds ratio based on a retrospective sampling scheme (i.e., sampling representative groups of cancer and patients free of cancer and recording smoking status). The odds ratio can also be computed by taking the ratio of the point prevalence (PP) or cumulative incidence (CI) of disease to (1 – PP) or (1 – CI), respectively. The odds ratio is the ratio of the odds of developing disease for persons exposed as compared to those unexposed. Using cumulative incidences, the odds ratio is defined as: Odds ratio = CIexposed / (1− CIexposed) CIunexposed / (1− CIunexposed) The odds ratio will approximate the relative risk when the disease under study is rare, usually defined as a prevalence or cumulative incidence less than 10%. For this reason, the interpretation of an odds ratio is often taken to be identical to that of a relative risk when the prevalence or cumulative incidence is low. 3.4.3 Issues with Person-Time Data There are some special characteristics of person-time data that need attention, one of which is censoring. Censoring occurs when the event of interest (e.g., disease status) is not observed on every individual, usually due to time constraints (e.g., the study follow-up period ends, subjects are lost to follow-up, or they withdraw from the study). In epidemiological studies, the most common type of censoring that ccurs is called right censoring. Suppose that we conduct a o longitudinal study and monitor subjects prospectively over time for the development of CVD. For participants who develop CVD, their time to event is known; for the remaining subjects, all we know is that they did not develop the event during the study observation period. For these participants, their time-to-event (also called their survival time) is longer than the observation time. For analytic purposes, these times are censored, and are called Type I censored or right-censored times. Methods to handle survival time, also called time-to-event data, are discussed in detail in Allison3 and in Chapter 11. 3.5 SUMMARY Prevalence and incidence measures are important measures that quantify the extent of disease and the rate of development of disease in study populations. Understanding the difference between prevalence and incidence is critical. Prevalence refers to the extent of a disease at a point in time, whereas incidence refers to the development of disease over a specified time. Because it can be difficult to ascertain disease status in every participant in longitudinal studies—particularly when the follow-up p eriod is long—measures that take into account all available data are needed. Incidence rates that account for v arying follow-up times are useful measures in epidemiological analysis. The formulas to estimate and compare prevalence and incidence are summarized in Table 3–7. In the next chapter, we present descriptive statistics. Specifically, we discuss how to estimate prevalence and incidence in study samples. We then move into statistical inference procedures, where we discuss estimating unknown population parameters based on sample statistics. 3.6 PRACTICE PROBLEMS 1. A cohort study is conducted to assess the association between clinical characteristics and the risk of stroke. The study involves n = 1250 participants who are free of stroke at the study start. Each participant is assessed at study start (baseline) and every year thereafter for 5 years. Table 3–8 displays data on hypertension s tatus measured at baseline and hypertension status measured 2 years later. a. Compute the prevalence of hypertension at baseline. b. C ompute the prevalence of hypertension at 2 years. c. Compute the cumulative incidence of hypertension over 2 years. Practice Problems 31 TABLE 3–7 Summary of Key Formulas Measure Formula Number of persons with disease Point prevalence (PP)* }}}}} Number of persons examined at baseline Cumulative incidence (CI)* Number of persons who develop disease during a specified period }}}}}} Number of persons at risk (at baseline) Incidence rate (IR) Number of persons who develop disease during a specified period }}}}}}} Sum of the lengths of time during which persons are disease-free Risk difference (RD) PPexposed – PPunexposed, CIexposed – CIunexposed, IRexposed – IRunexposed PPoverall − PPunexposed CIoverall − CIunexposed IR overall − IR unexposed , , PPoverall CIoverall IR overall Population attributable risk (PAR) PPexposed CIexposed , PPunexposed CIunexposed Relative risk (RR) PPexposed / (1−PPexposed) CIexposed / (1− CIexposed) , PPunexposed / (1−PPunexposed) CIunexposed / (1− CIunexposed) Odds ratio (OR) * Can also be expressed as a percentage. TABLE 3–8 Hypertension at Baseline and Two Years later TABLE 3–9 Hypertension at Baseline and Stroke Five Years Later Two Years Later: Not Hypertensive Two Years Later: Hypertensive Baseline: Not hypertensive 850 148 Baseline: Not hypertensive Baseline: hypertensive 45 207 Baseline: hypertensive 2. The data shown in Table 3–9 were collected in the study described in Problem 1 relating hypertensive status measured at baseline to incident stroke over 5 years. a. Compute the cumulative incidence of stroke in this study. b. Compute the cumulative incidence of stroke in patients classified as hypertensive at baseline. c. Compute the cumulative incidence of stroke in patients free of hypertension at baseline. Free of Stroke at Five Years Stroke 952 46 234 18 d. Compute the risk difference of stroke in patients with hypertension as compared to patients free of hypertension. e. Compute the relative risk of stroke in patients with hypertension as compared to patients free of hypertension. f. Compute the population attributable risk of stroke due to hypertension. 3. A case-control study is conducted to assess the relationship between heavy alcohol use during 32 CHAPTER 3 Quantifying the E xtent of Disease TABLE 3–10 Alcohol Use and Outcome of Pregnancy Miscarriage Delivered Full Term Heavy alcohol use 14 4 No heavy alcohol use 36 46 TABLE 3–12 Total Follow-Up Time by Treatment Number with Coronary Artery Disease Total Follow-Up (years) Cholesterol medication 28 3451 Placebo 42 2984 TABLE 3–11 Incident Coronary Artery Disease by Treatment Number of Participants Number with Coronary Artery Disease Cholesterol medication 400 28 Placebo 400 42 the first trimester of pregnancy and miscarriage. Fifty women who suffered miscarriage are enrolled, along with 50 who delivered full-term. Each p articipant's use of alcohol during pregnancy is ascertained. Heavy drinking is defined as four or more drinks on one occasion. The data are shown in Table 3–10. a. Compute the odds of miscarriage in women with heavy alcohol use during pregnancy. b. Compute the odds of miscarriage in women with no heavy alcohol use during pregnancy. c. Compute the odds ratio for miscarriage as a function of heavy alcohol use. 4. A randomized trial is conducted to evaluate the efficacy of a new cholesterol-lowering medication. The primary outcome is incident coronary artery - disease. Participants are free of coronary artery disease at the start of the study and randomized to receive either the new medication or a placebo. Participants are followed for a maximum of 10 years for the development of coronary artery disease. The observed data are shown in Table 3–11. a. Compute the relative risk of coronary artery disease in patients receiving the new cholesterol medication as compared to those receiving a placebo. b. Compute the odds ratio of coronary artery disease in patients receiving the new cholesterol medication as compared to those receiving a placebo. c. W hich measure is more appropriate in this design, the relative risk or odds ratio? Justify briefly. 5. In the study described in Problem 4, some patients were not followed for a total of 10 years. Some suffered events (i.e., developed coronary artery d isease during the course of follow-up), whereas others dropped out of the study. Table 3–12 displays the total number of person-years of follow-up in each group. a. Compute the incidence rate of coronary

artery disease in patients receiving the new cholesterol medication. b. Compute the incidence rate of coronary artery disease in patients receiving a placebo. 6. A small cohort study is conducted in 13 patients with an aggressive cellular disorder linked to cancer. The clinical courses of the patients are depicted graphically in Figure 3–2. a. Compute
the prevalence of cancer at 12 months. b. Compute the cumulative incidence of cancer at 12 months. c. C ompute the incidence rate (per month) of cancer. d. Compute the incidence rate (per month) of death. 7. Five hundred people are enrolled in a 10-year cohort study. At the start of the study, 50 have diagnosed CVD. Over the course of the
study, 40 people who were free of CVD at baseline develop CVD. a. What is the cumulative incidence of CVD over 10 years? b. What is the prevalence of CVD at baseline? c. What is the prevalence of CVD at 10 years? 8. A total of 150 participants are selected for a study of risk factors for cardiovascular disease. At b aseline (study start), 24 are
classified as hypertensive. At 1 year, an additional 12 have developed h ypertension, Practice Problems 33 FIGURE 3–2 Clinical Course for Patients with Aggressive Cellular Disorder Patient 1 Cancer 2 3 Death 4 Cancer 5 Cancer 6 7 8 Lost to Follow-up 9 10 Cancer 11 Cancer 12 Cancer 13 3 6 9 12 15 18 21 24 Months and at 2 years another 8 have
developed hypertension. What is the prevalence of hypertension at 2 years in the study? 9. Consider the study described in Problem 8. What is the 2-year cumulative incidence of hypertension? 10. A national survey is conducted to assess the association between hypertension and stroke in persons over 75 years of age with a family history of stroke.
Development of stroke is monitored over a 5-year follow-up period. The data are summarized in Table 3–13 and the numbers are in millions. a. Compute the cumulative incidence of stroke in persons over 75 years of age. b. Compute the relative risk of stroke in hypertensive as compared to non-hypertensive persons. c. Compute the odds ratio of
stroke in hypertensive as compared to non-hypertensive persons. 11. In a nursing home, a program is launched in 2005 to assess the extent to which its residents are affected by diabetes. Each resident has a blood test, and 48 TABLE 3–13 Hypertension and Development of Stroke Developed Stroke Did Not Develop Stroke Hypertension 12 37 No
hypertension 4 26 of the 625 residents have diabetes in 2005. Residents who did not already have diabetes were again tested in 2010, and 57 residents had diabetes. a. What is the prevalence of diabetes in 2005? b. What is the cumulative incidence of diabetes over 5 years? c. What is the prevalence of diabetes in 2010 (assume that none of the
residents in 2005 have died or left the nursing home)? 34 CHAPTER 3 Quantifying the E xtent of Disease TABLE 3–14 Incidence of Stroke in Men and Women Number of Strokes Number of Stroke-Free Person-Years Men (n = 125) 9 478 Women (n = 200) 21 97 12. A prospective cohort study is run to estimate the incidence of stroke in persons 55
years of age and older. All participants are free of stroke at study start. Each participant is followed for a maximum of 5 years. The data are summarized in Table 3–14. a. What is the annual incidence rate of stroke in men? b. What is the annual incidence rate of stroke in women? c. What is the annual incidence rate of stroke (men and women
combined)? 13. A clinical trial is run to assess the efficacy of a new drug to reduce high blood pressure. Patients with a diagnosis of hypertension (high blood pressure) are recruited to participate in the trial and randomized to receive either the new drug or placebo. Participants take the assigned drug for 12 weeks and their blood TABLE 3–15
Hypertension Status by Treatment Number Randomized Number Free of Hypertension at 12 Weeks Placebo 50 6 New drug 50 14 Group pressure status is recorded. At the end of the trial, participants are classified as still having hypertension or not. The data are shown in Table 3–15. a. What is the prevalence of hypertension at the start of the trial?
b. What is the prevalence of hypertension at the end of the trial? c. Estimate the relative risk comparing the proportions of patients who are free of hypertension at 12 weeks between groups. REFERENCES 1. Hennekens, C.H. and Buring, J.E. Epidemiology in Medicine. Philadelphia: Lippincott Williams & Wilkins, 1987. 2. Kleinbaum, D.G., Kupper,
L.L., and Morgenstern, H. Epidemiologic Research. New York: Van Nostrand Reinhold Company Inc., 1982. 3. Allison, P. Survival Analysis Using SAS: A Practical Guide. Cary, NC: SAS Institute, 1995. 4 © Sergey Nivens/Shutterstock chapter Summarizing Data Collected in the Sample W hen and W hy Key Questions •• What is the best way to make a
case for action using data? •• Are investigators being deceptive or just confusing when they report relative differences instead of absolute differences? •• How can we be sure that we are comparing like statistics (apples to apples) when we attempt to synthesize data from various sources? In the News Summary statistics on key indicators in different
groups and over time can make powerful statements. Simple tables or graphical displays of means, counts, or rates can shine a light on an issue that might be otherwise ignored. Some examples of current issues and a few key statistics are outlined here. As of 2014, more than 21 million Americans 12 years and older had a substance use disorder. Of
these disorders, nearly 2 million involved prescription painkillers and more than half a million involved heroin.1 The National Institute on Drug Abuse reports a 2.8-fold increase in overdose deaths in the United States from prescription drugs from 2001 to 2014, a 3.4-fold increase in deaths from opioid pain relievers, and a 6-fold increase in deaths due
to heroin over the same period.2 Dig In •• How would you summarize the extent of prescription drug use in your community? •• What would you measure and how? What are the challenges in collecting these data? •• If you were to compare the extent of prescription drug use in your community with that in another community, how could you ensure
that the data are comparable? American Society of Addiction Medicine. Opioid addiction 2016 facts and figures. Available at advocacy/opioid-addiction-disease-facts-figures.pdf. Accessed July 10, 2016. 2 National Institute on Drug Abuse. Overdose death rates. Available at . 1 L earning O bjectives By the end of this chapter, the reader will be able to
•• D istinguish between dichotomous, ordinal, categorical, and continuous variables •• Identify appropriate numerical and graphical summaries for each variable type •• Compute a mean, median, standard deviation, quartiles, and range for a continuous variable •• Construct a frequency distribution table for dichotomous, categorical, and ordinal
variables •• Provide an example of when the mean is a better measure of location than the median •• Interpret the standard deviation of a continuous variable •• Generate and interpret a box plot for a continuous variable •• Produce and interpret side-by-side box plots •• Differentiate between a histogram and a bar chart Before any biostatistical
analyses are performed, we must define the population of interest explicitly. The composition of the population depends on the investigator's research question. It is important to define the population explicitly as inferences based on the study sample will only be generalizable to the specified population. The population is the collection of all -
individuals about whom we wish to make generalizations. For example, if we wish to assess the prevalence of cardiovascular disease (CVD) among all adults 30 to 75 years of age living in the United States, then all adults in that age range living in the United States at the specified time of the study constitute the population of interest. If we wish to
assess the prevalence of CVD among all adults 30 to 75 years of age living in the state of Massachusetts, then all adults in that age range l iving in Massachusetts at the specified time of 36 CHAPTER 4 Summarizing Data Collected in the Sample the study constitute the population of interest. If we wish to assess the prevalence of CVD among all
adults 30 to 75 years of age living in the city of Boston, then all adults in that age range living in Boston at the specified time of the study constitute the population of interest. In most applications, the population is so large that it is impractical to study the entire population. Instead, we select a sample (a subset) from the population and make
inferences about the population based on the results of an analysis on the sample. The sample is a subset of individuals from the population. Ideally, individuals are selected from the population into the sample at random. (We discuss this procedure and other concepts related to sampling in detail in Chapter 5.) There are a number of techniques that
can be used to select a sample. Regardless of the specific techniques used, the sample should be representative of the population (i.e., the characteristics of individuals in the sample should be similar to those in the population). By definition, the number of individuals in the sample is smaller than the number of individuals in the population. There are
formulas to determine the appropriate number of individuals to include in the sample that depend on the characteristic being measured (i.e., exposure, risk factor, and outcome) and the desired level of p recision in the estimate. We present details about sample size computations in Chapter 8. Once a sample is selected, the characteristic of interest
must be summarized in the sample using appropriate techniques. This is the first step in an analysis. Once the sample is appropriately summarized, statistical inference procedures are then used to generate inferences about the population based on the sample. We discuss statistical inference procedures in Chapters 6, 7, 9, 10, and 11. In this chapter,
we present techniques to summarize data collected in a sample. The appropriate numerical summaries and graphical displays depend on the type of characteristic under study. Characteristics—sometimes called variables, outcomes, or endpoints—are classified as one of the following types: dichotomous, ordinal, categorical, or continuous.
Dichotomous variables have only two possible responses. The response options are usually coded "yes" or "no." Exposure to a particular risk factor (e.g., smoking) is an example of a dichotomous variable. Prevalent disease status is another example of a dichotomous variable, where each individual in a sample is classified as having or not having the -
disease of interest at a point in time. Ordinal and categorical variables have more than two possible responses but the response options are ordered and unordered, respectively. Symptom severity is an example of an ordinal variable with possible responses of minimal, TABLE 4–1 Blood Pressure Categories Classification of Blood Pressure SBP
and/or DBP Normal 100 moderate, and severe. The National Heart, Lung, and Blood Institute (NHLBI) issues guidelines to classify blood pressure as normal, pre-hypertension, Stage I hypertension, or Stage II hypertension.1 The classification scheme is shown in Table 4–1 and is based on specific levels of systolic blood pressure (SBP) and diastolic
blood pressure (DBP). Participants are classified into the highest category, as defined by their SBP and DBP. Blood pressure category is an ordinal variable. Categorical variables, sometimes called nominal variables, are similar to ordinal variables except that the responses are u nordered. Race/ethnicity is an example of a categorical variable. It is
often measured using the following response options: white, black, Hispanic, American Indian or Alaskan native, Asian or Pacific Islander, or other. Another example of a categorical variable is blood type, with response options A, B, AB, and O. Continuous variables, sometimes called quantitative or measurement variables, in theory take on an
unlimited number of responses between defined minimum and maximum values. Systolic blood pressure, diastolic blood pressure, total cholesterol level, CD4 cell count, platelet count, age, height, and weight are all examples of continuous variables. For example, systolic blood pressure is measured in millimeters of mercury (mmHg), and an individual
in a study could have a systolic blood pressure of 120, 120.2, or 120.23, depending on the precision of the instrument used to measure systolic blood pressure. In Chapter 11 we present statistical techniques for a specific continuous variable that measures time to an event of interest, for example time to development of heart disease, cancer, or death.
Almost all numerical summary measures depend on the specific type of variable under consideration. One exception is the sample size, which is an important summary measure for any variable type (dichotomous, ordinal, categorical, or continuous). The sample size, denoted as n, reflects the number of independent or distinct units (participants) in
the sample. For example, if a study is conducted to assess the total cholesterol in a population and a random sample of 100 individuals is Dichotomous Variables s elected for participation, then n = 100 (assuming all individuals selected agree to participate). In some applications, the unit of analysis is not an individual participant but might be a blood
sample or specimen. Suppose in the example study that each of the 100 participants provides blood samples for cholesterol testing at three distinct points in time (e.g., at the start of the study, and 6 and 12 months later). The unit of analysis could be the blood sample, in which case the sample size would be n = 300. It is important to note that these
300 blood samples are not 300 independent or unrelated observations because multiple blood samples are taken from each participant. Multiple measurements taken on the same individual are referred to as clustered or repeated measures data. Statistical methods that account for the clustering of measurements taken on the same individual must be
used in analyzing the 300 total cholesterol measurements taken on participants over time. Details of these techniques can be found in Sullivan.2 The sample size in most of the analyses discussed in this textbook refers to the number of individuals participating in the study. In the examples that follow, we indicate the sample size. It is always important
to report the sample size when summarizing data as it gives the reader a sense of the precision of the analysis. The notion of precision is discussed in subsequent chapters in detail. Numerical summary measures computed on samples are called statistics. Summary measures computed on populations are called parameters. The sample size is an
example of an important statistic that should always be reported when summarizing data. In the following sections, we present sample statistics as well as graphical displays for each type of variable. 4.1 DICHOTOMOUS VARIABLES Dichotomous variables take on one of only two possible responses. Sex is an example of a dichotomous variable, with
response options of "male" or "female," as are current smoking status and diabetes status, with response options of "yes" or "no." 4.1.1 Descriptive Statistics for Dichotomous Variables Dichotomous Variables Dichotomous variables are often used to classify participants as possessing or not possessing a particular characteristic, having or not having a particular attribute.
For example, in a study of cardiovascular risk factors we might collect information on participants such as whether or not they have diabetes, whether or not they smoke, and whether or not they are on treatment for high blood pressure or high cholesterol. The response options for each of these variables are "yes" or "no." When analyzing
dichotomous variables, responses are often classified as success or failure, with success denoting 37 TABLE 4–2 Frequency Distribution Table for Sex Frequency Relative Frequency (%) Male 1625 45.9 Female 1914 54.1 Total 3539 100.0 the response of interest. The success response is not necessarily the positive or healthy response but rather the
response of interest. In fact, in many medical applications the focus is often on the unhealthy or "at-risk" response. Example 4.1. The seventh examination of the offspring in the Framingham Heart Study was conducted between 1998 and 2001. A total of n = 3539 participants (1625 men and 1914 women) attended the seventh examination and
completed an extensive physical examination. At that examination, numerous variables were measured including demographic characteristics, such as sex, educational level, income, and marital status; clinical characteristics, such as height, weight, systolic and diastolic blood pressure, and total cholesterol; and behavioral characteristics, such as
smoking and exercise. Dichotomous variables are often summarized in frequency distribution tables. Table 4–2 displays a frequency distribution table for the variable sex measured in the seventh examination of the Framingham Offspring Study. The first column of the frequency distribution table indicates the specific response options of the
dichotomous variable (in this example, male and female). The second column contains the frequencies (counts or numbers) of individuals in each response category (the numbers of men and women, respectively). The third column contains the relative frequencies, which are computed by dividing the frequency in each response category by the sample
size (e.g., 1625 / 3539 = 0.459). The relative frequencies are often expressed as percentages by multiplying by 100 and are most often used to summarize dichotomous v ariables. For example, in this sample 45.9% are men and 54.1% are women. Another example of a frequency distribution table is presented in Table 4–3, showing the distribution of
treatment with anti-hypertensive medication in persons attending the seventh examination of the Framingham Offspring Study. Notice that there are only n = 3532 valid responses, although the sample size is actually n = 3539. There are seven individuals with missing data on this particular question. Missing data occur in studies for a variety of
reasons. When there is very little missing data (e.g., less than 5%) and there is no apparent 38 CHAPTER 4 Summarizing Data Collected in the Sample pattern to the missingness (e.g., there is no s ystematic reason for missing data), then statistical analyses based on the available data are generally appropriate. However, if there is extensive missing
data or if there is a pattern to the missingness, then caution must be exercised in performing statistical analyses. Techniques for handling missing data are beyond the scope of this book; more details can be found in Little and Rubin.3 From Table 4–3, we can see that 34.5% of the participants are currently being treated for hypertension. Sometimes it
is of interest to compare two or more groups on the basis of a dichotomous outcome variable. For e xample, suppose we wish to compare the extent of treatment with anti-hypertensive medication in men and women. Table 4–4 summarizes treatment with anti-hypertensive medication in men and women attending the seventh examination of the
Framingham Offspring Study. The first column of the table indicates the sex of the participant. Sex is a dichotomous variable, and in this example it is used to distinguish the comparison groups (men and women). The outcome variable is also a dichotomous variable and represents treatment with antihypertensive medication or not. A total of n = 611
men and n = 608 women are on anti-hypertensive treatment. Because there are different numbers of men and women (1622 versus 1910) in the study sample, comparing frequencies (611 versus 608) is not the most appropriate comparison. The frequencies TABLE 4–3 Frequency Distribution Table for Treatment with Anti-Hypertensive Medication
Frequency Relative Frequency (%) No treatment 2313 65.5 Treatment 1219 34.5 Total 3532 100.0 TABLE 4–4 Treatment with Anti-Hypertensive Medication in Men and Women Attending the Seventh Examination of the Framingham Offspring Study n Number on Treatment Relative Frequency (%) Male 1622 611 37.7 Female 1910 608 31.8 Total
3532 1219 34.5 i ndicate that almost equal numbers of men and women are on treatment. A more appropriate comparison is based on relative frequencies, 37.7% versus 31.8%, which incorporate the different numbers of men and women in the sample. Notice that the sum of the rightmost column is not 100%, as it was in previous examples. In this
example, the b ottom row contains data on the total sample and 34.5% of all participants are being treated with anti-hypertensive medication. In Chapter 6 and Chapter 7, we will discuss formal m ethods to compare relative frequencies between groups. 4.1.2 Bar Charts for Dichotomous Variables Graphical displays are very useful for summarizing
data. There are many options for graphical displays, and many widely available software programs offer a variety of displays. However, it is important to choose the graphical display that accurately conveys information in the sample. We discuss data visualization in detail in Chapter 12. The appropriate graphical display depends on the type of
variable being analyzed. Dichotomous variables are best summarized using bar charts. The response options (yes/no, present/absent) are shown on the horizontal axis, and either the frequencies or relative frequencies are plotted on the vertical axis, producing a frequency bar chart or relative frequency bar chart, r espectively. Figure 4–1 is a
frequency bar chart depicting the distribution of men and women attending the seventh examination of the Framingham Offspring Study. The horizontal axis displays the two response options (male and female), and the vertical axis displays the frequencies (the numbers of men and women who attended the seventh examination). Figure 4–2 is a
relative frequency bar chart of the distribution of treatment with anti-hypertensive medication measured in the seventh examination of the Framingham Offspring Study. Notice that the vertical axis in Figure 4–2 displays relative frequencies and not frequencies, as was the case in Figure 4–1. In Figure 4–2, it is not necessary to show both responses as
the relative frequencies, expressed as percentages, sum to 100%. If 65.5% of the sample is not being treated, then 34.5% must be on treatment. These types of bar charts are very useful for comparing relative frequencies between groups. Figure 4–3 is a relative frequency bar chart describing treatment with anti-hypertensive medication in men
versus women attending the seventh examination of the Framingham Offspring Study. Notice that the vertical axis displays relative frequencies and in this example, 37.7% of men were using anti-hypertensive medications as compared to 31.8% of women. Figure 4–4 is an alternative display of the same data. Notice the scaling of the vertical axis. How
do Dichotomous Variables FIGURE 4–1 Frequency Bar Chart of Sex Distribution 1950 1900 1850 Frequency 1800 1750 1700 1650 1600 1550 1500 1450 Male Female Gender Notice that there is a space between the two response options (male and female). This is important for dichotomous and categorical variables. FIGURE 4–2 Relative Frequency
Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication 70 Relative Frequency % 60 50 40 30 20 10 0 No Yes Treatment with Anti-hypertensives 39 CHAPTER 4 Summarizing Data Collected in the Sample FIGURE 4–3 Relative Frequency Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication by Sex % Using
Anti-Hypertensive Medication 40 35 30 25 20 15 10 5 0 Male Female Gender FIGURE 4–4 Relative Frequency Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication by Sex 100 % Using Anti-Hypertensive Medication 40 90 80 70 60 50 40 30 20 10 0 Male Femaie Gender Ordinal and Categorical Variables 41 FIGURE 4–5 Relative
Frequency Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication by Sex % Using Anti-Hypertensive Medication 40 39 38 37 36 35 34 33 32 31 30 Female Male Gender the r elative frequencies compare visually? Finally, consider a third d isplay of the same data, shown in Figure 4–5. How do the relative frequencies c ompare? The
axes in any graphical display should be scaled to accommodate the range of the data. While relative frequencies can, in theory, range from 0% to 100%, it is not necessary to always scale the axes from 0% to 100%. It is also potentially misleading to restrict the scaling of the vertical axis, as was done in Figure 4–3, to exaggerate the difference in the
use of anti-hypertensive medication between men and women, at least from a visual standpoint. In this example, the relative frequencies are 31.8% and 37.7%, and thus scaling from 0% to 40% is appropriate to accommodate the data. It is always important to label axes clearly so that readers can appropriately interpret the data. 4.2 ORDINAL AND
CATEGORICAL VARIABLES Ordinal and categorical variables have a fixed number of response options that are ordered and unordered, respectively. Ordinal and categorical variables typically have more than two distinct response options, whereas dichotomous variables have exactly two response options. Summary statistics for ordinal and
categorical variables again focus primarily on relative frequencies (or percentages) of responses in each response c ategory. 4.2.1 Descriptive Statistics for Ordinal and Categorical Variables Consider again a study of cardiovascular risk factors such as the Framingham Heart Study. In the study, we might collect information on participants such as
their blood pressure, total cholesterol, and body mass index (BMI). Often, clinicians classify patients into categories. Each of these variables—blood pressure, total cholesterol, and BMI—are continuous variables. In this section, we organize continuous measurements into ordinal categories. For example, the NHLBI and the American Heart Association
use the classification of blood pressure given in Table 4–1.1,4,5 The American Heart Association uses the following classification for total cholesterol levels: desirable, less than 200 mg/dl; borderline high risk, 200–239 mg/dl; and high risk, 240 mg/dl or more.4 Body mass index (BMI) is computed as the ratio of weight in kilograms to height in meters
squared and the following categories are often used: underweight, less than 18.5; normal weight, 18.5–24.9; overweight, 25.0–29.9; and obese, 30.0 or greater. These are all examples of ordinal variables. In each case, it is healthier to be in a lower category. Example 4.2. Using data from the seventh examination of the offspring in the Framingham
Heart Study (n = 3539), we create the categories as defined previously for blood 42 CHAPTER 4 Summarizing Data Collected in the Sample pressure, total cholesterol, and BMI. Frequency distribution tables, similar to those presented for dichotomous data, are also used to summarize categorical and ordinal variables. Table 4–5 is a frequency
distribution table for the ordinal blood pressure variable. The mutually exclusive (nonoverlapping) and exhaustive (covering all possible options) categories are shown in the first column of the table. The frequencies, or numbers of participants in each response category, are shown in the middle column, and the relative frequencies, as percentages, are
shown in the rightmost column. Notice that there are only n = 3533 valid responses, whereas the sample size is n = 3539. There are six individuals with missing blood pressure data. More than a third of the sample (34.1%) has normal blood pressure, 41.1% are classified as pre-hypertension, 18.5% have Stage I hypertension, and 6.3% have Stage II
hypertension. With ordinal variables, two additional columns are often displayed in the frequency distribution table, called the cumulative frequency and cumulative relative frequency, respectively (see Table 4–6). The cumulative frequencies in this example reflect the number of patients at the particular blood pressure level or TABLE 4–5
Frequency Distribution Table for Blood Pressure Categories Frequency Relative Frequency (%) Normal 1206 34.1 Pre-hypertension 1452 41.1 Stage I hypertension 653 18.5 Stage II hypertension 222 6.3 3533 100.0 Total below. For example, 2658 patients have normal blood pressure or pre-hypertension. There are 3311 patients with normal, pre-
hypertension, or Stage I hypertension. The cumulative relative frequencies are very useful for summarizing ordinal variables and indicate the percent of patients at a particular level or below. In this example, 75.2% of the patients are not classified as hypertensive (i.e., they have normal blood pressure or pre-hypertension). Notice that for the last
(highest) blood pressure category, the cumulative frequency is equal to the sample size (n = 3533) and the cumulative relative frequency is 100%, indicating that all of the patients are at the highest level or below. Table 4–7 shows the frequency distribution table for total cholesterol. The total cholesterol categories are defined as follows and again are
based on the measured cholesterol values. Persons are classified as having desirable total cholesterol if the measured total cholesterol is less than 200 mg/dl; borderline high total cholesterol, 200–239 mg/dl; and high total cholesterol, 240 mg/dl or more. At the seventh examination of the Framingham Offspring Study, 51.6% of the patients are
classified as having desirable total cholesterol levels and another 34.3% have borderline high total cholesterol. Using the cumulative relative frequencies, we can summarize the data as follows: 85.9% of patients have total cholesterol levels that are desirable or borderline high. The remaining 14.1% of patients are classified as having high cholesterol.
Table 4–8 is a frequency distribution table for the ordinal BMI variable. Both weight and height are measured directly and BMI is computed as described previously. In the seventh examination of the Framingham Offspring Study s ample, 28% of the patients are classified as normal weight, 41.3% are classified as overweight, and 30.1% are classified
as obese. Using the cumulative relative frequencies, almost 70% are underweight, normal weight, or overweight. This is equivalent to reporting that 70% of the sample is not obese. TABLE 4–6 Frequency Distribution Table for Blood Pressure Categories with Cumulative Frequency and Cumulative Relative Frequency Cumulative Relative Frequency
(%) Frequency Relative Frequency (%) Cumulative Frequency Normal 1206 34.1 1206 34.1 Pre-hypertension 1452 41.1 2658 75.2 Stage I hypertension 653 18.5 3311 93.7 Stage II hypertension 222 6.3 3533 100.0 3533 100.0 Total Ordinal and Categorical Variables 43 TABLE 4–7 Frequency Distribution Table for Total Cholesterol Categories
Frequency Relative Frequency (%) Cumulative Frequency Desirable 1712 51.6 1712 Borderline high Cumulative Relative Frequency (%) 51.6 1139 34.3 2851 85.9 High 469 14.1 3320 100.0 Total 3320 100.0 TABLE 4–8 Frequency Distribution Table for Body Mass Index Categories Frequency Underweight Normal weight Relative Frequency (%) 20
0.6 Cumulative Frequency Cumulative Relative Frequency (%) 20 0.6 932 28.0 952 28.6 Overweight 1374 41.3 2326 69.9 Obese 1000 30.1 3326 100.0 Total 3326 100.0 Table 4–5 through Table 4–8 contain summary statistics for ordinal variables. The key summary statistics for ordinal variables are relative frequencies and cumulative relative -
frequencies. Table 4–9 through Table 4–11 contain summary statistics for categorical variables. Categorical variables are variables with two or more distinct responses but the responses are unordered. Some examples of categorical variables measured in the Framingham Heart Study include marital status, handedness, and smoking status. For
categorical variables, frequency distribution tables with frequencies and relative frequencies provide appropriate summaries. Cumulative frequencies and cumulative relative frequencies are generally not useful for categorical variables, as it is usually not of interest to combine categories as there is no inherent ordering. Table 4–9 is a frequency
distribution table for the categorical marital status variable. The mutually exclusive and exhaustive categories are shown in the first column of the table. The frequencies, or numbers of participants in each response category, are shown in the middle column, and the relative frequencies, as percents, are shown in the rightmost column. There are n =
3530 valid responses to the marital status question. A total of 9 participants did not TABLE 4–9 Frequency Distribution Table for Marital Status Frequency Single Married Relative Frequency (%) 203 5.8 2580 73.1 Widowed 334 9.5 Divorced 367 10.4 Separated 46 1.3 3530 100.0 Total provide marital status data. The majority of the sample is
married (73.1%) and approximately 10% of the s ample is divorced, another 10% is widowed, 6% is single, and 1% is separated. The relative frequencies are the most relevant statistics used to describe a categorical variable. Cumulative frequencies and cumulative relative frequencies are not generally informative descriptive statistics for categorical
variables. 44 CHAPTER 4 Summarizing Data Collected in the Sample TABLE 4–10 Frequency Distribution Table for Dominant Hand TABLE 4–12 Current Smoking in the Framingham Offspring Study by Date of Exam Frequency Relative Frequency (%) Exam Cycle Dates Current Smokers (%) 3143 89.5 1 Aug 1971–Sept 1975 59.7 Left 370 10.5 2
Oct 1979–Oct 1983 28.5 Total 3513 100.0 3 Dec 1983–Sept 1987 23.9 4 Apr 1987–Sept 1991 21.7 5 Jan 1991–June 1995 17.4 6 Jan 1995–Sept 1998 13.8 7 Sept 1998–Oct 2001 13.6 Right TABLE 4–11 Frequency Distribution Table for Smoking Status Frequency Relative Frequency (%) Nonsmoker 1330 37.6 Former 1724 48.8 Current 482 13.6 3536
100.0 Total Marital status is a categorical variable, so there is no rdering to the responses and therefore the first column can o be organized differently. For example, sometimes responses are presented from the most frequently to least frequently occurring in the sample, and sometimes responses are presented alphabetically. Any ordering is
appropriate. In contrast, with ordinal variables there is an ordering to the responses and therefore response options can only be presented either from highest to lowest (healthiest to unhealthiest) or vice versa. The response options within an ordinal scale cannot be rearranged. Table 4–10 is a frequency distribution table for a dichotomous categorical
variable. Dichotomous variables are a special case of categorical variables with exactly two response options. Table 4–10 displays the distribution of the dominant hand of participants who attended the seventh examination of the Framingham Offspring Study. The response options are "right" or "left." There are n = 3513 valid responses to the
dominant hand assessment. A total of 26 participants did not provide data on their dominant hand. The majority of the Framingham sample is right-handed (89.5%). Table 4–11 is a frequency distribution table for a categorical variable reflecting smoking status. Smoking status here is measured as nonsmoker, former smoker, or current smoker. There
are n = 3536 valid responses to the smoking status questions. Three participants did not provide adequate data to be classified. Almost half of the sample is former smokers (48.8%), over a third (37.6%) has never smoked, and approximately 14% are current smokers. The a dverse health effects of smoking have been a major focus of public health
messages in recent years, and the percentage of participants reporting current smokers must be interpreted relative to the time of the study. Table 4–12 shows the proportions of participants reporting currently smoking at the time of each examination of the Framingham offspring. The dates of each exam are also provided. In the next two sections,
we present graphical displays for ordinal and categorical variables, respectively. Whereas the numerical summaries for ordinal and categorical variables are identical (at least in terms of the frequencies and relative frequencies), graphical displays for ordinal and categorical variables are different in a very important way. 4.2.2 Histograms for
Ordinal Variables Histograms are appropriate graphical displays for ordinal variables. A histogram is different from a bar chart in one important feature. The horizontal axis of a histogram shows the distinct ordered response options of the ordinal variable. The vertical axis can show either frequencies or relative f requencies, producing a frequency
histogram or relative frequency histogram, respectively. The bars are centered over each response option and scaled according to frequencies or relative frequencies as desired. The difference between a histogram and a bar chart is that the bars in a histogram run together; there is no space between adjacent responses. This reinforces the idea that
the response categories are ordered and based on an underlying continuum. This underlying continuum may or may not be measurable. Ordinal and Categorical Variables Figure 4–6 is a frequency histogram for the blood pressure data displayed in Table 4–5. The horizontal axis displays the ordered blood pressure categories and the vertical axis
displays the frequencies or numbers of participants classified in each category. The histogram immediately conveys the message that the majority of participants are in the lower (healthier) two categories of the distribution. A small number of participants are in the Stage II h ypertension category. The histogram in Figure 4–7 is a relative frequency
histogram for the same data. Notice that the figure is the same except for the FIGURE 4–6 Frequency Histogram for Blood Pressure Categories 1600 1400 Frequency 1200 1000 800 600 400 200 0 Normal Pre-hypertension Stage I hypertension Stage II hypertension Blood Pressure FIGURE 4–7 Relative Frequency Histogram for Blood Pressure
Categories 45 40 Relative Frequency % 35 30 25 20 15 10 5 0 Normal 45 Pre-hypertension Stage I hypertension Blood Pressure Stage II hypertension 46 CHAPTER 4 Summarizing Data Collected in the Sample vertical axis, which is scaled to accommodate relative frequencies instead of frequencies. Usually, relative frequency histograms are
preferred over frequency histograms, as the relative frequencies are most appropriate for summarizing the data. From Figure 4–7, we can see that approximately 34% of the participants have normal blood pressure, 41% have pre-hypertension, just under 20% have Stage I hypertension, and 6% have Stage II hypertension. Figure 4–8 is a relative
frequency histogram for the total cholesterol variable summarized in Table 4–7. The bars of the histogram run together to reflect the fact that there is an underlying continuum of total cholesterol measurements. From Figure 4–8, we see that over 50% of the participants have desirable total cholesterol levels and just under 15% have high total
cholesterol levels. The horizontal axis could be scaled differently. Figure 4–9 makes the continuum of total cholesterol underlying the categories used here to summarize the data more obvious. Another alternative is to mark the transition points. In Figure 4–9, the horizontal axis could be labeled 200 and 240 at the points of intersection of adjacent
bars. Figure 4–10 is a relative frequency histogram for the BMI data summarized in Table 4–8. The ordered BMI categories are shown in text along the horizontal axis and relative frequencies, as percents, are displayed along the vertical axis. From Figure 4–10, it is immediately evident that a small percentage of the participants are underweight and
that the majority of the participants are overweight or obese, with the former more likely than the latter. The horizontal axis of Figure 4–10 could be scaled differently to show the numerical values of BMI that define the ordinal categories or with labels to indicate the BMI values that separate adjacent bars (e.g., 18.5, 25.0, 30.0). 4.2.3 Bar Charts
for Categorical Variables Bar charts are appropriate graphical displays for categorical variables. Bar charts for categorical variables with more than two responses are constructed in the same fashion as bar charts for dichotomous variables. The horizontal axis of the bar chart again displays the distinct responses of the categorical variable. Because
the responses are unordered, they can be arranged in any order (e.g., from the most frequently to least frequently occurring in the sample, or alphabetically). The vertical axis can show either frequencies or relative frequencies, producing a frequency bar chart or relative frequency bar chart, respectively. The bars are centered over each response
option and scaled according to frequencies or relative frequencies as desired. Figure 4–11 is an example of a frequency bar chart displaying the distribution of marital status in the participants who attended the seventh examination of the Framingham FIGURE 4–8 Relative Frequency Histogram for Total Cholesterol Categories 60 Relative Frequency
% 50 40 30 20 10 0 Desirable Borderline high Total Cholesterol High Ordinal and Categorical Variables FIGURE 4–9 Relative Frequency Histogram for Total Cholesterol Categories 60 Relative Frequency % 50 40 30 20 10 0 200–239 200 240 Total Cholesterol FIGURE 4–10 Relative Frequency Histogram for BMI Categories 45 40 Relative Frequency
% 35 30 25 20 15 10 5 0 Underweight Normal Overweight BMI Obese 47 48 CHAPTER 4 Summarizing Data Collected in the Sample Offspring Study. The distinct response options are shown on the horizontal axis, arranged as they were in Table 4–9. Frequencies, or numbers of respondents selecting each response, are shown on the vertical axis. It
is immediately evident from the bar chart that the majority of respondents are married. Notice the spaces between the bars, indicating a clear separation between distinct response options. Because there is no inherent ordering to the response options, as is always the case with a categorical variable, the horizontal axis can be scaled differently.
Figure 4–12 and Figure 4–13 show FIGURE 4–11 Frequency Bar Chart for Marital Status 3000 2500 Frequency 2000 1500 1000 500 0 Single Married Widowed Divorced Separated Single Separated Marital Status FIGURE 4–12 Frequency Bar Chart for Marital Status 3000 2500 Frequency 2000 1500 1000 500 0 Married Divorced Widowed Marital
Status Ordinal and Categorical Variables 49 FIGURE 4–13 Frequency Bar Chart for Marital Status 3000 2500 Frequency 2000 1500 1000 500 0 Divorced Married Separated Single Widowed Marital Status FIGURE 4–14 Relative Frequency Bar Chart for Dominant Hand 100 90 Relative Frequency % 80 70 60 50 40 30 20 10 0 Left Right Dominant
Hand alternative arrangements of the response options. Each of the bar charts displays identical data. All three presentations are appropriate as the categorical responses can be ordered in any way. Figure 4–12 arranges the responses from most frequently to least frequently occurring, and Figure 4–13 arranges the responses alphabetically. It is
more typical to present relative frequency bar charts with relative frequencies displayed as proportions or percentages on the vertical axis. Figure 4–14 is a relative frequency bar chart displaying the distribution of the dominant hand and is summarized in Table 4–10. Figure 4–14 clearly displays that the vast majority of participants in the seventh
examination of 50 CHAPTER 4 Summarizing Data Collected in the Sample FIGURE 4–15 Relative Frequency Bar Chart for Smoking Status 60 Relative Frequency % 50 40 30 20 10 0 Former Nonsmoker Current Smoking Status the Framingham Heart Study are right-handed. Because the dominant hand is a dichotomous variable, we only need to -
display the relative frequency for one bar, as the second is determined from the first. Figure 4–15 is a relative frequency bar chart displaying the distribution of smoking status measured at the seventh examination of the Framingham Offspring Study and summarized in Table 4–11. Figure 4–15 gives the clear impression that a small percentage of the
participants are current smokers. Almost 50% are former smokers. Again, the response options in Figure 4–14 and Figure 4–15 may be rearranged, as there is no inherent ordering to these ordinal scales. 4.3 CONTINUOUS VARIABLES Continuous variables, sometimes called measurement or quantitative variables, take on an unlimited number of -
distinct responses between a theoretical minimum value and maximum value. In a study of cardiovascular risk factors, we might measure participants' ages, heights, weights, systolic and dia stolic blood pressures, total serum cholesterol levels, and so on. The measured values for each of these continuous variables depend on the scale of
measurement. For example, in adult studies such as the Framingham Heart Study, age is usually measured in years. Studies of infants might measure age in days or even hours, whichever is more appropriate. Heights can be measured in inches or centimeters, weights can be measured in pounds or kilograms. Assuming weight is measured in pounds,
measurements might be to the nearest pound, the nearest tenth of a pound, or the nearest hundredth (e.g., 145, 145.1, 145.13), depending on the precision of the scale. 4.3.1 Descriptive Statistics for Continuous Variables To illustrate the computations of descriptive statistics in detail, we selected a small subset of the Framingham Heart Study data.
After performing computations by hand on the small subset, we provide descriptive statistics for the full sample that were generated by computer. Example 4.3. At the seventh examination of the offspring in the Framingham Heart Study (n = 3539), numerous continuous variables were measured, including systolic and diastolic blood pressure, total
serum cholesterol, height, and weight. Using the heights and weights measured on each participant, we can compute their BMI. In this study, height is m easured in inches and weight in pounds. The following formula is used to compute BMI using these metrics: BMI 5 703.03 3 Weight in pounds (He ight in inches)2 To illustrate the computation of
descriptive statistics for continuous variables, we randomly selected a subset of Continuous Variables 51 TABLE 4–13 Subsample of n 5 10 Participants Attending the Seventh Examination of the Framingham Offspring Study Participant ID Systolic Blood Pressure Diastolic Blood Pressure Total Serum Cholesterol Weight (lbs) Height (in.) BMI 1 141
76 199 138 63.00 24.4 2 119 64 150 183 69.75 26.4 3 122 62 227 153 65.75 24.9 4 127 81 227 178 70.00 25.5 5 125 70 163 161 70.50 22.8 6 123 72 210 206 70.00 29.6 7 105 81 205 235 72.00 31.9 8 113 63 275 151 60.75 28.8 9 106 67 208 213 69.00 31.5 10 131 77 159 142 61.00 26.8 10 participants who completed the seventh examination of the
Framingham Offspring Study. The data values are shown in Table 4–13. The first column contains a unique identification number for each participant, the second through sixth columns contain the actual measurements on participants, and the rightmost column contains the BMI computed using the formula shown. Descriptive statistics for each
continuous variable are now computed. Formulas for the computations are presented within examples and then summarized at the end of this chapter. The first summary statistic for any continuous variable (as well as for dichotomous, categorical, and ordinal variables) is the sample size. The sample size here is n = 10. It is always important to report
the sample size to convey the size of the study. Larger studies are generally viewed more favorably, as larger sample sizes generally produce more precise results. However, there is a point at which increasing the sample size does not materially increase the precision of the analysis. (Sample size computations are discussed in detail in Chapter 8.)
Because this sample is small (n 5 10), it is relatively easy to summarize the sample by inspecting the observed values. Suppose we first consider the diastolic blood pressures. To facilitate interpretation, we order the diastolic blood pressures from smallest to largest: 62 63 64 67 70 72 76 77 81 81 Diastolic blood pressures less than
80 are considered normal (see Table 4–1); thus the participants in this sample can be summarized as generally having normal diastolic pressures. There are two participants with diastolic blood pressures of 81, but they hardly exceed the upper limit of the "normal" classification. The diastolic blood pressures in this sample are not all identical (with
the exception of the two measured values of 81) but they are relatively similar. In general, the participants in this sample can be described as having healthy diastolic blood pressures from a clinical standpoint. For larger samples, such as the full seventh examination of the Framingham Offspring Study with n = 3539, it is i mpossible to inspect
individual values to generate a summary, so summary statistics are necessary. There are two general aspects of a useful summary for a continuous v ariable. The first is a description of the center or average of the data (i.e., what is a typical value), and the second addresses variability in the data. Using the diastolic blood pressures, we now illustrate
the computation of several statistics that describe the average value and the variability of the data. In biostatistics, the term "average" is a very general term. There are several statistics that describe the average value of a continuous variable. The first is probably the most familiar—the sample mean. The sample mean is computed by summing all of
the values and dividing by the sample size. For the sample of diastolic blood pressures, the sample mean is computed as follows: 62+63+64+67+70+72+76+77+81+81 Sample Mean 5 }}}} 10 713 5 }} 5 71.3 10 52 CHAPTER 4 Summarizing Data Collected in the Sample To simplify the formulas for sample statistics (and for population
parameters), we usually denote the variable of interest as X. X is simply a placeholder for the variable being analyzed. Here X = diastolic blood pressure. The sample mean is denoted X w (read "X bar"), and the formula for the sample mean is X= ΣX n TABLE 4–14 Means and Medians of Variables in Subsample of Size n 5 10 Mean Median Diastolic
blood pressure 71.3 71.0 Systolic blood pressure 121.2 122.5 Total serum cholesterol 202.3 206.5 Weight (lbs) 176.0 169.5 where o indicates summation (i.e., the sum of the diastolic blood pressures in this sample). The mean diastolic blood pressure is X w = 71.3. When reporting summary statistics for a continuous variable, the convention is to
report one more decimal place than the number of decimal places measured. Here, systolic and diastolic blood pressures, total serum cholesterol, and weight are measured to the nearest integer, therefore the summary statistics are reported to the nearest tenths place. Height is measured to the nearest quarter inch (hundredths); therefore the
summary statistics are reported to the nearest thousandths place. BMI is computed to the nearest tenth, so summary statistics are reported to the nearest hundredths place. The sample mean is one measure of the average diastolic blood pressure. A second measure of the average value is the sample median. The sample median is the middle value in
the ordered dataset, or the value that separates the top 50% of the values from the bottom 50%. When there is an odd number of observations in the sample, the median is the value that holds as many values above it as below it in the ordered dataset. When there is an even number of observations in the sample, the median is defined as the mean of
the two middle values in the ordered dataset. In the sample of n = 10 diastolic blood pressures, the two middle values are 70 and 72, and thus the median is (70 + 72)/2 = 71. Half of the diastolic blood pressures are above 71 and half are below. The mean and median provide different information about the average value of a continuous variable.
Suppose the sample of 10 diastolic blood pressures looked like this: is preferred over the mean when there are extreme v alues (values either very small or very large relative to the others). When there are no extreme values, the mean is the preferred measure of a typical value, in part because each observation is considered in the computation of the
mean. When there are no extreme values in a sample, the mean and median of the sample will be close in value. Table 4–14 displays the sample means and medians for each of the continuous measures in the sample of n = 10. For each continuous variable measured in this subsample of participants, the means and medians are not identical but are
relatively close in value, suggesting that the mean is the most a ppropriate summary of a typical value for each of these variables. (If the mean and median are very different, it suggests that there are outliers affecting the mean.) A third measure of a typical value for a continuous variable is the mode. The mode is defined as the most frequent value.
The mode of the diastolic blood pressures is 81, the mode of the total cholesterol levels is 227, and the mode of the heights is 70 because these values each appear twice, whereas the other values only appear once. For each of the other continuous variables, there are 10 distinct values and thus there is no mode (because no value appears more
frequently than any other). Suppose the diastolic blood pressures looked like: 62 63 64 67 70 72 76 77 81 140 62 63 64 64 70 72 76 77 81 81 The sample mean for this sample is X w = 772/10 = 77.2. This does not represent a typical value, as the majority of d iastolic blood pressures in this sample are below 77.2.
The extreme value of 140 is affecting the computation of the mean. For this same sample, the median is 71. The median is unaffected by extreme or outlying values. For this reason, the median In this sample, there are two modes, 64 and 81. The mode is a useful summary statistic for a continuous variable. It is presented not instead of either the mean
or the median but rather in addition to the mean or median. The second aspect of a continuous variable that must be summarized is the variability in the sample. A relatively crude Height (in.) 67.175 69.375 Body mass index (BMI) 27.26 26.60 Continuous Variables TABLE 4–15 Ranges of Variables in Subsample of Size n 5 10 Minimum Maximum
Range Diastolic blood pressure 62 81 19 Systolic blood pressure 105 141 36 Total serum cholesterol 150 275 125 Weight (lbs) 138 235 97 TABLE 4–16 Deviations from the Mean Diastolic Blood Pressure (X) Deviation from the Mean (X2X w) 76 4.7 64 −7.3 62 −9.3 81 9.7 70 –1.3 72 0.7 Height (in.) 60.75 72.00 11.25 81 9.7 Body mass index
(BMI) 63 –8.3 22.8 31.9 9.1 67 –4.3 77 5.7 SX = 713 S(X2X w) = 0 yet important measure of variability in a sample is the sample range. The sample range is computed as follows: Sample range = Maximum value – Minimum value Table 4–15 displays the sample ranges for each of the c ontinuous measures in the subsample of n = 10 observations. The
range of a variable depends on the scale of measure ment. The blood pressures are measured in millimeters of mercury, total cholesterol is measured in milligrams per deciliter, weight in pounds, and so forth. The range of total serum cholesterol is large, with the minimum and maximum in the sample of size n = 10 differing by 125 units. In contrast, the
heights of participants are more homogeneous, with a range of 11.25 inches. The range is an important descriptive statistic for a continuous variable, but it is based on only two v alues in the dataset. Like the mean, the sample range can be affected by extreme values and thus it must be interpreted with caution. The most widely used measure of
variability for a continuous variable is called the standard deviation, which is described now. Assuming that there are no extreme or outlying values of the variable, the mean is the most appropriate summary of a typical value. To summarize variability in the data, we specifically estimate the variability in the sample around the sample mean. If all of
the observed values in a sample are close to the sample mean, the standard deviation is small (i.e., close to zero), and if the observed values vary widely around the sample mean, the standard deviation is large. If all of the values in the sample are identical, the sample standard deviation is zero. 53 In the sample of n = 10 diastolic blood pressures, we
found X w = 71.3. Table 4–16 displays each of the observed values along with the respective deviations from the sample mean. Thedeviations from the mean reflect how far each individual's diastolic blood pressure is from the mean diastolic blood pressure. The first participant's diastolic blood pressure is 4.7 units above the mean, whereas the second
participant's diastolic blood pressure is 7.3 units below the mean. What we need is a summary of these deviations from the mean, in p articular a measure of how far (on average) each participant is from the mean diastolic blood pressure. If we compute the mean of the deviations by summing the deviations and dividing by the sample size, we run into
a problem: the sum of the deviations from the mean is zero. This will a lways be the case as it is a property of the sample mean—the sum of the d eviations below the mean will always equal the sum of the deviations above the mean. The goal is to capture the magnitude of these deviations in a summary measure. To address this problem of the
deviations summing to zero, we could take the absolute values or the squares of each deviation from the mean. Both methods address the problem. The more popular method to summarize the deviations from the mean involves squaring the d eviations. (Absolute values are difficult in mathematical proofs, which are beyond the scope of this book.)
Table 4–17 displays each of the observed values, the respective deviations from the sample mean, and the squared deviations from the mean. The squared deviations are interpreted as follows. The first participant's squared deviation is 22.09, meaning that 54 CHAPTER 4 Summarizing Data Collected in the Sample his or her diastolic blood pressure
is 22.09 units squared from the mean diastolic blood pressure. The second participant's diastolic blood pressure is 53.29 units squared from the mean diastolic blood pressure. A quantity that is often used to measure variability in a sample is called the sample variance, and it is essentially the mean of the squared deviations. The sample variance is
denoted s2 and is computed as follows: s 2 ∑(X − X) = 2 n −1 The sample variance is not actually the mean of the squared deviations because we divide by (n – 1) instead of n. In s tatistical inference (which is described in detail in Chapters 6, 7, 9, 10, and 11), we make generalizations or estimates of population parameters based on sample statistics.
If we were to compute the sample variance by taking the mean of the TABLE 4–17 Squared Deviations from the Mean Diastolic Blood Pressure (X) Deviation from the Mean (X2X w) Squared Deviation from the Mean (X2X w)2 76 4.7 22.09 64 −7.3 53.29 62 −9.3 86.49 81 9.7 94.09 70 –1.3 1.69 72 0.7 0.49 81 9.7 94.09 63 –8.3 68.89 67 –4.3 18.49 77
5.7 32.49 SX = 713 S(X2X w) = 0 S(X2X w)2 5 472.10 squared deviations and divide by n, we would consistently underestimate the true population variance. Dividing by (n – 1) produces a better estimate of the population variance. The sample variance is nonetheless usually interpreted as the average squared deviation from the mean. In this sample
of n = 10 diastolic blood pressures, the sample variance is s2 = 472.10 / 9 = 52.46. Thus, on average, diastolic blood pressures are 52.46 units squared from the mean diastolic blood pressure. Because of the squaring, the variance is not particularly interpretable. The more common measure of variability in a sample is the sample standard deviation,
defined as the square root of the sample variance: 2 s= s = ∑(X − X) n −1 The sample standard deviation of the diastolic blood ressures is s = 52.46 = 7.2. On average, diastolic blood presp sures are 7.2 units from (above or below) the mean diastolic blood pressure. When a dataset has outliers, or extreme values, we summarize a typical value using
the median as opposed to the mean. When a dataset has outliers, variability is often summarized by a statistic called the interquartile range (IQR). The interquartile range is the difference between the first and third quartiles. The first quartile, denoted Q1, is the value in the dataset that holds 25% of the values below it. The third quartile, denoted Q3,
is the value in the dataset that holds 25% of the values above it. The IQR is defined as IQR 5 Q3 2 Q1 For the sample of n 5 10 diastolic blood pressures, the median is 71 (50% of the values are above 71 and 50% are below). The quartiles can be computed in the same way we computed the median, but we consider each half of the dataset separately
(see Figure 4–16). FIGURE 4–16 Computing the Quartiles Upper half Lower half 62 63 64 Lower quarter Q1 = 64 67 2 70 72 Median = 71 76 77 81 Upper quarter Q3 = 77 81 Continuous Variables There are five values below the median (lower half) and the middle value is 64, which is the first quartile. There are five values above the median (upper
half) and the middle value is 77, which is the third quartile. The IQR is 77 2 64 5 13; the IQR is the range of the middle 50% of the data. When the sample size is odd, the median and quartiles are determined in the same way. Suppose in the previous example that the l owest value (62) was excluded and the sample size was n 5 9. The median and
quartiles are indicated graphically in Figure 4–17. When the sample size is 9, the median is the m iddle number, 72. The quartiles are determined in the same way, looking at the lower and upper halves, respectively. There are four values in the lower half, so the first quartile is the mean of the two middle values in the lower half, (64 1 67) / 2 5 65.5.
The same approach is used in the upper half to determine the third quartile, (77 1 81) / 2 5 79. Some statistical computing packages use slightly different algorithms to compute the quartiles. Results can be different, especially for small samples. When there are no outliers in a sample, the mean and standard deviation are used to summarize a typical
value and the variability in the sample, respectively. When there are o utliers in a sample, the median and IQR are used to summarize a typical value and the variability in the sample, r espectively. An important issue is determining whether a sample has outliers or not. There are several methods to determine outliers in a sample. A very popular
method is based on the following: Outliers are values below Q1 2 1.5 3 (Q3 2 Q1) or above Q3 1 1.5 3 (Q3 2 Q1), or equivalently, values below Q1 2 1.5 3 IQR or above Q3 1 1.5 3 IQR These are referred to as Tukey fences.6 For the diastolic blood pressures, the lower limit is 64 2 1.5 3 (77 2 64) 5 44.5 and the upper limit is 77 1 1.5 3 (77 2 64) 5 96.5.
The diastolic blood pressures range from 62 to 81; therefore there are no outliers. The best summary of a typical diastolic blood pressure is the mean (X w = 71.3) and the best summary of v ariability is given by the standard deviation (s = 7.2). Table 4–18 displays the means, standard deviations, medians, quartiles, and IQRs for each of the continuous
FIGURE 4–17 Median and Quartiles for n = 9 63 64 67 70 72 76 77 81 Median = 72 Lower quarter 55 81 Upper quarter Q1 = (64 + 67)/2 = 65.5 Q3 = (77 + 81)/2 = 79 TABLE 4–18 Summary Statistics on n 5 10 Participants Attending the Seventh Examination of the Framingham Offspring Study Mean (X) Standard Deviation (s) Median Q1 Q3 IQR
Systolic blood pressure 121.2 11.1 122.5 113.0 127.0 14.0 Diastolic blood pressure 71.3 7.2 71.0 64.0 77.0 13.0 Total serum cholesterol 202.3 37.7 206.5 163.0 227.0 64.0 Weight (lbs) 176.0 33.0 169.5 151.0 206.0 55.0 Height (in.) 67.175 4.205 69.375 63.0 70.0 7.0 Body mass index (BMI) 27.26 3.10 26.60 24.9 29.6 4.7 56 CHAPTER 4 Summarizing
Data Collected in the Sample TABLE 4–19 Limits for Assessing Outliers in Characteristics Measured in n 5 10 Participants Attending the Seventh Examination of the Framingham Offspring Study Minimum Systolic blood pressure 105 Maximum Lower Limita 141 Upper Limitb 92 148 Diastolic blood pressure 62 81 Total serum cholesterol 150 275 67
44.5 323 96.5 Weight (lbs) 138 235 68.5 288.5 Height (in.) 60.75 72.00 52.5 80.5 Body mass index (BMI) 22.8 31.9 17.85 36.65 a b Determined by Q1 – 1.5 × (Q3 – Q1). Determined by Q3 + 1.5 × (Q3 – Q1). TABLE 4–20 Summary Statistics on Sample of Participants Attending the Seventh Examination of the Framingham Offspring Study (n 5 3539)
Mean (X) Standard Deviation (s) Median Q1 Q3 IQR Systolic blood pressure 127.3 19.0 125.0 114.0 138.0 24.0 Diastolic blood pressure 74.0 9.9 74.0 67.0 80.0 13.0 Total serum cholesterol 200.3 36.8 198.0 175.0 223.0 48.0 Weight (lbs) 174.4 38.7 170.0 146.0 198.0 52.0 Height (in.) 65.957 3.749 65.750 63.000 68.750 5.75 Body mass index (BMI)
28.15 5.32 27.40 24.5 30.8 6.3 v ariables displayed in Table 4–13 in the subsample of n = 10 participants who attended the seventh examination of the Framingham Offspring Study. Table 4–19 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the v ariables in the
subsample of n = 10 participants. Are there outliers in any of the variables? Which statistics are most appropriate to summarize the average or typical value and the dispersion or variability? Because there are no suspected outliers in the subsample of n = 10 participants, the mean and standard deviation are the most appropriate statistics to
summarize average values and dispersion, respectively, of each of these characteristics. Table 4–20 displays the means, standard deviations, medians, quartiles, and IQRs for each of the continuous variables displayed in Table 4–13 in the full sample (n = 3539) of participants who attended the seventh examination of the Framingham Offspring Study.
Looking just at the means and medians, does it appear that any of the characteristics are subject to outliers in the full sample? Table 4–21 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the full sample (n = 3539) of participants who attended the
seventh examination of the Framingham Offspring Study. Are there outliers in any of the variables? Which statistics are most Continuous Variables 57 TABLE 4–21 Limits for Assessing Outliers in Characteristics Measured in Participants Attending the Seventh Examination of the Framingham Offspring Study Minimum Maximum Systolic blood
pressure 81.0 216.0 Diastolic blood pressure 41.0 114.0 Total serum cholesterol 83.0 357.0 Weight (lbs) 90.0 375.0 Height (in.) 55.00 Body mass index (BMI) 15.8 a b Lower Limita 78 47.5 103 Upper Limitb 174 99.5 295 68.0 276.0 78.75 54.4 77.4 64.0 15.05 40.25 Determined by Q1 – 1.5 × (Q3 – Q1). Determined by Q3 + 1.5 × (Q3 – Q1). a-
ppropriate to summarize the average or typical values and the dispersion or variability for each variable? In the full sample, each of the characteristics has outliers on the upper end of the distribution, as the maximum values exceed the upper limits in each case. There are also outliers on the low end for diastolic blood pressure and total cholesterol,
as the m inimums are below the lower limits. For some of these characteristics, the difference between the upper limit and the maximum (or the lower limit and the minimum) is small (e.g., height, systolic and diastolic blood pressures), whereas for others (e.g., total cholesterol, weight, and BMI) the difference is much larger. This method for
determining outliers is a popular one but is not generally applied as a hard and fast rule. In this application, it would be reasonable to present means and standard deviations for height and systolic and diastolic blood pressures, and medians and IQRs for total cholesterol, weight, and BMI. Another method for assessing whether a distribution is subject
to outliers or extreme values is through graphical displays. 4.3.2 Box-Whisker Plots for Continuous Variables Box-whisker plots are very useful plots for displaying the distribution of a continuous variable. In Example 4.3, we considered a subsample of n = 10 participants who attended the seventh examination of the Framingham Offspring Study. We
computed the following summary statistics on diastolic blood pressures. These statistics are sometimes referred to as quantiles, or percentiles, of the distribution. A specific quantile or percentile is a value in the dataset that holds a specific percentage of the values at or below it. For example, the first quartile is the 25th percentile, meaning that it
holds 25% of the values at or below it. The median is the 50th percentile, the third quartile is the 75th percentile, and the maximum is the 100th percentile (i.e., 100% of the values are at or below it). Minimum 62 Q1 64 Median 71 Q3 77 Maximum 81 A box-whisker plot is a graphical display of these percentiles. Figure 4–18 is a box-whisker plot of the
diastolic blood pressures measured in the subsample of n 5 10 participants described in Example 4.3. The horizontal lines represent (from the top) the maximum, the third quartile, the median (also indicated by the dot), the first quartile, and the minimum. The shaded box represents the middle 50% of the distribution (between the first and third
quartiles). A boxwhisker plot is meant to convey the distribution of a variable at a quick glance. Figure 4–19 is a box-whisker plot of the diastolic blood pressures measured in the full sample of participants who attended the seventh examination of the Framingham Offspring Study. In the full sample, we determined that there were outliers both at the
low and the high end (see Table 4–21). In Figure 4–19, the outliers are displayed as horizontal lines at the top and bottom of the distribution. At the low end of the distribution, there are five values that are considered outliers (i.e., values below 47.5, which was the lower limit for determining outliers). At the high end of the distribution, there are 12
values that are considered CHAPTER 4 Summarizing Data Collected in the Sample FIGURE 4–18 Box-Whisker Plot of Diastolic Blood Pressures in Subsample of n 5 10 Diastolic Blood Pressure 80 75 70 65 60 FIGURE 4–19 Box-Whisker Plot of Diastolic Blood Pressures in Participants Attending the Seventh Examination of the Framingham Offspring
Study 120 100 Diastolic Blood Pressure 58 80 60 40 Continuous Variables outliers (i.e., values above 99.5, which was the upper limit for determining o utliers). The "whiskers" of the plot (the notched horizontal lines) are the limits we determined for detecting outliers (47.5 and 99.5). Figure 4–20 is a box-whisker plot of the total serum cholesterol
levels measured in the full sample of participants who attended the seventh examination of the Framingham Offspring Study. In the full sample, we determined that there were outliers both at the low and the high end (see Table 4–21). Again in Figure 4–20, the outliers are displayed as horizontal lines at the top and bottom of the distribution. The
outliers of total cholesterol are more numerous than those we observed for diastolic blood pressure, particularly on the high end of the distribution. Box-whisker plots are very useful for comparing distributions. Figure 4–21 shows side-by-side box-whisker plots of the distributions of weight (in pounds) for men and women attending the seventh
examination of the Framingham Offspring Study. The figure clearly shows a shift in the distributions, with men having much higher weights. In fact, the 25th percentile of weight in men is approximately 180 lbs, equal to the 75th percentile in women. Specifically, 25% of the men weigh 180 lbs or less as compared to 75% of the women. There are a s-
ubstantial n umber of outliers at the high end of the distribution among both men and women. There are two outlying low values among men. Because men are generally taller than women (see Figure 4–22), it is not surprising that men have greater weights than women. A more appropriate comparison is of BMI (see Figure 4–23). The distributions of
BMI are similar for men and women. There are again a substantial number of outliers in the distributions in both men and women. However, when taking height into account (by comparing BMI instead of comparing weight), we see that the most extreme outliers are among the women. What statistics would be most appropriate to summarize typical
BMI for men and women? In the box-whisker plots, outliers are values that either exceed Q3 1 1.5 3 IQR or fall below Q1 2 1.5 3 IQR. Some statistical computing packages use the following to determine outliers: values which either exceed Q3 1 3 3 IQR FIGURE 4–20 Box-Whisker Plot of Total Serum Cholesterol Levels in Participants Attending the
Seventh Examination of the Framingham Offspring Study Total Cholesterol 350 250 150 50 59 CHAPTER 4 Summarizing Data Collected in the Sample FIGURE 4–21 Side-by-Side Box-Whisker Plots of Weight in Men and Women Attending the Seventh Examination of the Framingham Offspring Study Weight 350 250 150 50 Female Male Sex FIGURE
4–22 Side-by-Side Box-Whisker Plots of Height in Men and Women Attending the Seventh Examination of the Framingham Offspring Study 76 71 Height 60 66 61 55 51 Female Male Sex Summary 61 FIGURE 4–23 Side-by-Side Box-Whisker Plots of Body Mass Index in Men and Women Attending the Seventh Examination of the Framingham

Offspring Study Body Mass Index 60 50 40 30 20 10 Female Male Sex or fall below Q1 – 3 3 IQR, which would result in fewer observations being classified as outliers.7,8 The rule using 1.5 3 IQR is the more commonly applied rule to determine outliers. 4.4 SUMMARY The first important aspect of any statistical analysis is an appropriate summary of
the key variables. This involves first identifying the type of variable being analyzed. This step is extremely important, as the appropriate numerical and graphical summaries depend on the type of variable being analyzed. Variables are dichotomous, o rdinal, categorical, or continuous. The best numerical summaries for dichotomous, ordinal, and
categorical variables are relative frequencies. The best numerical summaries for continuous variables include the mean and standard deviation or the median and interquartile range, d epending on whether or not there are outliers in the sample. The mean and s tandard deviation, or the median and interquartile range, s ummarize location and
dispersion, respectively. The best graphical summary for dichotomous and categorical v ariables is a bar chart, and the best graphical summary for an ordinal variable is a histogram. Both bar charts and h istograms can be designed to display f requencies or relative frequencies, with the latter being the more popular display. Box-whisker plots
provide a very useful and informative summary for continuous variables. Boxwhisker plots are also useful for comparing the distributions of a continuous variable among mutually exclusive (i.e., non-overlapping) comparison groups. Figure 4–24 summarizes key statistics and graphical displays organized by variable type. 62 CHAPTER 4 Summarizing
Data Collected in the Sample? FIGURE 4–24 Key Statistics and Graphical Displays Variable Type Statistic/Graphical Display Dichotomous, Ordinal, or Categorical Relative Frequency Dichotomous or Categorical Frequency or Relative Frequency Bar Chart Ordinal Frequency or Relative Frequency Histogram Continuous Mean Definition Frequency/n
X= Standard Deviation s= ∑X n ∑(X − X)2 n − 1 Median Middle value in ordered dataset First Quartile Q1 5 Value holding 25% below it Third Quartile Interquartile Range Criteria for Outliers Box-Whisker Plot Q3 5 Value holding 25% above it IQR = Q3 2 Q1 Values below Q1 2 1.5 × (Q3 2 Q1) or above Q3 1 1.5 × (Q3 2 Q1) Practice Problems 4.5
PRACTICE PROBLEMS d. What proportion of the girls has total cholesterol less than 205? 4. The following data were collected as part of a study of coffee consumption among graduate students. The following reflect cups per day consumed: 1. A study is run to estimate the mean total cholesterol level in children 2 to 6 years of age. A sample of 9
participants is selected and their total cholesterol levels are measured as follows. 185 225 240 196 175 180 194 147 223 a. Compute the sample mean. b. Compute the sample standard deviation. c. Compute the median. d. Compute the first and third quartiles. e. Which measure, the mean or median, is a better measure of a
typical value? Justify. f. Which measure, the standard deviation or the interquartile range, is a better measure of dispersion? Justify. 2. Generate a box-whisker plot for the data in Problem 1. 3. The box-whisker plots in Figure 4–25 show the distributions of total cholesterol levels in boys and girls 10 to 15 years of age. a. What is the median total
cholesterol level in boys? b. Are there any outliers in total cholesterol in boys? Justify briefly. c. What proportion of the boys has total cholesterol less than 205? 3 4 6 8 2 1 0 2 a. Compute the sample mean. b. Compute the sample standard deviation. c. Compute the median. d. Compute the first and third quartiles. e. Which
measure, the mean or median, is a better measure of a typical value? Justify. f. Which measure, the standard deviation or the interquartile range, is a better measure of dispersion? Justify. 5. In a study of a new anti-hypertensive medication, systolic blood pressures are measured at baseline. The data are as follows: 120 112 138 145 135
150 145 163 148 128 143 156 160 142 150 a. Compute the sample mean. b. Compute the sample median. c. Compute the sample standard deviation. FIGURE 4–25 Box-Whisker Plots of Cholesterol Level in Boys and Girls Boys Girls 120 130 140 150 63 160 170 180 190 200 210 220 230 240 250 260 270 64 CHAPTER 4
Summarizing Data Collected in the Sample d. Compute the sample range. e. Are there any outliers? Justify. 6. Organize the systolic blood pressures in Problem 5 into categories and develop a frequency distribution table. 7. Generate a relative frequency histogram using the data in Problem 6. 8. Figure 4–26 shows birth weight (in grams) of infants
who were delivered full-term, classified by their mother's weight gain during pregnancy: a. What is the median birth weight for mothers who gained 50 to 74 pounds? b. What is the interquartile range for mothers who gained 75 pounds or more? c. Are there any outliers in birth weight among mothers who gained less than 50 pounds? d. What are
the best measures of a typical birth weight and the variability in birth weight among mothers who gained less than 50 pounds? 9. The following are baby height measurements (in centimeters) for a sample of infants participating in a study of infant health: 28 30 41 48 29 48 62 49 51 39 a. Compute the sample mean. b. Compute
the sample standard deviation. c. Compute the median. d. Compute the first and third quartiles. e. Which measure, the mean or median, is a better measure of a typical value? Justify. f. Which measure, the standard deviation or the interquartile range, is a better measure of dispersion? Justify. 10. Construct a frequency distribution table using the
data in Problem 9 and the following categories: less than 35, 35–44, 45–54, 55 or more. 11. Generate a relative frequency histogram using the data in Problem 10. FIGURE 4–26 Birth Weight by Maternal Weight Gain 75+pounds 50–74 pounds Less than 50 pounds 2500 2750 3000 3250 3500 Birth Weight (grams) 3750 4000 4250 4500 Practice
Problems 12. The following data were collected from 10 randomly selected patients undergoing physical therapy following knee surgery. The data represent the percent gain in range of motion after 3 weeks of physical therapy. 24% 32% 50% 62% 21% 45% 80% 24% 30% 10% a. Compute the sample mean percent gain in range of motion.
b. Compute the sample standard deviation. c. Compute the median. d. Compute the first and third quartiles. 13. The following are HDL levels measured in healthy females. 65 60 58 52 70 54 72 80 38 a. C ompute the sample mean HDL in healthy females. b. Compute the sample standard deviation of HDL in healthy females. c. Compute
the median HDL in healthy females. 14. In a clinical trial for a new cancer treatment, participants are asked to rate their quality of life at baseline. Data for all participants are summarized (in Table 4–22). Generate a graphical display of the quality of life at baseline. 15. The following are summary statistics on triglyceride levels measured in a sample
of n = 200 participants. Are there outliers? Justify your answer. n = 200 X̄= 203 s = 64 Min = 38 Q1 = 59 Median = 180 Q3 = 243 Max = 394 16. F igure 4–27 is a graphical display of the ages of participants in a study of risk factors for d ementia. TABLE 4–22 Quality of Life Quality of Life Number of Participants Excellent 25 Very
good 29 Good 36 Fair 53 Poor 57 65 Identify the median age, and the first and third quartiles of age. 17. A clinical trial is conducted to evaluate the efficacy of a new drug to increase HDL cholesterol ("good" cholesterol) in men at high risk for cardiovascular disease. Participants are randomized to receive the new drug or placebo and followed for 12
weeks, at which time HDL cholesterol is measured. The data are summarized in Figure 4–28. a. What is the 25th percentile of HDL in patients receiving the placebo? b. What is median HDL in patients receiving the new drug? c. Are there any outliers in HDL in patients receiving the new drug? Justify. 18. A pilot study is run to investigate the
feasibility of recruiting pregnant women into a study of risk factors for preterm delivery. Women are invited to participate at their first clinical visit for prenatal care. The following represent the gestational ages in weeks of women who consent to participate in the study. 11 14 21 22 9 10 13 18 a. Compute the sample mean gestational age.
b. Compute the sample standard deviation of gestational age. c. Compute the median gestational age. d. What proportion of the sample enroll in the first trimester of pregnancy (i.e., between 1 week and 13 weeks, inclusive, of pregnancy)? 19. If there are outliers in a sample, which of the following is always true? a. Mean > Median b. S tandard
deviation is smaller than expected (smaller than if there were no outliers) c. Mean < Median d. Standard deviation is larger than expected (larger than if there were no outliers) 20. The following data are a sample of white blood counts in thousands of cells per cubic millimeter for nine patients entering a hospital in Boston, MA, on a given day. 7
35 5 9 8 3 10 12 8 a. Are there any outliers in this data? Justify. b. Appropriately summarize the data. 66 CHAPTER 4 Summarizing Data Collected in the Sample FIGURE 4–27 Age of Participants in a Study of Risk Factors for Dementia 65 70 75 80 85 90 95 FIGURE 4–28 HDL Cholesterol New Drug Placebo 30 35 40 REFERENCES
1. National Heart, Lung, and Blood Institute. Available at http:// www.nhlbi.nih.gov. 2. Sullivan, L.M. "Repeated measures." Circulation 2008; 117(9): 1238–1243. 3. Little, R.J. and Rubin, D.B. Statistical Analysis with Missing Data. New York: John Wiley & Sons, 1987. 4. American Heart Association. Available at . 45 50 55 5. The Expert Panel.
"Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults: summary of the second report of the NCEP expert panel (Adult Treatment Panel II)." Journal of the American Medical Association 1993; 269: 3015–3023. 6. Hoaglin, D.C. "John W. Tukey and data analysis." Statistical Science 2003; 18(3): 311–318. 7. SAS
version 9.1. © 2002–2003 by SAS Institute, Cary, NC. 8. S-PLUS version 7.0. © 1999–2006 by Insightful Corp., Seattle, WA. © Sergey Nivens/Shutterstock © holbox/Shutterstock 5 chapter The Role of Probability W hen and W hy Key Questions •• How much variability in a measure is reasonable? •• What does it mean when you have a medical test
and your result is "within normal limits"? •• Should you be concerned if you take a medical test and your result is not "normal"? In the News Hemoglobin is a protein found in red blood cells that carries oxygen. People with low levels might have anemia, whereas people with high levels might have heart or lung problems. In men, the "normal" range is
13.8–17.2 g/dL; in women, the "normal" range is 12.1–15.1 g/dL.1 The American Heart Association categorizes systolic (when the heart muscle contracts) and diastolic (when the heart is at rest and refilling) blood pressure levels as normal, pre-hypertensive, stage 1 hypertension, stage 2 hypertension, or hypertensive crisis.2 Normal systolic blood
pressure is less than 120 mm Hg and normal diastolic blood pressure is less than 80 mmHg. Cholesterol levels—for example, total serum cholesterol, high-density lipoprotein (HDL) cholesterol, and low-density lipoprotein (LDL) cholesterol—are not categorized as normal or not normal, but rather as desirable, borderline high, and high. For some of the
cholesterol subfractions, different values define desirable versus borderline high categories in men versus women. Desirable total serum cholesterol for both men and women is less than 200 mg/dL.3 Normal ranges of characteristics are usually determined by measurements in a reasonably large sample of healthy volunteers and reflect the range of
values that extend two standard deviations Available at /003645.htm. 2 Available at Pressure/AboutHighBloodPressure/Understanding-Blood-Pressure- Readings_UCM_301764_Article.jsp#.V4pYkrgrIps. 3 Available at articles/summer12pg6-7.html. 1 from the mean in either direction. Approximately 5% of values outside of that range are also observed
in healthy volunteers. Dig In •• Are there any psychological risks in defining "normal ranges" for tests that, by design, include healthy values that are outside of those ranges? •• Can you think of a currently available test that is particularly problematic in terms of psychological risk? How might we minimize the risks for patients interested in taking
that test? •• Can you think of a better approach to defining "normal ranges"? L earning O bjectives By the end of this chapter, the reader will be able to •• Define the terms "equally likely" and "at random" •• Compute and interpret unconditional and conditional probabilities •• Evaluate and interpret independence of events •• Explain the key
features of the binomial distribution model •• Calculate probabilities using the binomial formula •• Explain the key features of the normal distribution model •• Calculate probabilities using the standard normal distribution table •• Compute and interpret percentiles of the normal distribution •• Define and interpret the standard error •• Explain
sampling variability •• Apply and interpret the results of the Central Limit Theorem Probabilities are numbers that reflect the likelihood that ap articular event occurs. We hear about probabilities in many everyday situations, ranging from weather forecasts 68 CHAPTER 5 The Role of Probability (probability of rain or snow) to the lottery
(probability of hitting the big jackpot). In biostatistical applications, it is probability theory that underlies statistical inference. Statistical inference involves making generalizations or inferences about unknown population parameters based on sample statistics. A population parameter is any summary measure computed on a population (e.g., the
population mean, which is denoted m; the p opulation variance, which is denoted s 2; or the population standard deviation, which is denoted s). In Chapter 4, we presented a number of different statistics to summarize s ample data. In Chapters 6, 7, 9, 10, and 11, we discuss statistical inference in detail. Specifically, we present formulas and
procedures to make inferences about a population based on a single sample. There are many details to consider but, in general, we select a sample from the population of interest, measure the characteristic under study, summarize this characteristic in our sample, and then make inferences about the population based on what we observe in the
sample. This process can be very powerful. In many statistical applications, we make inferences about a large number of individuals in a population based on a study of only a small fraction of the population (i.e., the sample). This is the power of biostatistics. If a study is replicated or repeated on another sample from the population, it is possible that
we might observe slightly different results. This can be due to the fact that the second study involves a slightly different sample, the sample again being only a subset of the population. A third replication might produce yet different results. From any given population, there are many different samples that can be selected. The results based on each
sample can vary, and this variability is called sampling variability. In practice, we select only one sample. Understanding how that study sample is produced is important for statistical inference. In our discussion of probability, we focus on probability as it applies to selecting individuals from a population into a sample. We also focus on quantifying -
sampling variability. Whenever we make statistical inferences, we must recognize that we are making these inferences based on incomplete information—specifically, we are making inferences about a population based only on a single sample. 5.1 SAMPLING Probability is important in selecting individuals from a population into a sample, and again
in statistical inference when we make generalizations about the population based on that sample. When we select a sample from a population, we want that sample to be representative of the population. Specifically, we want the sample to be similar to the population in terms of key characteristics. For example, suppose we are interested in studying
obesity in residents of the state of California. The Centers for Disease Control and Prevention (CDC) reports data on demographic information on the residents of each state, and some of these demographic characteristics are summarized for the state of California in Table 5–1.1 In California, the residents are almost equally split by sex, their median
age is 33.3 years, and 72.7% are 18 years of age and older. The state has 63.4% whites, 7.4% African Americans, and the remainder are other races. Almost 77% of the residents have a high school diploma, and over 26% have a bachelors degree or higher. If we select a sample of the residents of California for a study, we would not want the sample to
be 75% men with a median age of 50. That sample would not be representative of the p opulation. Table 5–2 contains the same demographic i nformation for California, Massachusetts, Florida, and Alabama.1 Notice the differences in demographic characteristics among these states. Previous studies have shown that obesity is related to sex, race, and
educational level; therefore, when we select a sample to study obesity, we must be careful to accurately reflect these characteristics.2–4 Studies have shown that prevalent obesity is inversely related to educational level (i.e., persons with higher levels of education are less likely to be obese). If we select a sample to assess obesity in a state, we would
want to take care not to over-represent (or underrepresent) persons with lower educational levels as this could inflate (or diminish) our estimate of obesity in that state. Sampling individuals from a population into a sample is a critically important step in any biostatistical analysis. There are two popular types of sampling, probability sampling and non-
probability sampling. In probability sampling, each member of the population has a known probability of being TABLE 5–1 Demographic Characteristics for California Men (%) 49.8 Median age (years) 33.3 18 years of age and older (%) 72.7 White (%) 63.4 African-American (%) 7.4 High school graduate or higher (%) 76.8 Bachelors degree or higher
(%) 26.6 What Is PublicSampling Health? 69 TABLE 5–2 Demographic Characteristics for California, Massachusetts, Florida, and Alabama California Massachusetts Florida Alabama Men (%) 49.8 48.2 48.8 48.3 Median age (years) 33.3 36.5 38.7 35.8 18 years of age and older (%) 72.7 76.4 77.2 74.7 White (%) 63.4 86.2 79.7 72.0 7.4 6.3 15.5 26.3
High school graduate or higher (%) 76.8 84.8 79.9 75.3 Bachelors degree or higher (%) 26.6 33.2 22.3 19.0 African-American (%) selected. In non-probability sampling, each member of the population is selected without the use of probability. Popular methods of each type of sampling are described here. More details can be found in Cochran and
Kish.5,6 5.1.1 Probability Sampling: Simple Random Sampling In simple random sampling, we start with what is called the sampling frame, a complete list or enumeration of the entire population. Each member of the population is assigned a unique identification number, and then a set of numbers is selected at random to determine the individuals
to be included in the sample. Many introductory statistical textbooks contain tables of random numbers that can be used, and there are also statistical computing packages that can be used to generate random numbers.7,8 Simple random sampling is a technique against which many other sampling techniques are compared. It is most useful when the
population is relatively small because it requires a complete enumeration of the population as a starting point. In this sampling scheme, each individual in the population has the same chance of being selected. We use N to represent the number of individuals in the population, or the population size. Using simple random sampling, the probability that
any individual is selected into the sample is 1 / N. 5.1.2 Probability Sampling: Systematic Sampling In systematic sampling, we again start with the complete sampling frame and members of the population are assigned unique identification numbers. However, in systematic s ampling every third or every fifth person is selected. The s pacing or
interval between selections is determined by the ratio of the population size to the sample size (N / n). For e xample, if the population size is N 5 1000 and a sample size of n 5 100 is d esired, then the sampling interval is 1000 / 100 5 10; so every tenth person is selected into the sample. The first p erson is selected at random from among the first ten
in the list, and the first selection is made at random using a random numbers table or a computer-generated random number. If the desired sample size is n 5 175, then the sampling fraction is 1000 / 175 5 5.7. Clearly, we cannot take every 5.7th person, so we round this down to 5 and take every fifth person. Once the first person is selected at
random from the first five in the list, every fifth person is s elected from that point on through the end of the list. There is a possibility that a systematic sample might not be representative of a population. This can occur if there is a systematic arrangement of individuals in the population. Suppose that the population consists of married couples and
that the sampling frame is set up to list each husband and then his wife. Selecting every tenth person (or any even-numbered multiple) would result in selecting all men or women depending on the starting point. This is a very extreme example, but as a general principle, all potential sources of systematic bias should be considered in the sampling
process. 5.1.3 Probability Sampling: Stratified Sampling In stratified sampling, we split the population into non- overlapping groups or strata (e.g., men and women; people under 30 years of age and people 30 years of age and older) and then sample within each strata. Sampling within each strata can be by simple random sampling or systematic
sampling. The idea behind stratified sampling is to ensure adequate r epresentation of individuals within each strata. For example, if a population contains 70% men and 30% women and we want to ensure the same representation in the sample, we can stratify and sample the requisite numbers of men and women to ensure the same representation.
For example, if the desired sample size is n 5 200, then n 5 140 men and n 5 60 70 CHAPTER 5 The Role of Probability women could be sampled either by simple random sampling or by s ystematic sampling. 5.1.4 Non-Probability Sampling: Convenience Sampling Non-probability samples are often used in practice because in many applications, it is
not possible to generate a sampling frame. In non-probability samples, the probability that any individual is selected into the sample is unknown. Whereas it is informative to know the likelihood that any individual is selected from the population into the sample (which is only possible with a complete sampling frame), it is not possible here. However,
what is most important is selecting a sample that is representative of the population. In convenience sampling, we select individuals into our sample by any convenient contact. For example, we might approach patients seeking medical care at a particular hospital in a waiting or reception area. Convenience samples are useful for collecting
preliminary data. They should not be used for statistical inference as they are generally not constructed to be representative of any specific population. 5.1.5 Non-Probability Sampling: Quota Sampling In quota sampling, we determine a specific number of individuals to select into our sample in each of several non-overlapping groups. The idea is
similar to stratified sampling in that we develop non-overlapping groups and sample a predetermined number of individuals within each group. For example, suppose we wish to ensure that the distribution of participants' ages in the sample is similar to that in the population. Suppose our desired sample size is n 5 300 and we know from census data
that in the population, approximately 30% are under age 20, 40% are between 20 and 49, and 30% are 50 years of age and older. We then sample n 5 90 persons under age 20, n 5 120 between the ages of 20 and 49, and n 5 90 who are 50 years of age and older. Sampling proceeds until these totals, or quotas, are reached in each group. Quota
sampling is different from stratified sampling because in a stratified sample, individuals within each stratum are selected at random. Here we enroll participants until the quota is reached. 5.2 BASIC CONCEPTS We now discuss how probabilities are determined and we consider sampling by simple random sampling. A probability is a number that
reflects the likelihood that a particular event, such as sampling a particular individual from a population into a sample, will occur. Probabilities range from 0 to 1. Sometimes probabilities are converted to percentages, in which case the range is 0% to 100%. A probability of 0 indi- cates that there is no chance that a particular event will occur, whereas
a probability of 1 indicates that an event is certain to occur. In most a pplications, we are dealing with probabilities between 0 and 1, not inclusive. Example 5.1. Suppose we wish to conduct a study of obesity in children 5 to 10 years of age who are seeking medical care at a particular pediatric practice. The population includes all children who were
seen in the practice in the past 12 months and is summarized in Table 5–3. If we select a child at random (by simple random sampling), then each child has the same probability of being selected. In statistics, the phrase "at random" is synonomous with "equally likely." In this application, each child has the same probability of being selected and that
probability is determined by 1 / N, where N is the population size. Thus, the probability that any child is selected is 1 / 5290 5 0.0002. In most sampling situations, we are generally not concerned with sampling a specific individual but instead concern ourselves with the probability of sampling certain "types" of individuals. For example, what is the
probability of selecting a boy or a child 7 years of age? The following formula can be used to compute probabilities of selecting individuals with specific attributes or characteristics. P(characteristic) 5 Number of persons with characteristic }}}}} Total number of persons in the population (N) If we select a child at random, the probability that we s-
elect a boy is computed as follows: P(boy) 5 2560 / 5290 5 0.484. This probability can be interpreted in two ways. The probability that we select a boy at random from the population is 0.484, or 48.4%. A second interpretation is that the proportion of boys in the population is 48.4%. The probability that we select a child who is 7 years of age is P(7
years of age) 5 913 / 5290 5 0.173. This formula can also be used to compute probabilities of characteristics defined more narrowly or more broadly. For example, what is the probability of selecting a boy who is TABLE 5–3 Children Seen over the Past 12 Months Boys Girls Total 5 432 408 840 6 379 513 892 Age (years) 7 8 501 410 412 436 913 846
Total 9 420 461 881 10 418 500 918 2560 2730 5290 Conditional What Is Public Probability Health? 71 5.3.1 Evaluating Screening Tests 10 years of age? P(boy who is 10 years of age) 5 418 / 5290 5 0.079. What is the probability of selecting a child (boy or girl) who is at least 8 years of age? P(at least 8 years of age) 5 (846 1 881 1 918) / 5290 5
2645 / 5290 5 0.500. To use this formula to compute probabilities, we must enumerate the population as in Table 5–3. Once the population is enumerated, we can compute probabilities relatively easily by counting the numbers of individuals who possess the characteristic of interest and dividing by the population size. Screening tests are often used in
clinical practice to assess the likelihood that a person has a particular medical condition. Many screening tests are based on laboratory tests that detect particular markers that are related to a specific disease. For example, a popular screening test for prostate cancer is called the prostate-specific antigen (PSA) test and is recommended for men over
50 years of age. The test measures PSA, which is a protein produced by the prostate gland. The PSA screening test measures the amount of PSA in the blood. Low levels of PSA are normal but elevated levels have been shown to be a ssociated with prostate cancer. The PSA test alone does not provide a diagnosis of prostate cancer. A more invasive
test, such as a biopsy—in which tissue is extracted and examined under a microscope—is needed to make a diagnosis. Data such as that shown in Example 5.2 are often compiled to evaluate the usefulness of a screening test. Example 5.2. Suppose that we wish to investigate whether the PSA test is a useful screening test for prostate cancer. To be
useful clinically, we want the test to distinguish b etween men with and without prostate cancer. Suppose that a population of N 5 120 men over 50 years of age who are considered high-risk for prostate cancer have both the PSA screening test and a biopsy. The PSA results are reported as low, slightly to moderately elevated, or highly elevated. The
classifications are based on the following levels of measured protein, respectively: 022.5, 2.6219.9, and 20 or more nanograms per milliliter (ng/ml).9 The biopsy provides the diagnosis of prostate cancer. The results of the study are shown in Table 5–4. The probability that a man has prostate cancer given that he has a low level of PSA is P(prostate
cancer | low PSA) 5 3 / 64 5 0.047. The probability that a man has prostate cancer given that he has a slightly to moderately elevated level of PSA is P(prostate cancer | slightly to moderately elevated PSA) 5 13 / 41 5 0.317. The probability that a man has prostate cancer given that he has a highly elevated level of PSA is 5.3 CONDITIONAL
PROBABILITY All of the probabilities computed in the previous section are unconditional probabilities. In each case, the denominator is the total population size, N 5 5290, reflecting the fact that everyone in the entire population is eligible to be selected. Sometimes it is of interest to focus on a particular subset of the population (e.g., a
subpopulation). For example, suppose we are interested just in the girls and ask the question, what is the probability of selecting a 9-year-old from the subpopulation of girls? There are a total of NG 5 2730 girls (here NG refers to the population size of girls) and the probability of selecting a 9-year-old from the subpopulation of girls is written as -
follows: P(9-year-old | girls), where "|" refers to the fact that we are conditioning on or referring to a specific subgroup, and that subgroup is specified to the right of "|". The conditional probability is computed using the same approach we used to compute unconditional probabilities. P(9-year-old | girls) 5 461 / 2730 5 0.169. This means that 16.9% of
the girls are 9 years of age. Note that this is not the same as the probability of selecting a 9-year-old girl, which is P(girl who is 9 years of age) 5 461 / 5290 5 0.087. What is the probability of selecting a boy from among the 6-year-olds? P(boy | 6 years of age) 5 379 / 892 5 0.425. Thus, 42.5% of the 6-year-olds are boys (and 57.5% of the 6-year-olds are
girls). Some popular applications of conditional probability in public health and medicine are described here. TABLE 5–4 Evaluation of PSA Test PSA Level Biopsy Results Prostate Cancer Low 3 Slightly to moderately elevated Highly elevated Total Total No Prostate Cancer 61 64 13 28 41 12 3 15 28 92 120 72 CHAPTER 5 The Role of
Probability P(prostate cancer | highly elevated PSA) 5 12 / 15 5 0.80. The probability or likelihood that a man has prostate cancer is highly dependent on his PSA level. Based on these data, is the PSA test a clinically useful screening test? Some screening tests are based on laboratory tests, whereas others are based on a set of self-administered
questions that address a specific state, such as depression, anxiety, or stress. Patients who screen positive (e.g., score high or low on a set of questions) might be referred to u ndergo subsequent screening or more involved testing to make a diagnosis. Example 5.3. In pregnancy, women often undergo screening to assess whether their fetus is likely
to have Down Syndrome. The screening test evaluates levels of specific hormones in the blood. The screening tests are reported as positive or negative, indicating that a woman is more or less likely to be carrying an affected fetus. Suppose that a population of N 5 4810 pregnant women undergoes the screening test and the women scored as either
positive or negative depending on the levels of hormones in the blood. In addition, suppose that each woman in the study has an amniocentesis. Amniocentesis is an invasive procedure that provides a more definitive assessment as to whether a fetus is affected with Down Syndrome. Amniocentesis is called a diagnostic test, or the gold standard (i.e.,
the true state). The results of the screening tests and the amniocentesis (diagnostic test) are summarized in Table 5–5. Using the data from Table 5–5, the probability that a woman with a positive test has an affected fetus is P(affected fetus | screen positive) 5 9 / 360 5 0.025, and the probability that a woman with a negative test has an affected fetus
is P(affected fetus | negative screen positive) 5 1 / 4450 5 0.0002. Is the screening test a clinically useful test? 5.3.2 Sensitivity and Specificity are a number of widely used screening tests available today, including the PSA test for prostate cancer, mammograms for breast cancer, and serum and ultrasound assessments for prenatal diagnosis. When
a screening test is proposed, there are two m easures that are often used to evaluate its performance, the sensitivity and specificity of the test. Suppose that the results of the screening test are dichotomous—specifically, each person is classified as either positive or negative for having the condition of interest. Sensitivity is also called the true
positive fraction and is defined as the probability that a diseased person screens positive. Specificity is also called the true negative fraction and is defined as the probability that a disease-free person screens negative. To evaluate the screening test, each participant u ndergoes the test and is classified as positive or negative based on criteria that are
specific to the test (e.g., high levels of a marker in a serum test, or the presence of a mass on a mammogram). Each participant also undergoes a diagnostic test, which provides a definitive diagnosis (e.g., an amniocentesis, or a biopsy). A total of N patients complete both the screening test and the diagnostic test. The data are often organized as
shown in Table 5–6. The results of the screening test are shown in the rows of the table, and the results of the diagnostic test are shown in the columns. The definitions of sensitivity and specificity are given as: Sensitivity 5 True Positive Fraction 5 P(screen positive | disease) 5 a / (a 1 c). Specificity 5 True Negative Fraction 5 P(screen negative |
disease free) 5 d / (b 1 d). False Positive Fraction 5 P(screen positive | disease free) 5 b / (b 1 d). False Negative Fraction 5 P(screen negative | disease) 5 c / (a 1 c). Screening tests are not used to make medical diagnoses but instead to identify individuals most likely to have a certain condition. Ideally, screening tests are not excessively costly and
pose little risk to the patients undergoing the test. There TABLE 5–6 Data Layout for Evaluating Screening Tests Diseased TABLE 5–5 Evaluation of Prenatal Screening Test Affected Fetus Unaffected Fetus Total Positive 9 351 360 Negative 1 4449 Total 10 4800 Disease Free Total Screen positive a b a1b 4450 Screen negative c d c1d 4810
Total a1c b1d N Conditional What Is Public Probability Health? The false positive fraction is 1–specificity and the false negative fraction is 1–sensitivity. Therefore, knowing sensitivity and specificity captures the information in the false positive and false negative fractions. These are simply alternate ways of expressing the same information. Often,
sensitivity and the false positive fraction are reported for a test. Example 5.4. Consider again the study presented in Example 5.3 in which a population of N 5 4810 pregnant women has a screening test and an amniocentesis to assess the likelihood of carrying a fetus with Down Syndrome. The results are summarized in Table 5–5. The performance
characteristics of the test are: Sensitivity 5 P(screen positive | affected fetus) 5 9 / 10 5 0.900. Specificity 5 P(screen negative | unaffected fetus) 5 4449 / 4800 5 0.927. False Positive Fraction 5 P(screen positive | unaffected fetus) 5 351 / 4800 5 0.073. False Negative Fraction 5 P(screen negative | affected fetus) 5 1 / 10 5 0.100. These results are
interpreted as follows. If a woman is c arrying an affected fetus, there is a 90% chance that the screening test will be positive. If the woman is carrying an unaffected fetus (defined here as a fetus free of Down Syndrome, because this test does not assess other abnormalities), there is a 92.7% chance that the screening test will be negative. If a woman
is carrying an unaffected fetus, there is a 7.3% chance that the screening test will be positive, and if the woman is carrying an affected fetus, there is a 10% chance that the test will be negative. The false positive and false negative fractions quantify e rrors in the test. The errors are often of greatest concern. For example, if a woman is carrying an
unaffected fetus, there is a 7.3% chance that the test will incorrectly come back positive. This is potentially a s erious problem, as a positive test result would likely produce great anxiety for the woman and her family. A false negative result is also problematic. If a woman is carrying an affected fetus, there is a 10% chance that the test will come back
negative. The woman and her family might feel a false sense of assurance that the fetus is not affected when, in fact, the screening test missed the abnormality. The sensitivity and false positive fractions are often reported for screening tests. However, for some tests the specificity and false negative fractions might be the most important. The most
important characteristics of any screening test depend on the implications of an error. In all cases, it is important to under- 73 stand the performance characteristics of any screening test to appropriately interpret results and their implications. 5.3.3 Positive and Negative Predictive Value Patients undergoing medical testing often ask the following
questions: What is the probability that I have the d isease if my screening test comes back positive? What is the probability that I do not have the disease if my test comes back negative? These questions again can be answered with conditional probabilities. These quantities are called the positive predictive value and negative predictive value of a test
and are defined in this section. Consider again the data layout of Table 5–6 summarizing the results of screening and diagnostic tests applied to a population of N individuals. The definitions of positive and negative predictive values are given as follows: Positive predictive value 5 P(disease | screen positive) 5 a / (a 1 b). Negative predictive value 5
P(disease free | screen negative) 5 d / (c 1 d). Notice the difference in the definitions of the positive and negative predictive values as compared to the sensitivity and specificity provided previously. Example 5.5. Consider again the study presented in Example 5.3 and Example 5.4, in which a population of N 5 4810 pregnant women has a screening
test and an amniocentesis to assess the likelihood of carrying a fetus with Down Syndrome. The results are summarized in Table 5–5. The positive predictive value P(affected fetus | screen positive) 5 9/360 5 0.025 and negative predictive value P(unaffected | screen negative) 5 4449/4450 5 0.999. These results are interpreted as follows. If a woman
screens positive, there is a 2.5% chance that she is carrying an affected fetus. If a woman screens n egative, there is a 99.9% chance that she is carrying an unaffected fetus. Patients often want to know the probability of having the disease if they screen positive (positive predictive value). As with this example, the positive predictive value can be very
low (2.5%) because it depends on the prevalence of the disease in the population. Thankfully, the prevalence of Down Syndrome in the population is low. Using the data in Example 5.5, the prevalence of Down Syndrome in the population of N 5 4810 women is 10 / 4810 5 0.002 (i.e., Down Syndrome affects 2 fetuses per 1000). Whereas this screening
test has good performance characteristics (sensitivity of 90% and specificity of 92.7%) because the prevalence of the condition is so low, even a test with a high probability of detecting an affected fetus does not translate into a high positive predictive value. Because positive and negative predictive values depend on the prevalence of the disease, they
cannot be estimated in case control designs (see Section 2.2.4 for more details). 74 CHAPTER 5 The Role of Probability 5.4 INDEPENDENCE In probability, two events are said to be independent if the probability of one is not affected by the occurrence or non-occurrence of the other. The definition alone is difficult to digest. Example 5.6. In
Example 5.2, we analyzed data from a population of N 5 120 men who had both a PSA test and a biopsy for prostate cancer. Suppose we have a different test for prostate cancer. This prostate test produces a numerical risk that classifies a man as at low, moderate, or high risk for prostate cancer. A sample of 120 men undergo the new test and also
have a biopsy. The data are summarized in Table 5–7. The probability that a man has prostate cancer given he has a low risk is P(prostate cancer | low risk) 5 10 / 60 5 0.167. The probability that a man has prostate cancer given he has a moderate risk is P(prostate cancer | moderate risk) 5 6 / 36 5 0.167. The probability that a man has prostate cancer
given he has a high risk is P(prostate cancer | high risk) 5 4 / 24 5 0.167. The probability or likelihood that a man has prostate cancer here is unrelated to or independent of his risk based on the prostate test. Knowing a man's prostate test result does not affect the likelihood that he has prostate cancer in this e xample. Thus, the likelihood that a man
has prostate cancer is independent of his prostate test result. Independence can be demonstrated in several ways. Consider two events—call them A and B (A could be a low risk based on the prostate test and B a diagnosis of prostate cancer). Two events are independent if P(A | B) 5 P(A) or if P(B | A) 5 P(B). For example, define A 5 low risk and B 5
prostate cancer. To check independence, we must compare a conditional and an unconditional probability: P(A | B) 5 P(low risk | prostate cancer) 5 10 / 20 5 0.50, and P(A) 5 P(low risk) 5 60 / 120 5 0.50. The equality of the conditional and unconditional probabilities here indicates independence. Independence can also be checked using P(B | A) 5
P(prostate cancer | low risk) 5 10 / 60 5 0.167, and P(B) 5 P(prostate cancer) 5 20 / 120 5 0.167. Both versions of the definition of independence will give the same result. Example 5.7. Table 5–8 contains information on a population of N 5 6732 individuals who are classified as having or not having prevalent cardiovascular disease (CVD). Each
individual is also classified in terms of having a family history of cardiovascular disease or not. In this analysis, family history is defined as a first-degree relative (parent or sibling) with diagnosed CVD before age 60. Are family history and prevalent CVD independent? Another way to ask the question is as follows: Is there a relationship between family
history and prevalent CVD? This is a question of independence of events. Let A 5 prevalent CVD and B 5 family history of CVD. (Note that it does not matter how we define A and B; the result will be identical.) We now must check whether P(A | B) 5 P(A) or if P(B | A) 5 P(B). Again, it makes no difference which definition is used, the results will be
identical. We compute P(A | B) 5 P(prevalent CVD | family history of CVD) 5 491 / 859 5 0.572 and compare it to P(A) 5 P(prevalent CVD) 5 643 / 6732 5 0.096. These probabilities are not equal; therefore family history and prevalent CVD are not independent. In the population, the probability of prevalent CVD is 0.096%, or 9.6% of the population has
prevalent CVD. Individuals with a family history of CVD are much more likely to have prevalent CVD. The chance of prevalent CVD given a family history is 57.2%, as compared to 2.6% (152/5873 = 0.026) among patients with no family history. 5.5 BAYES' THEOREM Bayes' Theorem is a probability rule that can be used to compute a conditional
probability based on specific available information. TABLE 5–7 Prostate Test and Biopsy Results Prostate Test Risk Biopsy Results Prostate Cancer Total No Prostate Cancer Low 10 50 60 Moderate 6 30 36 High 4 20 24 Total 20 100 120 WhatProbability Is Public Health? Models TABLE 5–8 Family History and Prevalent CVD Prevalent CVD
Family history of CVD 491 Free of CVD Total 368 859 No family history of CVD 152 5721 5873 Total 643 6089 6732 There are several versions of the theorem, ranging from simple to more involved. The following is the simple statement of the rule: P(A|B) = P(B | A) P(A) P(B) Example 5.8. Suppose a patient exhibits symptoms raising concern with
his physician that he may have a particular disease. Suppose the disease is relatively rare with a prevalence of 0.2% (meaning it affects 2 out of every 1000 persons). The physician recommends testing, starting with a screening test. The screening test is noninvasive, based on a blood sample and costs $250. Before agreeing to the screening test, the
patient wants to know what will be learned from the test—specifically, he wants to know his chances of having the disease if the test comes back positive. The physician reports that the screening test is widely used and has a reported sensitivity of 85%. In addition, the test comes back positive 8% of the time and negative 92% of the time. The patient
wants to know the positive predictive value or P(disease | screen positive). Using Bayes' Theorem, we can compute this as follows: P(disease | screen positive) 5 P(screen positive | disease) P(disease)/P(screen positive). We know that P(disease) 5 0.002, P(screen positive | disease) 5 0.85 and P(screen positive) 5 0.08. We can now substitute the values
into the above equation to compute the desired probability: 75 Another important question that the patient might ask is, What is the chance of a false positive result? Specifically, what is P(screen positive | no disease)? We can compute this conditional probability with the available information using Bayes' Theorem: P(screen positive | no disease) 5
P(no disease | screen positive)P(screen positive) }}}}} P(no disease) Note that if P(disease) 5 0.002, then P(no disease) 51 – 0.002. The events disease and no disease are called complementary events. The no disease group includes all members of the population not in the disease group. The probabilities of complementary events must sum to 1, i.e.,
P(disease) 1 P(no disease) 5 1. Similarly, P(no disease | screen positive) 1 P(disease | screen positive) 5 1. Hence P(screen positive | no disease) 5 (1 – 0.021)(0.08) / (1 – 0.002) 5 0.078. Using Bayes' Theorem, there is a 7.8% chance that the screening test will be positive in patients free of disease, which is the false positive fraction of the test. 5.6
PROBABILITY MODELS To compute the probabilities in the previous sections, we counted the number of participants that had the characteristic of interest and divided by the population size. For conditional probabilities, the population size (denominator) was modified to reflect the subpopulation of interest. In each of the examples in the previous
sections, we had a tabulation of the population (the sampling frame) that allowed us to compute the desired probabilities. There are instances where a complete enumeration or tabulation is not available. In some of these instances, probability models or mathematical equations can be used to generate probabilities. There are many probability models,
and the model appropriate for a specific application depends on the specific attributes of the application. If an application satisfies specific attributes of a probability model, the model can be used to generate probabilities. In the following two sections, we describe two very popular probability models. These probability models are extremely important
in statistical inference, which is discussed in detail in Chapters 6, 7, 9, 10, and 11. P(disease | screen positive) 5 (0.85)(0.002) / (0.08) 5 0.021. 5.6.1 A Probability Model for a Discrete Outcome: The Binomial Distribution If the patient undergoes the test and it comes back positive, there is a 2.1% chance that he has the disease. Without the test, there
is a 0.2% chance that he has the disease (the prevalence in the general population). Should the patient have the screening test? The binomial distribution model is an important probability distribution model that is appropriate when a particular experiment or process has two possible outcomes. The name itself, binomial, reflects the dichotomy of
responses. 76 CHAPTER 5 The Role of Probability There are many different probability models and if a particular process results in more than two distinct outcomes, a multinomial probability model might be appropriate. Here we focus on the situation in which the outcome is dichotomous. For example, adults with allergies might report relief with
medication or not, children with a bacterial infection might respond to antibiotic therapy or not, adults who suffer a myocardial infarction might survive or not, or a medical device such as a coronary stent might be successfully implanted or not. These are just a few examples of applications or processes where the outcome of interest has two possible
values. The two outcomes are often labeled "success" and "failure," with "success" denoting the outcome of interest. In some medical and public health applications, we are interested in quantifying the extent of disease (the unhealthy response). Clearly, the unhealthy response is not a "success"; this nomenclature is simply used for the application of
the binomial distribution model. In any application of the binomial distribution, we must clearly specify which outcome is the success and which is the failure. The binomial distribution model allows us to compute the probability of observing a specified number of successes when the process is repeated a specific number of times (e.g., in a set of
patients) and the outcome for a given patient is either a success or a failure. We must first introduce some notation that is necessary for the binomial distribution model. First, we let n denote the number of times the application or process is repeated and x denote the number of successes, out of n, of interest. We let p denote the probability of success
for any individual. The binomial distribution model is defined as P(x successes) = n! p x (1 − p)n− x , x !(n − x)! where ! denotes factorial, defined as k! 5 k(k – 1)(k – 2)...1. For example, 4! 5 4(3)(2)1 5 24, 2! 5 2(1) 5 2 and 1! 5 1. There is one special case where 0! 5 1. Appropriate use of the binomial distribution model depends on satisfying the
following three assumptions about the application or process under investigation: each replication of the process results in one of two possible outcomes (success or failure); the probability of success is the same for each replication; and the replications are independent, meaning here that a success in one patient does not influence the probability of
success in another. Example 5.9. Consider the example where adults with allergies report relief from allergic symptoms with a specific medication. Suppose we know that the medication is effective in 80% of patients with allergies who take it as prescribed. If we provide the medication to 10 patients with allergies, what is the probability that it is
effective in exactly 7 patients? Do we satisfy the three assumptions of the binomial distribution model? Each replication of this application i nvolves providing medication to a patient and assessing whether he or she reports relief from symptoms. The outcome is relief from symptoms ("yes" or "no"), and here we will call a reported relief from
symptoms a success. The probability of success for each person is 0.8. The final assumption is that the replications are independent. In this setting, this essentially means that the probability of success for any given patient does not depend on the success or failure of any other patient (For more details, see D'Agostino, Sullivan and Beiser.10) This is a
reasonable assumption in this case. We now need to set the notation: n 5 10, x 5 7, and p 5 0.80. The desired probability is P(7 successes) = 10! 0.807 (1 − 0.80)10 − 7 . 7!(10 − 7)! The first (factorial) portion of the formula is computed as follows: 10(9)(8)(7)(6)(5)(4)(3)(2)1 10! = = 7!(10 − 7)! {7(6)(5)(4)(3)(2)(1)}{(3)(2)(1)} 10(9)(8) = 120. 3(2)
Substituting, P(7 successes) 5 (120)(0.2097)(0.008) 5 0.2013. There is a 20.13% chance that exactly 7 of 10 patients will report relief from symptoms when the probability that anyone reports relief is 0.80. What is the probability that none report relief? We can again use the binomial distribution model with n 5 10, x 5 0, and p 5 0.80: P(0 successes) =
10! 0.800 (1 − 0.80)10−0 . 0!(10 − 0)! The first (factorial) portion of the formula is computed as follows: 10! 10! = = 1. 0!(10 − 0)! {1}{10!} Substituting, P(0 successes) 5 (1)(1)(0.0000001024) 5 0.0000001024. WhatProbability Is Public Health? Models There is practically no chance that none of the 10 report relief from symptoms when the
probability of reporting relief for any individual patient is 0.80. What is the most likely number of patients who will report relief when the medication is given to 10 patients? If 80% report relief and we consider 10 patients, we e xpect 8 to report relief. What is the probability that 8 of 10 report relief? P(8 successes) = 10! 0.808 (1 − 0.80)10−8 . 8!(10
− 8)! The first (factorial) portion of the formula is computed as follows: 10! 10(9)(8)(7)(6)(5)(4)(3)(2)1 = = 8!(10 − 8)! {8(7)(6)(5)(4)(3)(2)(1)}{(2)(1)} 10(9) = 45. 2 Substituting, P(8 successes) 5 (45)(0.1678)(0.04) 5 0.3020. There is a 30.20% chance that exactly 8 of 10 patients will report relief from symptoms when the probability that anyone
reports relief is 0.80. The probability that exactly 8 report relief will be the highest probability of all possible outcomes (0 through 10). Many statistical textbooks provide tables of binomial probabilities for specific combinations of n, x, and p.10 These tables can be used to determine the probability of observing x successes out of n replications when
the probability of success is p. The tables can be of limited use as they often contain only some combinations of n, x, and p. The binomial formula can always be used as long as assumptions are appropriately satisfied. Example 5.10. The likelihood that a patient with a heart attack dies of the attack is 4% (i.e., 4 of 100 die of the attack). Suppose we
have 5 patients who suffer a heart attack. What is the probability that all will survive? Here we are provided with the probability that an attack is fatal. For this example, we call a success a fatal attack, and thus p 5 0.04. We have n 5 5 patients and want to know the probability that all survive or, in other words, that none are fatal (0 successes). We
again need to assess the assumptions. Each attack is fatal or non-fatal, the likelihood of a fatal attack is 4% for all patients, and the outcome of individual patients is independent. It should be noted that the assumption that the probability of success is constant must be evaluated carefully. The probability that a patient dies from a heart attack depends
on many factors including age, the severity of the a ttack, and other co-morbid conditions. 77 To apply the binomial formula, we must be convinced that all patients are at the same risk of a fatal attack. The assumption of the independence of events must be evaluated carefully. As long as the patients are unrelated, the assumption is usually
appropriate. Prognosis of disease could be higher in members of the same family or in individuals who are co-habitating. In this example, suppose that the 5 patients being analyzed are unrelated, of similar age, and free of comorbid conditions: 5! P(0 successes) = 0.04 0 (1 − 0.04)5−0 . 0!(5 − 0)! The first (factorial) portion of the formula is computed
as follows: 5! 5! = = 1. 0!(5 − 0)! {1}{5} Substituting, P(0 successes) 5 (1)(1)(0.8154) 5 0.8154. There is an 81.54% chance that all patients will survive the attack when the chance that any one dies is 0.04. In this example, the possible outcomes are 0, 1, 2, 3, 4, or 5 s uccesses (fatalities). Because the probability of fatality is so low, the most likely
response is 0 (all patients survive). The binomial formula generates the probability of observing exactly x successes out of n. If we want to compute the probability of a range of outcomes, we need to apply the formula more than once. Suppose in this example we wanted to compute the probability that no more than 1 person dies of the attack.
Specifically, we want P(no more than 1 success) 5 P(0 or 1 successes) 5 P(0 successes) 1 P(1 success). To compute this probability, we apply the binomial formula twice. We already computed P(0 successes), so we now compute P(1 success): 5! P(1 success) = 0.041 (1 − 0.04)5−1 . 1!(5 − 1)! The first (factorial) portion of the formula is computed as
follows: 5(4)(3)(2)1 5! = = 5. 1!(5 − 1)! {1}{4(3)(2)(1)} Substituting, P(1 success) 5 (5)(0.04)(0.8493) 5 0.1697. P(no more than 1 success) 5 P(0 or 1 successes) 5 P(0 successes) 1 P(1 success) 5 0.8154 1 0.1697 5 0.9851. The chance that no more than 1 of 5 (or equivalently, that at most 1 of 5) die from the attack is 98.51%. 78 CHAPTER 5 The Role
of Probability What is the probability that 2 or more of 5 die from the attack? Here we want to compute P(2 or more successes). The possible outcomes are 0, 1, 2, 3, 4, or 5, and the sum of the probabilities of each of these outcomes is 1 (i.e., we are certain to observe either 0, 1, 2, 3, 4, or 5 successes). We just computed P(0 or 1 successes) 5 0.9851,
so P(2, 3, 4, or 5 successes) 5 1 2 P(0 or 1 successes) 5 1 – 0.9851 5 0.0149. There is a 1.49% chance that 2 or more of 5 will die from the attack. The mean or expected number of successes of a binomial population is defined as m 5 np, and the standard deviation is σ = n(p)(1 − p). In Example 5.9, we considered a binomial distribution with n 5 10 and
p 5 0.80. The mean, or expected, number of successes is m 5 np 5 10(0.80) 5 8, and the standard deviation is σ = n(p)(1 − p) = 10(0.8)(0.2) 5 1.3. In Example 5.10, we considered a binomial distribution with n 5 5 and p 5 0.04. The mean, or expected, number of successes is m 5 np 5 5(0.04) 5 0.2 (i.e., the most likely number of successes is 0 out of
5). 5.6.2 A Probability Model for a Continuous Outcome: The Normal Distribution The normal distribution model is an important probability distribution model that is appropriate when a particular experiment or process results in a continuous outcome. There are many different probability models for continuous outcomes and the appropriate model
depends on the distribution FIGURE 5–1 Normal Distribution of the outcome of interest. The normal probability model applies when the distribution of the continuous outcome follows what is called the Gaussian distribution, or is well described by a bell-shaped curve (Figure 5–1). The horizontal or x-axis is used to display the scale of the
characteristic being analyzed (e.g., height, weight, systolic blood pressure). The vertical axis reflects the probability of observing each value. Notice that the curve is highest in the middle, suggesting that the middle values have higher probabilities or are more likely to occur. The curve tails off above and below the middle, suggesting that values at -
either extreme are much less likely to occur. Similar to theb inomial distribution model, there are some assumptions for appropriate use of the normal distribution model. The normal distribution model is appropriate for a continuous outcome if the following conditions are true. First, in a normal d istribution the mean is equal to the median and also
equal to the mode, which is defined as the most frequently observed value. As we discussed in Chapter 4, it is not always the case that the mean and median are equal. For example, if a particular characteristic is subject to outliers, then the mean will not be equal to the median and therefore the c haracteristic will not follow a normal distribution. An -
example might be the length of stay in the hospital (measured in days) following a specific procedure. Length of stay often follows a skewed distribution, as illustrated in Figure 5–2. WhatProbability Is Public Health? Models 79 FIGURE 5–2 Non-Normal Distribution Many characteristics are approximately normally distributed, such as height and
weight for specific age and sex groups, as are many laboratory and clinical measures such as cholesterol level and systolic blood pressure. The first p roperty of the normal distribution implies the following: P(x > m) 5 P(x , m) 5 0.5, where x denotes the continuous variable of interest and m is the population mean. The probability that a value e xceeds
the mean is 0.5 and equivalent to the probability that a value is below the mean, which is the definition of the median. In a normal distribution, the mean, median, and mode (the most frequent value) are equal. A continuous variable that follows a normal distribution is one for which the following three statements are also true. i. Approximately 68% of
the values fall between the mean and one standard deviation (in either direction), i.e., P(m 2 s , x , m 1 s) 5 0.68, where m is the population mean and s is the population standard deviation. ii. Approximately 95% of the values fall between the mean and two standard deviations (in either direction), i.e., P(m 2 2s , x , m 1 2s) 5 0.95. iii. Approximately
99.9% of the values fall between the mean and three standard deviations (in either direction), i.e., P(m 2 3s , x , m 1 3s) 5 0.999. Part (iii) of the preceding indicates that for a continuous v ariable with a normal distribution, almost all of the observations fall between m2 3s and m 1 3s ; thus the m inimum value is approximately m 2 3s and the
maximum is a pproximately m 1 3s. In the following examples, we will illustrate how these probabilities 0.68, 0.95, and 0.999 were derived. Another attribute of a normal distribution is that it is symmetric about the mean. The curve to the right of the mean is a mirror image of that to the left. A continuous variable with a distribution like that displayed
in Figure 5–1—whose mean, median, and mode are equal—is symmetric and satisfies the preceding conditions (i) through (iii) follows a normal distribution. Similar to the binomial case, there is a normal distribution model that can be used to compute probabilities. The normal probability model is shown below and computing probabilities with the
normal distribution model requires calculus, P(x) = 2 2 1 e − (x − μ) /(2σ) , σ 2π where m is the population mean and s is the population standard deviation. There is an alternative to using calculus to compute probabilities for normal variables, and it involves the use of probability tables. This is the approach we use. 80 CHAPTER 5 The Role of
Probability Example 5.11. Body mass index (BMI) for specific sex and age groups is approximately normally distributed. The mean BMI for men aged 60 is 29 with a standard deviation of 6, and for women aged 60 the mean BMI is 28 with a standard deviation of 7. Suppose we consider the distribution of BMI among men aged 60. Knowing that the
distribution is normal and having the mean and standard deviation allow us to completely generate the distribution. The distribution of BMI among men aged 60 is shown in Figure 5–3. Notice that the mean (m 5 29) is in the center of the distribution, the horizontal axis is scaled in units of the s tandard deviation (s 5 6), and the distribution essentially
ranges from m 2 3s to m 1 3s. This is not to say that there are not BMI values below 11 or above 47; there are such values, but they occur very infrequently. To compute probabilities about normal d istributions, we compute areas under the curve. For e xample, suppose a man aged 60 is selected at random—what is the probability his BMI is less than
29? The probability is displayed graphically and represented by the area under the curve to the left of the value 29 in Figure 5–4. The probability that a male has a BMI less than 29 is equivalent to the area under the curve to the left of the line drawn at 29. For any probability distribution, the total area under the curve is 1. For the normal distribution,
we know that the mean is equal to the median, and thus half (50%) of the area under the curve is above the mean and half is below, so P(BMI , 29) 5 0.50. We might want to compute the probability that a male has a BMI of 29 or less, P(x # 29). This can be thought of as P(x # 29) 5 P(x , 29) 1 P(x 5 29). We know that P(x , 29) 5 0.50. With the normal
distribution, P(x 5 29) is defined as 0. In fact, the probability of being exactly equal to any value is a lways defined as 0. That is not to say that there are no men with a BMI of 29. There is no area in a single line, and thus P(X 5 29) is defined as 0. Thus, P(x # 29) 5 P(x , 29). This c oncept will be illustrated further. Suppose we want to know the
probability that a male has a BMI less than 35. The probability is displayed graphically and represented by the area under the curve to the left of the value 35 in Figure 5–5. The probability that a male has a BMI less than 35 is equivalent to the area under the curve to the left of the line drawn at 35. For the normal distribution, we know that
approximately 68% of the area under the curve lies between the mean plus or minus one standard deviation. For men aged 60, 68% of the area under the curve lies between 23 and 35. We also know that thenormal distribution is symmetric about the mean; therefore P(29 , x , 35) 5 P(23 , x , 29) 5 0.34. Thus, P(x , 35) 5 0.5 1 0.34 5 0.84. What is the
probability that a male aged 60 has a BMI less than 41? Using similar logic and the fact that approximately FIGURE 5–3 Distribution of BMI Among Men Aged 60: Mean 5 29, Standard Deviation 5 6 11 17 23 29 35 41 47 WhatProbability Is Public Health? Models 81 FIGURE 5–4 P(BMI < 29) P(BMI 29) 11 17 23 29 35 41 47 35 41 47 FIGURE 5–5
P(BMI < 35) P(BMI 35) 11 17 23 95% of the area under the curve lies between the mean plus or minus two standard deviations—i.e., P(29 , x , 41) 5 P(17 , x , 29) 5 0.475—we can compute P(x , 41) 5 0.5 1 0.475 5 0.975. 29 Suppose we now want to compute the probability that a male aged 60 has a BMI less than 30 (the threshold for classifying
someone as obese). The area of interest, reflecting the probability, is displayed graphically in Figure 5–6. Because 82 CHAPTER 5 The Role of Probability FIGURE 5–6 P(BMI < 30) P(BMI 30) 11 17 23 30 is not the mean nor a multiple of standard d eviations above or below the mean, we cannot use the p roperties of a normal distribution to
determine P(x , 30). The P(x , 30) is certainly between 0.5 and 0.84, based on the previous computations, but we can determine a more exact value. To do so, we need a table of probabilities for the normal distribution. Because every application we face could potentially involve a normal distribution with a different mean and standard deviation, we will
use a table of probabilities for the standard normal distribution. The standard normal distribution is a normal distribution with a mean of 0 and standard deviation of 1. We will always use z to refer to a standard normal variable. Up to this point, we have been using x to denote the variable of interest (e.g., x 5 BMI, x 5 height, x 5 weight). z will be
reserved to refer to the standard normal distribution. The standard normal distribution is displayed in Figure 5–7. The mean of the standard normal distribution is 0; thus the distribution is centered at 0. Multiples of the standard deviation above and below the mean are by units of the standard deviation (s 5 1). The range of the standard normal d -
istribution is approximately 23 to 3. Table 1 in the Appendix contains probabilities for the standard normal distribution. 29 35 41 47 The body of Table 1 contains probabilities for the standard normal distribution, which correspond to areas under the s tandard normal curve. Specifically, Table 1 is organized to p rovide the area under the curve to the
left of or less than the specified z value. Table 1 can accommodate two decimal places of z. The units place and the first decimal place are shown in the left column and the second decimal place is displayed across the top row. For example, suppose we want to compute P(z , 0). Because Table 1 contains z to two decimal places, this is equivalent to P(z ,
0.00). We locate 0.0 in the left column (units and tenths place) and 0.00 across the top row (hundredths place). P(z , 0.00) 5 0.5000. Similarly, P(z , 0.52) 5 0.6985. The question of interest is P(x , 30). We now have Table 1, which contains all of the probabilities for the standard normal distribution. BMI follows a normal distribution with a mean of 29
and a standard deviation of 6. We can use the standard normal distribution to solve this problem. Figure 5–8 shows the distributions of BMI for men aged 60 and the standard normal distribution side-by-side. The areas under the curve are identical; only the scaling of the x-axis is different. BMI ranges from 11 to 47 while the standard normal variable,
z, ranges from 23 to 3. We want to compute P(x , 30). We determine the z value that c orresponds to x 5 30 and then use Table 1 to find the probability or area under WhatProbability Is Public Health? Models 83 FIGURE 5–7 The Standard Normal Distribution: m 5 0, s 5 1 –3 –1 –2 1 0 3 2 FIGURE 5–8 Distribution of BMI and Standard Normal
Distribution X 11 17 23 29 35 BMI, 29, 41 47 –3 the curve. The following formula converts an x value into a z score, also called a standardized score: Z= x−μ , σ where m is the mean and s is the standard deviation of the variable x. We want to compute P(x , 30). Using the preceding 0, Z, 6 –2 –1 0 1 2 1 3 formula, we convert (x 5 30) to its
corresponding z score (this is called standardizing): Z= 30 − 29 1 = = 0.17. 6 6 Thus, P(x , 30) 5 P(z , 0.17). We can solve the latter using Table 1: P(x , 30) 5 P(z , 0.17) 5 0.5675. Notice in 84 CHAPTER 5 The Role of Probability Figure 5–9 that the area below 30 and the area below 0.17 in the x and z distributions, respectively, are identical. Using
Table 1, P(z , 0.17) 5 0.5675. Thus, the chance that a male aged 60 has a BMI less than 30 is 56.75%. Suppose we now want to compute the probability that a male aged 60 has a BMI of 30 or less. Specifically, we want P(x # 30). P(x # 30) 5 P(x , 30) 1 P(x 5 30). The second term reflects the probability of observing a male age 60 with a BMI of exactly
30. We are computing probabilities for the normal distribution as areas under the curve. There is no area in a single line and thus P(x 5 30) is defined as 0. For the normal distribution, and for other probability distributions for continuous variables, FIGURE 5–9 P(x < 30) 5 P(z < 0.17) P(x 30) 11 17 23 29 35 41 47 0 1 2 3 P(z 0.17) –3 –2 –1
WhatProbability Is Public Health? Models this will be the case. Therefore, P(x ▯ 30) 5 P(x , 30) 5 0.5675. Note that for the binomial distribution and for other probability distributions for discrete variables, the probability of taking on a specific value is not defined as 0 (see Section 5.6.1). Consider again Example 5.11. What is the probability that a
male aged 60 has a BMI exceeding 35? Specifically, what is P(x > 35)? Again we standardize: P(x > 35) = P(z > 35 − 29 6 = = 1). 6 6 We now need to compute P(z . 1). If we look up z 5 1.00 in Table 1, we find that P(z , 1.00) 5 0.8413. Table 1 always gives the probability that z is less than the specified value. We want P(z . 1) (see Figure 5–10). Table 1
gives P(z , 1) 5 0.8413; thus P(z . 1) 5 1 2 0.8413 5 0.1587. Almost 16% of men aged 60 have BMI over 35. What is the probability that a male aged 60 has a BMI between 30 and 35? Note that this is the same as asking what proportion of men aged 60 has a BMI between 30 and 35. Specifically, we want P(30 , x , 35). To solve this, we standardize and
use Table 1. From the preceding examples, P(30 , x , 35) 5 P(0.17 , z , 1). This can be computed as P(0.17 , z , 1) 5 0.8413 2 0.5675 5 0.2738. This probability can be thought of as P(0.17 , z , 1) 5 P(z , 1) 2 P(z , 0.17). Now consider BMI in women. What is the probability that a female aged 60 has a BMI less than 30? We use the same approach, but
recall that for women aged 60, the mean is 28 and the standard deviation is 7, P(x < 30) = P(z < 30 − 28 2 = = 0.29). 7 7 Using Table 1, P(z , 0.29) 5 0.6141. Therefore, 61.41% of women aged 60 have a BMI less than 30, and 38.59% of women have a BMI of 30 or more. What is the probability that a female aged 60 has a BMI exceeding 40?
Specifically, what is P(x > 40)? Again, we standardize: P(x > 40) = P(z > 40 − 28 12 = = 1.71). 7 7 We need now to compute P(z > 1.71). If we look up z 5 1.71 on Table 1, we find that P(z , 1.71) 5 0.9564. P(z . 1.71) 5 1 2 0.9564 5 0.0436. Less than 5% of the women aged 60 have a BMI exceeding 40. Table 1 is very useful for computing probabilities
about normal distributions. To do this, we first standardize or convert a problem about a normal distribution (x) into a problem about the standard normal distribution (z). Once we have the problem in terms of z, we use Table 1 in the Appendix to compute the desired probability. FIGURE 5–10 Using Table 1 to Compute P(z . 1) P(z 1) 0.8413 P(z 1)

? –3 –2 –1 85 0 1 2 3 86 CHAPTER 5   The Role of Probability The standard normal distribution can also be useful for computing percentiles. A percentile is a value in the distribution that holds a specified percentage of the population below it. The pth percentile is the value that holds p% of the values below it. For example, the median is the 50th
percentile, the first quartile is the 25th percentile, and the third quartile is the 75th percentile. In some instances, it may be of interest to compute other percentiles, for example, the 5th or 95th. The following formula is used to compute percentiles of a normal distribution. To compute the 90th percentile, we use the formula x 5 m 1 zs. The mean and
standard deviation are 29 and 6, respectively; what is needed is the z value reflecting the 90th percentile of the standard normal distribution. To compute this value, we use Table 1—however, we use Table 1 almost in “ reverse.” When computing percentiles, we know the area under the curve (or probability) and want to compute the z score. In
Example 5.11, we computed z scores and used Table 1 to d etermine areas under the curve, or probabilities. Here we know that the area under the curve below the desired z value is 0.90 (or 90%). What z score holds 0.90 below it? The i nterior of Table 1 contains areas under the curve below z. If the area under the curve below z is 0.90, we find 0.90
in the body (center) of Table 1. The value 0.90 is not there exactly; however, the values 0.8997 and 0.9015 are contained in Table 1. These correspond to z values of 1.28 and 1.29, respectively (i.e., 89.97% of the area under the standard normal curve is below 1.28). The exact z value holding 90% of the values below it is 1.282. This value is
determined by a statistical computing package (e.g., Microsoft Excel®) with more precision than that shown in Table 1. Using z 5 1.282, we can now compute the 90th p ercentile of BMI for men: x 5 29 1 1.282(6) 5 36.69. Ninety percent of the BMIs in men aged 60 are below 36.69, and 10% of the BMIs in men aged 60 are above 36.69. What is the
90th percentile of BMI among women aged 60? x 5 28 1 1.282(7) 5 36.97. Ninety percent of x 5 m 1 zs, where m is the mean, s is the standard deviation of the variable x, and z is the value from the standard normal distribution for the desired percentile. Example 5.12. Consider again Example 5–11, where we analyzed BMI in men and women aged 60.
The mean BMI for men aged 60 is 29 with a standard deviation of 6, and for women aged 60 the mean BMI is 28 with a standard deviation of 7. What is the 90th percentile of BMI for men? Figure 5–11 shows the distribution of BMI in men aged 60. The 90th percentile is the BMI that holds 90% of the BMI values below it. It therefore must be a value
in the high (right) end of the distribution if 90% of the values are below it, and therefore only 10% are above it. The vertical line in Figure 5–11 is an estimate of the value of the 90th percentile. FIGURE 5–11   90th Percentile of BMI Among Men Aged 60: Mean529, Standard Deviation56 90% 11 17 23 29 35 90th Percentile 41 47 WhatProbability Is
Public Health? Models the BMIs in women aged 60 are below 36.97, and 10% of the BMIs in women aged 60 are above 36.97. Table 1A in the Appendix was developed using Table 1 and contains the z values for popular percentiles. It can be used to compute percentiles for normal distributions. A popular application of percentiles is in
anthropometrics, which is the study of human measurements. These measures, such as weight and height, are used to study patterns in body size. For example, pediatricians often measure a child’s weight, length, or height, and head circumference. The observed values are often converted into percentiles to assess where a particular child falls
relative to his or her peers (i.e., children of the same sex and age). For example, if a child’s weight for his or her age is extremely low, it might be an indication of malnutrition. Growth charts for boys and girls including length-for-age, weight-for-age, BMI-for-age, head circumference-for-age, and other anthropometric charts are available for infants
from birth to 36 months and children 2 to 20 years old on the Centers for Disease Control (CDC) website at . Example 5.13. For infant girls, the mean length at 10 months is 72 cm with a standard deviation of 3 cm. Suppose a girl of 10 months has a measured length of 67 cm. How does her length compare to other girls of 10 months? We can compute
her percentile by determining the proportion of girls with lengths below 67. Specifically, P(x < 67) = P(z < 67 − 72 −5 = = − 1.67). 3 3 Using Table 1, P(z , 21.67) 5 0.0475. This girl is in the 4.75th percentile. Fewer than 5% of girls of 10 months are below 67 cm. This may be an instance where some intervention is needed. 5.6.3   Sampling
Distributions In Chapters 6, 7, 9, 10, and 11, we focus on statistical inference, where we make inferences or generalizations about population parameters based on observed sample statistics. In Chapter 4, we presented techniques and a number of statistics to summarize sample data. Suppose we want to generate an estimate of a continuous variable
in a population (e.g., weight, HDL cholesterol level). It is very typical to estimate the mean of a continuous variable in a population. The mean of a representative s ample is a very good estimate of the unknown p opulation mean. If a second sample is selected, that sample might produce a slightly different estimate (i.e., the mean of the second sample
might be slightly different than the mean of the first). Whenever we perform statistical inference, we must recognize that we are essentially working with incomplete information— 87 specifically, only a fraction of the population. When we make estimates about population parameters based on sample statistics, it is e xtremely important to quantify the
precision in our estimates. This is done using probability and, in particular, the probability models we have just discussed. Example 5.14. Consider the following small population consisting of N 5 6 patients who recently underwent total hip replacement. We are interested in a patient’s self-reported pain-free function, rated on a scale of 0 to 100, with
higher scores representing better function (e.g., 0 5 severely limited and painful functioning to 100 5 completely pain-free functioning), measured 3 months post-procedure. The data are shown below and are ordered from smallest to largest: 25   50   80   85   90   100 The population mean is m 5 SX }} N 5 71.7 and the s tandard ∑(X − μ)2
= 25.9. A box-whisker plot of N the population data is shown in Figure 5–12. deviation is σ = The distribution of pain-free function scores is slightly skewed, with the majority of patients reporting high scores. Suppose we did not have the population data and instead were interested in estimating the mean pain-free function score based on a sample.
Suppose we planned to take a sample of size n 5 4. Table 5–9 shows all possible samples of size n 5 4 from the population of N 5 6 when sampling without replacement. (Sampling without replacement means that we select an individual and with that person aside, we select a second from those remaining, and so on. In contrast, when sampling with
replacement, we make a selection, record that selection, and place that person back before making a second selection. When sampling with replacement, the same individual can be selected into the sample multiple times.) The right column shows the sample mean based on the four observations contained in that sample. The probability of selecting
any particular s ample in Table 5–9 is 1 / 15 5 0.07. Suppose by chance we select Sample 1. The mean of Sample 1 is 60. If we based our estimate of the unknown population mean on Sample 1, and particularly on the sample mean of Sample 1, we would underestimate the true population mean (m = 71.7). If we selected Sample 15, we would
overestimate the population mean because the mean of Sample 15 is X w = 88.8. The collection of all possible sample means (in this example, there are 15 distinct samples that are produced by sampling four individuals at random without replacement) is called the sampling distribution of the sample means. We 88 CHAPTER 5   The Role of Probability
FIGURE 5–12   Box-Whisker Plot of Population Data 20 40 60 80 100 X TABLE 5–9   All Samples of Size n 5 4 Sample 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 Observations in the Sample 25 50 80 85 25 50 80 90 25 50 80 100 25 50 85 90 25 25 25 25 25 25 25 25 50 50 50 50 80 50 50 80 80 80 80 85 80 80 80 80 85 85 85 90 85 85 90 85 90 90 90 85 85 90 90 90 100 100 90 100
100 100 90 100 100 100 100 Sample Mean (w X) 60.0 61.3 63.8 62.5 65.0 66.3 70.0 72.5 73.8 75.0 76.3 78.8 80.0 81.3 88.8 consider it a population because it includes all possible values produced in this case by a specific sampling scheme. If we compute the mean and standard deviation of this population of sample means, we get the following: m X 5
71.7 and a stanw dard deviation of s X 5 8.5. The subscripts here are to distinw guish these parameters from those based on the population data (x). To be consistent, the parameters based on the population data could include a subscript x. Notice that the mean of the sample means is m X 5 71.7, w which is precisely the value of the population mean
(m). This will a lways be the case. Specifically, the mean of the sampling distribution of the sample means will always be equivalent to the population mean. This is important as it indicates that, on average, the sample mean is equal to the population mean. This is the definition of an unbiased estimator. Unbiasedness is a desirable property in an
estimator. Notice also that the variability in the sample means is much smaller than the variability in the population; this will also always be the case. A box-whisker plot of the population of sample means is shown in Figure 5–13. WhatProbability Is Public Health? Models 89 FIGURE 5–13   Box-Whisker Plot of Population of Sample Means 55 60 65 70
75 80 85 90 Mean Notice that the distribution of the sample means is more symmetric and has a much more restricted range (60 to 88.8) than the distribution of the population data (25 to 100) shown in Figure 5–12. The importance of these observations is stated formally in the Central Limit Theorem. Central Limit Theorem. Suppose we have a
population with known mean and standard deviation, m and s, respectively. The distribution of the population can be normal or it can be non-normal (e.g., skewed toward the high or low end, or flat). If we take simple random samples of size n from the population with replacement, then for large samples (usually defined as samples with n > 30), the
sampling distri bution of the sample means is approximately normally distributed with a mean of m X 5 m and a standard deviation w σ . σX = n The importance of this theorem for applications is as f ollows. The theorem states that, regardless of the distribution of the population (normal or not), as long as the sample is sufficiently large (usually n .
30), then the distribution of the sample means is approximately normal. In Section 5.6.2, we learned that it is relatively straightforward to compute probabilities about a normal distribution. Therefore, when we make inferences about a population mean based on the sample mean, we can use the normal probability model to quantify uncertainty. This
will become explicit in Chapters 6, 7, 9, 10, and 11. Before illustrating the use of the Central Limit Theorem, we first illustrate the result. For the result of the Central Limit Theorem to hold, the sample must be sufficiently large (n > 30). There are two exceptions to this: If the outcome in the population is normal, then the result holds for samples of
any size (i.e., the sampling distribution of the sample means is approximately normal even for samples of size less than 30). If the o utcome in the population is dichotomous, then the result holds for samples that meet the following criterion: min[np, n(1 2 p)] . 5, where n is the sample size and p is the probability of success in the population. We next
illustrate the results of the Central Limit Theorem for populations with different distributions. Example 5.15. Suppose we measure a characteristic in a population and that this characteristic follows a normal distribution with a mean of 75 and standard deviation of 8. The 90 CHAPTER 5   The Role of Probability distribution of the characteristic in the
population is shown in Figure 5–14. If we take simple random samples with replacement of size n 5 10 from the population and for each sample we compute the sample mean, the Central Limit Theorem states that the distribution of sample means is approximately normal. Note that the sample size n 5 10 does not meet the criterion of n . 30, but in
this case, our original population is normal and therefore the result holds. The distribution of sample means based on samples of size n 5 10 is shown in Figure 5–15. The mean of the sample means is 75 and the standard deviation of the sample means is 2.5 (i.e., mX 5 m 5 75 w and a standard deviation of σ X = σ = 8 = 2.5). n 10 If we take simple
random samples with replacement of size n 5 5, we get a similar distribution. The distribution of sample means based on samples of size n 5 5 is shown in Figure 5–16. The mean of the sample means is again 75 and the standard deviation of the sample means is 3.6 (i.e., mX 5 m w σ 8 5 75 and a standard deviation of σ X = = = 3.6). Notice n 5 that the
variability in sample means is larger for samples of size 5 as compared to samples of size 10. Example 5.16. Suppose we measure a characteristic in a population and that this characteristic is dichotomous with 30% of the population classified as a success (i.e., p 5 0.30). The characteristic might represent disease status, the presence or absence of a
geneticabnormality, or the success of a medical procedure. The d istribution of the outcome in the population is shown in Figure 5–17. This population is clearly not normal. The results of the Central Limit Theorem are said to apply to binomial populations as long as the minimum of np and n(1 2 p) is at least 5, where n refers to the sample size. If we
take simple random samples with replacement of size n 5 20 from this binomial population and for each sample we compute the sample mean, the Central Limit Theorem states that the distribution of sample means should be approximately normal because min [np, n(1 2 p)] 5 min[20(0.3), 20(0.7)] 5 min(6, 14) 5 6. The distribution of sample means
based on samples of size n 5 20 is shown in Figure 5–18. The mean of the sample means is 6 and the standard deviation of the sample means is 0.4 (i.e., σ 2 = = 0.4). m X 5 m 5 6 and a standard deviation of σ X = w n 20 Suppose we take simple random samples with re placement of size n 5 10. In this scenario, we do not meet the s ample size
requirement for the results of the Central Limit Theorem to hold—i.e., min[np, n(1 2 p)] 5 min[10(0.3), 10(0.7)] 5 min(3, 7) 5 3. The distribution of sample means based on samples of size n 5 10 is shown in Figure 5–19. The distribution of sample means based on samples of size n 5 10 is not quite normally distributed. FIGURE 5–14   Normal Population
12,000 10,000 8000 6000 4000 2000 0 51 59 67 75 83 91 99 WhatProbability Is Public Health? Models FIGURE 5–15   Distribution of Sample Means Based on Samples of Size n 5 10 10,000 8000 6000 4000 2000 0 67.5 70.0 72.5 75 77.5 80 82.5 FIGURE 5–16   Distribution of Sample Means Based on Samples of Size n 5 5 12,000 10,000 8000 6000
4000 2000 0 64.2 67.8 71.4 75 78.6 82.2 85.8 91 92 CHAPTER 5   The Role of Probability FIGURE 5–17   Population with a Dichotomous Outcome 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0 Yes No FIGURE 5–18   Distribution of Sample Means Based on Samples of Size n 5 20 20,000 15,000 10,000 5000 0 4.8 5.2 5.6 6.0 6.4 6.8 7.2 WhatProbability Is Public
Health? Models 93 FIGURE 5–19   Distribution of Sample Means Based on Samples of Size n 5 10 25,000 20,000 15,000 10,000 5000 0 3.0 4.0 5.0 The sample size must be larger for the distribution to approach normality. Example 5.17. Suppose we measure a characteristic in a population and that this characteristic follows a Poisson distribution with
m 5 3 and s 5 1.7. The Poisson is another probability model for a discrete variable and its distribution is shown in Figure 5–20. This population is not normal. The results of the Central Limit Theorem are said to apply when n > 30. The distribution of sample means based on samples of size n 5 30 is shown in Figure 5–21. The mean of the sample means
is 3 and the standard deviation of the sample means is 0.3 (i.e., mX 5 m 5 3 and w σ 1.7 a standard deviation of σ X = = = 0.3). n 30 Samples of smaller size from this population do not meet the requirements of the Central Limit Theorem, and thus the result would not hold. For example, suppose we take simple random samples with replacement of
size n 5 10. The distribution of sample means based on samples of size n 5 10 is shown in Figure 5–22 and is not quite normally distributed. Samples of size 30 or greater will be approximately normally distributed. The distribution of sample means based 6.0 7.0 8.0 9.0 on samples of size n 5 50 is shown in Figure 5–23 and is normally distributed. The
mean of the sample means will always be equal to the population mean (mX 5 m). The standard deviation of the w σ σ sample means, defined as sX , equals (i.e., σ X =) w n n and is also called the standard error. The standard error decreases as the sample size increases. Specifically, the variability in the sample means is smaller for larger sample
sizes. This is intuitively sensible as extreme values will have less impact in samples of larger size. Notice in the previous examples how the standard errors decrease as the sample sizes increase. Example 5.18. High density lipoprotein (HDL) cholesterol—the “good” cholesterol—has a mean of 54 and a standard deviation of 17 in patients over age 50.
Suppose a physician has 40 patients over age 50 and wants to determine the probability that their mean HDL cholesterol is 60 or more. Specifically, we want to know P(X w . 60). Probability questions about a sample mean can be addressed with the Central Limit Theorem as long as the sample size is sufficiently large. In this example, we have 94
CHAPTER 5   The Role of Probability FIGURE 5–20   Poisson Population 20,000 15,000 10,000 5000 0 0 2 1 3 4 5 6 7 8 9 FIGURE 5–21   Distribution of Sample Means Based on Samples of Size n 5 30 15,000 10,000 5000 0 2.1 2.4 2.7 3.0 3.3 3.6 3.9 10 WhatProbability Is Public Health? Models FIGURE 5–22   Distribution of Sample Means Based on
Samples of Size n 5 10 20,000 15,000 10,000 5000 0 1.0 2.0 3.0 4.0 5.0 6.0 FIGURE 5–23   Distribution of Sample Means Based on Samples of Size n 5 50 12,000 8000 4000 0 2.28 2.52 2.76 3.0 3.24 3.48 3.72 95 96 CHAPTER 5   The Role of Probability n 5 40. We can therefore assume that the distribution of the sample mean is approximately normal
and can compute the desired probability by standardizing and using Table 1. Standardizing involves subtracting the mean and dividing by the standard deviation. The mean of X w is mX 5 m and the w σ σ = standard deviation of X w is X . Therefore, the formula n to standardize a sample mean is Z= X − μX σX = X−μ . σ/ n We want to compute P(X >
60) = P(z > 60 − 54 6 = = 2.22). 17 / 40 2.7 P(z . 2.22) can be solved with Table 1 with P(z . 2.22) 5 1 2 0.9868 5 0.0132. The chance that the mean HDL in 40 patients exceeds 60 is 1.32%. What is the probability that the mean HDL cholesterol in 40 patients is less than 50? We now want P(X < 50) = P(z > 50 − 54 −4 = = − 1.48). 17 / 40 2.7 P(z ,
21.48) can be solved with Table 1, P(z , 21.48) 5 0.0694. The chance that the mean HDL in 40 patients is less than 50 is 6.94%. Example 5.19. We want to estimate the mean low density lipoprotein (LDL)—the “bad” cholesterol—in the population of adults 65 years of age and older. Suppose that we know from studies of adults under age 65 that the
standard deviation is 13 and we assume that the variability in LDL in adults 65 years of age and older is the same. We select a sample of n 5 100 participants 65 years of age and older and use the mean of the sample as an estimate of the population mean. We want our estimate to be precise—specifically, we want it to be within 3 units of the true
mean LDL value. What is the probability that our estimate (i.e., the sample mean) is within 3 units of the true mean? – We can represent this question as P(m 2 3 , X , m 1 3). Because this is a probability about a sample mean, we appeal to the Central Limit Theorem. With a sample of size n 5 100, we satisfy the sample size criterion and can use the
Central Limit Theorem to solve the problem (i.e., convert to z and use Table 1). In the previous example, we asked questions around specific values of the sample mean (e.g., 50, 60) and converted those to z scores and worked with Table 1. Here the values of interest are m 2 3 and m 1 3. We use these values below: P(m 2 3 , X w , m 1 3) 5 (μ + 3) − μ
⎞⎛ (μ − 3) − μ , z< P⎜ ⎟= ⎝ 13/ 100 13/ 100 ⎠3⎞⎛ −3 P⎜ < z < ⎟= P(− 2.31 < z < 2.31). ⎝ 1.3 1.3 ⎠This we can solve with Table 1, P(22.31 , z , 2.31) 5 0.9896 2 0.0104 5 0.9792. There is a 97.92% chance that the sample mean, based on a sample of size n 5 100, will be within 3 units of the true population mean. This is a very powerful statement. When
looking only at 100 individuals aged 65 and older, there is almost a 98% chance that the sample mean is within 3 units of the population mean. Example 5.20. Alpha fetoprotein (AFP), a substance produced by a fetus, is often measured in pregnant women as a means of assessing whether there might be problems with fetal development. High levels of
AFP have been seen in babies with neural-tube defects. When measured at 15–20 weeks gestation, AFP is normally distributed with a mean of 58 and a standard deviation of 18. What is the probability that AFP exceeds 75 in a pregnant woman measured at 18 weeks gestation? Specifically, what is P(x > 75)? Because AFP is normally distributed, we
standardize: 75 – 58 17 P(x > 75) = P(z > = = 0.94). Using Table 1, 18 18 P(x > 75) = P(z > 0.94) = 1 – 0.8264 = 0.1736. There is a 17% chance that AFP exceeds 75 in a pregnant woman measured at 18 weeks gestation. In a sample of 50 women, what is the probability that – their mean AFP exceeds 75? Specifically, what is P(X > 75)? Using the
Central Limit Theorem, we standardize: 75 – 58 17 P(X > 75) = P(z > = = 6.67). It is ex18 / 50 2.55 tremely unlikely (probability very close to 0) to observe a z score exceeding 6.67. There is virtually no chance that in a sample of 50 women their mean AFP exceeds 75. Notice that the first part of the question addresses the probability of observing a
single woman with an AFP exceeding 75, whereas the second part of the question addresses the probability that the mean AFP in a sample of 50 women exceeds 75. The latter requires the application of the Central Limit Theorem. 5.7   SUMMARY In Chapters 6, 7, 9, 10, and 11, we discuss statistical inference in detail. We present formulas and
procedures to make inferences about populations based on a single sample. The relationship between the sample statistic and the population parameter is based on the sampling distribution of that statistic and What Practice Is PublicProblems Health? probability theory. Here we discussed probability as it applies to selecting individuals from a
population into a sample. There are some basic concepts of probability that can be applied when the entire population can be enumerated. When the population enumeration is not available, probability models can be used to determine probabilities as long as specific attributes are satisfied. The binomial and normal distribution models are popular
models for discrete and continuous outcomes, respectively. A key theorem is the Central Limit Theorem, which brings together the concepts of probability and inference. We rely heavily on the Central Limit Theorem in the next chapters, where we discuss statistical inference in detail. Table 5–10 summarizes key f ormulas and concepts in probability.
5.8   PRACTICE PROBLEMS 1. A recent study reported that the prevalence of hyperlipidemia (defined as total cholesterol over 200) is 30% in children 2 to 6 years of age. If 12 children are analyzed: a. What is the probability that at least 3 are hyperlipidemic? b. What is the probability that exactly 3 are hyperlipidemic? c. How many would be
expected to meet the criteria for hyperlipidemia? 2. Hyperlipidemia in children has been hypothesized to be related to high cholesterol in their parents. The data in Table 5–11 were collected on parents and children. a. What is the probability that one or both parents are hyperlipidemic? b. What is the probability that a child and both parents are
hyperlipidemic? c. What is the probability that a child is hyperlipidemic if neither parent is hyperlipidemic? TABLE 5–10   Summary of Key Formulas Concept Formula Basic probability Conditional probability rule Sensitivity Specificity False positive fraction False negative fraction Positive predictive value Negative predictive value Independent
events Bayes Theorem Binomial distribution Standard normal distribution Percentiles of the normal distribution Application of Central Limit Theorem 97 Number of persons with characteristic P(Characteristic) 5 }}}} N P(A B) = P(A and B) P(B) P(Screen positive) Disease) P(Screen negative) Disease free) P(Screen positive) Disease free) P(Screen
negative) Disease) P(Disease) Screen positive) P(Disease free) Screen negative) P(A) B) 5 P(A) or P(B) A) 5 P(B) P(B A)P(A) P(B) n! P(x successes) = p x (1 − p)n− x x !(n − x)! P(A B) = X2m z 5 }} (Table 1) s X 5 m 1 zs (Table 1A) z= X −μ (Table 1) σ/ n 98 CHAPTER 5   The Role of Probability TABLE 5–11   Hyperlipidemia in Parents and Children
Child Both Parents Hyperlipidemic One Parent Hyperlipidemic Neither Parent Hyperlipidemic Not hyperlipidemic 13 34 83 Hyperlipidemic 45 42   6 d. What is the probability that a child is hyperlipidemic if both parents are hyperlipidemic? 3. Total cholesterol in children aged 10 to 15 is assumed to follow a normal distribution with a mean of 191 and
a standard deviation of 22.4. a. What proportion of children 10 to15 years of age has total cholesterol between 180 and 190? b. What proportion of children 10 to 15 years of age would be classified as hyperlipidemic? (Assume that hyperlipidemia is defined as a total cholesterol level over 200.) c. If a sample of 20 children is selected, what is the
probability that the mean cholesterol level in the sample will exceed 200? 4. A national survey of graduate students is conducted to assess their consumption of coffee. Table 5–12 summarizes the data. a. What proportion of students drinks decaffeinated coffee only? b. What proportion of coffee drinkers (caffeinated and decaffeinated) is female?
c. What proportion of the females does not drink coffee? 5. Among coffee drinkers, men drink a mean of 3.2 cups per day with a standard deviation of 0.8 cup. Assume the number of cups per day follows a normal distribution. a. What proportion drinks 2 cups per day or more? b. What proportion drinks no more than 4 cups per day? c. If the top 5% of
coffee drinkers are considered “heavy” coffee drinkers, what is the minimum number of cups consumed by a heavy coffee drinker? d. If a sample of 20 men is selected, what is the probability that the mean number of cups per day is greater than 3? 6. A study is conducted to assess the impact of caffeine consumption, smoking, alcohol consumption,
and physical activity on cardiovascular disease. Suppose that 40% of participants consume caffeine and smoke. If 8 participants are evaluated, what is the probability that: a. E xactly half of them consume caffeine and smoke? b. More than 6 consume caffeine and smoke? c. Exactly 4 do not consume caffeine or smoke?   7. As part of the study
described in Problem 6, investigators wanted to assess the accuracy of self-reported smoking status. Participants are asked whether they currently smoke or not. In addition, laboratory tests are performed on hair samples to determine the presence or absence of nicotine. The laboratory assessment is considered the gold standard, or the truth about
nicotine consumption. The data are shown in Table 5–13. a. What is the sensitivity of the self-reported smoking status? b. What is the specificity of the self-reported smoking status? TABLE 5–12   Coffee Consumption in Graduate Students Do Not Drink Coffee Drink Decaffeinated Only Drink Caffeinated Coffee Male 145   94 365 Female   80 121 430
What Practice Is PublicProblems Health? TABLE 5–13   Self-Reported Smoking Status Nicotine Absent Nicotine Present Self-reported nonsmoker 82 14 Self-reported smoker 12 52   8. A recent study of cardiovascular risk factors r eported that 30% of adults meet criteria for hypertension. If 15 adults are assessed, what is the probability that:
a. Exactly 5 are the criteria for hypertension? b. None meet the criteria for hypertension? c. How many would you expect to meet the criteria for hypertension?   9. Table 5–14 displays blood pressure status by gender. a. What proportion of the participants has optimal blood pressure? b. W hat proportion of men has optimal blood pressure? c. What
proportion of participants with hypertension is men? d. Are hypertensive status and male gender independent? 10. Diastolic blood pressures are assumed to follow a normal distribution with a mean of 85 and a standard deviation of 12. a. What proportion of people has diastolic blood pressures less than 90? b. What proportion has diastolic blood
pressures between 80 and 90? c. If someone has a diastolic blood pressure of 100, what percentile does this represent? 11. Consider the data described in Problem 10. If 15 participants are sampled, what is the probability that their mean diastolic blood pressure exceeds 90? TABLE 5–14   Blood Pressure by Gender 99 12. A large national study finds
that 10% of pregnant women deliver prematurely. A local obstetrician is seeing 16 pregnant women in his next clinic session. a. What is the probability that none will deliver prematurely? b. What is the probability that fewer than 3 will deliver prematurely? c. What is the probability that none will deliver prematurely if, in fact, the true percentage
who deliver prematurely is 5.5%? d. If the true percentage is 10% and this obstetrician has 146 pregnant women under his care, how many would be expected to deliver prematurely? 13. Table 5–15 cross-classifies pregnant women in a study by their body mass index (BMI) at 16 weeks gestation and whether they had a pre-term delivery. a. What is
the probability that a woman delivers preterm? b. What is the probability a woman has a BMI less than 30 and delivers pre-term? c. What proportion of women with a BMI greater than 35 delivers pre-term? d. Are BMI and pre-term delivery independent? Justify. 14. In the study described in Problem 13, suppose the mean BMI at 16 weeks gestation
is 28.5 with a standard deviation of 3.6, and that BMI is assumed to follow a normal distribution. Find the following: a. The proportion of women with a BMI greater than 30. b. The proportion of women with a BMI greater than 40. c. The BMI that separates the top 10% from the rest. 15. Suppose we want to estimate the mean BMI for women in
pregnancy at 20 weeks gestation. If we have a sample of 100 women and measure their BMI at 20 weeks gestation, what is the probability that the sample mean is within 1 unit of the true BMI if the standard deviation in the BMI is taken to be 3.6? TABLE 5–15   Body Mass Index and Preterm Delivery Optimal Normal Hypertension Total Male 22   73
  55 150 Female 43 132   65 240 Pre-term   320   80 120 Total 65 205 120 390 Full term 4700 480 300 BMI < 30 BMI 30–34.9 BMI 351 100 CHAPTER 5   The Role of Probability 16. Diastolic blood pressures are approximately normally distributed with a mean of 75 and a standard deviation of 10. a. What is the 90th percentile of diastolic blood -
pressure? b. If we consider samples of 20 patients, what is the 90th percentile of the mean diastolic blood pressure? 17. Table 5–16 displays the number of children in a local town classified as normal weight, overweight, and obese by year in school. a. What proportion of the children is obese? b. What proportion of the elementary school children is
overweight or obese? c. What proportion of the normal weight children is in high school? 18. In a primary care clinic, 25% of all patients who have appointments fail to show up. Each clinic session has 10 scheduled appointments. a. What is the probability that half of the patients fail to show up? b. What is the probability that all patients show up?
c. In a week, there are 10 clinic sessions. How many patients would you expect to fail to show up? 19. High density lipoprotein (HDL) in healthy males follows a normal distribution with a mean of 50 and a standard deviation of 8. a. W hat proportion of healthy males has HDL exceeding 60? b. What proportion of healthy males has HDL lower than 40?
c. What is the 90th percentile of HDL in healthy males? 20. Table 5–17 summarizes data collected in a study to evaluate a new screening test for ovarian cancer. TABLE 5–16   Obesity Status in Children Normal Weight Overweight Obese Total Elementary   50 120   80 250 Middle   30   45   45 120 Junior high   50   50   40 140 High school   30   85   85
200 Total 160 300 250 710 TABLE 5–17   New Screening Test for Ovarian Cancer Screening Test Ovarian Cancer Free of Ovarian Cancer Positive 28 23 Negative 22 127 Total 50 150 There are 200 women are involved in the study. Fifty have ovarian cancer and 150 do not. The results are tabulated below. a. Find the sensitivity of the screening test.
b. Find the false positive fraction of the screening test. c. What proportion of women who screen positive actually has ovarian cancer? 21. An experimental drug has been shown to be 75% effective in eliminating symptoms of allergies in animal studies. A small human study involving six participants is conducted. What is the probability that the drug
is effective on more than half of the participants? REFERENCES 1. National Center for Health Statistics. Health, United States, 2003, with Chartbook on Trends in the Health of Americans. Hyattsville, MD: US Government Printing Office, 2003. 2. Hedley, A.A., Ogden, C.L., Johnson, C.L., Carroll, M.D., Curtin, L.R., and Flegal, K.M. “Prevalence of
overweight and obesity among U.S. children, adolescents, and adults, 1999–2002.” Journal of the American Medical Association 2004; 291: 2847–2850. 3. Cope, M.B. and Allison, D.B. “Obesity: Person and population.” Obesity 2006; 14: S156–S159. 4. Kim, J., Peterson, K.F., Scanlon, K.S., Fitzmaurice, G.M., Must, A., Oken, E., Rifas-Shiman, S.L., Rich-
Edwards, J.W., and Gillman, M.W. “Trends in overweight from 1980 through 2001 among preschooled children enrolled in a health maintenance organization.” Obesity 2006; 14: 1107–1112. 5. Cochran, W.G. Sampling Techniques (3rd ed.). New York: John Wiley & Sons, 1977. 6. Kish, L. Survey Sampling (Wiley Classics Library). New York: John
Wiley & Sons, 1995. 7. Rosner, B. Fundamentals of Biostatistics. Belmont, CA: Duxbury-Brooks/ Cole, 2006. 8. SAS Version 9.1. © 2002–2003 by SAS Institute Inc., Cary, NC. 9. Thompson, I.M., Pauler, D.K., Goodman P.J., et al. “Prevalence of prostate cancer among men with a prostate-specific antigen level # 4.0 ng per milliliter.” The New England
Journal of Medicine 2004; 350(22): 2239–2246. 10. D’Agostino, R.B., Sullivan, L.M., and Beiser, A. Introductory Applied Biostatistics. Belmont, CA: Duxbury–Brooks/Cole, 2004.   6 chapter Confidence Interval Estimates W hen and W hy Key Questions •• What is the difference between a 95%
confidence interval and the mid-range of a sample defined by the 2.5th and 97.5th percentiles? •• What do pollsters mean when they show different percentages of voters in favor of specific candidates and call the results “a statistical dead heat”? •• Why not generate a 99% (even 100%) confidence interval instead of 95% to be sure about results? In
the News Traumatic brain injury (TBI) and concussion, a special type of TBI, are the source of extensive research and debate. The Centers for Disease Control and Prevention (CDC) collects data on TBIs and produces fact sheets on this topic regularly.1 There are nearly 2 million reported TBIs in the United States each year, and approximately
300,000 of these injuries are sports related. Among men, such an injury is most likely to be sustained while playing football. Among women, TBI is most likely to be sustained while playing soccer. A study of almost 9 million insurance records reported that concussions are on the rise, particularly for adolescents, and that prior estimates of concussion
incidence were actually underestimates.2 There are a number of challenges in collecting accurate data on concussions, as approximately 47% of people who suffer concussions do not experience symptoms immediately after the injury occurs.3 Available at .html. 2 “Concussions on the rise for adolescents.” ScienceDaily. Available at . htm. Accessed
July 10, 2016. 3 Sports Concussion Institute. Available at treatment.com/concussionfacts.html. 1 Dig In •• How would you go about estimating the proportion of athletes in your school who suffered a concussion in the past year? •• How would you be sure that you captured all of the concussion events? •• Suppose that you collect data and estimate
that 7% of students engaged in organized sports at your school experienced a concussion during the academic year, with a 95% confidence interval of (1%−13%). What is the a ppropriate interpretation of this confidence interval? L earning O bjectives By the end of this chapter, the reader will be able to •• Define point estimate, standard error,
confidence level, and margin of error •• Compare and contrast standard error and margin of error •• Compute and interpret confidence intervals for means and proportions •• Differentiate independent and matched or paired samples •• Compute confidence intervals for the difference in means and proportions in independent samples and for the
mean difference in paired samples •• Identify the appropriate confidence interval formula based on type of outcome variable and number of samples We now begin statistical inference. In Chapter 4, we presented descriptive statistics used to summarize sample data. In Chapter 5, we presented key concepts in probability and the Central Limit
Theorem. In statistical inference, we use all of these concepts to make inferences about unknown 102 CHAPTER 6   Confidence Interval Estimates opulation parameters based on sample statistics. There p are two broad areas of statistical inference: estimation and hypothesis testing. In estimation, sample statistics are used to generate estimates
about unknown population parameters. In hypothesis testing, a specific statement or hypothesis is generated about a population parameter and sample statistics are used to assess the likelihood that the hypothesis is true. We discuss estimation here and introduce hypothesis testing in Chapter 7. Estimation is the process of determining a likely value
for a population parameter (e.g., the true population mean or population proportion) based on a random sample. In practice, we select a sample from the population and use sample statistics (e.g., the sample mean or the sample proportion) to estimate the unknown parameter. The sample should be representative of the population, with participants
selected at random from the population. Because different samples can produce different results, it is necessary to quantify the precision—or lack thereof—that might exist among estimates from different samples. The techniques for estimation as well as for other procedures in statistical inference depend on the appropriate classification of the key
study variable (which we also call the outcome or endpoint) as continuous or dichotomous. (There are other types of variables, which are discussed in Chapters 7, 10, and 11; here we focus on continuous and dichotomous outcomes.) Another key issue is the number of comparison groups in the investigation. For example, in the two-comparison-group
case it is important to d etermine whether the samples from the groups are independent (i.e., physically separate, such as men versus women) or dependent (also called matching or pairing). These issues dictate the appropriate estimation t echnique. Table 6–1 outlines these issues and identifies the estimation techniques that we discuss here. Each is
discussed in detail. 6.1   INTRODUCTION TO ESTIMATION There are two types of estimates that can be produced for any population parameter, a point estimate and a confidence interval estimate. A point estimate for a population parameter is a single-valued estimate of that parameter. A confidence interval (CI) estimate is a range of values for a
population parameter with a level of confidence attached (e.g., 95% confidence that the interval contains the unknown parameter). The level of confidence is similar to a probability. The CI starts with the point e stimate and builds in what is called a margin of error. The margin of error incorporates the confidence level (e.g., 90% or 95%, which is
chosen by the investigator) and the sampling variability or the standard error of the point estimate. A CI is a range of values that is likely to cover the true population parameter, and its general form is point estimate 6 margin of error. The point estimate is determined first. The point estimates for the population mean and proportion are the sample
mean and sample proportion, respectively. These are our best single-valued estimates of the unknown population parameters. Recall from Chapter 5 that the sample mean is an unbiased estimator of the population mean. The same holds true for the sample proportion with regard to estimating the population proportion. Thus the starting place, or
point estimate, for the CI for the population mean is the s ample mean, and the point estimate for the population proportion is the sample proportion. Next, a level of confidence is selected that reflects the likelihood that the CI contains the true, unknown parameter. Usually, confidence levels of 90%, 95%, and 99% are chosen, although theoretically
any confidence level between 0% and 100% can be selected. Suppose we want to generate a CI estimate for an unknown population mean. Again, the form of the CI is point estimate 6 margin of error, or X w6 margin of error. Suppose we select a 95% TABLE 6–1   Estimation Techniques Number of Samples Outcome Variable Parameter to be
Estimated One sample Continuous Mean Two independent samples Continuous Difference in means Two dependent, matched samples Continuous Mean difference One sample Dichotomous Proportion (e.g., prevalence, cumulative incidence) Two independent samples Dichotomous Difference or ratio of proportions (e.g., attributable risk, relative risk,
odds ratio) 103 Introduction to Estimation confidence level. This means that there is a 95% probability that a CI will contain the true population mean. Thus, P(X w 2 margin of error , m , X w 1 margin of error) 5 0.95. In Chapter 5, we presented the Central Limit Theorem, which stated that for large samples, the distribution of the sample means is
approximately normal with a mean m Xw 5 m σ and standard deviation σ X = . We use the Central Limit n Theorem to develop the margin of error. For the standard normal distribution, the following is a true statement: P(21.96 , z , 1.96) 5 0.95, i.e., there is a 95% chance that a standard normal variable (z) will fall between 21.96 and 1.96. The Central
Limit Theorem states that for large X–μ samples, z 5 . If we make this substitution, the followσ n X–μ ing statement is true: P(21.96 , σ n , 1.96) 5 0.95. Using algebra, we can rework this inequality such that the mean (μ) is the middle term. The steps are outlined below: X–μ , 1.96) 5 0.95 σ n σ σ P(21.96 ,X w2m , 1.96) 5 0.95 n n σ σ P(2X w 21.96 ,2m
,2X w 1 1.96) 5 0.95 n n σ σ P(X w21.96 ,m , X w 1 1.96) 5 0.95 n n P(21.96 , The 95% CI for the population mean is the interval in the last probability statement and is given by: X w 6 1.96 σ n. The margin of error is 1.96 σ n , where 1.96 reflects the fact that a 95% confidence level is selected and σ n is the standard error (or the standard deviation of
the point estimate, X w). The general form of a CI can be rewritten as follows: In practice, we often do not know the value of the population standard deviation (s). If the sample size is large (n . 30), then the sample standard deviation (s) can be used to estimate the population standard deviation. Note that the prior derivation is based on the Central
Limit Theorem, which requires a large sample size. There are instances in which the sample size is not sufficiently large (e.g., n , 30), and therefore the general result of the Central Limit Theorem does not apply. In this case, we cannot use the standard normal distribution (z) in the confidence interval. Instead we use another probability d istribution,
called the t distribution, which is appropriate for small samples. The t distribution is another probability model for a continuous variable. The t distribution is similar to the s tandard normal distribution but takes a slightly different shape depending on the exact sample size. Specifically, the t values for CIs are larger for smaller samples, resulting in
larger margins of error (i.e., there is more imprecision with small samples). t values for CIs are contained in Table 2 in the Appendix. t values are indexed by degrees of freedom (df) in Table 2, which is defined as n 2 1. Table 6–2 is an excerpt of Table 2 showing the t values for small samples ranging in size from 5 to 10 (thus, the degrees of freedom
range from 4 to 9, as df 5 n 2 1). Specific guidelines for using the standard normal (z) or t distributions are provided in subsequent sections as we discuss the CI formulas for specific applications. It is important to note that appropriate use of the t distribution assumes that the outcome of interest is approximately normally distributed. Before providing
specific formulas, we first discuss the interpretation of CIs in general. Suppose we want to estimate a population mean using a 95% confidence level. If we take 100 different samples (in practice, we take only one) and for each sample we compute a 95% CI, in theory 95 out of the 100 CIs will contain the true mean value (m). This leaves 5 of 100 CIs
that will not include the true mean value. In practice, we select one random sample and generate one CI. This interval may or may point estimate 6 z SE (point estimate), where z is the value from the standard normal distribution reflecting the selected confidence level (e.g., for a 95% confidence level, z 5 1.96). Table 1B in the Appendix contains the z
values for popular confidence levels such as 90%, 95%, and 99%. In Table 1B, we find for 90%, z 5 1.645; for 95%, z 51.96; and for 99%, z 52.576. Higher confidence levels have larger z values, which translate to larger margins of error and wider CIs. For example, to be 99% confident that a CI contains the true unknown parameter, we need a wider
interval. In many applications, a confidence level of 95% is used. This is a generally accepted, but not prescribed, value. TABLE 6–2   t Values for Confidence Intervals Confidence Level df 80% 90% 95% 98% 99% 4 1.533 2.132 2.776 3.747 4.604 5 1.476 2.015 2.571 3.365 4.032 6 1.440 1.943 2.447 3.143 3.707 7 1.415 1.895 2.365 2.998 3.499 8
1.397 1.860 2.306 2.896 3.355 9 1.383 1.833 2.262 2.821 3.250 104 CHAPTER 6   Interval Confidence Confidence Estimates Interval Estimates not contain the true mean; the observed interval may overestimate m or underestimate m. The 95% CI is the likely range of the true, unknown parameter. It is important to note that a CI does not reflect the
variability in the unknown parameter but instead provides a range of values that are likely to include the unknown parameter. 6.2   CONFIDENCE INTERVALS FOR ONE SAMPLE, CONTINUOUS OUTCOME We wish to estimate the mean of a continuous outcome variable in a single population. For example, we wish to estimate the mean systolic
blood pressure, body mass index (BMI), total cholesterol level, or white blood cell count in a single population. We select a sample and compute descriptive statistics on the sample data using the techniques described in Chapter 4. Specifically, we compute the sample size (n), the sample mean (X w), and the sample standard deviation (s). The formulas
for CIs for the population mean depend on the sample size and are given in Table 6–3. Example 6.1. In Chapter 4, we presented data on n 5 3539 participants who attended the seventh examination of the offspring in the Framingham Heart Study. Descriptive statistics on variables measured in the sample are shown in Table 6–4 (these and other
statistics were presented in Table 4–20). The numbers of participants (n) who provided information on each characteristic are shown in the second column of Table 6–4. We wish to generate a 95% CI for systolic blood pressure using data collected in the Framingham Offspring Study. Because the sample size is large, we use the following formula, s . X
±z n The z value for 95% confidence is z 5 1.96. Substituting the sample statistics and the z value for 95% confidence, we have 127.3 ± 1.96 19.0 . 3534 TABLE 6–3   Confidence Intervals for m n $ 30 X ± z s n (Find z in Table 1B) n , 30 X ±t s n (Find t in Table 2, df 5 n 2 1) TABLE 6–4   Summary Statistics on Participants Attending the Seventh
Examination of the Framingham Offspring Study (n 5 3539) n Systolic blood pressure Diastolic blood pressure Total serum cholesterol Weight (lbs) Height (in.) Body mass index (BMI) Mean (w X) Standard Deviation (s) 3534 127.3 19.0 3532 74.0 9.9 3310 3506 3326 200.3 174.4 65.957 36.8 38.7 3.749 3326 28.15 5.32 Working through the
computations, we have 127.3 6 0.63. Adding and subtracting the margin of error, we get (126.7, 127.9). A point estimate for the true mean systolic blood pressure in the population is 127.3, and we are 95% confident that the true mean is between 126.7 and 127.9. The margin of error is very small here because of the large sample size. A 90% CI for
BMI is given below. Notice that z 5 1.645 to reflect the 90% confidence level: 28.15 ± 1.645 5.32 , 3326 28.15 6 0.152, (28.00, 28.30). We are 90% confident that the true mean BMI in the population is between 28.00 and 28.30. Again, the CI is very precise or narrow due to the large sample size. Example 6.2. In Chapter 4, we also presented data on a
subsample of n 5 10 participants who attended the seventh examination of the Framingham Offspring Study. Descriptive statistics on variables measured in the subsample are shown in Table 6–5 (these and other statistics were presented in Table 4–18). Suppose we compute a 95% CI for the true systolic blood pressure using data in the subsample.
Because the sample size is small, we must now use the CI formula that involves t rather than z, s X ±t . n Confidence Intervals for One Sample, Dichotomous Outcome TABLE 6–5   Summary Statistics on n 5 10 Participants Attending the Seventh Examination of the Framingham Offspring Study n Systolic blood pressure Mean (w X) Standard
Deviation (s) 10 121.2 11.1 Diastolic blood pressure 10 71.3 7.2 Total serum cholesterol 10 202.3 37.7 Weight (lbs) 10 176.0 33.0 Height (in.) 10 67.175 4.205 Body mass index (BMI) 10 27.26 3.10 We first need to determine the appropriate t value from Table 2. To do this, we need df 5 n – 1 5 10 – 1 5 9. The t value for 95% confidence with df 5 9 is t
5 2.262. Substituting the sample statistics and the t value for 95% confidence, we have 121.2 ± 2.262 11.1 . 10 Working through the computations, we have 121.2 6 7.94. Adding and subtracting the margin of error, we get (113.3, 129.1). Based on this sample of size n 5 10, our best estimate of the true mean systolic blood pressure in the population is
121.2. Based on this sample, we are 95% confident that the true mean systolic blood pressure in the population is between 113.3 and 129.1. Notice that the margin of error is larger here primarily due to the smaller sample size. Using the subsample, we now compute a 90% CI for the mean BMI. Because the sample size is small, we again need to
determine an appropriate value from the t distribution. For 90% confidence and df 5 9, t 5 1.833. 27.26 ± 1.833 3.10 , 10 105 6.3   CONFIDENCE INTERVALS FOR ONE SAMPLE, DICHOTOMOUS OUTCOME There are many applications where the outcome of interest is dichotomous. The parameter of interest is the unknown population proportion,
denoted p. For example, suppose we wish to estimate the proportion of people with diabetes in a population, or the proportion of people with hypertension or obesity. The latter are defined by specific levels of blood pressure and BMI, respectively. Again, we select a sample and compute descriptive statistics on the sample data using the techniques
described in Chapter 4. When the outcome of interest is dichotomous, we record on each member of the sample whether they have the characteristic of interest or not. The sample size is denoted by n and we let x denote the number of successes in the sample. Recall that for dichotomous outcomes, we define one of the outcomes a success and the
other a failure. The specific response that is considered a s uccess is defined by the investigator. For example, if we wish to estimate the proportion of people with diabetes in a population, we consider a diagnosis of diabetes (the outcome of interest) a success and lack of diagnosis a failure. In this example, x represents the number of people with a
diagnosis of diabetes in the sample. The sample proportion is denoted ˆp, and is computed by taking the ratio of the number of successes in the sample to the sample size, ˆp 5 }nx}. The formula for the CI for the population proportion is given in Table 6–6. The CI for the population proportion takes the same form as the CI for the population mean
(i.e., point estimate 6 margin of error). The point estimate for the population proportion is the sample proportion. The margin of error is the product of the z value for the desired confidence level (e.g., z 5 1.96 for 95% confidence) and the standard error of the point estimate, SE(ˆp) 5 p ˆ (1− p ˆ) . n The preceding formula is appropriate for large
samples, defined as at least five successes (nˆp) and at least five failures TABLE 6–6   Confidence Interval for p 27.26 6 1.80, (25.46, 29.06). We are 90% confident that the true mean BMI in the population is between 25.46 and 29.06. Again, because of the small sample size, the CI is less precise. min[nˆp, n(1 2 ˆp)] ≥ 5 ˆp 6 z p ˆ (1 − p ˆ) n (Find z in
Table 1B) 106 CHAPTER 6   Confidence Interval Estimates [n(12ˆp)] in the sample. If there are fewer than five successes or failures, then alternative procedures—called exact methods— must be used to estimate the population proportion.1, 2 Example 6.3. In Chapter 4, we presented data on n 5 3539 participants who attended the seventh
examination of the Offspring in the Framingham Heart Study. One particular characteristic measured was treatment with antihypertensive medication. There were a total of 1219 participants on treatment and 2313 participants not on treatment (Table 4–3). If we call treatment a success, then x 5 1219 and n 5 3532. The sample proportion is x 1219
ˆp 5 }} 5 }} 5 0.345. n 3532 Thus, a point estimate for the population proportion is 0.345, or 34.5%. Our best estimate of the proportion of participants in the population on treatment for hypertension is 0.345. Suppose we now wish to generate a 95% CI. To use the preceding formula, we need to satisfy the sample size criterion—specifically, we need
at least five successes and five failures. Here we more than satisfy that requirement, so the CI formula in Table 6–6 can be used. ˆp 6 z TABLE 6–7   Prevalent CVD in Men and Women Free of CVD Prevalent CVD Men 1548 244 1792 Women 1872 135 2007 Total 3420 379 3799 women is 135 / 2007 5 0.0673. These are point estimates. Following, we
generate 95% CI estimates for prevalence in the total population and in the populations of men and women. To use the preceding formula, we need to satisfy the sample size criterion—specifically, we need at least five successes and five failures in each sample. In this example, a success is prevalent CVD and a failure is freedom from CVD. Here we
more than satisfy that requirement for men, women, and the pooled or total sample. For the total sample, 0.0998 ± 1.96 pˆ (1− pˆ) , n 0.345(1 − 0.345) 0.345 6 1.96 , 3532 0.345 6 0.016, (0.329, 0.361). Thus, we are 95% confident that the true proportion of persons on antihypertensive medication is between 0.329 and 0.361, or between 32.9% and
36.1%. Specific applications of estimation for a single population with a dichotomous outcome involve estimating prevalence and cumulative incidence. In Chapter 3, we generated point estimates for prevalence and incidence data; in the following examples, we add CI estimates. Example 6.4. In Example 3.1, we presented data on n 5 3799
participants who attended the fifth examination of the offspring in the Framingham Heart Study. Table 6–7 shows the numbers of men and women with diagnosed, or prevalent, cardiovascular disease (CVD) at the fifth examination. The prevalence of CVD for all participants attending the fifth examination of the Framingham Offspring Study is 379 /
3799 5 0.0998. The prevalence of CVD among men is 244 / 1792 5 0.1362, and the prevalence of CVD among Total 0.0998(1 − 0.0998) , 3799 0.0998 6 0.0095, (0.090, 0.109). We are 95% confident that the true prevalence of CVD in the population is between 9% and 10.9%. The CI estimates for men and women are as follows. For men: 0.1362 ± 1.96
0.1362(1 − 0.1362) , 1792 0.1362 6 0.016, (0.120, 0.152). For women: 0.0673 ± 1.96 0.0673(1 − 0.0673) , 2007 0.0673 6 0.011, (0.056, 0.078). We are 95% confident that the true prevalence of CVD in men is between 12% and 15.2%, and in women, between 5.6% and 7.8%. Each of these CI estimates is very precise (small margins of error) because
of the large sample sizes. Confidence Intervals for Two Independent Samples, Continuous Outcome 6.4   CONFIDENCE INTERVALS FOR TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME There are many applications where it is of interest to compare two groups with respect to their mean scores on a continuous outcome. For example, the
comparison groups might be men versus women, patients assigned to an experimental treatment versus a placebo in a clinical trial, or patients with a history of cardiovascular disease versus patients free of cardiovascular disease. We can compare mean systolic blood pressures in men versus women, or mean BMI or total cholesterol levels in patients
a ssigned to experimental treatment versus placebo. A key feature here is that the two comparison groups are independent, or physically separate. The two groups might be determined by a particular attribute or characteristic (e.g., sex, diagnosis of cardiovascular disease) or might be set up by the investigator (e.g., participants assigned to receive
an experimental drug or placebo). Similar to the approach we used to estimate the mean of a continuous variable in a single population (Section 6.2), we first compute descriptive statistics on each of the two samples using the techniques described in Chapter 4. Specifically, we compute the sample size, mean, and standard deviation in each sample.
We denote these summary statistics as n1, X w1, and s1 for Sample 1 and n2, X w2, and s2 for Sample 2. (The designation of Sample 1 and Sample 2 is essentially arbitrary. In a clinical trial setting, the convention is to call the experimental treatment Group 1 and the control treatment Group 2. However, when comparing groups such as men and
women, either group can be 1 or 2.) The interpretation of the CI estimate depends on how the samples are assigned, but this is just for interpretation and does not affect the computations. This is illustrated in the following examples. In the two-independent-samples application with a con tinuous outcome, the parameter of interest is the difference in
population means, m1 2 m2. The point estimate for the difference in population means is the difference in sample means, 107 X w12X   w2. The form of the CI is again point estimate 6 margin of error, and the margin of error incorporates a value from either the z or t distribution reflecting the selected confidence level and the standard error of the
point estimate. The use of z or t depends on whether the sample sizes are large or small. (The specific guidelines follow.) The standard error of the point estimate incorporates the variability in the outcome of interest in each of the comparison groups. Table 6–8 contains the formulas for CIs for the difference in population means. The formulas in Table
6–8 assume equal variability in the two populations (i.e., that the population variances are equal, or s 12 5 s 22). This means that the outcome is equally variable in each of the comparison populations. For analysis, we have samples from each of the comparison populations. If the sample variances are similar, then the assumption about variability in
the populations is reasonable. As a guideline, if the ratio of the sample variances, s 12/s 22, is between 0.5 and 2 (i.e., if one variance is no more than double the other), then the formulas in Table 6–8 are appropriate. If the ratio of the sample variances is greater than 2 or less than 0.5, then a lternative formulas must be used to account for the
heterogeneity in variances. Specific details can be found in D’Agostino, Sullivan, and Beiser and in Rosner.3,4 In the formulas in Table 6–8, w X1 and w X2 are the means of the outcome in the independent samples, z or t are values from the z or t distributions reflecting the desired confidence level, 1 1 + and Sp is the standard error of the point
estimate, n1 n2 SE(X w12 w X2). Here, Sp is the pooled estimate of the common standard deviation (again, assuming that the variances in the populations are similar) computed as the weighted average of the standard deviations in the samples: Sp = (n1 − 1)s12 + (n2 − 1)s22 . n1 + n2 − 2 TABLE 6–8   Confidence Intervals for (m12m2) n1 $ 30 and
n2 $ 30 (X w12X w2) 6 zSp 1 1 + n1 n2 (Find z in Table 1B) n1 , 30 or n2 , 30 (X w12X w2) 6 tSp 1 1 + n1 n2 (Find t in Table 2, df 5 n1 1 n2 2 2) 108 CHAPTER 6   Confidence Interval Estimates Because we are assuming equal variances in the groups, we pool the information on variability (sample variances) to generate an estimate of the population
variability. Example 6.5. In Example 6.1, we recapped the data presented in Chapter 4 on n 5 3539 participants who attended the seventh examination of the offspring in the Framingham Heart Study. Table 6–9 contains descriptive statistics on the same continuous characteristics, stratified (or generated separately) by sex. Suppose we want to
compare mean systolic blood pressures in men versus women using a 95% CI. The summary statistics shown in Table 6–9 indicate that we have large samples (more than 30) of both men and women, and therefore we use the CI formula from Table 6–8 with z as opposed to t. However, before implementing the formula, we first check whether the
assumption of equality of the population variances is reasonable. The guideline suggests investigating the ratio of the sample variances, s 12/s 22. Suppose we call men Group 1 and women Group 2. (Again, this is arbitrary. It only needs to be noted when interpreting the results.) The ratio of the sample variances is 17.52 / 20.12 5 0.76, which falls
between 0.5 and 2, suggesting that the assumption of equality of the population variances is reasonable. The appropriate CI formula for the difference in mean systolic blood pressures between men and women is: 1 1 (X w12 X w2) 6 zSp n + n 1 2 Before substituting, we first compute Sp, the pooled estimate of the common standard deviation: Sp =
(n1 − 1)s12 + (n2 − 1)s22 , n1 + n2 − 2 Sp = (1623 − 1)(17.5)2 + (1911 − 1)(20.1)2 , 1623 + 1911 − 2 Sp = 359.12 = 19.0. Notice that the pooled estimate of the common standard deviation, Sp, falls between the standard deviations in the comparison groups (i.e., 17.5 and 20.1). Sp is slightly closer in value to the standard deviation in the women
(20.1) as there are slightly more women in the sample. Recall that Sp is a weighted average of the standard deviations in the comparison groups, weighted by the respective sample sizes. The 95% CI for the difference in mean systolic blood pressures is: (128.2 − 126.5) ± 1.96(19.0) 1 1 + , 1623 1911 1.7 6 1.26, (0.44, 2.96). The CI is interpreted as
follows: We are 95% confident that the difference in mean systolic blood pressures between men and women is between 0.44 and 2.96 units. Our best estimate of the difference, the point estimate, is 1.7 units. The standard error of the difference is 0.641, and the margin of error is 1.26 units. Note that when we generate estimates for a population
parameter in a single sample (e.g., the mean m or population proportion p), the resulting CI provides a range of likely values for that parameter. In contrast, when there are two independent samples and the goal is to compare means (or proportions, discussed in Section 6.6), the resultant CI does not provide a range of values for the parameters in the
TABLE 6–9   Summary Statistics on Men and Women Attending the Seventh Examination of the Framingham Offspring Study Men Women n w X s n w X s Systolic blood pressure 1623 128.2 17.5 1911 126.5 20.1 Diastolic blood pressure 1622 75.6 9.8 1910 72.6 9.7 Total serum cholesterol 1544 192.4 35.2 1766 207.1 36.7 Weight (lbs) 1612 194.0
33.8 1894 157.7 34.6 Height (in.) 1545 68.9 2.7 1781 63.4 2.5 Body mass index (BMI) 1545 28.8 4.6 1781 27.6 5.9 Confidence Intervals for Two Independent Samples, Continuous Outcome comparison populations; instead, the CI provides a range of values for the difference. In this example, we estimate that the difference in mean systolic blood
pressures is between 0.44 and 2.96 units, with men having the higher values. The last aspect of the interpretation is based on the fact that the CI is positive and that we called men Group 1 and women Group 2. Had we designated the groups the other way (i.e., women as Group 1 and men as Group 2), the CI would have been 22.96 to 20.44,
suggesting that women have lower systolic blood pressures (anywhere from 0.44 to 2.96 units lower than men). Table 6–10 includes 95% CIs for the difference in means in each characteristic, computed using the same formula we used for the CI for the difference in mean systolic blood pressures. Notice that the 95% CI for the difference in mean total
cholesterol levels between men and women is 217.16 to 212.24. Men have lower mean total cholesterol levels than women, anywhere from 12.24 to 17.16 units lower. The men have higher mean values on each of the other characteristics considered (indicated by the positive CIs). Again, the CI for the difference in means provides an estimate of the
absolute difference in means of the outcome variable of interest between the comparison groups. It is often of interest to make a judgment as to whether there is a statistically meaningful difference between comparison groups. This judgment is based on whether the observed difference is beyond that expected by chance. The CI for the difference in
means provides a range of likely values for (m12 m2). It is important to note that all values in the CI are equally likely estimates of the true value of (m12 m2). If there is no difference between the population means, then the difference is zero (i.e., m12 m2 5 0). Zero is the no difference or null value of the parameter (in this case, no difference in
means). If a 95% CI for the difference in 109 means includes 0 (the null value), then we say that there is no statistically meaningful or statistically significant difference in the means. If the CI does not include 0, then we conclude that there is a statistically significant difference in means. (We discuss statistical significance in much greater detail in
Chapter 7.) For each of the characteristics considered in Example 6.5, there is a statistically significant difference in means between men and women because none of the CIs include the null value, 0. However, notice that some of the means are not very different between men and women (e.g., systolic blood pressure and BMI). Because the 95% CIs
do not include 0, we conclude that there are s tatistically meaningful differences between means. This is primarily due to the large sample sizes. We discuss this in further detail in Chapter 7. Example 6.6. In Example 6.2, we recapped the data presented in Chapter 4 on the subsample of n 5 10 participants who attended the seventh examination of the
offspring in the Framingham Heart Study. Table 6–11 contains descriptive statistics on the same continuous characteristics in the subsample, stratified (or generated separately) by sex. Using these data, suppose we wish to construct a 95% CI for the difference in mean systolic blood pressures between men and women. Again, we call men Group 1
and women Group 2. Because the sample sizes are small (i.e., n1 , 30 and n2 , 30), the CI formula with t is appropriate. However, before implementing the formula we first check whether the assumption of equality of the population variances is reasonable. The ratio of the sample variances is 9.72 / 12.02 5 0.65, which falls between 0.5 and 2,
suggesting that the assumption of equality of the population variances is reasonable. The appropriate CI formula for comparing mean systolic blood pressures between men and women is: (X w12X w2) 6 tSp 1 1 + . n1 n2 TABLE 6–10   Confidence Intervals for Differences in Means Men Women 95% CI w X s w X s Systolic blood pressure 128.2 17.5
126.5 20.1 (0.44, 2.96) Diastolic blood pressure 75.6 9.8 72.6 9.7 (2.38, 3.67) Total serum cholesterol 192.4 35.2 207.1 36.7 (217.16, 212.24) Weight (lbs) 194.0 33.8 157.7 34.6 (33.98, 38.53) Height (in.) 68.9 2.7 63.4 2.5 (5.31, 5.66) Body mass index (BMI) 28.8 4.6 27.6 5.9 (0.82, 1.56) 110 CHAPTER 6   Confidence Interval Estimates TABLE 6–11
Summary Statistics on the Subsample of n 5 10 Men and Women Attending the Seventh Examination of the Framingham Offspring Study Men n Systolic blood pressure 6 Diastolic blood pressure 6 Total serum cholesterol 6 Weight (lbs) 6 Height (in.) Body mass index (BMI) w X n w X s 117.5 9.7 4 126.8 12.0 72.5 7.1 4 69.5 7.9 193.8 34.2 4 215.0 42.8
196.9 26.9 4 146.0 27.2 6 70.2 2.7 4 62.6 2.3 6 28.0 3.6 4 26.2 2.9 Before substituting, we first compute Sp, the pooled estimate of the common standard deviation: Sp = (6 − 1)(9.7)2 + (4 − 1)(12.0)2 6+4−2 5 112.81 5 10.6. Notice that again the pooled estimate of the common standard deviation, Sp, falls between the standard deviations in the
comparison groups (i.e., 9.7 and 12). We next need to find the appropriate t value for 95% confidence. The degrees of freedom for the two sample procedures are defined as df 5 n1 1 n2 22 5 6 1 4 22 5 8. From Table 2, t 5 2.306. The 95% CI for the difference in mean systolic blood pressures is: (117.52126.8) 6 2.306(10.6) 1 6 Women s + 1 , 4 29.3 6
15.78, (225.08, 6.48). We are 95% confident that the difference in mean systolic blood pressures between men and women is between 225.08 and 6.48 units. Our best estimate of the difference, the point estimate, is 29.3 units. The standard error of the difference is 6.84 units, and the margin of error is 15.78 units. In this sample, the men have lower
mean systolic blood pressures than women by 9.3 units. However, based on this interval, we cannot conclude that there is a statistically significant difference in mean systolic blood pressures between men and women because the 95% CI includes the null value, 0. Again, the CI is a range of likely values for the difference in means. Because the
interval contains some positive values (suggesting men have higher mean blood pressures) and some negative values (suggesting that women have higher values) and therefore also 0 (no difference), we cannot conclude that there is a statistically significant difference. This 95% CI for the difference in mean blood pressures is much wider than the one
based on the full sample derived in Example 6.5. The small sample size produces a very imprecise estimate of the difference in mean systolic blood pressures. 6.5   CONFIDENCE INTERVALS FOR MATCHED SAMPLES, CONTINUOUS OUTCOME There is a study design alternative to that described in the previous section, again where it is of interest
to compare two groups with respect to their mean scores on a continuous outcome. Here the two comparison groups are dependent (or matched, or paired). One such possible scenario involves a single sample of participants and each participant is measured twice, possibly before and after an intervention or under two experimental conditions (e.g., in
a crossover trial). The goal of the analysis is to compare the mean score measured before the intervention with the mean score measured afterward. Another s cenario is one in which matched samples are analyzed. For example, we might be interested in the difference in an outcome between twins or siblings. Again, we have two samples and the goal
is to compare the two means; however, the samples are related or dependent. In the first scenario, before and after measurements are taken in the same individual. In the second scenario, measures are taken in pairs of individuals from the same family. When the samples are dependent, we cannot use the techniques described in Section 6.4 to
compare means. Because the samples are dependent, statistical techniques that account for the dependency must be used. The technique here focuses on difference scores Confidence Intervals for Matched Samples, Continuous Outcome (e.g., the difference between measures taken before versus after the intervention, or the difference between

measures taken in twins or siblings). In estimation and other statistical inference applications, it is critically important to appropriately identify the unit of analysis. Units of analysis are independent entities. In onesample and two-independent-samples applications, participants are the units of analysis. In the two-dependent-samples application, the
pair is the unit and not the number of mea surements, which is twice the number of pairs or units. The parameter of interest is the mean difference, md. Again, the first step is to compute descriptive statistics on the sample data. Specifically, descriptive statistics are computed on difference scores. We compute the sample size (which, in this case, is
the number of distinct participants or distinct pairs) and the mean and standard deviation of the difference scores. We denote these summary statistics as n, X wd, and sd, respectively. The appropriate formula for the CI for the mean difference depends on the sample size. The formulas are shown in Table 6–12 and are identical to those we presented
for estimating the mean of a single sample (see Table 6–3), except here we focus on difference scores. Here n is the number of participants or pairs, X wd and sd are the mean and standard deviation of the difference scores (where differences are computed on each participant or between members of a matched pair), and z or t are the values from the
z or t distributions reflecting the desired confidence level. sd n is the standard error of the point estimate, SE(X wd). Example 6.7. In the Framingham Offspring Study, participants attend clinical examinations approximately every four years. The data in Table 6–13 are systolic blood pressures measured at the sixth and seventh examinations in a
subsample of n 5 15 randomly selected participants. The first column c ontains unique identification numbers, assigned only to distinguish individual participants. Suppose we want to compare systolic blood pressures between examinations (i.e., changes over four years). Because the data in the two samples (Examination 6 and TABLE 6–12
Confidence Intervals for md n $ 30 w Xd 6 z sd (Find z in Table 1B) n n , 30 w Xd 6 t sd (Find t in Table 2, df 5 n21) n 111 TABLE 6–13 Systolic Blood Pressures Measured at the Sixth and Seventh Examinations of the Framingham Offspring Study Subject Identification Number 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 Examination 6 168 111 139 127 155
115 125 123 130 137 130 129 112 141 122 Examination 7 141 119 122 127 125 123 113 106 131 142 131 135 119 130 121 Examination 7) are matched, we compute difference scores. The difference scores can be computed by subtracting the blood pressure measured at Examination 7 from that measured at Examination 6, or vice versa. If we
subtract the blood pressure measured at Examination 6 from that measured at Examination 7, then positive differences represent increases over time and negative differences represent decreases over time. Table 6–14 contains the difference scores for each participant. Notice that several participants' systolic blood pressures de creased over four
years (e.g., Participant 1's blood pressure decreased by 27 units from 168 to 141), whereas others increased (e.g., Participant 2's blood pressure increased by 8 units from 111 to 119). We now estimate the mean difference in blood pressures over four years. This is similar to a one-sample problem with a continuous outcome (Section 6.2), except here
we focus on the difference scores. In this sample, we have n 5 15, X wd 5 25.3, and sd 5 12.8, respectively. The data for the calcu−79.0 lations are shown in Table 6–15, and Xd = = − 5.3 and 15 2296.95 = 164.07 = 12.8. sd = (15 − 1) We now use these descriptive statistics to compute a 95% CI for the mean difference in systolic blood pressures in
the 112 CHAPTER 6 Confidence Interval Estimates TABLE 6–14 Difference Scores Subject Identification Number Examination 6 Examination 7 Difference 1 168 141 227 2 111 119 8 3 139 122 217 4 127 127 0 5 155 125 230 6 115 123 8 7 125 113 212 8 123 106 217 9 130 131 1 10 137 142 5 11 130 131 1 12 129 135 6 13 112 119 7 14 141 130
211 15 122 121 21 TABLE 6–15 Summary Statistics on Difference Scores Subject Identification Number Difference Difference − w Xd (Difference − w Xd)2 1 227 221.7 470.89 2 8 13.3 176.89 3 217 211.7 136.89 4 0 5.3 28.09 5 230 224.7 610.09 6 8 13.3 176.89 7 212 26.7 44.89 8 217 211.7 136.89 9 1 6.3 39.69 10 5 10.3 106.09 11 1 6.3
39.69 12 6 11.3 127.69 13 7 12.3 151.29 14 211 25.7 32.49 15 21 4.3 279.0 0.5* *The sum of the differences from the mean should sum to 0 but here sum to 0.5 due to rounding. 18.49 2296.95 Confidence Intervals for Matched Samples, Continuous Outcome population. Because the sample size is small (n 5 15), we use the following formula: Xd 6 t sd
. n We first need the t value for 95% confidence. The degrees of freedom are df 5 n 2 1 5 14. From Table 2, t 5 2.145. We now substitute the descriptive statistics on the difference scores and the t value for 95% confidence: −5.3 ± 2.145 12.8 , 15 are measured in each patient. The d ifference in depressive symptoms is measured in each p atient by
subtracting the depressive symptom score after taking the placebo from the depressive symptom score after taking the new drug. A total of 100 participants completed the trial and the data are summarized in Table 6–16. The mean difference in the sample is X wd 5 212.7, meaning on average patients scored 12.7 points lower on the depressive
symptoms scale after taking the new drug as compared to placebo (i.e., improved by 12.7 points on average). We now construct a 95% CI for the mean difference in the population. Because the sample size is large, we use the following formula: 25.3 1 7.1, Xd ± z (212.4, 1.8). We are 95% confident that the mean difference in systolic blood pressures
between Examination 6 and Examination 7 (approximately four years apart) is between 212.4 and 1.8. The null (or no effect) value of the CI for the mean difference is 0. Therefore, based on the 95% CI we cannot conclude that there is a statistically significant difference in blood pressures over time because the CI for the mean difference includes 0.
In Chapter 2, we discussed various study designs, particularly a special case of the randomized trial called the crossover trial (Section 2.3.2). In a crossover trial with two treatments (e.g., an experimental treatment and a control), each participant receives both treatments. Outcomes are measured after each treatment in each participant (see Figure
2.6). A major advantage to the crossover trial is that each participant acts as his or her own control, and therefore fewer participants are required to demonstrate an effect. When the outcome is continuous, the assessment of a treatment effect in a crossover trial is performed using the techniques described here. Example 6.8. A crossover trial is
conducted to evaluate the effectiveness of a new drug designed to reduce the symptoms of depression in adults over 65 years of age following a stroke. Symptoms of depression are measured on a scale of 0 to 100, with higher scores indicative of more frequent and s evere symptoms of depression. Patients who suffered a stroke are eligible for the
trial. The trial is run as a crossover trial where each patient receives both the new drug and a placebo. Patients are blind to the treatment assignment, and the order of treatments (e.g., placebo and then new drug or new drug and then placebo) are randomly assigned. After each treatment, depressive symptoms 113 sd , n −12.7 ± 1.96 8.9 , 100 212.7
6 1.74, (214.4,210.7). We are 95% confident that the mean improvement in depressive symptoms after taking the new drug as compared to placebo is between 10.7 and 14.4 units (or alternatively, the depressive symptoms scores are 10.7 to 14.4 units lower after taking the new drug as compared to the placebo). Because we computed the differences
by subtracting the scores after taking the placebo from the scores after taking the new drug, and because higher scores are indicative of worse or more severe depressive symptoms, negative differences reflect improvement (i.e., lower depressive symptoms scores after taking the new drug as compared to the placebo). Because the 95% CI TABLE 6–
16 Summary Statistics on Differences in Depressive Symptoms n Depressive symptoms after taking the new drug – Depressive symptoms after taking placebo 100 Mean Difference 212.7 Std Dev Difference 8.9 114 CHAPTER 6 Confidence Interval Estimates for the mean difference does not include 0, we can conclude that there is a statistically
significant difference (in this case, a significant improvement) in depressive symptom scores after taking the new drug as compared to the placebo. 6.6 CONFIDENCE INTERVALS FOR TWO INDEPENDENT SAMPLES, DICHOTOMOUS OUTCOME It is very common to compare two groups in terms of the presence or absence of a particular
characteristic or attribute. There are many instances in which the outcome variable is dichotomous (e.g., prevalent cardiovascular disease or diabetes, current smoking status, incident coronary heart diesease, cancer remission, successful device implant). Similar to the applications described in Section 6.4 for c ontinuous outcomes, we focus here on
the case where there are two comparison groups that are independent or physically separate and the outcome is dichotomous. The two groups might be determined by a particular attribute or characte ristic of the participant (e.g., sex, age less than 65 v ersus age 65 and older) or might be set up by the investigator (e.g., participants assigned to
receive an experimental drug or a placebo, a pharmacological versus a surgical treatment). When the outcome is dichotomous, the analysis involves comparing the proportions of successes between the two groups. We discussed several methods that are used to compare proportions in two independent groups in Chapter 3, including the risk
difference, which is computed by taking the difference in proportions between comparison groups and is similar to the estimate of the difference in means for a continuous outcome described in Section 6.4. The relative risk, also called the risk ratio, is another useful measure to compare proportions between two independent populations, and it is
computed by taking the ratio of proportions. Generally, the reference group (e.g., unexposed persons, persons without a risk factor, or persons assigned to the control group in a clinical trial setting) is considered in the denominator of the ratio. The relative risk is often felt to be a better measure of the strength of an effect than the risk difference
because it is a relative as opposed to an absolute measure. A third popular measure is the odds ratio. The odds ratio is computed by taking the ratio of the odds of success in the comparison groups. The odds is defined as the ratio of the number of successes to the number of failures (see Section 3.4.2 for more details). CI estimates for the risk
difference, the relative risk, and the odds ratio are described in the following. 6.6.1 Confidence Intervals for the Risk Difference The risk difference (RD) is similar to the difference in means when the outcome is continuous. The parameter of interest is the difference in proportions in the population, RD 5 p12 p2. The point estimate is the difference
in sample proportions, ˆ 5 ˆp 2 ˆp . The sample proportions are computed by taking RD 1 2 the ratio of the number of successes (x) to the sample size (n) in each group, ˆp1 5 x1 / n1 and ˆp2 5 x2 / n2, respectively. The formula for the CI for the difference in proportions, or the RD, is given in Table 6–17. The formula in Table 6–17 is appropriate for
large samples, defined as at least five successes (nˆp) and at least five failures [n(1 2 ˆp)] in each sample. If there are fewer than five successes or five failures in either comparison group, then alternative procedures called exact methods must be used to estimate the difference in population proportions.5 Example 6.9. In Example 3.1, we presented
data measured in participants who attended the fifth examination of the offspring in the Framingham Heart Study. A total of n 5 3799 participants attended the fifth examination, and Table 6–18 contains data on prevalent CVD among participants who were and were not currently smoking cigarettes at the time of the fifth examination in the
Framingham Offspring Study. The outcome is prevalent CVD and the two comparison groups are defined by current smoking status. The point estimate of prevalent CVD among nonsmokers is 298 / 3055 5 0.0975, and the point estimate of prevalent CVD among current smokers is 81 / 744 5 0.1089. When constructing CIs for the RD, the convention is
to call the exposed or treated Group 1 and the unexposed or untreated Group 2. Here smoking status TABLE 6–17 Confidence Interval for (p1 2 p2) min[n1ˆp1, n1(1 2 ˆp1), n2 ˆp2, n2(1 2 ˆp2)] $ 5 pˆ1 − pˆ 2 6 z pˆ1 (1 − pˆ1) pˆ (1 − pˆ 2) + 2 n1 n2 (Find z in Table 1B) Confidence Intervals for Two Independent Samples, Dichotomous Outcome
TABLE 6–18 Prevalent CVD in Smokers and Table 6–19 Pain Reduction by Treatment Nonsmokers Free of CVD Nonsmoker Current smoker Total History of CVD Total 2757 298 3055 663 81 744 3420 379 3799 defines the comparison groups, and we will call the current smokers Group 1 and the nonsmokers Group 2. A CI for the difference in
prevalent CVD (or RD) between smokers and nonsmokers is given below. In this example, we have more than enough successes (cases of prevalent CVD) and failures (persons free of CVD) in each comparison group. pˆ (1 − pˆ1) pˆ2 (1 − pˆ2) pˆ1 − pˆ2 6 z 1 + , n1 n2 (0.1089 − 0.0975) 1.96 115 0.1089(1 − 0.1089) 0.0975(1 − 0.0975) + , 744 3055
0.0114 0.0247, (−0.0133, 0.0361). We are 95% confident that the difference in proportions of smokers as compared to nonsmokers with prevalent CVD is between 20.0133 and 0.0361. The null, or no difference, value for the RD is 0. Because the 95% CI includes 0, we cannot conclude that there is a statistically significant difference in prevalent CVD
between smokers and nonsmokers. Example 6.10. A randomized trial is conducted to evaluate the effectiveness of a newly developed pain reliever designed to reduce pain in patients following joint replacement surgery. The trial compares the new pain reliever to the pain reliever currently used (called the standard of care). A total of 100 patients
undergoing joint replacement surgery agree to participate in the trial. Patients are randomly assigned to receive either the new pain reliever or the standard pain reliever following surgery. The patients are blind to the treatment assignment. Before receiving the assigned treatment, patients are asked to rate their pain on a scale of 0 to 10, with
higher scores indicative of more pain. Each patient is then given the assigned treatment and after 30 minutes is again asked to rate his or her pain on the same scale. The primary outcome is a reduction in pain of Treatment Group n Reduction of 31 Points Number Proportion New pain reliever 50 23 0.46 Standard pain reliever 50 11 0.22 3 or more
scale points (defined by clinicians as a clinically meaningful reduction). The data shown in Table 6–19 are observed in the trial. A point estimate for the difference in proportions of patients reporting a clinically meaningful reduction in pain between treatment groups is 0.46 2 0.22 5 0.24. (Notice that we call the experimental or new treatment Group
1 and the standard Group 2.) There is a 24 percentage point increase in patients reporting a meaningful reduction in pain with the new pain reliever as compared to the standard pain reliever. We now construct a 95% CI for the difference in proportions. The sample sizes in each comparison group are adequate—i.e., each treatment group has at least
five successes (patients reporting reduction in pain) and at least five failures—therefore the formula for the CI is (0.46 − 0.22) ± 1.96 0.46(1 − 0.46) + 50 0.22(1 − 0.22) , 50 0.24 6 0.180, (0.06, 0.42). We are 95% confident that the difference in proportions of patients reporting a meaningful reduction in pain is between 0.06 and 0.42 comparing the
new and standard pain relievers. Our best estimate is an increase of 24 percentage points with the new pain reliever. Because the 95% CI does not contain 0 (the null value), we can conclude that there is a statistically significant difference between pain relievers—in this case, in favor of the new pain reliever. 6.6.2 Confidence Intervals for the
Relative Risk The RD quantifies the absolute difference in risk (or proportions), whereas the relative risk is, as the name indicates, a relative measure. Both the RD and the RR are useful but they convey different information. A relative risk (RR) is computed 116 CHAPTER 6 Confidence Interval Estimates by taking the ratio of proportions, p1/p2.
Again, the convention is to place the proportion of successes or the risk for the control or unexposed group in the denominator. The parameter of interest is the relative risk in the population, RR 5 p1/p2, and the point estimate is the relative risk in the sample, RˆR 5 ˆp1/ˆp2. The RR is a ratio and does not follow a normal distribution, regardless of the
sample sizes in the comparison groups. However, the natural log (ln) of the RR is approximately normally distributed and is used to produce a CI for the RR. First, a CI is generated for ln(RR), and then the antilog of the upper and lower limits of the CI for ln(RR) are taken to give the upper and lower limits of the CI for the RR. The two steps are
detailed in Table 6–20. The null or no difference value of the CI for the RR is 1. If a 95% CI for the RR does not include 1, then the risks (or proportions) are said to be statistically significantly different. Example 6.11. Consider again the data shown in Table 6–18, reflecting the prevalence of CVD among participants who were and were not currently
smoking cigarettes at the time of the fifth examination in the Framingham Offspring Study. The outcome is prevalent CVD and the two comparison groups are defined by current smoking status. In Example 6.9, we generated a point estimate for the difference in prevalence of CVD between smokers and nonsmokers of 0.0114. Smokers had a higher
prevalence by 1.14%. We then generated a 95% CI for the difference in prevalent CVD between smokers and nonsmokers of (20.0133, 0.0361). Because the 95% CI included 0, we could not conclude that there was a statistically significant difference in prevalent CVD between smokers and nonsmokers. We now use these data to generate a point
estimate and 95% CI estimate for the RR. The point estimate for the relative risk is pˆ 0.1089 RˆR 5 1 = = 1.12 pˆ2 0.0975 and the interpretation is as follows: Smokers have 1.12 times the prevalence of CVD compared to nonsmokers. The 95% CI estimate for the RR is computed using the two-step procedure outlined in Table 6–20: (n1 − x1)/ x1 (n −
x 2)/x 2 ln(RˆR) z + 2 , n1 n2 ln(1.12) 1.96 2757/298 663/81 , + 744 3055 0.113 6 0.232, (20.119, 0.345). A 95% CI for ln(RR) is (20.119, 0.345). To generate the CI for the RR, we take the antilog (exp) of the lower and upper limits: (exp(20.119), exp(0.345)), (0.89, 1.41). Thus, we are 95% confident that the RR of prevalent CVD in smokers as
compared to nonsmokers is between 0.89 and 1.41. The null, or no difference, value for a relative risk is 1. Because this CI includes the null value, we cannot conclude that there is a statistically significant difference in prevalent CVD between smokers and nonsmokers. Example 6.12. Consider again the data shown in Table 6–19 from the randomized
trial assessing the effectiveness of a newly developed pain reliever as compared to the standard pain reliever. The primary outcome is reduction in pain of 3 or more scale points (defined by clinicians as a clinically meaningful reduction). We generated a point estimate for the difference in proportions of patients reporting a clinically meaningful
reduction in pain between pain relievers as 0.46 2 0.22 5 0.24, and the 95% CI was (0.06, 0.42). Because the 95% CI for the RD did not contain 0 (the null value), we concluded that there was a statistically significant difference between pain relievers. TABLE 6–20 Confidence Interval for RR 5 p1/p2 CI for ln(RR) CI for RR (n − x1)/ x1 (n2 − x 2)/ x 2
(Find z in Table 1B) ln(RˆR) 6 z 1 + n1 n2 exp(Lower limit), exp (Upper limit) Confidence Intervals for Two Independent Samples, Dichotomous Outcome ratio is similar to the RR in that it is a ratio, as opposed to a difference. The odds ratio (OR) is computed by taking the ratio of odds, where the odds in each group is computed by dividing the
number of successes (x) by the number of failures (n 2 x) in that group. Similar to the RR, the convention is to place the odds in the control or unexposed group in the denominator. The parameter of interest is the odds ratio in the population, OR, and the point estimate is the odds ratio in the sample, OˆR. We now use these data to generate a point estimate
and 95% CI estimate for the RR. The point estimate for the relative risk is pˆ 0.46 RˆR 5 1 = = 2.09. ˆp2 0.22 Patients receiving the new pain reliever have 2.09 times the likelihood of reporting a meaningful reduction in pain as compared to patients receiving the standard pain reliever. The 95% CI estimate for the RR is computed using the twostep
procedure outlined in Table 6–20. ln(2.09) 6 1.96 pˆ1 / (1 − pˆ1) OˆR 5 ˆ . p2 / (1 − pˆ2) The odds ratio can also be expressed as 27/23 39/11 + , 50 50 OˆR 5 0.737 6 0.602, x1 / (n1 − x1) . x 2 / (n2 − x 2) Similar to the RR, the OR does not follow a normal distribution and so we use the log transformation to promote normality. We first generate a CI
for ln(OR) and then take the antilog of the upper and lower limits of the CI for ln(OR) to determine the upper and lower limits of the CI for the OR. The two steps are detailed in Table 6–21. The null, or no difference, value for the OR is 1. If a 95% CI for the OR does not include 1, then the odds are said to be statistically significantly different. We again
reconsider Example 6.9 and Example 6.10 and produce estimates of ORs and compare these to our estimates of RDs and RRs. Example 6.13. Consider again the data shown in Table 6–18 reflecting the prevalence of CVD among participants who were and were not currently smoking cigarettes at the time of the fifth examination in the Framingham
Offspring Study. The outcome is prevalent CVD and the two comparison groups are defined by current smoking status. In Example 6.9, we generated a point estimate for the difference in prevalent CVD between smokers and nonsmokers of 0.0114 and a 95% CI of (20.0133, 0.0361). Because the 95% CI included 0, we could not conclude that there
was a statistically significant difference in prevalent (0.135, 1.339). A 95% CI for ln(RR) is (0.135, 1.339). To generate the CI for the RR, we take the antilog (exp) of the lower and upper limits: (exp(0.135), exp(1.339)), (1.14, 3.82). Thus, we are 95% confident that patients receiving the new pain reliever have between 1.14 and 3.82 times the likelihood
of reporting a meaningful reduction in pain as compared to patients receiving the standard pain reliever. Because this CI does not include 1 (the null value), we can conclude that this difference is statistically significant. 6.6.3 Confidence Intervals for the Odds Ratio There are some study designs in which it is not possible to estimate an R R (e.g., a
case-control study), and therefore an odds ratio is computed as a relative measure of effect.6 An odds TABLE 6–21 Confidence Interval for OR CI for ln(OR) CI for OR 117 1 1 1 1 ln(OˆR) 6 z x1 + (n1 − x1) + x 2 + (n2 − x 2) (Find z in Table 1B) exp(Lower limit), exp (Upper limit) 118 CHAPTER 6 Confidence Interval Estimates CVD between
smokers and nonsmokers. In Example 6.11, we generated a point estimate for the relative risk, pˆ 0.1089 RˆR 5 1 = = 1.12 , ˆp2 0.0975 and a 95% CI of (0.89, 1.41), and because the CI for the RR included 1, we concluded that there was no statistically significant difference in prevalent CVD between smokers and nonsmokers. We now use these data
to estimate an OR. A point estimate for the odds ratio is 81/663 x1 / (n1 − x1) = = 1.13, OˆR 5 x 2 / (n2 − x 2) 298/2757 and the interpretation is that smokers have 1.13 times the odds of prevalent CVD as compared to nonsmokers. The 95% CI estimate for the OR is computed using the two-step procedure outlined in Table 6–21, ln(OˆR) 6 z 1 1 1 1 +
+ + , x1 n1 − x1 x 2 n2 − x 2 ln(1.13) 6 1.96 1 1 1 1 + + + , 81 663 298 2757 We generated a point estimate for the difference in proportions of patients reporting a clinically meaningful reduction in pain between pain relievers as 0.46 2 0.22 5 0.24, and the 95% CI was (0.06, 0.42). Because the 95% CI for the RD did not contain 0 (the null value), we
concluded that there was a statistically significant difference between pain r elievers. We then generated a point estimate for the relative risk, pˆ 0.46 RˆR 5 1 = = 2.09. pˆ2 0.22 and a 95% CI of (1.14, 3.82). Because this CI for the RR did not include 1, we concluded that this difference was statistically significant. We now use these data to generate a
point estimate and 95% CI estimate for the OR. The point estimate for the odds ratio is x /(n − x) 23/27 = 3.02. OˆR 5 1 1 1 = x 2 /(n2 − x 2) 11/39 This suggests that the odds that a patient on the new pain reliever reports a meaningful reduction in pain is 3.02 times the odds that a patient on the standard pain reliever reports a meaningful reduction.
The 95% CI estimate for the OR is computed using the two-step procedure outlined in Table 6–21, 0.122 6 0.260, (20.138, 0.382). ln(3.02) 6 1.96 1 1 1 1 + + + , 23 27 11 39 A 95% CI for ln(OR) is (20.138, 0.382). To generate the CI for the OR, we take the antilog (exp) of the lower and upper limits, 1.105 6 0.870, (exp(20.138), exp(0.382)), A 95% CI
for ln(OR) is (0.235, 1.975). To generate the CI for the OR, we take the antilog (exp) of the lower and upper limits, (0.87, 1.47). Thus, we are 95% confident that the odds of prevalent CVD in smokers is between 0.87 and 1.47 times the odds of prevalent CVD in nonsmokers. The null value is 1 and because this CI includes 1, we cannot conclude that
there is a statistically significant difference in the odds of prevalent CVD between smokers and nonsmokers. Example 6.14. Consider again the data shown in Table 6–19 from the randomized trial assessing the effectiveness of a newly developed pain reliever as compared to the standard of care. The primary outcome is reduction in pain of 3 or more
scale points (defined by clinicians as a clinically meaningful reduction). (0.235, 1.975). (exp(0.235), exp(1.975)), (1.26, 7.21). Thus, we are 95% confident that the odds that a patient on the new pain reliever reports a meaningful reduction in pain is between 1.26 and 7.21 times the odds that a patient on the standard pain reliever reports a meaningful
reduction. Because this CI does not include 1, we conclude that this difference is statistically significant. When the study design allows for the calculation of an RR, it is the preferred measure as it is far more interpretable than an OR. However, the OR is extremely important as it is the only measure of effect that can be computed in a Summary case-
control study design. When the outcome of interest is relatively rare (e.g., fewer than 10% successes), then the OR and RR will be very close in magnitude. In such a case, investigators often interpret the OR as if it were a RR—i.e., as a comparison of risks rather than a comparison of odds, which is less intuitive. 6.7 SUMMARY In this chapter, we
presented various formulas for estimating different unknown population parameters. In each application, a random sample or two independent random samples were selected from the population and sample statistics (e.g., sample sizes, means, and standard deviations, or sample sizes and proportions) were generated. Point estimates are the best
single-valued estimates of unknown population parameters. Because these can vary from sample to sample, most investigations start with a point estimate and build in a margin of error. The margin of error quantifies sampling variability and includes a value from the z or t distribution reflecting the selected confidence level as well as the standard
error of the point estimate. It is important to remember that the confidence interval contains a range of likely values for the unknown population parameter—a range of values for the population parameter consistent with the data. It is also possible that the confidence interval does not contain the true population parameter. This is important to
remember in interpreting intervals. Confidence intervals are also very useful for comparing means or proportions and can be used to assess whether there is a statistically meaningful difference in parameters. This is based on whether the confidence interval includes the null value (e.g., 0 for the difference in means, mean difference, and risk
difference, or 1 for the relative risk and odds ratio). The precision of a confidence interval is defined by the margin of error (or the width of the interval). A larger margin of error (wider interval) is indicative of a less precise estimate. Table 6–22 summarizes key formulas for confidence interval estimates. TABLE 6–22 Summary of Key Formulas
Number of Groups, Outcome: Parameter Confidence Interval* s w X 6 z n One sample, continuous: CI for m (X w1 2 w X2) 6 zS p Two independent samples, continuous: CI for (m12m2) Two matched samples, continuous: Xd 6 z CI for md One sample, dichotomous: CI for p Two independent samples, dichotomous: CI for RD 5 (p12 p2) Two
independent samples, dichotomous: p CI for RR 5 }p1} 2 Two independent samples, dichotomous: CI for OR 5 x1 / (n1 − x1) x 2 / (n2 − x 2) 119 pˆ 6 z pˆ1 − pˆ 2 6 z 1 1 + n1 n2 sd n pˆ(1 − pˆ) n pˆ1 (1 − pˆ1) pˆ (1 − pˆ 2) + 2 n1 n2 (n − x1)/ x1 (n2 − x 2)/ x 2 ln(RˆR) 6 z 1 + n1 n2 exp(Lower limit), exp (Upper limit) ln(OˆR) 6 z 1 1 1 1 + +
+ x1 x2 (n1 − x1) (n2 − x 2) exp(Lower limit), exp (Upper limit) *See Tables 6–3, 6–8, and 6–12 for alternative formulas that are appropriate for small samples. 120 CHAPTER 6 Confidence Interval Estimates 6.8 PRACTICE PROBLEMS The following values reflect cups per day consumed: 1. A study is run to estimate the mean total cholesterol
level in children 2 to 6 years of age. A sample of nine participants is selected and their total cholesterol levels are measured as follows: 185 225 240 196 175 180 194 147 223 Generate a 95% CI for the true mean total cholesterol levels in children. 2. A clinical trial is planned to compare an experimental medication designed to lower blood pressure
to a placebo. Before starting the trial, a pilot study is conducted involving ten participants. The objective of the study is to assess how systolic blood pressure changes over time untreated. Systolic blood pressures are measured at baseline and again 4 weeks later. Compute a 95% CI for the difference in blood pressures over 4 weeks. Baseline: 120 145
130 160 152 143 126 121 115 135 4 Weeks: 122 142 135 158 155 140 130 120 124 130 3. The main trial for the medication in Problem 2 is conducted and involves a total of 200 patients. Patients are enrolled and randomized to receive either the experimental medication or the placebo. The data shown in Table 6–23 are data collected at the end of
the study after 6 weeks on the assigned treatment. a. Generate a 95% CI for the difference in mean systolic blood pressures between groups. b. Generate a 95% CI for the difference in proportions of patients with hypertension between groups. c. Generate a point estimate for the RR of side effects in patients assigned to the experimental group as
compared to placebo. Generate a 95% CI for the RR. 4. The following data are collected as part of a study of coffee consumption among undergraduate students. TABLE 6–23 Outcome Data by Treatment Experimental (n5100) Mean (SD) systolic blood pressure Placebo (n5100) 120.2 (15.4) 131.4 (18.9) Hypertensive (%) 14 22 Side effects (%) 6 8
3 4 6 8 2 1 0 2 a. Compute the sample mean. b. Compute the sample standard deviation. c. Construct a 95% CI for the mean number of cups of coffee consumed among all undergraduates. 5. A clinical trial is conducted to evaluate a new pain medication for arthritis. Participants are randomly assigned to receive the new
medication or a placebo. The outcome is pain relief within 30 minutes, and the data are shown in Table 6–24. a. Estimate the RD in pain relief between treatments. b. Estimate the RR in pain relief between treatments. c. Estimate the OR in pain relief between treatments. d. Construct a 95% CI for the OR. 6. Data are collected in a clinical trial
evaluating a new compound designed to improve wound healing in trauma patients. The new compound is compared against a placebo. After treatment for 5 days with the new compound or placebo, the extent of wound healing is measured and the data are shown in Table 6–25. Suppose that clinicians feel that if the percent reduction in the size of the
wound is greater than 50%, then the treatment is a success. a. Generate a 95% CI for the percent success in patients receiving the new compound. TABLE 6–24 Pain Relief by Treatment Pain Relief No Pain Relief New medication 44 76 Placebo 21 99 TABLE 6–25 Wound Healing by Treatment Treatment New compound (n 5 125) Placebo (n 5 125)
Number of Patients with Percent Reduction in Size of Wound None 1–25 26–50 51–75 76–100 4 11 37 32 41 12 24 45 34 10 Practice Problems b. Generate a 95% CI for the difference in the percent success between the new compound and placebo. c. Generate a 95% CI for the RR of treatment success between treatments. d. Generate a 95% CI for
the OR of treatment success between treatments. 7. A clinical trial is conducted to compare an experimental medication to placebo to reduce the symptoms of asthma. Two hundred participants are enrolled in the study and randomized to receive either the experimental medication or placebo. The primary outcome is a self-reported reduction of
symptoms. Among 100 participants who receive the experimental medication, 38 report a reduction of symptoms as compared to 21 participants of 100 assigned to the placebo. a. Generate a 95% CI for the difference in proportions of participants reporting a reduction of symptoms between the experimental and placebo groups. b. Estimate the RR
for reduction in symptoms between groups. c. Estimate the OR for reduction in symptoms between groups. d. Generate a 95% CI for the RR. 8. Table 6–26 displays descriptive statistics on the participants involved in the study described in Problem 7. a. Generate a 95% CI for the mean age among participants assigned to the placebo. b. Generate a
95% CI for the difference in mean ages in participants assigned to the experimental versus the placebo groups. c. Generate a 95% CI for the difference in mean BMI in participants assigned to the experimental versus the placebo groups. TABLE 6–26 Descriptive Statistics by Treatment Experimental Medication (n5100) Mean (SD) age, (years)
Placebo (n5100) 47.2 (4.3) 46.1 (5.1) 46% 58% Mean (SD) educational level (years) 13.1 (2.9) 14.2 (3.1) Mean (SD) annual income $36,560 ($1054) $37,470 ($998) 24.7 (2.7) 25.1 (2.4) Men (%) Mean (SD) body mass index (BMI) 121 9. A crossover trial is conducted to compare an experimental medication for migraine headaches to a currently
available medication. A total of 50 patients are enrolled in the study and each patient receives both treatments. The outcome is the time, in minutes, until the headache pain resolves. Following each treatment, patients record the time it takes until pain is resolved. Treatments are assigned in r andom order (i.e., some patients receive the currently -
available medication first and then the experimental medication, and others receive the experimental medication first and then the currently available medication). The mean difference in times between the experimental and currently available medication is 29.4 minutes with a standard deviation of 2.8 minutes. Construct a 95% CI for the mean -
difference in times between the experimental and currently available medication. 10. Suppose in the study described in Problem 9 each participant is also asked if the assigned medication causes any stomach upset. Among the 50 participants, 12 reported stomach upset with the experimental medication. Construct a 90% CI for the proportion of
participants who experience stomach upset with the experimental medication. 11. A 95% confidence interval for the mean diastolic blood pressure in women is 62 to 110. What is the margin of error? What is the standard error? 12. The following data are collected from 10 randomly selected patients undergoing physical therapy following knee
surgery. The data represent the percentage gain in range of motion after 3 weeks of therapy. Generate a 95% confidence interval for the mean percentage gain in range of motion. 24% 32% 50% 62% 21% 45% 80% 24% 30% 10% 13. The data in Table 6–27 are collected in a randomized trial to test the efficacy of a new drug for
migraine headaches. The following are characteristics of study participants overall and then organized by the treatment to which they are assigned. a. Generate a 95% confidence interval for the difference in mean ages between groups. b. Generate a 95% confidence interval for the d ifference in proportions of males between groups. c. Generate a
95% confidence interval for the difference in proportions of patients with severe migraine headaches between groups. 122 CHAPTER 6 Confidence Interval Estimates TABLE 6–27 Characteristics in Participants in Study of Treatment for Migraine Headaches Mean (SD) age, years in Pacemaker Study New Drug Placebo (n = 100) (n = 100) All (n =
200) Baseline Characteristics Mean (SD) age, years 32.8 (4.7) 31.9 (5.1) 32.0 (4.9) % Male 54% 48% 51% % High school graduate 76% 80% 78% Severity of migraine headaches % Mild 22% 20% 21% % Moderate 38% 42% 39% % Severe 40% 38% 39% 5 (3–12) 6 (2–18) 6 (3–17) Median (Q1–Q3) number of days missed work in past year due to
migrane Min–Max number of days missed work in past year due to migraine TABLE 6–28 Baseline Characteristics of Participants New Pacemaker (n = 100) Available Pacemaker (n = 100) 67.3 (5.9) 66.9 (5.6) 48% 52% Mean (SD) number of days with AF event 8.4 (3.2) 14.9 (3.9) % hospitalized for AF 4% 9% % Male Outcomes Compute a 95%
confidence interval for the mean gestational age of women enrolling in the study. 0–35 0–48 0–48 14. A clinical trial is run to assess the efficacy of a new pacemaker device in patients with atrial fibrillation (AF). Two hundred participants are randomized to receive the new pacemaker or a currently available pacemaker. There are two primary
outcomes of interest: the number of days in a 3-month period with an atrial fibrillation event, and hospitalization for atrial fibrillation over the 3-month follow-up period. Data on baseline characteristics and the outcomes are shown in Table 6–28. a. Compute a 95% confidence interval for the difference in mean number of days with an AF event
between participants receiving the new pacemaker as compared to the available pacemaker. b. Compute a 95% confidence interval for the mean number of days with an AF event among participants receiving the new pacemaker. 15. A pilot study is run to investigate the feasibility of recruiting pregnant women into a study of risk factors for preterm
delivery. Women are invited to participate at their first clinical visit for prenatal care. The following represent the gestational ages in weeks of women who consented to participate in the study. 11 14 21 22 9 10 13 18 16. A clinical trial is run comparing a new drug for high cholesterol to a placebo. A total of 40 participants are randomized
(with equal assignment to treatments) to receive either the new drug or placebo. Their total serum cholesterol levels are measured after eight weeks on the assigned treatment. Participants receiving the new drug reported a mean total serum cholesterol level of 209.5 (SD = 21.6) and participants receiving the placebo reported a mean total serum
cholesterol level of 228.1 (SD = 19.7). a. Construct a 95% confidence interval for the difference in mean total serum cholesterol levels between participants receiving the new drug versus placebo. b. Is the new drug effective? Justify briefly. REFERENCES 1. Newcomb, R.G. "Two-sided confidence intervals for the single proportion: Comparison of
seven methods." Statistics in Medicine 1998; 17(8): 857–872. 2. StatXact Version 7. © 2006 by Cytel, Inc., Cambridge, MA. 3. D'Agostino, R.B., Sullivan, L.M., and Beiser, A. Introductory Applied Biostatistics. Belmont, CA: Duxbury-Brooks/Cole, 2004. 4. Rosner, B. Fundamentals of Biostatistics. Belmont, CA: DuxburyBrooks/ Cole, 2006. 5. Agresti,
A. Categorical Data Analysis (2nd ed.). New York: John Wiley & Sons, 2002. 6. Rothman, K.J. and Greenland, S. Modern Epidemiology (2nd ed.). Philadelphia: Lippincott-Raven, 1998. © Sergey Nivens/Shutterstock © holbox/Shutterstock 7 chapter Hypothesis Testing Procedures W hen and W hy Key Questions •• What does it mean when a study is
statistically significant but not clinically meaningful? •• What does a test of hypothesis tell you that a confidence interval does not? •• When testing a hypothesis, is it better to run a matched pairs test or an independent samples test? In the News The Physicians' Health Study, which was sponsored by the National Cancer Institute and the National,
Heart, Lung, and Blood Institute, was a randomized, controlled trial involving more than 20,000 male physicians that investigated the effects of low-dose aspirin and beta-carotene on cardiovascular mortality and cancer incidence, respectively.1 The aspirin component of the trial was stopped early because there was statistically significant evidence of
a reduction in myocardial infarction (heart attack) risk in participants who took aspirin. Participants who took aspirin had a 44% reduction in risk of heart attack as compared to their counterparts who received placebo (relative risk = 0.56; 95% confidence interval, 0.45 to 0.70; p < 0.0001). More than 11,000 participants were assigned to each arm of
the trial (aspirin or placebo), and 139 participants taking aspirin experienced heart attack as compared to 239 who took placebo.2 For some time, low-dose aspirin has been considered a safe and effective preventive measure.3 However, there may be risks for some groups of patients.4 Physicians' Health Study. Available at /index.html. 2 Steering
Committee of the Physicians' Health Study research group. "Final report on the aspirin component of the ongoing Physicians' Health Study." New England Journal of Medicine 1989; 321: 129–135. 3 CardioSmart. American College of Cardiology. "Heart disease: aspirin." Available at /media /Documents/Fact%20Sheets/en/tb1597.ashx. Accessed July
21, 2016. 4 Everyday Health guest columnist. "6 essential facts about aspirin therapy for your heart." Everyday Health. Available at .everydayhealth.com/columns/health-answers/essential-facts-about -aspirin-therapy-your-heart/. Accessed July 21, 2016. 1 Dig In •• What percentage of patients suffered heart attacks in the Physicians' Health Study in
each arm of the trial? •• How substantial is the evidence supporting the use of low-dose aspirin in the Physicians' Health Study? •• Are the recommendations for use of low-dose aspirin clear to consumers (all of whom may not understand statistics)? L earning O bjectives By the end of this chapter, the reader will be able to •• D efine null and
research hypothesis, test statistic, level of significance, and decision rule •• Distinguish between Type I and Type II errors and discuss the implications of each •• Explain the difference between one- and two-sided tests of hypothesis •• Estimate and interpret p-values •• Explain the relationship between confidence interval estimates and p-values in
drawing inferences •• Perform analysis of variance by hand •• Appropriately interpret results of analysis of variance tests •• Distinguish between one- and two-factor analysis of variance •• Perform chi-square tests by hand •• Appropriately interpret results of chi-square tests •• Identify the appropriate hypothesis testing procedure based on
type of outcome variable and number of samples The second area of statistical inference is hypothesis testing. In hypothesis testing, a specific statement or hypothesis is generated about a population parameter, and sample statistics are used to assess the likelihood that the hypothesis is true. 124 CHAPTER 7 Hypothesis Testing Procedures This
statement or hypothesis is based on available information and the investigator's belief about the parameter. The process of hypothesis testing involves setting up two competing hypotheses: one reflects no difference, no association, or no effect (called the null hypothesis) and the other reflects the investigator's belief (called the research or alternative
hypothesis). We select a random sample (or multiple samples when there are more comparison groups) and generate summary statistics. We then assess the likelihood that the sample data support the research or alternative hypothesis. Similar to estimation, the process of hypothesis testing is based on probability theory and the Central Limit
Theorem. The techniques for hypothesis testing again depend on the appropriate classification of the key study outcome variable or endpoint. The number of comparison groups in the investigation must also be specified. It is again important to determine whether the comparison groups are independent (i.e., physically separate, such as men versus
women or participants assigned to receive a new drug or placebo in a clinical trial) or dependent (i.e., matched or paired, such as pre- and post-assessments on the same participants). These issues dictate the appropriate hypothesis testing technique. In e stimation, we focused explicitly on techniques for one and two samples. We discussed estimation
for a specific parameter (e.g., the mean m or proportion p of a population), for differences (e.g., difference in means m1 – m2, the risk difference p1 – p2) and for ratios [(e.g., the relative risk RR = p1/p2 and odds ratio OR = (p1 /(1 – p1))/(p2 /(1 – p2))]. Here we focus on procedures for one, two, and more than two samples. Hypothesis testing can
generalize to the situation of more than two groups, whereas estimation is not intuitive when there are more than two groups. Table 7–1 outlines hypothesis testing procedures that we consider here. 7.1 INTRODUCTION TO HYPOTHESIS TESTING Before discussing the different procedures that are appropriate for each scenario outlined in Table
7–1, we first present the general approach using a simple example. The Centers for Disease Control and Prevention (CDC) reported on trends in weight, height, and body mass index (BMI) from the 1960s through 2002.1 The data for the report were collected in the National Health Examination and the National Health and Nutrition Examination
Surveys, which were conducted between 1960 and 2002. The general trend was that Americans were much heavier and slightly taller in 2002 as compared to 1960. The report indicated that both men and women gained approximately 24 pounds, on average, between 1960 and 2002. In 2002, the mean weight for American men was reported at 191
pounds and the mean weight for American women was reported at 163 pounds. TABLE 7–1 Hypothesis Testing Techniques Number of Samples Outcome Variable One sample Continuous Two independent samples Continuous Two dependent, matched samples Continuous More than two independent samples Continuous One sample Dichotomous
Two independent samples Dichotomous More than two independent samples Dichotomous One sample Categorical or ordinal (more than 2 response options) Two or more independent samples Categorical or ordinal Suppose we focus on the mean weight in men. In 2002, the mean weight for men was reported at 191 pounds. Suppose that an
investigator hypothesizes that weights are even higher in 2006 (i.e., that the trend continued over the next 4 years). In hypothesis testing, we set up competing hypotheses about the unknown parameter. One hypothesis is called the null hypothesis and the other is called the alternative or research hypothesis. The research hypothesis is that the mean
weight in men in 2006 is more than 191 pounds. The null hypothesis is that there is no change in weight and therefore the mean weight is still 191 pounds in 2006. The null and research hypotheses are denoted as Null hypothesis, H0: m 5 191 (no change), Research hypothesis, H1: m . 191 (investigator's belief). To test the hypotheses, we select a
random sample of American males in 2006 and measure their weights. Suppose we have resources available to recruit n 5 100 men into our sample. We weigh each participant and compute summary statistics on the sample data. Suppose in the sample we determine the following: n 5 100, X w5 197.1, s 5 25.6. Do the sample data support the null or
research hypothesis? The sample mean of 197.1 pounds is numerically higher than 191 pounds. However, is this difference more than would be expected by chance? In hypothesis testing, we assume that the null h ypothesis holds until proven otherwise. IntroductionWhat to Hypothesis Is Public Health? Testing We therefore need to determine the
likelihood of observing a sample mean of 197.1 or higher when the true population mean is 191 (i.e., if the null hypothesis is true or under the null hypothesis). We compute this probability using the Central Limit Theorem. Specifically, 197.1 – 191 5 25.6/ 100 P(z . 2.38) 5120.9913 5 0.0087. 1 P(X w . 197.1) 5 P z > 2 (Notice that we use the sample
standard deviation in computing the z score. This is generally an appropriate substitution as long as the sample size is large, n . 30.) Thus, there is less than a 1% chance of observing a sample mean as large as 197.1 when the true population mean is 191. Do you think that the null hypothesis is likely true? Based on how unlikely it is to observe a
sample mean of 197.1 under the null hypothesis (i.e., less than 1% chance), we might infer from our data that the null hypothesis is probably not true. Suppose that the sample data had turned out differently. Suppose that we instead observed the following in 2006: n 5 100, X w 5 192.1, s 5 25.6. How likely is it to observe a sample mean of 192.1 or
higher when the true population mean is 191 (i.e., if the null hypothesis is true)? We again compute this probability using the Central Limit Theorem. Specifically, ⎛ 192.1 − 191 ⎞P X > 192.1 = P⎜⎜ z > ⎟⎟= 25.6 100 ⎠⎝ ( ) P(z . 0.43) 5 1 2 0.6664 5 0.3336. There is a 33.4% chance of observing a sample mean as large as 192.1 when the true population
mean is 191. Do you think that the null hypothesis is likely true? We need to determine a threshold or cutoff point (called the critical value) to decide when to believe the null hypothesis and when to believe the research hypothesis. It is important to note that although it is possible to observe any sample mean when the true population mean is 191,
some values are very unlikely. Based on the previous two samples, it would seem reasonable to believe the research hypothesis when X w 5 197.1, but to believe the null hypothesis when X w 5 192.1. What we need is a threshold value such that if X w is above that threshold, we believe that H1 is true and if X w is below that threshold, we believe
that H0 is true. The difficulty in determining a threshold for X w is that it depends on the scale of measurement. In this example, the critical value might be 195 pounds (i.e., if the sample mean is 195 pounds or more, we believe that H1 is true and if the sample mean is less than 195 pounds, we believe that H0 is true). Suppose we are interested in
assessing an increase in blood pressure over time. The critical value would be different because blood pressures 125 are measured in millimeters of mercury (mmHg) as opposed to in pounds. In the following, we explain how the critical value is determined and how we handle the issue of scale. First, to address the issue of scale in determining the
critical value we convert our sample data (in particular, the sample mean) into a z score. We know from Chapter 5 that the center of the z distribution is 0 and extreme values are those that exceed 2 or fall below 22 (values above 2 and below 22 represent approximately 5% of all z values). If the observed sample mean is close to the mean specified in
H0 (m 5191), then z is close to 0. If the observed sample mean is much larger than the mean specified in H0, then z is large. In hypothesis testing, we select a critical value from the z distribution. This is done by first determining what is called the level of significance, denoted a. Remember that if the null hypothesis is true, it is possible to observe any
sample mean. What we are doing here is drawing a line at extreme values. The level of significance is the probability that we r eject the null hypothesis (in favor of the alternative) when it is actually true: a 5 Level of significance 5 P(Reject H0 u H0 is true). Because a is a probability, it ranges between 0 and 1. The usual value for a is 0.05, or 5%. If an
investigator selects a 5 0.05, he or she is allowing a 5% probability of incorrectly rejecting the null hypothesis in favor of the alternative when the null is true. The typical values for a are 0.01, 0.05, and 0.10, with a 5 0.05 the most commonly used value. Suppose in our weight study we select a 5 0.05. We need to determine the value of z that holds 5%
of the values above it (Figure 7–1). The critical value of z for a 5 0.05 is z 5 1.645 (i.e., 5% of the distribution is above 1.645). With this value, we can set up what is called our decision rule for the test. The rule is to reject H0 if the z score is 1.645 or higher. With the first sample, we have X w 5 197.1 and z 5 2.38. Because 2.38 . 1.645, we reject the
null hypothesis. (The same conclusion can be drawn by comparing the 0.0087 probability of observing a sample mean as extreme as 197.1 to the level of significance of 0.05. If the observed probability is smaller than the level of significance, we reject H0.) Because the z score exceeds the critical value (2.38 > 1.645), we conclude that the mean weight
for men in 2006 is more than 191 pounds, the value reported in 2002. If we observed the second sample (X w5 192.1), we would not reject the null hypothesis because the z score is 0.43, which is not in the r ejection region (i.e., the region in the tail end of the curve at or above 1.645). With the second sample we do not have evidence to conclude that
weights have increased. (Again, the same conclusion can be reached by comparing probabilities. 126 CHAPTER 7 Hypothesis Testing Procedures FIGURE 7–1 Critical Value of Z for a 5 0.05 =0.05 –2 –3 –1 The probability of observing a sample mean as e xtreme as 192.1 is 33.4%, which is not below our 5% level of significance.) The procedure for
hypothesis testing is based on the ideas described previously. Specifically, we set up competing hypotheses, select a random sample from the population of interest, and compute summary statistics. We then determine whether the sample data supports the null or alternative hypotheses. The procedure can be broken down into the following five steps.
We use this five-step approach in performing tests of hypotheses for all of the scenarios in Table 7–1. Step 1: Set up hypotheses and determine the level of significance. H0: Null hypothesis (no change, no difference), H1: Research hypothesis (investigator's belief), a 5 0.05. The research or alternative hypothesis can take one of three forms. An
investigator might believe that the parameter has increased, decreased, or changed. For example, an investigator might hypothesize: 1. H1: m . m0, where m0 is the comparator or null value (e.g., m0 5191 in our example about weight in men in 2006) and an increase is hypothesized—this type of test is called an upper-tailed test. 0 1 1.645 3 2 2. H1:
m , m0, where a decrease is hypothesized—this is called a lower-tailed test. 3. H1: m m0, where a difference is hypothesized—this is called a two-tailed test. The exact form of the research hypothesis depends on the investigator's belief about the parameter of interest and whether it has possibly increased, decreased, or is different from the null value.
The research hypothesis is set up by the investigator before any data are collected. Step 2: Select the appropriate test statistic. The test statistic is a single number that summarizes the sample information. An example of a test statistic is the z statistic computed as z= X − µ0 . s n When the sample size is small, we use t statistics (just as we did in
estimation in Chapter 6). As we present each scenario, alternative test statistics are provided along with conditions for their appropriate use. Step 3: Set up the decision rule. The decision rule is a statement that tells under what circumstances to reject the null hypothesis. The decision rule is based on specific values of the test statistic (e.g., reject
IntroductionWhat to Hypothesis Is Public Health? Testing Figures 7–2 through Figure 7–4 illustrate the rejection regions defined by the decision rule for upper-, lower-, and twotailed z tests with a 5 0.05. Notice that the rejection regions are in the upper, lower, and both tails of the curves, respectively. Notice in the two-tailed test that the rejection r-
egion is split into two equal parts. The total area in the rejection region is still equal to a. In Figure 7–4, a 5 0.05 and the area in each tail is 0.025. Critical values of z for upper-, lower-, and two-tailed tests can be found in Table 1C in the Appendix. Critical values of t for upper, lower, and two-tailed tests can be found in Table 2 in the Appendix. H0 if z
$ 1.645). The decision rule for a specific test depends on three factors: the research hypothesis, the test statistic, and the level of significance. The decision rule depends on whether an upper-tailed, lower-tailed, or two-tailed test is proposed. In an upper-tailed test, the decision rule has investigators reject H0 if the test statistic is greater than or equal
to the critical value. In a lowertailed test, the decision rule has investigators reject H0 if the test statistic is less than or equal to the critical value. In a two-tailed test, the decision rule has investigators reject H0 if the test statistic is extreme—either greater than or equal to an upper critical value or less than or equal to a lower critical value. The
exact form of the test statistic is also important in determining the decision rule. If the test statistic follows the standard normal distribution (z), then the decision rule is based on the standard normal distribution. If the test statistic follows the t distribution, then the decision rule is based on the t distribution. The appropriate critical value is selected
from the t distribution again depending on the specific alternative hypothesis and the level of significance. The third factor is the level of significance, which is selected in Step 1 (e.g., a 5 0.05). For example, in an uppertailed z test, if a 5 0.05, then the critical value is z 5 1.645. Step 4: Compute the test statistic. Here we compute the test statistic by
substituting the observed sample data into the test statistic identified in Step 2. Step 5: Conclusion. The final conclusion is made by comparing the test statistic (which is a summary of the information observed in the sample) to the decision rule. The final conclusion is e ither to reject the null hypothesis (because the sample data are FIGURE 7–2
Rejection Region for Upper-Tailed Z Test (H1: m . m0) with a 5 0.05 =0.05 –3 –2 –1 127 0 1 1.645 2 3 128 CHAPTER 7 Hypothesis Testing Procedures FIGURE 7–3 Rejection Region for Lower-Tailed Z Test (H1: m , m0) with a 5 0.05 =0.05 –3 –2 –1.645 –1 0 1 3 2 FIGURE 7–4 Rejection Region for Two-Tailed Z Test (H1: m m0) with a 5 0.05
2=0.025 2=0.025 –3 –2 –1.960 –1 0 1 2 1.960 3 IntroductionWhat to Hypothesis Is Public Health? Testing 129 Step 3: Set up the decision rule. TABLE 7–2 Critical Values for Upper-Tailed Z Tests Upper-Tailed Test a Critical Value of z 0.10 1.282 0.05 1.645 0.025 1.960 0.010 2.326 0.005 2.576 0.001 3.090 0.0001 3.719 very unlikely if the null
hypothesis is true) or not to r eject the null hypothesis (because the sample data are not very unlikely). If the null hypothesis is rejected, an exact significance level is computed to describe the likelihood of observing the sample data assuming that the null hypothesis is true. The exact level of significance is called the p-value and it will be less than the
chosen level of significance. Statistical computing packages provide exact p-values as part of their standard output for hypothesis tests. We approximate p-values using Table 1C for tests involving z statistics and Table 2 for tests involving t statistics. We now use the five-step procedure to test the research hypothesis that the mean weight in men in
2006 is more than 191 pounds. We assume the observed sample data are as follows: n 5 100, X w 5 197.1, and s 5 25.6. Step 1: Set up hypotheses and determine the level of significance. H0: m 5 191, H1: m . 191, a 5 0.05. The research hypothesis is that weights have increased, and therefore an upper-tailed test is used. Step 2: Select the
appropriate test statistic. Because the sample size is large (n . 30), the appropriate test statistic is X − µ0 z= . s n (In Section 7.2, we present alternative test statistics appropriate for small samples.) In this example, we are performing an upper-tailed test (H1: m . 191), with a z test statistic and selected a 5 0.05. The decision rule is shown in Figure 7–
2, Reject H0 if z $ 1.645. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2, z= X − µ0 197.1 − 191 = = 2.38 s n 25.6 100 Step 5: Conclusion. We reject H0 because 2.38 . 1.645. We have statistically significant evidence at a 5 0.05 to show that the mean weight in men in
2006 is more than 191 pounds. Because we reject the null hypothesis, we now approximate the p-value, which is the likelihood of observing data as more extreme under the assumed statistical model. An alternative definition of the p-value is the smallest level of significance where we still reject H0. In this example, we observe z 5 2.38, and for a 5
0.05 the critical value is 1.645. Because 2.38 exceeds 1.645, we reject H0. In our conclusion, we report a statistically significant increase in mean weight at a 5% level of significance. The data actually provide stronger evidence. Table 7–2 is a copy of Table 1C in the Appendix that contains critical values for upper-tailed tests. What is the smallest level
of significance we could choose and still reject H0? If we select a 5 0.025, the critical value is 1.96, and we still reject H0 because 2.38 . 1.960. If we select a 5 0.010, the critical value is 2.326, and we still reject H0 because 2.38 . 2.326. However, if we select a 5 0.005, the critical value is 2.576 and we cannot reject H0 because 2.38 , 2.576. Therefore,
the smallest a where we still reject H0 is 0.010. This is the p-value. A statistical computing package produces a more precise p-value, which would be between 0.005 and 0.010. Here we are approximating the p-value using Table 1C in the Appendix, and we report p , 0.010. p-values reflect the exact significance of tests of hypothesis. In this example,
we find p , 0.010, indicating that there is less than a 1% chance that we are incorrectly rejecting the null hypothesis if the null hypothesis is true. Suppose in this example that the test statistic is z 5 1.70. We still reject H0 at a 5 0.05 because 1.70 . 1.645. However, with z 5 1.70, the p-value would be reported as p , 0.05. The p-value is shown
graphically in Figure 7–5. 130 CHAPTER 7 Hypothesis Testing Procedures FIGURE 7–5 p-Value, Exact Significance Level p -value –3 –2 –1 Smaller p-values are indicative of more incompatibility of the data with the assumed statistical model. In the literature, investigators often report p-values to summarize the significance of tests of hypothesis. The
following rule can be used to interpret p-values: Reject H0 if p # a, where p is the p-value and a is the level of significance selected by the investigator. For example, suppose we wish to test H0: m 5 100 versus H1: m . 100 at a 5 0.05. Data are collected and analyzed with a statistical computing package that reports p 5 0.0176. Because p 5 0.0176 , a
5 0.05, we reject H0. However, if we had selected a 5 0.01, we would not reject H0. In the examples that follow, we approximate p-values using Table 1C in the Appendix for tests involving z statistics and Table 2 in the Appendix for tests involving t statistics. We also discuss appropriate interpretation. In all tests of hypothesis (for continuous,
dichotomous, categorical and ordinal outcomes, with one, two, or more than two samples), there are two errors that can be 0 1 1.70 2 3 c ommitted. The first is called a Type I error and refers to the situation where we incorrectly reject H0 when, in fact, it is true. This is also called a false positive result (as we incorrectly conclude that the research
hypothesis is true when it is not). When we run a test of hypothesis and decide to reject H0 (e.g., because the test statistic is greater than or equal to the critical value in an upper-tailed test), e ither we make a correct decision because the research hypothesis is true or we commit a Type I error. The different conclusions are summarized in Table 7–3.
Note that we never know whether the null hypothesis is really true or false (i.e., we never know which row of Table 7–3 r eflects reality). TABLE 7–3 Errors in Tests of Hypothesis Conclusion in Test of Hypothesis Do Not Reject H0 Reject H0 H0 is true Correct decision Type I error H0 is false Type II error Correct decision Tests with One Sample,
What Continuous Is PublicOutcome Health? In the first step of a hypothesis test, we select a level of significance, a, and a 5 P(Type I error). Because we purposely select a small value for a, we control the probability of committing a Type I error. For example, if we select a 5 0.05 and our test tells us to reject H0, then there is a 5% probability that we
commit a Type I error. Most investigators are very comfortable with this and are confident when rejecting H0 that the research hypothesis is true, as it is the more likely scenario when we reject H0. When we run a test of hypothesis and decide not to reject H0 (e.g., because the test statistic is below the critical value in an upper-tailed test), then
either we make a correct decision because the null hypothesis is true or we commit a Type II error. b represents the probability of a Type II error and is defined as b 5 P(Type II error) 5 P(Do not reject H0 u H0 is false). Unfortunately, we cannot choose b to be small (e.g., 0.05) to control the probability of committing a Type II error because b depends
on several factors, including the sample size, the level of significance (a), and the research hypothesis. These issues will be discussed in much further detail in Chapter 8. For now, we must recognize that when we do not reject H0, it may be very likely that we are committing a Type II error (i.e., failing to reject H0 when it is false). Therefore, when
tests are run and the null hypothesis is not rejected, we often make a weak concluding statement allowing for the possibility that we might be committing a Type II error. If we do not reject H0, we conclude that we do not have significant evidence to show that H1 is true. We do not conclude that H0 is true. The most common reason for a Type II error
is a small sample size. 7.2 TESTS WITH ONE SAMPLE, CONTINUOUS OUTCOME Hypothesis testing applications with a continuous outcome variable in a single population are performed according to the five-step procedure outlined earlier. A key component is setting up the null and research hypotheses. The objective is to compare the mean in a
single population (m) to a known mean (m0). The known value is generally derived from another study or report—for example, a study in a similar but not identical population or a study performed some years ago. The latter is called a historical control ("control" here referring to the fact that the historical study is the comparator). It is important in
setting up the hypotheses in a one-sample test that the mean 131 specified in the null hypothesis is a fair and reasonable comparator. This is discussed in the examples that follow. In one-sample tests for a continuous outcome, we set up our hypotheses against an appropriate comparator. We select a sample and compute descriptive statistics on the
sample data using the techniques described in Chapter 4. Specifically, we compute the sample size (n), the sample mean (X w), and the sample standard deviation (s). We then determine the appropriate test statistic (Step 2) for the hypothesis test. The formulas for test statistics depend on the sample size and are given in Table 7–4. Appropriate use of
the t distribution assumes that the outcome of interest is approximately normally distributed. The National Center for Health Statistics (NCHS) published a report in 2005 entitled "Health, United States," and it contains extensive information on major trends in the health of Americans.2 Data are provided for the U.S. population as a whole and for
specific ages, by sex and race. We use some of the national statistics here as comparative values in one-sample tests in this section for means of continuous outcomes and in Section 7.3 and Section 7.4 for proportions of dichotomous, categorical, and ordinal variables respectively. Example 7.1. The NCHS report indicated that in 2002, Americans paid
an average of $3302 per year on health care and prescription drugs. An investigator hypothesizes that in 2005, expenditures are lower primarily due to the availability of generic drugs. To test the hypothesis, a sample of 100 Americans is selected and their expenditures on health care and prescription drugs in 2005 are measured. The sample data are
summarized as follows: n 5 100, X w5 $3190, and s 5 $890. Is there statistical evidence of a reduction in expenditures on health care and prescription drugs in 2005? Is the sample mean of $3190 evidence of a true reduction in the mean or is it within chance fluctuation? We run the test using the five-step approach. TABLE 7–4 Test Statistics for
Testing H0:m m0 n $ 30 n , 30 z= X − µ0 s n (Find critical value in Table 1C) t= X − µ0 s n (Find critical value in Table 2, df 5 n 2 1) 132 CHAPTER 7 Hypothesis Testing Procedures Step 1: Set up hypotheses and determine the level of significance. H0: m 5 3302, H1: m , 3302, a 5 0.05. The research hypothesis is that expenditures have decreased,
and therefore a lower-tailed test is used. Step 2: Select the appropriate test statistic. Because the sample size is large (n . 30), the appropriate test statistic is z= X − µ0 . s n presented data on n 5 3539 participants who attended the seventh examination of the Offspring in the Framingham Heart Study. Descriptive statistics on variables measured in
the sample were presented in Table 4.20 and included the following statistics on total cholesterol levels of participants: n 5 3310, w X5 200.3, and s 5 36.8. Is there statistical evidence of a difference in mean cholesterol levels in the Framingham Offspring as compared to the national mean? Here we want to assess whether the sample mean of 200.3 in
the Framingham sample is statistically significantly different from 203 (i.e., beyond what we would expect by chance). We run the test using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: m 5 203, Step 3: Set up the decision rule. This is a lower-tailed test, using a z statistic and a 5% level of significance.
The appropriate critical value can be found in Table 1C in the Appendix and the decision rule is Reject H0 if z # 21.645. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2: H1: m 203, a 5 0.05. The research hypothesis is that cholesterol levels are different in the Framingham
Offspring, and therefore a two-tailed test is used. Step 2: Select the appropriate test statistic. Because the sample size is large (n . 30), the appropriate test statistic is z= X − µ0 3190 − 3302 z= = = −1.26 s n 890 100 Step 5: Conclusion. We do not reject H0 because 21.26 . 21.645. We do not have statistically significant evidence at a 5 0.05 to show
that the mean expenditures on health care and prescription drugs are lower in 2005 than the mean of $3302 reported in 2002. Recall that when we fail to reject H0 in a test of hypothesis that either the null hypothesis is true (the mean expenditures in 2005 are the same as those in 2002 and equal to $3302) or we are committing a Type II error (we
fail to reject H0 when in fact it is false). In summarizing this test, we conclude that we do not have sufficient evidence to reject H0. We do not conclude that H0 is true because there may be a moderate to high probability that we are committing a Type II error. It is possible that the sample size is not large enough to detect a difference in mean
expenditures (see Chapter 8 for more details). Example 7.2. The NCHS reported that the mean total cholesterol level in 2002 for all adults was 203. In Chapter 4, we X − µ0 . s n Step 3: Set up the decision rule. This is a two-tailed test, using a z statistic and a 5% level of significance. The appropriate critical values can be found in Table 1C in the
Appendix and the decision rule is Reject H0 if z # 21.960 or if z $ 1.960. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2: z= X − µ0 200.3 − 203 = = −4.22. s n 36.8 3310 Step 5: Conclusion. We reject H0 because 24.22 ,21.960. We have statistically significant evidence at
a 5 0.05 to show that the mean total cholesterol level in the Framingham Offspring is different from the national mean of 203 reported in 2002. Because we reject Tests with One Sample, What Dichotomous Is PublicOutcome Health? H0, we also approximate a p-value. Using the two-sided signi ficance levels in Table 1C in the Appendix, p , 0.0001.

Example 7.2 raises an important concept of statistical versus clinical or practical significance. From a statistical standpoint, the mean total cholesterol level in the Framingham sample is highly statistically significantly different from the national mean with p , 0.0001 (i.e., there is less than a 0.01% chance that we are incorrectly rejecting the null hypothesis given the observed data). However, the sample mean in the Framingham Offspring study is 200.3, less than 3 units different from the national mean of 203. The reason the data are so highly statistically significant is the very large sample size. It is always important to assess both statistical and practical significance of data. This is particularly relevant when the sample size is large. The five-step procedure allows for an assessment of statistical significance. Investigators must also assess practical or clinical significance. Is a 3-unit difference in total cholesterol a meaningful difference? Example 7.3. Consider again the NCHS-reported mean total cholesterol level of 203 in 2002 for all adults. Suppose a new drug is proposed to lower total cholesterol, and a study is designed to evaluate the efficacy of the drug in lowering cholesterol. Fifteen patients are enrolled in the study and asked to take the new drug for 6 weeks. At the end of 6 weeks, each patient's total cholesterol level is measured and the sample statistics are as follows: n 5 15, X w 5 195.9, and s 5 28.7. Is there statistical evidence of a reduction in mean total cholesterol in patients after using the new drug for 6 weeks? We run the test using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: m 5 203, H1: m 203, a 5 0.05. Step 2: Select the appropriate test statistic. Because the sample size is small (n , 30), the appropriate test statistic is t= X − µ0 . s n Step 3: Set up the decision rule. This is a lower-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in Table 2 in the Appendix. To determine the critical value 133 of t, we need degrees of freedom, df, defined as df 5 n 2 1. In this example, df 5 15 2 1 5 14. The critical value for a lower-tailed test with df 5 14 and a 5 0.05 is 21.761 and the decision rule is Reject H0 if t #21.761. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. t= X − µ0 195.9 − 203 = = −0.96. s n 28.7 15 Step 5: Conclusion. We do not reject H0 because 20.96 . 21.761. We do not have statistically significant evidence at a 5 0.05 to show that the mean total cholesterol level in patients taking the new drug for 6 weeks is lower than the national mean. Again, because we fail to reject the null hypothesis, we make a weaker concluding statement, allowing for the possibility that we may be committing a Type II error (i.e., fail to reject H0 when the drug is efficacious). Example 7.3 raises an important issue in terms of study design. In this example, we assume in the null hypothesis that the mean cholesterol level is 203. This is taken to be the mean cholesterol level in patients without treatment. Is this an appropriate comparator? Alternative and potentially more efficient study designs to evaluate the effect of the new drug could involve two treatment groups, where one group receives the new drug and the other does not, or we could measure each patient's baseline or pre-treatment cholesterol level and then assess changes from baseline to 6 weeks post-treatment. These designs are discussed in Section 7.5 and Section 7.6, respectively. 7.3 TESTS WITH ONE SAMPLE, DICHOTOMOUS OUTCOME Hypothesis testing applications with a dichotomous outcome variable in a single population are also performed according to the five-step procedure. Similar to tests for means described in Section 7.2, a key component is setting up the null and research hypotheses. The objective is to compare the proportion of successes in a single population to a known proportion (p0). That known proportion is generally derived from another study or report and is sometimes called a historical control. It is important in setting up the hypotheses in a onesample test that the proportion specified in the null hypothesis is a fair and reasonable comparator. 134 CHAPTER 7 Hypothesis Testing Procedures In one-sample tests for a dichotomous outcome, we set up our hypotheses against an appropriate comparator. We select a sample and compute descriptive statistics on the sample data using the techniques described in Chapter 4. Specifically, we compute the sample size (n) and the sample proportion (pˆ), which is computed by taking the ratio of the number of successes to the sample size, pˆ 5 x/n. We then determine the appropriate test statistic (Step 2) for the hypothesis test. The formula for the test statistic is given in Table 7–5. The preceding formula is appropriate for large samples, defined when the smaller of np0 and n(1 2 p0) is at least 5. This is similar, but not identical, to the condition required for appropriate use of the confidence interval formula for a population proportion specified in Table 6–6—i.e., min[npˆ, n(1 2 pˆ)] $ 5. Here we use the proportion specified in the null hypothesis (p0) as the true proportion of successes rather than the sample proportion (pˆ). If we fail to satisfy the condition, then alternative procedures called exact methods must be used to test the hypothesis about the population proportion.3,4 Example 7.4. The NCHS report indicated that in 2002 the prevalence of cigarette smoking among American adults was 21.1%. In Chapter 4, we presented data on prevalent smoking in n 5 3536 participants who attended the seventh examination of the offspring in the Framingham Heart Study. Suppose we want to assess whether the prevalence of smoking is lower in the Framingham offspring sample due to the focus on cardiovascular health in that community. Data from the Framingham Offspring Study were summarized in Table 4–11 and indicated that 482 / 3536 5 13.6% of the respondents were currently smoking at the time of the exam. Is there evidence of a statistically lower prevalence of smoking in the Framingham Offspring Study as compared to the prevalence among all Americans? Step 1: Set up hypotheses and determine the level of significance. H0: p 5 0.211, H1: p , 0.211, a 5 0.05. TABLE 7–5 Test Statistic for Testing H0: p 5 p0 min[np0, z= n(1 2 p0)] $ 5 pˆ − p0 (Find critical value p0 (1 − p0) n in Table 1C) Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–5. We must first check that the sample size is adequate. Specifically, we need to check min[np0, n(12p0)] 5 min[3536(0.211), 3536(1 2 0.211)] 5 min(746, 2790) 5 746 The sample size is more than adequate so the following formula can be used: z= pˆ − p0 p0 (1 − p0) n . Step 3: Set up the decision rule. This is a lower-tailed test, using a z statistic and a 5% level of significance. The appropriate critical value can be found in Table 1C in the Appendix, and the decision rule is Reject H0 if z #21.645. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2: z= pˆ − p0 p0 (1 − p0) n = 0.136 − 0.211 0.211(1 − 0.211) 3536 = −10.93. Step 5: Conclusion. We reject H0 because 210.93 , 21.645. We have statistically significant evidence at a 5 0.05 to show that the prevalence of smoking in the Framingham Offspring is lower than the national prevalence (21.1%). Using Table 1C in the Appendix, the p-value is p , 0.0001. Example 7.5. The NCHS report indicated that in 2002, 75% of children aged 2 to 17 saw a dentist in the past year. An investigator wants to assess whether use of dental services is different in children living in the city of Boston. A sample of 125 children aged 2 to 17 living in Boston is surveyed and 64 reported seeing a dentist over the past 12 months. Is there a significant difference in the proportion of children living in Boston who use dental services as compared to the national proportion? Step 1: Set up hypotheses and determine the level of significance. H0: p 5 0.75, H1: p 0.75, a 5 0.05. Tests with One Sample, CategoricalWhat and Ordinal Is Public Outcomes Health? Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–5. We must first check that the sample size is adequate. Specifically, we need to check min[np0, n(1 2 p0)] 5 min[125(0.75), 125(1 2 0.75)] 5 min(94, 31) 5 31. The sample size is more than adequate, so the following formula can be used: pˆ − p0 z= . p0 (1 − p0) n Step 3: Set up the decision rule. This is a two-tailed test, using a z statistic and a 5% level of significance. The appropriate critical value can be found in Table 1C in the Appendix and the decision rule is Reject H0 if z # 21.960 or if z $ 1.960. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. The sample proportion is pˆ 5 64 / 125 5 0.512, and pˆ − p0 0.512 − 0.75 = −6.15. = z= 0.75(1 − 0.75) 125 p0 (1 − p0) n Step 5: Conclusion. We reject H0 because 26.15 ,21.960. We have evidence at a 5 0.05 to show that there is a statistically significant difference in the proportion of children living in Boston who use dental services as compared to the national proportion. Using Table 1C in the Appendix, the p-value is p , 0.0001. 7.4 TESTS WITH ONE SAMPLE, CATEGORICAL AND ORDINAL OUTCOMES Hypothesis testing with a categorical or ordinal outcome variable in a single population is again performed according to the five-step procedure. Similar to tests for means and proportions described in Section 7.2 and Section 7.3, a key component is setting up the null and research hypotheses. Categorical and ordinal variables are variables that take on more than two distinct responses or categories, and responses can be ordered or unordered (i.e., ordinal or categorical). The procedure we describe here can be used for ordinal or categorical outcomes, and the objective is to compare the distribution of responses—or the proportions of participants in each response category—to a known distribution. The known 135 distribution is derived from another study or report and it is again important in setting up the hypotheses that the comparator distribution specified in the null hypothesis is a fair comparison. In one-sample tests for a categorical or ordinal outcome, we set up our hypotheses against an appropriate comparator. We select a sample and compute descriptive statistics on the sample data using the techniques described in Chapter 4. Specifically, we compute the sample size (n) and the proportions of participants in each response category (pˆ 1, pˆ 2, ..., pˆ k) where k represents the number of response categories. We then determine the appropriate test statistic (Step 2) for the hypothesis test. The formula for the test statistic is in Table 7–6. With the x 2 statistic, we compare the observed frequencies in each response category (O) to the frequencies we would expect (E) if the null hypothesis were true. These expected frequencies are determined by allocating the sample to the response categories according to the distribution specified in H0. This is done by multiplying the observed sample size (n) by the proportions specified in the null hypothesis (p10, p20, ..., pk0). To ensure that the sample size is appropriate for the use of the test statistic in Table 7–6, we need to ensure that the expected frequency in each response category is at least 5, or min(np10, np20,..., npk0) $ 5. The test of hypothesis with a categorical or ordinal outcome measured in a single sample, where the goal is to assess whether the distribution of responses follows a known distribution, is called the x2 goodness-of-fit test. As the name indicates, the idea is to assess whether the distribution of responses in the sample “fits” a specified population distribution. In the next example, we illustrate the test using the five-step approach. As we work through the example, we provide additional details related to the use of this new test statistic. TABLE 7–6 Test Statistic for Testing H0: p1 5 p10, p2 5 p20, …, pk 5 pk0 χ2 = ∑ (O − E)2 E (Find critical value in Table 3, df 5 k 2 1) 136 CHAPTER 7 Hypothesis Testing Procedures Example 7.6. A university conducted a survey of its recent graduates to collect demographic and health information for future planning purposes, as well as to assess students' satisfaction with their undergraduate experiences. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many felt their nutrition was poor, and a substantial number were smoking. In response to a question on regular exercise, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically, and 15% reported exercising regularly as undergraduates. The next year, the university launched a health promotion campaign on campus in an attempt to increase health behaviors among undergraduates. The program included modules on exercise, nutrition, and smoking cessation. To evaluate the impact of the program, the university again surveyed graduates and asked the same questions. The survey is completed by 470 graduates and the data shown in Table 7–7 are collected on the exercise question. Based on the data, is there evidence of a shift in the distribution of responses to the exercise question following the implementation of the health promotion campaign on campus? Run the test at a 5% level of significance. In this example, we have one sample and an ordinal outcome variable (with three response options). We specifically want to compare the distribution of responses in the sample to the distribution reported the previous year (i.e., 60%, 25%, and 15% reporting no, sporadic, and regular exercise, respectively). We now run the test using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. The null hypothesis again represents the “no change” or “no difference” situation. If the health promotion campaign has no impact, then we expect the distribution of responses to the exercise question to be the same as that measured prior to the implementation of the program. H0: p1 5 0.60, p2 5 0.25, p3 5 0.15 or, equivalently, TABLE 7–7 Survey Results on Regular Exercise Number of students No Regular Exercise Sporadic Exercise Regular Exercise Total 255 125 90 470 H0: Distribution of responses is 0.60, 0.25, 0.15, H1: H0 is false, a 5 0.05. Notice that the research hypothesis is written in words rather than in symbols. The research hypothesis as stated captures any difference in the distribution of responses from that specified in the null hypothesis. We do not specify a specific alternative distribution; instead, we are testing whether the sample data “fit” the distribution in H0 or not. With the x2 goodness-of-fit test, there is no upper- or lowertailed version of the test. Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–6 and is χ2 = ∑ (O − E)2 E . We must first assess whether the sample size is adequate. Specifically, we need to check min(np10,...,npk0) $ 5. The sample size here is n 5 470 and the proportions specified in the null hypothesis are 0.60, 0.25, and 0.15. Thus, min[470(0.60), 470(0.25), 470(0.15)] 5 min(282, 117.5, 70.5) 5 70.5. The sample size is more than adequate, so the test statistic can be used. Step 3: Set up the decision rule. The decision rule for the x2 test is set up in a similar way to decision rules we established for z and t tests. The decision rule depends on the level of significance and the degrees of freedom, defined as df 5 k 2 1, where k is the number of response categories. Again, with x2 tests there are no upperor lower-tailed versions of the test. If the null hypothesis is true, the observed and expected frequencies are close in value and the x2 statistic is close to 0. If the null hypothesis is false, then the x2 statistic is large. The rejection region for the x2 test is always in the upper-tail (right), as shown in Figure 7–6. Table 3 in the Appendix contains critical values for the x2 test, indexed by degrees of freedom and the desired level of significance. Here we have df 5 k21 5 321 5 2 and a 5% level of significance. The appropriate critical value from Table 3 is 5.99, and the decision rule is Reject H0 if x2 $ 5.99. Tests with One Sample, CategoricalWhat and Ordinal Is Public Outcomes Health? 137 FIGURE 7–6 Rejection Region for x2 Test with a 5 0.05 and df 5 2 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 =0.05 0 2 0 4 6 Step 4: Compute the test statistic. We now compute the expected frequencies using the sample size and the proportions specified in the null hypothesis. We then substitute the sample data (observed frequencies) and the expected frequencies into the formula for the test statistic identified in Step 2. The computations can be organized as shown in Table 7–8. Notice that the expected frequencies are taken to one decimal place and that the sum of the observed frequencies is equal to the sum of the expected frequencies. The test statistic is computed as χ2 = (255 − 282)2 282 + (125 − 177.5)2 117.5 + (90 − 70.5)2 70.5 , x2 5 2.59 1 0.48 1 5.39 5 8.46. 8 10 12 14 Step 5: Conclusion. We reject H0 because 8.46 . 5.99. We have statistically significant evidence at a 5 0.05 to show that H0 is false, or that the distribution of responses is not 0.60, 0.25, and 0.15. Using Table 3 in the Appendix, we can approximate the p-value. We need to look at critical values for smaller levels of significance with df 5 2. Using Table 3 in the Appendix, the p-value is p , 0.025. In the x2 goodness-of-fit test, we conclude that either the distribution specified in H0 is false (when we reject H0) or that we do not have sufficient evidence to show that the distribution specified in H0 is false (when we fail to reject H0). In Example 7.6, we reject H0 and conclude that the distribution of r esponses to the exercise question following the implementation of the health promotion campaign is not the same as TABLE 7–8 Computing Expected Frequencies No Regular Exercise Sporadic Exercise Regular Exercise Total Observed frequencies (O) 255 125 90 470 Expected frequencies (E) 470(0.60) 5 282 470(0.25) 5 117.5 470(0.15) 5 70.5 470 138 CHAPTER 7 Hypothesis Testing Procedures the distribution prior. The test itself does not provide details of how the distribution has shifted. A comparison of the observed and expected frequencies provides some insight into the shift (when the null hypothesis is rejected). In Example 7.6, we observe the data shown in the first row of Table 7–8. If the null hypothesis is true, we would have expected more students to fall in the No Regular Exercise category and fewer in the Regular Exercise category. In the sample, 255 / 470 5 54% report no regular exercise and 90 / 470 5 19% report regular exercise. Thus, there is a shift toward more regular exercise following the implementation of the health promotion campaign. Example 7.7. The NCHS provided data on the distribution of weight (in categories) among Americans in 2002. The distribution was based on specific values of body mass index (BMI) computed as weight in kilograms over height in meters squared. Underweight was defined as a BMI less than 18.5, normal weight as a BMI between 18.5 and 24.9, overweight as a BMI between 25 and 29.9, and obese as a BMI of 30 or greater. Americans in 2002 were distributed as follows: 2% underweight, 39% normal weight, 36% overweight, and 23% obese. Suppose we want to assess whether the distribution of BMI is different in the Framingham offspring sample. Using data from the n 5 3536 participants who attended the seventh e xamination of the offspring in the Framingham Heart Study, we create the BMI categories as defined and are the data shown in Table 7–9. We must assess whether the sample size is adequate. Specifically, we need to check min(np10, ..., npk0) $ 5. The sample size here is n 5 3326 and the proportions specified in the null hypothesis are 0.02, 0.39, 0.36, and 0.23. Thus, min[3326(0.02), 3326(0.39), 3326(0.36), 3326(0.23)] 5 min(66.5, 1297.1, 1197.4, 765.0) 5 66.5. The sample size is more than adequate, so the test statistic can be used. Step 3: Set up the decision rule. Here we have df 5 k – 1 5 4 – 1 5 3 and a 5% level of significance. The appropriate critical value from Table 3 in the Appendix is 7.81, and the decision rule is Reject H0 if x2 $ 7.81. Step 4: Compute the test statistic. We now compute the expected frequencies using the sample size and the proportions specified in the null hypothesis. We then substitute the sample data (observed frequencies) into the formula for the test statistic identified in Step 2. We organize the computations in Table 7–10. The test statistic is computed as Step 1: Set up hypotheses and determine the level of significance. χ2 = H0: p1 5 0.02, p2 5 0.39, p3 5 0.36, p4 5 0.23 or, e quivalently, (20 − 66.5)2 66.5 (1374 − 1197.4)2 1197.4 H0: Distribution of responses is 0.02, 0.39, 0.36, 0.23, + + (932 − 1297.1)2 1297.1 (1000 − 765.0)2 765.0 + , x2 5 32.52 1 102.77 1 26.05 1 72.19 5 233.53. H1: H0 is false, Step 5: Conclusion. a 5 0.05. We reject H0 because 233.53 . 7.81. We have statistically significant evidence at a 5 0.05 to show that H0 is false or that the distribution of BMI in the Framingham offspring is different from the national data reported in 2002. Using Table 3 in the Appendix, we can approximate the p-value. We need Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–6 and is (O − E)2 χ2 = ∑ . E TABLE 7–9 Distribution of BMI in Framingham Offspring Number of participants Underweight BMI < 18.5 Normal Weight BMI 18.5–24.9 Overweight BMI 25.0–29.9 Obese BMI 301 Total 20 932 1374 1000 3326 Tests with Two Independent Samples, What Continuous Is PublicOutcome Health? 139 TABLE 7–10 Observed and Expected Frequencies Underweight BMI < 18.5 Normal Weight BMI 18.5–24.9 Overweight BMI 25.0–29.9 Obese BMI 301 Total Observed frequencies (O) 20 932 1374 1000 3326 Expected frequencies (E) 66.5 1297.1 1197.4 765.0 3326 to look at critical values for smaller levels of significance for df 5 3. Using Table 3, the p-value is p , 0.005. Again, the x2 goodness-of-fit test allows us to assess whether the distribution of responses “fits” a specified distribution. In Example 7.7, we show that the distribution of BMI in the Framingham Offspring Study is different from the national distribution. To understand the nature of the difference, we compare observed and expected frequencies or observed and expected proportions (or percentages). In Example 7.7, the frequencies are large because of the large sample size—the observed percentages of patients in the Framingham sample are 0.6% underweight, 28% normal weight, 41% overweight, and 30% obese. In the Framingham Offspring sample, there are higher percentages of overweight and obese persons (41% and 30% in Framingham as compared to 36% and 23% nationally) and lower percentages of underweight and normal weight persons (0.6% and 28% in Framingham as compared to 2% and 39% nationally). 7.5 TESTS WITH TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME There are many applications where it is of interest to compare two independent groups with respect to their mean scores on a continuous outcome. In Chapter 6, we presented techniques to estimate the difference in means. Here we again compare means between groups, but rather than generating an estimate of the difference, we test whether the observed difference (increase, decrease, or difference) is statistically significant or not. In this section, we discuss the comparison of means when the two comparison groups are independent or physically separate. The two groups might be determined by a particular attribute (e.g., sex, history of cardiovascular disease) or might be set up by the investigator (e.g., par- ticipants assigned to receive an experimental drug or placebo). The first step in the analysis involves computing descriptive statistics on each of the two samples using the techniques d escribed in Chapter 4. Specifically, we compute the sample size, mean, and standard deviation in each sample and we d enote these summary s tatistics as follows: n 1, X w1 , and s1 for Sample 1 and n2, X w2 , and s2 for Sample 2. The designation of Sample 1 and Sample 2 is essentially arbitrary. In a clinical trial setting, the convention is to call the treatment Group 1 and the control Group 2. However, when comparing men and women, either group can be 1 or 2. In the two independent samples application with a continuous outcome, the parameter of interest in the test of h ypothesis is the difference in population means, m 12m 2. The null hypothesis is always that there is no difference between groups with respect to means, i.e., H0: m12m2 50. The null h ypothesis can also be written as H0: m15m2. In the research hypothesis, an investigator can hypothesize that the first mean is larger than the second (H 1: m 1.m 2), that the first mean is smaller than the second (H 1: m 1,m 2), or that the means are different (H 1: m 1 m 2). The three different alternatives represent upper-, lower-, and two-tailed tests, respectively. Table 7–11 contains the formulas for test statistics for the difference in population means. In the formulas in Table 7–11, X w1 and X w2 are the means of the outcome in the independent samples, and Sp is the pooled estimate of the common standard deviation (again assuming that the variances in the populations are similar) computed as the weighted average of the standard deviations in the samples, Sp = (n1 − 1) s12 + (n2 − 1) s22 . n1 + n2 − 2 140 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–11 Test Statistics for Testing H0: m1 5 m2 n1 $ 30 and n2 $ 30 n1 , 30 or n2 , 30 z= X1 − X 2 S p 1 n1 + 1 n2 (Find critical value of z in Table 1C) t= X1 − X 2 S p 1 n1 + 1 n2 (Find critical value of t in Table 2, df 5 n1 1 n22 2) Because we are assuming equal variances between groups, we pool the information on variability (sample variances) to generate an estimate of the variability in the population. As a guideline, if the ratio of the sample variances is between 0.5 and 2, the assumption of equality of population variances is taken to be appropriate. (Note that because Sp is a weighted average of the standard deviations in the sample, Sp is always between s1 and s2.) Example 7.8. In Example 6.5, we used data presented in Chapter 4 on n 5 3539 participants who attended the seventh examination of the offspring in the Framingham Heart Study and constructed a 95% confidence interval for the difference in mean systolic blood pressures between men and women. Table 7–12 contains summary statistics on the characteristics measured in men and women. Suppose we now wish to assess whether there is a statistically significant difference in mean systolic blood pressures between men and women using a 5% level of significance. Step 1: Set up hypotheses and determine the level of significance. H0: m1 5 m2, H1: m1 m2, a 5 0.05. Step 2: Select the appropriate test statistic. Because both samples are large (n1 $ 30 and n2 $ 30), we use the z test statistic as opposed to t. Before implementing the TABLE 7–12 Summary Statistics in Men and Women Men Women n X w s n X w s Systolic blood pressure 1623 128.2 17.5 1911 126.5 20.1 Diastolic blood pressure 1622 75.6 9.8 1910 72.6 9.7 Total serum cholesterol 1544 192.4 35.2 1766 207.1 36.7 Weight (lbs) 1612 194.0 33.8 1894 157.7 34.6 Height (in.) 1545 68.9 2.7 1781 63.4 2.5 Body mass index (BMI) 1545 28.8 4.6 1781 27.6 5.9 Tests with Two Independent Samples, What Continuous Is PublicOutcome Health? formula, we first check whether the assumption of equality of population variances is reasonable. The guideline suggests investigating the ratio of the sample variances, s12 / s 22 . Suppose we call the men Group 1 and the women Group 2. Again, this is arbitrary; it needs to be noted only when interpreting the results. The ratio of the sample variances is 17.52 / 20.12 5 0.76, which falls in between 0.5 and 2, suggesting that the assumption of equality of population variances is reasonable. The appropriate test statistic is z= X1 − X 2 . S p 1 n1 + 1 n2 Step 3: Set up the decision rule. This is a two-tailed test, using a z statistic and a 5% level of significance. The appropriate critical values can be found in Table 1C in the Appendix and the decision rule is Reject H0 if z # 21.960 or if z $ 1.960. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. Before substituting, we first compute Sp, the pooled estimate of the common standard deviation. Sp = Sp = (n1 − 1) s12 + (n2 − 1) s22 , n1 + n2 − 2 (1623 − 1)(17.5)2 + (1911 − 1)(20.1)2 1623 + 1911 − 2 = 359.12 = 19.0. Notice that the pooled estimate of the common standard deviation, Sp, falls between the standard deviations in the comparison groups (i.e., 17.5 and 20.1). Sp is slightly closer in value to the standard deviation in women (20.1) as there are slightly more women in the sample. Recall that Sp is a weighted average of the standard deviations in the comparison groups, weighted by the respective sample sizes. We now calculate the test statistic, z = 128.2 − 126.5 1.7 = = 2.66. 0.64 19.0 1 / 1623 + 1 / 1911 Step 5: Conclusion. We reject H0 because 2.66 . 1.960. We have statistically significant evidence at a 5 0.05 to show that there is a difference in mean systolic blood pressures between men 141 and women. The p-value can be found in Table 1C in the Appendix and is equal to p , 0.010. In Example 7.8, we find that there is a statistically significant difference in mean systolic blood pressures between men and women at p , 0.010. Notice that there is a very small difference in the sample means (128.22126.5 5 1.7 units) but this difference is beyond what would be expected by chance. The large sample sizes in this example are driving the statistical significance. In Example 6.5, we computed a 95% confidence interval for the difference in mean systolic blood pressures as 1.7 6 1.26, or (0.44, 2.96). The confidence interval provides an assessment of the magnitude of the difference between means, whereas the test of hypothesis and p-value provides an assessment of the statistical significance of the difference. From the confidence interval in Example 6.5, we see that the difference in means is significant at the 5% level of significance because the 95% confidence interval does not include the null value of 0. The formal test is needed to compute the exact statistical significance of the difference or the p-value. In Example 7.3, we analyzed data from a study to evaluate a new drug designed to lower total cholesterol. The study involved one sample of patients, each patient took the new drug for 6 weeks and had his or her cholesterol measured. As a means of evaluating the efficacy of the new drug, the mean total cholesterol following 6 weeks of treatment was compared to a national mean of 203 based on the NCHS. At the end of the example, we discussed the appropriateness of the historical comparator as well as an alternative study design to evaluate the effect of the new drug involving two treatment groups, where one group receives the new drug and the other does not. In Example 7.9, we revisit this example with a concurrent or parallel control group, which is very typical in randomized controlled trials or clinical trials (for more details, see Section 2.3.1). Example 7.9. A new drug is proposed to lower total cholesterol. A randomized controlled trial is designed to evaluate the efficacy of the drug in lowering cholesterol. Thirty participants are enrolled in the trial and are randomly assigned to receive either the new drug or a placebo. The participants do not know which treatment they are assigned. Each participant is asked to take the assigned treatment for 6 weeks. At the end of 6 weeks, each patient's total cholesterol level is measured and the sample statistics are shown in Table 7–13. Is there statistical evidence of a reduction in mean total cholesterol in patients taking the new drug for 6 weeks as compared to participants taking a placebo? We run the test using the five-step approach and call the new drug Group 1 and the placebo Group 2. 142 CHAPTER 7 Hypothesis Testing Procedures Now calculate the test statistic, TABLE 7–13 Cholesterol Levels by Treatment Sample Size Mean Standard Deviation New drug 15 195.9 28.7 Placebo 15 227.4 30.3 Step 1: Set up hypotheses and determine the level of significance. H0: m1 5 m2, H1: m1 , m2, a 5 0.05. Step 2: Select the appropriate test statistic. Because both samples are small (n1 , 30 and n2 , 30), we use the t test statistic. Before implementing the formula, we first check whether the assumption of equality of population variances is reasonable. The ratio of the sample variances, s12 / s22 5 28.72 / 30.32 5 0.90, which falls between 0.5 and 2, suggesting that the assumption of equality of population variances is reasonable. The appropriate test statistic is t= X1 − X 2 . S p 1 n1 + 1 n2 Step 3: Set up the decision rule. This is a lower-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in Table 2 in the Appendix. To determine the critical value of t, we need degrees of freedom, df, defined as df 5 n1 1 n2 2 2. In this example, df 5 15 1 15 2 2 5 28. The critical value for a lower-tailed test with df 5 28 and a 5 0.05 is 21.701, and the decision rule is Reject H0 if t # 21.701. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. Before substituting, we first compute Sp, the pooled estimate of the common standard deviation: Sp = (15 − 1)(28.7)2 + (15 − 1)(30.3)2 15 + 15 − 2 = 870.89 = 29.5. t= 195.9 − 227.4 −31.5 = = = −2.92. 29.5 1 15 + 1 15 10.77 Step 5: Conclusion. We reject H0 because 22.92 , 21.701. We have statistically significant evidence at a 5 0.05 to show that the mean total cholesterol level is lower in patients taking the new drug for 6 weeks as compared to patients taking a placebo. Using Table 2 in the Appendix, the p-value is p , 0.005. The clinical trial in Example 7.9 finds a statistically significant reduction in total cholesterol, whereas in Example 7.3 we did not demonstrate the efficacy of the new drug. Notice that the mean total cholesterol level in patients taking a placebo is 227.4, which is different from the mean cholesterol of 203 reported among all Americans in 2002 and used as the comparator. The historical control value may not have been the most appropriate comparator as cholesterol levels have been increasing over time. In Section 7.6, we present another design that can be used to assess the efficacy of the new drug. 7.6 TESTS WITH MATCHED SAMPLES, CONTINUOUS OUTCOME An alternative study design to that described in the previous section (to compare two groups with respect to their mean scores on a continuous outcome) is one based on matched or paired samples. The two comparison groups are said to be dependent (matched or paired) and the data can arise from a single sample of p articipants where each participant is measured twice, possibly before and after an intervention, or from two samples that are matched or paired on one or more specific characteristics (e.g., siblings). When the samples are dependent, we focus on difference scores in each participant or between members of a pair, and the test of hypothesis is based on the mean difference, md. The null hypothesis again reflects “no difference” and is stated as md 5 0. Note that there are some instances where it is of interest to test whether there is a difference of a particular magnitude (e.g., md 5 5) but in most instances, the null hypothesis reflects no d ifference (i.e., md 5 0). The appropriate formula for the test of hypothesis depends on the sample size. The formulas are shown in Table 7–14 and are identical to those we presented for hypothesis testing with one sample and a continuous outcome presented in Section 7.2, except here we focus on difference scores. Tests with Matched Samples, Continuous What Is PublicOutcome Health? 143 Step 4: Compute the test statistic. TABLE 7–14 Test Statistics for Testing H0: md 5 0 n $ 30 n , 30 z= t= X d − µd sd n (Find critical value of z in Table 1C) X d − µd sd n (Find critical value of t in Table 2, df 5 n 2 1) t= Example 7.10. In Example 6.7, we compared systolic blood pressures measured at the sixth and seventh examinations (approximately 4 years apart) of the Framingham Offspring Study in a subsample of n 5 15 randomly selected participants. The data are shown in Example 6.7, where we generated a 95% confidence interval for the mean difference in systolic blood pressures over a 4-year period. Using the same data, we now test whether there is a statistically significant difference in systolic blood pressures over 4 years using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: md 5 0, H1: md 0, a 5 0.05. Step 2: Select the appropriate test statistic. Because the sample is small (n , 30), we use the t test statistic, t= We now substitute the sample data into the formula for the test statistic identified in Step 2. In Example 6.7, we had: n 5 15, X wd 5 25.3, and sd 5 12.8. The test statistic is X d − µd sd n Step 3: Set up the decision rule. This is a two-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in Table 2 in the Appendix with degrees of freedom, df, defined as df 5 n 2 1 5 15 2 1 5 14. The critical value is 2.145 and the decision rule is Reject H0 if t # 22.145 or if t $ 2.145. X d − µd −5.3 − 0 = = −1.60. sd n 12.8 15 Step 5: Conclusion. We do not reject H0 because 22.145 , 21.60 , 2.145. We do not have statistically significant evidence at a 5 0.05 to show that there is a difference in systolic blood pressures over time. In Example 6.7, we estimated a confidence interval and were 95% confident that the mean difference in systolic blood pressures between Examination 6 and Examination 7 (approximately 4 years apart) was between 212.4 and 1.8. Because the null value of the confidence interval for the mean difference is 0, we concluded that there was no statistically significant difference in blood pressures over time because the confidence interval for the mean difference included 0. The test of hypothesis gives the same result. In Example 7.11, we revisit Example 7.3 and Example 7.9 where we evaluated a new drug designed to lower total cholesterol. In Example 7.3, we collected data on a sample of patients who took the new drug and we compared their mean total cholesterol level to a historical control. In Example 7.9, we evaluated the efficacy of the new drug using a clinical trial with a concurrent or parallel placebo control group. In Example 7.11, we again evaluate the efficacy of the new drug using a matched design. Example 7.11. A new drug is proposed to lower total cholesterol and a study is designed to evaluate the efficacy of the drug in lowering cholesterol. Fifteen patients agree to participate in the study and each is asked to take the new drug for 6 weeks. However, before starting the treatment, each patient's total c holesterol level is measured. The initial measurement is a pre-treatment or baseline value. After taking the drug for 6 weeks, each patient's total cholesterol level is measured again and the data are shown in Table 7–15. The right column contains difference scores for each patient, computed by subtracting the 6-week cholesterol level from the baseline level. The differences represent the reduction in total cholesterol over 6 weeks. (The differences could have been computed by subtracting the baseline total cholesterol level from the level measured at 6 weeks. The way in which the differences are computed does not affect the outcome of the analysis, only the interpretation.) 144 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–16 Summary Statistics on TABLE 7–15 Differences in Cholesterol Difference Scores over 6 Weeks Difference Subject Identification Number Difference Difference2 205 10 1 10 100 190 156 34 2 34 1156 230 190 40 3 40 1600 4 220 180 40 4 40 1600 5 214 201 13 5 13 169 6 240 227 13 6 13 169 7 210 197 13 7 13 169 8 193 173 20 8 20 400 9 210 204 6 9 6 36 10 230 217 13 10 13 169 11 180 142 38 11 38 1444 12 260 262 22 12 22 4 13 210 207 3 13 3 9 14 190 184 6 14 6 36 15 200 193 7 15 7 49 254 7110 Subject Identification Number Baseline 6 Weeks 1 215 2 3 Total Because the differences are computed by subtracting the cholesterol levels measured at 6 weeks from the baseline values, positive differences indicate reductions and negative differences indicate increases (e.g., Participant 12 increases by 2 units over 6 weeks). The goal here is to test whether there is a statistically significant reduction in cholesterol. Because of the way in which we compute the differences, we want to look for an increase in the mean difference (i.e., a positive reduction). To conduct the test, we need to summarize the differences. In this sample, we have n 5 15, X wd 5 16.9, and sd5 14.2, respectively. The data necessary to compute the sample statistics is shown in Table 7–16. ∑ Differences 254 = = 16.9 Xd = n 15 and sd = ∑ Differences 2 − (∑ Differences) n , n −1 sd = 7110 − (254) 15 = 200.64 = 14.2 15 − 1 2 2 Is there statistical evidence of a reduction in mean total cholesterol in patients after using the new medication for 6 weeks? We run the test using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: md 5 0, H1: md . 0, a 5 0.05. Note that if we had computed differences by subtracting the baseline level from the level measured at 6 weeks, then negative differences would have reflected reductions and the research hypothesis would have been H1: md , 0. Step 2: Select the appropriate test statistic. Because the sample size is small (n , 30), the appropriate test statistic is X d − µd t= . sd n Tests with Two Independent Samples, What Dichotomous Is PublicOutcome Health? Step 3: Set up the decision rule. This is an upper-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in Table 2 in the Appendix. To determine the critical value of t we need degrees of freedom, df, defined as df 5 n 2 1. In this example, df 5 15 2 1 5 14. The critical value for an upper-tailed test with df 5 14 and a 5 0.05 is 1.761, and the decision rule is Reject H0 if t $ 1.761. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2: t= X d − µd 16.9 − 0 = = 4.61. sd n 14.2 15 Step 5: Conclusion. We reject H0 because 4.61 . 1.761. We have statistically significant evidence at a 5 0.05 to show that there is a reduction in cholesterol levels over 6 weeks. Using Table 2 in the Appendix, the p-value is p , 0.005. In Example 7.9 and Example 7.11, using parallel samples and matched designs, respectively, we find statistically significant reductions in total cholesterol. In Example 7.3, using a historical comparator, we do not. It is extremely important to design studies that are best suited to detect a meaningful difference when one exists. There are often several alternatives, and investigators work with biostatisticians to determine the best design for each application. It is worth noting that the matched design used in Example 7.11 can be problematic in that observed differences may only reflect a “placebo” effect. All participants took the assigned medication but is the observed reduction attributable to the medication or a result of participation in a study? 7.7 TESTS WITH TWO INDEPENDENT SAMPLES, DICHOTOMOUS OUTCOME Techniques presented in Section 7.5 are appropriate when there are two independent comparison groups and the outcome of interest is continuous (e.g., blood pressure, total cholesterol, weight loss). Here we consider the situation where there are two independent comparison groups and the outcome of interest is dichotomous (e.g., success/failure). The goal of the analysis is to compare proportions of successes between the two groups. The relevant sample data are the 145 sample sizes in each comparison group (n1 and n2) and the sample proportions (pˆ1and pˆ2), which are computed by taking the ratios of the numbers of successes to the sample sizes in each group, e.g., pˆ1 5 x1 /n1 and pˆ2 5 x2/n2. In Chapter 3, we introduced the risk difference, relative risk, and odds ratio, which are different measures to compare proportions in two independent groups. We developed confidence intervals for each of these measures in Chapter 6 in Section 6.6.1, Section 6.6.2, and Section 6.6.3, respectively. The confidence interval formulas were different and depended on the specific measure. In tests of hypothesis comparing proportions between two independent groups, one test is performed and it can be interpreted relative to a risk difference, relative risk, or odds ratio. As a reminder, the risk difference is computed by taking the difference in proportions between comparison groups, the relative risk is computed by taking the ratio of proportions, and the odds ratio is computed by taking the ratio of the odds of success in the comparison groups (see Section 3.4.2 for more details). Because the null values for the risk difference, the relative risk, and the odds ratio are different, the hypotheses in tests of hypothesis look slightly different depending on which measure is used. When performing tests of hypothesis for the risk difference, relative risk, or odds ratio, the convention is to label the exposed or treated Group 1 and the unexposed or control Group 2. For example, suppose a study is designed to assess whether there is a significant difference in proportions in two independent comparison groups. The test of interest is H0: p1 5 p2 versus H1: p1 p2. The following are the hypotheses for testing for a difference in proportions using the risk difference (RD), the relative risk (RR), and the odds ratio (OR) as measures of effect. First, the previous hypotheses are equivalent to H0: p1 2 p2 5 0 versus H1: p1 2 p2 0, which are, by definition, equal to H0: RD 5 0 versus H1: RD 0. If an investigator wants to focus on the relative risk, the equivalent hypotheses are H0: RR 5 1 versus H1: RR 1, 146 CHAPTER 7 Hypothesis Testing Procedures and if the investigator wants to focus on the odds ratio, the equivalent hypotheses are or that the proportions are different, H1: p1 p2, which is equivalent to H1: RD 0, H1: RR 1, and H1: OR 1. H0: OR 5 1 versus H1: OR 1. Suppose a test is performed to test The three different alternatives represent upper-, lower-, and two-tailed tests, respectively. The formula for the test statistic for the difference in proportions is given in Table 7–17. Note that pˆ 1 and pˆ 2 are the proportions of successes in Groups 1 and 2, respectively. pˆ is the overall proportion of successes, which is computed as x + x2 pˆ = 1 . The preceding formula is appropriate for large n1 + n2 samples, defined as at least five successes (npˆ) and at least five failures [n(1 2 pˆ)] in each of the two samples. If there are fewer than five successes or failures in e ither comparison group, then alternative procedures called exact methods must be used to test whether there is a difference in p opulation proportions.3,4 Example 7.12. In Example 6.9, we analyzed data from n 5 3799 participants who attended the fifth examination of the offspring in the Framingham Heart Study. The outcome of interest was prevalent CVD and we compared prevalent CVD between participants who were and were not currently smoking cigarettes at the time of the fifth examination of the Framingham Offspring Study. The data are shown in Table 7–18. We now use the data to test if the prevalence of CVD is significantly different in smokers as compared to nonsmokers. The prevalence of CVD (or proportion of participants with prevalent CVD) among nonsmokers is 298 / 3055 5 0.0975, and the prevalence of CVD among current s mokers is 81 / 744 5 0.1089. Here smoking status defines the c omparison groups, and we call the current smokers Group 1 (exposed) and the nonsmokers (unexposed) Group 2. The test of hypothesis is conducted on the following page using the five-step approach. H0: RD 5 0 versus H1: RD 0, and the test rejects H0 at a 5 0.05. Based on this test, we can conclude that there is significant evidence (a 5 0.05) of a difference in proportions, significant evidence that the risk difference is not 0, and significant evidence that the relative risk and odds ratio are not 1. The risk difference is analogous to the difference in means when the outcome is continuous. The test for the difference in means was described in Section 7.5. Here the parameter of interest is the difference in proportions in the population, RD 5 p1 2 p2, and the null value for the risk difference is 0. In a test of hypothesis for the risk difference, the null hypothesis is always H0: RD 5 0. This is equivalent to H0: RR 5 1 and H0: OR 5 1. In the research hypothesis, the investigator can hypothesize that the first proportion is larger than the second, H1: p1 > p2, which is equivalent to H1: RD . 0, H1: RR . 1, and H1: OR . 1, that the first proportion is smaller than the second, H1: p1, p2, which is equivalent to H1: RD , 0, H1: RR , 1, and H1: OR , 1, TABLE 7–17 Test Statistic for Testing H0: p1 5 p2 min[n1 ˆp1, n1(1 2 ˆp1), n2 ˆp2, n2(1 2 ˆp2)] $ 5 z= pˆ1 − pˆ2 (pˆ (1 − pˆ) 1 n + 1 n 1 2) (Find critical value of z in Table 1C) 147 Tests with Two Independent Samples, What Dichotomous Is PublicOutcome Health? z= TABLE 7–18 Prevalent CVD in Smokers and Nonsmokers Nonsmoker Current smoker Total Free of CVD History of CVD Total 2757 298 3055 663 81 744 3420 379 3799 Step 1: Set up hypotheses and determine the level of significance. H0: p1 5 p2, H1: p1 p2, a 5 0.05. Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–17. We must first check that the sample size is adequate. Specifically, we need to ensure that we have at least five successes and five failures in each comparison group. In this example, we have more than enough successes (cases of prevalent CVD) and failures (persons free of CVD) in each comparison group. The sample size is more than adequate, so the following formula can be used: z= pˆ1 + pˆ2 (pˆ (1 − pˆ) 1n + 1n 1 2) . Step 3: Set up the decision rule. This is a two-tailed test, using a z statistic and a 5% level of significance. The appropriate critical value can be found in Table 1C in the Appendix and the decision rule is Reject H0 if z # 21.960 or if z $ 1.960. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. We first compute the overall proportion of successes: pˆ 5 x1 1 x2 81 1 298 379 5 5 5 0.0998. n 1 1 n2 744 1 3055 3799 We now substitute to compute the test statistic, 0.1089 − 0.0975 0.0988(1 − 0.0988)(1 744 + 1 3055) = 0.0114 5 0.927. 0.0123 Step 5: Conclusion. We do not reject H0 because 21.960 , 0.927 , 1.960. We do not have statistically significant evidence at a 5 0.05 to show that there is a difference in prevalent CVD between smokers and nonsmokers. In Example 6.9, we estimated the 95% confidence interval for the difference in prevalent CVD (or risk difference) between smokers and nonsmokers as 0.0114 6 0.0247, or between 20.0133 and 0.0361. Because the 95% confidence interval for the risk difference included 0, we could not conclude that there was a statistically significant difference in prevalent CVD between smokers and nonsmokers. This is consistent with the test of hypothesis result. In Example 6.11, we used the same data and estimated the relative risk along with a 95% confidence interval for the relative risk. We estimated that among smokers the risk of CVD was 1.12 times that among nonsmokers with a 95% confidence interval of 0.89 to 1.41. Because this confidence interval included the null value of 1, we could not conclude that there was a difference in prevalent CVD between smokers and nonsmokers. In Example 6.13, we again used the same data and estimated the odds ratio along with a 95% confidence interval for the odds ratio. We estimated that among smokers the odds of CVD was 1.13 times that among nonsmokers with a 95% confidence interval estimate of 0.87 to 1.47. Because this confidence interval included the null value of 1, we could not conclude that there was a difference in prevalent CVD between smokers and nonsmokers. Smoking has been shown to be a risk factor for cardiovascular disease. What might explain the fact that we did not observe a statistically significant difference using data from the Framingham Heart Study? (Hint: Here we consider prevalent CVD. Would the results have been different if we considered incident CVD?) Example 7.13. In Example 6.10, we analyzed data from a randomized trial designed to evaluate the effectiveness of a newly developed pain reliever designed to reduce pain in patients following joint replacement surgery. The trial compared the new pain reliever to the pain reliever currently in use (called the standard of care). A total of 100 patients undergoing joint replacement surgery agreed to participate 148 CHAPTER 7 Hypothesis Testing Procedures in the trial. Patients were randomly assigned to receive either the new pain reliever or the standard pain reliever following surgery and were blind to the treatment assignment. Before receiving the assigned treatment, patients were asked to rate their pain on a scale of 0 to 10, with higher scores indicative of more pain. Each patient was then given the assigned treatment and after 30 minutes was again asked to rate his or her pain on the same scale. The primary outcome was a reduction in pain of 3 or more scale points (defined by clinicians as a clinically meaningful reduction). The data shown in Table 7–19 were observed in the trial. We now test whether there is a statistically significant difference in the proportions of patients reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: p1 5 p2, The sample size is adequate, so the following formula can be used: pˆ1 + pˆ2 z= . pˆ (1 − pˆ) 1n + 1n (2) Step 3: Set up the decision rule. This is a two-tailed test, using a z statistic and a 5% level of significance. The appropriate critical value can be found in Table 1C in the Appendix and the decision rule is Reject H0 if z # 21.960 or if z $ 1.960. Step 4: Compute the test statistic. We now substitute the sample data into the formula for the test statistic identified in Step 2. We first compute the overall proportion of successes: H1: p1 p2, ˆp 5 a 5 0.05. 1 x1 1 x2 23 1 11 34 5 5 }} 5 0.34. n 1 1 n2 50 1 50 100 We now substitute to compute the test statistic, Here the new pain reliever is represented as Group 1 and the standard pain reliever as Group 2. Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–17. We must first check that the sample size is adequate. Specifically, we need to ensure that we have at least five successes and five failures in each comparison group or that min[n1pˆ1, n1(1 2 pˆ1), n2 pˆ2, n2(1 2 pˆ2)] $ 5. In this example, we have min[50(0.46), 50(1 2 0.46), 50(0.22), 50(1 2 0.22)] 5 min(23, 27, 11, 39) 5 11. TABLE 7–19 Pain Reduction by Treatment Reduction of 31 Points Treatment Group n Number Proportion New pain reliever 50 23 0.46 Standard pain reliever 50 11 0.22 z= 0.46 − 0.22 (0.34 (1 − 0.34) 1 50 + 1 50 Step 5: Conclusion.) = 0.24 = 2.526. 0.095 We reject H0 because 2.526 . 1960. We have statistically significant evidence at a 5 0.05 to show that there is a difference in the proportions of patients on the new pain reliever reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) as compared to patients on the standard pain reliever. In Example 6.10, we estimated a 95% confidence interval for the difference in proportions of patients on the new pain reliever reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) as compared to patients on the standard pain reliever, and it was 0.24 6 0.18, or between 0.06 and 0.42. Because the 95% confidence interval did not include 0, we concluded that there was a statistically significant difference in proportions, which is consistent with the test of hypothesis result. In Example 6.12, we used the same data and estimated the relative risk along with a 95% confidence interval for the relative risk. We estimated that patients receiving the new pain reliever had 2.09 times the likelihood of reporting a meaningful reduction in pain as compared to patients receiving the standard pain reliever, with a 95% confidence interval of 1.14 to 3.82. Because this confidence interval did not include the null value of 1, we concluded that there was a significant difference 149 Tests with More Than Two Independent Samples, What Continuous Is PublicOutcome Health? in the relative risk of a meaningful reduction in pain in patients on the new and standard pain relievers. In Example 6.14, we used the same data and estimated the odds ratio along with a 95% confidence interval for the odds ratio. We estimated that the odds that patients receiving the new pain reliever reported a meaningful reduction in pain was 3.02 times the odds that patients on standard pain reliever reported a meaningful reduction in pain, with a 95% confidence interval of 1.26 to 7.21. Because this confidence interval did not include the null value of 1, we concluded that there was a significant difference in the odds of a meaningful reduction in pain in patients on the new and standard pain relievers. The procedures discussed here apply to applications where there are two independent comparison groups and a dichotomous outcome. There are other applications in which it is of interest to compare a dichotomous outcome in matched or paired samples. For example, in a clinical trial we might wish to test the effectiveness of a new antibiotic eyedrop for the treatment of bacterial conjunctivitis. Participants use the new antibiotic eyedrop in one eye and a comparator (placebo or active control treatment) in the other. The success of the treatment (yes/no) is recorded for each participant for each eye. Because the two assessments (success or failure) are paired, we cannot use the procedures discussed here. The appropriate test is called McNemar's test (sometimes called McNemar's test for dependent proportions). For more details, see Agresti3 or Rosner.6 groups. The ANOVA procedure is used to compare the means of the comparison groups and is conducted using the same five-step approach used in the scenarios discussed in previous sections. However, because there are more than two groups, the computation of the test statistic is more involved. The test statistic must take into account the sample sizes, sample means, and sample standard deviations in each of the comparison groups. Before illustrating the computation of the test statistic, we first present the logic of the procedure. Consider an example with four independent groups and a continuous outcome measure. The independent groups might be defined by a particular characteristic of the participants, such as BMI (e.g., underweight, normal weight, overweight, or obese), or by the investigator (e.g., by randomizing participants to one of four competing treatments, call them A, B, C, and D). Suppose that the outcome is systolic blood pressure and we wish to test whether there is a statistically significant difference in mean systolic blood pressures among the four groups. The sample data are summarized as shown in Table 7–20. The hypotheses of interest in an ANOVA are H0: m1 5 m2 5 ... 5 mk , H1: Means are not all equal, where k is the number of independent comparison groups. For the example just described, the hypotheses are H0: m1 5 m2 5 m3 5 m4, H1: Means are not all equal. 7.8 TESTS WITH MORE THAN TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME There are many applications where it is of interest to compare more than two independent groups with respect to their mean scores on a continuous outcome. For example, in some clinical trials there are more than two comparison groups. Suppose a clinical trial is designed to evaluate a new medication for asthma and investigators compare an experimental medication to a placebo and to a standard treatment (i.e., a medication currently being used). In an observational study such as the Framingham Heart Study, it might be of interest to compare mean blood pressures or mean cholesterol levels in persons who are underweight, normal weight, overweight, and obese. The technique to test for a difference in more than two independent means is an extension of the two independent samples procedure discussed in Section 7.5 and is called analysis of variance (ANOVA). The ANOVA technique applies when there are more than two independent comparison The null hypothesis in ANOVA is always that there is no difference in the means. The research or alternative hypothesis is always that the means are not all equal and is usually written in words rather than in mathematical symbols. The research hypothesis captures any difference in means and includes the situation where all four means are unequal, where one is different from the other three, where two are different, and so TABLE 7–20 Summary Statistics for ANOVA Group 1 Group 2 Group 3 Group 4 Sample size n1 n2 n3 n4 Sample mean w X1 w X2 w X3 w X4 Sample standard deviation s1 s2 s3 s4 150 CHAPTER 7 Hypothesis Testing Procedures on. The alternative hypothesis (as shown earlier) captures all possible situations other than equality of all means specified in the null hypothesis. The test statistic for ANOVA is given in Table 7–21. The F statistic is computed by taking the ratio of what is called the “between-treatment” variability to the “residual or error” variability. This is where the name of the procedure originates. In analysis of variance, we are testing for a difference in means (H0: means are all equal versus H1: means are not all equal) by evaluating variability in the data. The numerator captures between-treatment variability (i.e., differences among the sample means), and the denominator contains an estimate TABLE 7–21 Test Statistic for Testing H0: m1 5 m2 5 ... 5 mk ∑n (X − X) F5 ∑ ∑(X − X) j 2 (k − 1) 2 (N − k) j j (Find critical value in Table 4, df1 5 k 2 1, df2 5 N 2 k) where nj 5 the sample size in the jth group (e.g., j 5 1, – 2, 3, and 4 when there are 4 comparison groups), Xj is – the sample mean in the jth group, and X is the overall mean. k represents the number of independent groups (k . 2), and N represents the total number of observations in the analysis. Note that N does not refer to a population size, but instead to the total sample size in the analysis (the sum of the sample sizes in the comparison groups, e.g., N = n1 + n2 + n3 + n4). The test statistic is complicated because it incorporates all of the sample data. While it is not easy to see the extension, the F statistic is a generalization of the test statistic shown in Table 7–11, which is appropriate for exactly two groups. NOTE: Appropriate use of the F statistic is based on several assumptions. First, the outcome is assumed to follow a normal distribution in each of the comparison groups. Second, the data are assumed to be equally variable in each of the k populations (i.e., the population variances are equal, or s 12 5 s 22 5 ... 5 s k2). This means that the outcome is equally variable in each of the comparison populations. This assumption is the same as that assumed for appropriate use of the test statistics in Table 7–11 when there were two comparison groups. It is possible to assess the likelihood that the assumption of equal variances is true and the test can be conducted in most statistical computing packages.7 If the variability in the k comparison groups is not similar, then alternative techniques must be used.8 of the variability in the outcome. The test statistic is a measure that allows us to assess whether the differences among the sample means (numerator) are larger than would be expected by chance if the null hypothesis is true. Recall in the two-- independent samples test (see test statistics in Table 7–11), the test statistics (z or t) were computed by taking the ratio of the difference in sample means (numerator) to the variability in the 1 1 +). n1 n2 The decision rule for the F test in ANOVA is set up in a similar way to decision rules we established for z and t tests. The decision rule again depends on the level of significance and the degrees of freedom. The F statistic has two degrees of freedom. These are denoted df1 and df2, and called the numerator and denominator degrees of freedom, respectively. The degrees of freedom are defined as df1 5 k 2 1 and df2 5 N 2 k, where k is the number of comparison groups and N is the total number of observations in the analysis. Similar to x2 tests, there is no upper- or lower-tailed version of the test. If the null hypothesis is true, the between-treatment variation (numerator) is close in value to the residual or error variation (denominator) and the F statistic is small. If the null hypothesis is false, then the F statistic is large. The rejection region for the F test is always in the upper tail (right) of the distribution, as shown in Figure 7–7. Table 4 in the Appendix contains critical values for the F distribution for tests when a 5 0.05, indexed by df1 and df2. Figure 7–7 is an example for the situation with a 5 0.05, df1 5 3, and df2 5 36. The degrees of freedom are based on an application with four comparison groups (k 5 4) and a sample size of 40 (N 5 40). The appropriate critical value from Table 4 in the Appendix is 2.87, and the decision rule is outcome (estimated by S p Reject H0 if F $ 2.87. We next illustrate the ANOVA procedure using the fivestep approach. Because the computation of the test s tatistic is involved, the computations are often organized in an ANOVA table. The ANOVA table breaks down the components of variation in the data into variation between treatments and error or residual variation. The ANOVA table is set up as shown in Table 7–22. The ANOVA table is organized as follows. The first column is entitled Source of Variation and delineates the between-- treatment and error or residual variation. The total variation is the sum of the between-treatment and error variation. The second column is entitled Sums of Squares (SS). The between-treatment sums of squares is SSB 5 onj(X wj 2 X w)2, Tests with More Than Two Independent Samples, What Continuous Is PublicOutcome Health? 151 FIGURE 7–7 Rejection Region for F Test with a 5 0.05, df1 5 3 and df2 5 36 (k 5 4, N 5 40) 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 =0.05 0.1 0 1 0 2 3 and is computed by summing the squared differences between – – each treatment (or group) mean (Xj) and the overall mean (X). The squared differences are weighted by the sample sizes per group (nj). The error sums of squares is SSE 5 oo(X 2 X wj)2, and is computed by summing the squared differences be- tween each observation (X) and its group mean (X j) (i.e., the squared differences between each observation in Group 1 and the Group 1 mean, the squared differences between 4 5 6 7 8 each observation in Group 2 and the Group 2 mean, and so on). The double summation (oo) indicates summation of the squared differences within each treatment and then summation of these totals across treatments to produce a single value. (This is illustrated in the following examples.) The total sums of squares is SST 5 oo(X 2 X w)2, and is computed by summing the squared differences between each observation (X) and the overall sample mean TABLE 7–22 ANOVA Table Source of Variation Between treatments Error or residual Total Sums of Squares (SS) Degrees of Freedom (df) Mean Squares (MS) – – SSB 5 onj(Xj 2X)2 k21 – SSE 5 oo(X 2 Xj)2 N2k – SST 5 oo(X 2 X)2 N21 MSB 5 SSB k −1 MSE 5 SSB N−k F MSB F 5 }} MSE 152 CHAPTER 7 Hypothesis Testing Procedures (X w). In an ANOVA, data are organized by comparison or treatment groups. If all of the data are pooled into a single sample, SST is the numerator of the sample variance computed on the pooled or total sample. SST does not figure into the F statistic directly. However, SST 5 SSB 1 SSE; thus if two sums of squares are known, the third can be computed from the other two. The third column contains degrees of freedom. The between-treatment degrees of freedom is df1 5 k 2 1. The error degrees of freedom is df2 5 N 2 k. The total degrees of freedom is N 2 1, and it is also true that (k 2 1) 1 (N 2 k) 5 N 2 1. The fourth column contains Mean Squares (MS), which are computed by dividing sums of squares (SS) by degrees of freedom (df), row by row. Specifically, TABLE 7–23 Weight Loss in Each Treatment Low-Calorie Low-Fat Low-Carbohydrate Control 8 2 3 2 9 4 5 2 6 3 4 21 7 5 2 0 3 1 3 3 significant d ifference in the mean weight loss among the four diets? We run the ANOVA using the five-step a pproach. SSB SSE MSB 5 }} and MSE 5 }}. k21 N2k Step 1: Set up hypotheses and determine the level of significance. Dividing SST / (N 2 1) produces the variance of the total sample. The F statistic is in the right column of the ANOVA table and is computed by taking the ratio of MSB / MSE. Example 7.14. A clinical trial is run to compare weightloss programs, and participants are randomly assigned to one of the comparison programs and counseled on the details of the assigned program. Participants follow the assigned program for 8 weeks. The outcome of interest is weight loss, defined as the difference in weight measured at the start of the study (baseline) and weight measured at the end of the study (8 weeks), in pounds. Three popular weight-loss programs are considered. The first is a low-calorie diet. The second is a low-fat diet and the third is a low-carbohydrate diet. For comparison purposes, a fourth group is considered as a control group. Participants in the fourth group are told that they are participating in a study of healthy behaviors with weight loss being only one component of interest. The control group is included here to assess the placebo effect (i.e., weight loss due to simply participating in the study). A total of 20 patients agree to participate in the study and are randomly assigned to one of the four diet groups. Weights are measured at baseline and patients are counseled on the proper implementation of the assigned diet (with the exception of the control group). After 8 weeks, each participant's weight is again measured and the d ifference in weights is computed by subtracting the 8-week weight from the baseline weight. Positive differences indicate weight loss and negative differences indicate weight gain. For interpretation purposes, we refer to the difference in weights as weight loss, and the observed weight losses are shown in Table 7–23. Is there a statistically H0: m1 5 m2 5 m3 5 m4 , H1: Means are not all equal, a 5 0.05. Step 2: Select the appropriate test statistic. The test statistic is the F statistic for ANOVA, MSB F 5 }}. MSE Step 3: Set up the decision rule. The appropriate critical value can be found in Table 4 in the Appendix. To determine the critical value of F, we need d egrees of freedom, df1 5 k 2 1 and df2 5 N 2 k. In this example, df1 5 k 2 1 5 4 2 1 5 3 and df2 5 N 2 k 5 20 2 4 5 16. The critical value is 3.24, and the decision rule is Reject H0 if F $ 3.24. Step 4: Compute the test statistic. To organize our computations, we complete the ANOVA table. To compute the sums of squares, we first compute the sample means for each group (see Table 7–24). The overall mean based on the total sample (n 5 20) is X w 5 3.6. We can now compute SSB: 2 SSB 5 onj(X wj 2 X w) , SSB 5 5(6.6 – 3.6)2 1 5(3.0 – 3.6)2 1 5(3.4 – 3.6)2 1 5(1.2 – 3.6)2, SSB 5 45.0 1 1.8 1 0.2 1 28.8 5 75.8. Tests with More Than Two Independent Samples, What Continuous Is PublicOutcome Health? Next, we compute SSE. SSE requires computing the squared differences between each observation and its group mean. We compute SSE in parts (see Tables 7–25, 7–26, 7–27, 7–28), and then sum. SSE 5 oo(X 2 X wj)2 5 21.4 1 10.0 1 5.4 1 10.6 5 47.4. TABLE 7–24 Summary Statistics on Weight Loss by Treatment Low-Calorie Low-Fat Low-Carbohydrate Control n1 5 5 n2 5 5 n3 5 5 w X1 5 6.6 w X2 5 3.0 w X3 5 3.4 153 TABLE 7–27 Deviations from Mean Weight Loss on Low-Carbohydrate Diet Low-Carbohydrate (X 2 3.4) (X 2 3.4)2 3 20.4 0.2 5 1.6 2.6 4 0.6 0.4 2 21.4 2.0 3 20.4 0.2 Total 0 5.4 n4 5 5 w X4 5 1.2 TABLE 7–28 Deviations from Mean Weight Loss in Control Group Control TABLE 7–25 Deviations from Mean Weight Loss on Low-Calorie Diet Low-Calorie (X 2 6.6) (X 2 6.6)2 8 1.4 2.0 9 2.4 5.8 6 20.6 0.4 7 0.4 0.2 3 23.6 13.0 Total 0 21.4 on Low-Fat Diet Low-Fat (X 2 3.0) (X 2 3.0)2 2 21.0 1.0 4 1.0 1.0 3 0.0 0.0 5 2.0 4.0 1 22.0 4.0 Total 0 10.0 (X 2 1.2)2 2 0.8 0.6 2 0.8 0.6 21 22.2 4.8 0 21.2 1.4 3 1.8 3.2 Total 0 10.6 We can now construct the ANOVA table (see Table 7–29). Step 5: Conclusion. TABLE 7–26 Deviations from Mean Weight Loss (X 2 1.2) We reject H0 because 8.43 . 3.24. We have statistically significant evidence at a 5 0.05 to show that there is a difference in mean weight loss among the four diets. Note that Table 4 in the Appendix provides critical values for the F test only when a 5 0.05. Without tables for other levels of significance, we cannot approximate p-values for ANOVA by hand. Because ANOVA is a tedious technique to implement by hand, it is usually performed with the use of a statistical computing package that produces, along with the ANOVA table, an exact p-value. ANOVA is a test that provides a global assessment of the statistical significance in more than two independent means. In this example, we find that there is a statistically significant difference in mean weight loss among the four diets considered. In addition to reporting the results of the statistical test of hypothesis (i.e., that there is a statistically 154 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–29 ANOVA Table Source of Variation Sums of Squares (SS) Degrees of Freedom (df) Mean Squares (MS) F 25.3 / 3.0 5 8.43 Between treatments 75.8 42153 75.8 / 3 = 25.3

Error or residual 47.4 20 2 4 5 16 47.4 / 16 = 3.0 123.2 20 2 1 5 19 Total significant difference in mean weight loss at a 5 0.05), investigators should also r eport the observed sample means to facilitate interpretation of the results. In this example, participants in the low-calorie diet lost a mean of 6.6 pounds over 8 weeks, as compared to 3 and 3.4 pounds in the low-fat and low-carbohydrate groups, respectively. Participants in the control group lost a mean of 1.2 pounds, which could be called a placebo effect because these participants were not participating in an active arm of the trial specifically targeted for weight loss. Example 7.15. Calcium is an essential mineral that regulates the heart and is important for blood clotting and for building healthy bones. The National Osteoporosis Foundation recommends a daily calcium intake of 1000 to 1200 mg/day for adult men and women.9 Whereas calcium is contained in some foods, most adults do not get enough calcium in their diets and take supplements. Unfortunately, some of the supplements have side effects such as gastric d istress, making them difficult for some patients to take on a regular basis. A study is designed to test whether there is a difference in mean daily calcium intake in adults with normal bone density, adults with osteopenia (a low bone density, which may lead to osteoporosis), and adults with osteoporosis. Adults 60 years of age with normal bone density, osteopenia, and osteoporosis are selected at random from hospital records and invited to participate in the study. Each participant's daily calcium intake is measured based on reported food intake and supplements. The data are shown in Table 7–30. Is there a statistically significant difference in mean calcium intake in patients with normal bone density as compared to patients with osteopenia and osteoporosis? We run the ANOVA using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: m1 5 m2 5 m3, TABLE 7–30 Calcium Intake in Each Treatment Normal Bone Density Osteopenia 1200 1000 1000 1100 650 980 700 1100 900 800 900 750 500 400 800 700 350 Osteoporosis 890 H1: Means are not all equal, a 5 0.05. Step 2: Select the appropriate test statistic. The test statistic is the F statistic for ANOVA, MSB F 5 }}. MSE Step 3: Set up the decision rule. The appropriate critical value can be found in Table 4 in the Appendix. To determine the critical value of F, we need d egrees of freedom, df1 5 k 2 1 and df2 5 N 2 k. In this example, df1 5 k 2 1 5 3 2 1 5 2 and df2 5 N 2 k 5 18 2 3 5 15. The critical value is 3.68, and the decision rule is Reject H0 if F $ 3.68. Step 4: Compute the test statistic. To organize our computations, we complete the ANOVA table. To compute the sums of squares, we must first compute the sample means for each group (see Table 7–31). The overall mean is X w 5 817.8. We can now compute SSB: Tests with More Than Two Independent Samples, What Continuous Is PublicOutcome Health? TABLE 7–31 Summary Statistics on Calcium TABLE 7–33 Deviations from Mean Calcium Intake by Treatment Normal Bone Density Osteopenia Intake in Patients with Osteopenia Osteoporosis n1 5 6 n2 5 6 n3 5 6 w X1 5 938.3 w X2 5 800.0 w X3 5 715.0 2 SSB 5 onj(X wj 2 X w) , SSB 5 6(938.3 – 817.8) 1 6(715.0 – 817.8) 1 6(800.0 2 817.8)2, 2 155 2 Osteopenia (X 2 800.0) (X 2 800.0)2 1000 200.0 40,000.0 1100 300.0 90,000.0 700 2100.0 10,000.0 800 0.0 0.0 500 2300.0 90,000.0 700 200.0 10,000.0 Total 0 240,000.0 SSB 5 87,121.5 1 63,407.0 1 1,901.0 5 152,429.5. Next, we calculate SSE. SSE requires computing the squared differences between each observation and its group mean. We compute SSE in parts (see Tables 7–32, 7–33, and 7–34), and then sum. TABLE 7–34 Deviations from Mean Calcium Intake in Patients with Osteoporosis SSE 5 oo(X 2 X wj)2 5 130,083.4 1 240,000.0 1 449,750.0 5 819,833.4. TABLE 7–32 Deviations from Mean Calcium Intake in Patients with Normal Bone Density Normal Bone Density (X 2 938.3) 261.7 68,486.9 1000 61.7 3806.9 980 41.7 1738.9 900 238.3 1466.9 750 2188.3 35,456.9 800 2138.3 19,126.9 Total 0 130,083.4 (X 2 715.0) (X 2 715.0)2 890 175.0 30,625.0 650 265.0 4,225.0 1100 385.0 148,225.0 900 185.0 34,225.0 400 2315.0 99,225.0 350 2365.0 133,225.0 Total (X 2 938.3)2 1200 Osteoporosis 0 449,750.0 We can now construct the ANOVA table (see Table 7–35). Step 5: Conclusion. We do not reject H0 because 1.39 , 3.68. We do not have statistically significant evidence at a 5 0.05 to show that there is a difference in mean calcium intake in patients with normal bone density as compared to osteopenia and osteoporosis. TABLE 7–35 ANOVA Table Source of Variation Sums of Squares (SS) Degrees of Freedom (df) Mean Squares (MS) F Between treatments 152,429.5 2 76,214.8 1.39 Error or residual 819,833.4 15 54,655.6 Total 972,262.9 17 156 CHAPTER 7 Hypothesis Testing Procedures When sample sizes in each comparison group are equal, the design is called a balanced design. Balanced designs are preferred over unbalanced designs (unequal numbers of participants in the comparison groups) because they are more robust (e.g., they ensure a 5% Type I error rate when a = 0.05) when the assumptions, such as normality of the outcome, are violated. The ANOVA tests described here are called one-way or one-factor ANOVAs. There is one treatment or grouping factor with k . 2 levels, and we wish to compare the means of a continuous outcome across the different categories of this factor (also called different treatments). The factor might represent different diets, different classifications of risk for disease (e.g., osteoporosis), different medical treatments, different age groups, or different racial or ethnic groups. There are s ituations where it may be of interest to compare means of a continuous outcome across two or more factors. For example, suppose a clinical trial is designed to compare three d ifferent treatments for joint pain in patients with osteoarthritis. Investigators might also hypothesize that there are differences in the outcome by sex. This is an example of a two-factor ANOVA where the factors are treatment (with three levels) and sex (with two levels). In the two-factor ANOVA, investigators can assess whether there are differences in means due to the treatments, due to the sex of the participant, or due to the combination or interaction of treatment and s ex. If there are differences in the means of the outcome by treatment, we say there is a main effect of treatment. If there are differences in the means of the outcome by sex, we say there is a main effect of sex. If there are differences in the means of the outcome among treatments but these vary by sex, we say there is an interaction effect. Higher-order ANOVAs are conducted in the same way as one-factor ANOVAs presented here, and the computations are again organized in ANOVA tables, with more rows to distinguish the different sources of variation (e.g., between treatments, between men and women). An example of a two-factor ANOVA is given in Example 7.16. More details on higher-order ANOVA can be found in Snedecor and Cochran.10 Example 7.16. Consider the clinical trial outlined above, in which three competing treatments for joint pain are compared in terms of their mean time to pain relief in patients with osteoarthritis. Because investigators hypothesize that there may be a difference in time to pain relief in men versus women, they randomly assign 15 participating men to one of the three competing treatments and randomly assign 15 participating women to one of the three competing treatments (i.e., stratified randomization). Participating men and women do not know to which treatment they are assigned. They are instructed to take the assigned medication when they experience joint pain and to record the time, in minutes, until the pain subsides. The data (times to pain relief) are shown in Table 7–36 and are organized by the assigned treatment and gender of the participant. The analysis in two-factor ANOVA is similar to that illustrated in Section 7.8 for one-factor ANOVA. The computations are again organized in an ANOVA table, but the total variation is partitioned into that due to the main effect of treatment, the main effect of gender, and the interaction effect. The results of the analysis are shown in Table 7–37 (see Snedecor and Cochran for technical details.10) There are four statistical tests in the ANOVA table. The first test is an overall test to assess whether there is a difference among the six cell means (cells are defined by treatment and sex). The F-statistic is 20.7 and is highly statistically significant with p = 0.0001. When the overall test is significant, focus then turns to the factors that may be driving the significance (in this example, treatment, sex, or the interaction between the two). The next three statistical tests assess the significance of the main effect of treatment, the main effect of sex, and the interaction effect. In this example, there is a highly significant main effect of treat- TABLE 7–36 Time to Pain Relief by Treatment and Sex Treatment Male Female A 12 21 15 19 16 18 17 24 14 25 14 21 17 20 19 23 20 27 17 25 25 37 27 34 29 36 24 26 22 29 B C Tests for Two or More Independent Samples, CategoricalWhat and Ordinal Is Public Outcomes Health? 157 TABLE 7–37 ANOVA Table for Two-Factor ANOVA Sums of Squares SS Degrees of Freedom df Mean Squares MS F p-value Model 967.0 5 193.4 20.7 0.0001 Treatment 651.5 2 325.7 34.8 0.0001 Sex 313.6 1 313.6 33.5 0.0001 Treatment * Sex 1.9 2 0.9 0.1 0.9054 Error or Residual 224.4 24 9.4 Total 1191.4 29 Source of Variation ment (p = 0.0001) and a highly significant main effect of sex (p = 0.0001). The interaction between the two does not reach statistical significance. Table 7–38 contains the mean times to pain relief in each of the treatments for men and women (note that each sample mean is computed on the 5 observations measured under that experimental condition). Treatment A appears to be the most efficacious treatment for both men and women. The mean times to relief are lower in Treatment A for both men and women and highest in Treatment C for both men and women. In each treatment, women report longer times to pain relief. Suppose that the same clinical trial is replicated in a second clinical site and the data in Table 7–39 are observed. The ANOVA table for the data measured in clinical site 2 are summarized in Table 7–40. Notice that the overall test is significant (F = 19.4, p = 0.0001) and there is a significant treatment effect and sex effect and a highly significant interaction effect. Table 7–41 contains the mean times to relief in each of the treatments for men and women. Notice that now the differences in mean time to pain relief among the treatments depend on sex. Among men, the mean time to pain relief is highest in Treatment A and lowest in Treatment C. Among women, the reverse is true. This is TABLE 7–38 Mean Time to Pain Relief by Treatment and Sex Treatment Male Female A 14.8 21.4 B 17.4 23.2 C 25.4 32.4 an interaction effect. When interaction effects are present, some investigators do not examine main effects. We discuss interaction effects in more detail in Chapter 9. 7.9 TESTS FOR TWO OR MORE INDEPENDENT SAMPLES, CATEGORICAL AND ORDINAL OUTCOMES In Section 7.4, we presented the x2 goodness-of-fit test, which was used to test whether the distribution of responses to a cate gorical or ordinal variable measured in a single sample followed TABLE 7–39 Time to Pain Relief by Treatment and Sex: Clinical Site 2 Treatment Male Female A 22 21 25 19 26 18 27 24 24 25 14 21 17 20 19 23 20 27 17 25 15 37 17 34 19 36 14 26 12 29 B C 158 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–40 ANOVA Table for Two-Factor ANOVA: Clinical Site 2 Sums of Squares SS Degrees of Freedom df Mean Squares MS F p-value Model 907.0 5 181.4 19.4 0.0001 Treatment 71.5 2 35.7 3.8 0.0362 Sex 313.6 1 313.6 33.5 0.0001 Treatment * Sex 521.9 2 260.9 27.9 0.0001 Error or Residual 224.4 24 9.4 Total 1131.4 29 Source of Variation TABLE 7–41 Mean Time to Pain Relief by Treatment and Sex: Clinical Site 2 Treatment Male Female A 24.8 21.4 B 17.4 23.2 C 15.4 32.4 a known distribution. Here we extend that application to the two or more independent samples case. Specifically, the outcome of interest has two or more responses, and the responses are ordered or unordered (i.e., ordinal or categorical). We now consider the situation where there are two or more independent comparison groups, and the goal of the analysis is to compare the distribution of responses to the categorical or ordinal outcome variable among several independent comparison groups. The test is called the x2 test of independence and the null hypothesis is that there is no difference in the distribution of responses to the outcome across comparison groups. This is often stated as: The outcome variable and the grouping variable (e.g., the comparison treatments or comparison groups) are independent (hence the name of the test). Independence here implies homogeneity in the distribution of the outcome among comparison groups. The null hypothesis in the x2 test of independence is often stated in words as H0: The distribution of the outcome is independent of the groups. The research hypothesis is that there is a difference in the distribution of the outcome variable among the comparison groups (i.e., that the distribution of responses "depends" on the group). To test the hypothesis, we measure the categorical or ordinal outcome variable in each participant in each comparison group. The data of interest are the observed frequencies (or number of participants in each response category in each group). The formula for the test statistic for the x2 test of independence is given in Table 7–42. The data for the x2 test of independence are organized in a two-way table. The outcome and grouping variable are shown in the rows and columns of the table. Table 7–43 illustrates the sample data layout. The table entries (blank) are the numbers of participants in each group responding to each response category of the outcome variable. In Table 7–43, the grouping variable is shown in the rows of the table; r denotes the number of independent groups. The outcome variable is shown in the columns of the table; c denotes the number of response options in the outcome variable. Each combination of a row (group) and column (response) is called a cell of the table. The table has r 3 c cells and is sometimes called an r 3 c ("r by c") table. For example, if there are four groups and five categories in the outcome variable, the data are organized in a 4 3 5 table. The row and column totals are shown along the right margin and the bottom of the table, respectively. The total sample size, N, is computed by summing the row totals or the column totals. Similar to ANOVA, N here does not refer to a population size but rather to the total sample size in the analysis. The sample data are organized into a table like that shown in Table 7–43. TABLE 7–42 Test Statistic for Testing H0: Distribution of Outcome Is Independent of Groups (O 2 E)2 x2 5 ^ E (Find critical value in Table 3, df 5 (r 2 1)(c 2 1)) 159 Tests for Two or More Independent Samples, CategoricalWhat and Ordinal Is Public Outcomes Health? TABLE 7–43 Data Layout for Chi-Square Test of Independence Outcome Variable Grouping Variable Response Option 1 Response Option 2 ... Response Option c Row Totals Group 1 Group 2 ... Group r N Column totals The numbers of participants within each group who select each response option are shown in the cells of the table, and these are the observed frequencies used in the test statistic. The test statistic for the x2 test of independence involves comparing observed (sample data) and expected frequencies in each cell of the table. The expected frequencies are computed assuming that the null hypothesis is true. The null hypothesis states that the two variables (the grouping variable and the outcome) are independent. In Chapter 5, we introduced the concept of independence. The definition of independence is To convert the probability to a f requency, we multiply by N (the total sample size). Consider the following example. The data shown in Table 7–44 are measured in a sample of size N 5 150. The frequencies in the cells of the table are the observed frequencies. If Group and Response are independent, then we can compute the probability that a person in the sample is in Group 1 and Response Option 1 using Two events, A and B, are independent if P(A|B) 5 P(A) or if P(B|A) = P(B), or, equivalently, Thus, if Group and Response are independent, we would expect 6.9% of the sample to be in the top-left cell of the table (Group 1 and Response 1). The expected frequency is 150(0.069) 5 10.4. We could do the same for Group 2 and Response 1: Two events, A and B, are independent if P(A and B) 5 P(A) P(B). The last statement indicates that if two events, A and B, are independent, then the probability of their intersection can be computed by multiplying the probability of each individual event. To conduct the x2 test of independence, we must compute expected frequencies in each cell of the table. Expected frequencies are computed by assuming that the grouping variable and outcome are independent (i.e., under the null hypothesis). Thus, if the null hypothesis is true, using the definition of independence: P(Group 1 and Response Option 1) 5 P(Group 1) P(Response Option 1). The preceding states that the probability that an individual is in Group 1 and his or her outcome is Response Option 1 is computed by multiplying the probability that person is in Group 1 by the probability that a person gives Response Option 1. This is true if Group and Response are independent. To conduct the x2 test of independence, we need e xpected frequencies and not expected probabilities. P(Group 1 and Response 1) 5 P(Group 1) P(Response 1), P(Group 1 and Response 1) 5 (25 / 150)(62 / 150) 5 0.069. P(Group 2 and Response 1) 5 P(Group 2) P(Response 1), P(Group 2 and Response 1) 5 (50 / 150) (62 / 150) 5 0.138. The expected frequency in Group 2 and Response 1 is 150(0.138) 5 20.7. TABLE 7–44 Observed Responses by Group Response Response Response 1 2 3 Total Group 1 10 8 7 25 Group 2 22 15 13 50 Group 3 30 28 17 75 Total 62 51 37 150 160 CHAPTER 7 Hypothesis Testing Procedures The formula for determining the expected cell frequencies in the x2 test of independence is Step 1: Set up hypotheses and determine the level of significance. Row total 3 column total Expected cell frequency 5 }}} . N H0: Living arrangement and exercise are independent, The preceding equation produces the expected frequency in one step rather than computing the expected probability first and then converting to a frequency. Example 7.17. In Example 7.6, we examined data from a survey of university graduates that assessed (among other things) how frequently they exercised. The survey was completed by 470 graduates. We used the x2 goodness-of-fit test to assess whether there was a shift in the distribution of responses to the exercise question following the implementation of a health promotion campaign on campus. We specifically considered one sample (all students) and compared the observed distribution of responses to the exercise question to the distribution of responses the prior year (a historical control). Suppose we wish to assess whether there is a relationship between exercise on campus and s tudents' living a rrangement. As part of the same survey, graduates were asked where they lived their senior year. The r esponse options were dormitory, on-campus apartment, off-campus apartment, and at home (i.e., commuted to and from the university). The data are shown in Table 7–45. Based on the data, is there a relationship between exercise and a student's living arrangement? Here we have four independent comparison groups (living arrangements) and an ordinal outcome variable with three response options. We specifically want to test whether living arrangement and exercise are independent. We run the test using the five-step approach. a 5 0.05. TABLE 7–45 Exercise by Living Arrangement No Regular Sporadic Regular Exercise Exercise Exercise Total Dormitory 32 30 28 90 On-campus apartment 74 64 42 180 Off-campus apartment 110 25 15 150 At home 39 6 5 50 Total 255 125 90 470 H1: H0 is false, The null and research hypotheses are written in words rather than in symbols. The research hypothesis is that the grouping variable (living arrangement) and the outcome variable (exercise) are dependent or related. Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–42 and is given as (O2E)2 x2 5 ^ . E The condition for appropriate use of the preceding test statistic is that each expected frequency is at least five. In Step 4, we compute the expected frequencies and ensure that the condition is met. Step 3: Set up the decision rule. The decision rule for the x2 test of independence is set up in a similar way to decision rules we established for z and t tests. The decision rule depends on the level of significance and the degrees of freedom, defined as df 5 (r 2 1)(c 2 1), where r and c are the numbers of rows and columns in the two-way data table. The row variable is the living arrangement, and there are four arrangements considered; thus r 5 4. The column variable is exercise, and three responses are considered; thus c 5 3. For this test, df 5 (4 2 1)(3 2 1) 5 3(2) 5 6. Again, with x2 tests there are no upper- or lower-tailed versions of the test. If the null hypothesis is true, the observed and expected frequencies are close in value and the x2 statistic is close to 0. If the null hypothesis is false, then the x2 statistic is large. The rejection region for the x2 test of independence is always in the upper tail (right) of the distribution, as shown in Figure 7–6. Table 3 in the Appendix contains critical values for the x2 test indexed by degrees of freedom and the desired level of significance. For df 5 6 and a 5% level of significance, the appropriate critical value from Table 3 is 12.59, and the decision rule is Reject H0 if x2 $ 12.59. 161 Tests for Two or More Independent Samples, CategoricalWhat and Ordinal Is Public Outcomes Health? Step 4: Compute the test statistic. We now compute the expected frequencies using the formula, Row total 3 column total Expected cell frequency 5 }}} . N The computations are organized in a two-way table. The expected frequencies are taken to one decimal place and the sums of the observed frequencies are equal to the sums of the expected frequencies in each row and column of the table (see Table 7–46). Recall in Step 2 that a condition for the appropriate use of the test statistic was that each expected frequency is at least five. This is true for this sample (the smallest expected frequency is 9.6), and therefore it is appropriate to use the test statistic. The test statistic is computed as follows: x2 = (32 − 48.8)2 (30 − 23.9)2 (28 − 17.2)2 + + + 48.8 23.9 17.2 (74 − 97.7)2 (64 − 47.9)2 (42 − 34.5)2 + + + 97.7 47.9 34.5 (110 − 81.4)2 (25 − 39.9)2 (15 − 28.7)2 + + + 81.4 39.9 28.7 (39 − 27.1)2 (6 − 13.3)2 (5 − 9.6)2 , + + 27.1 13.3 9.6 x2 5 5.78 1 1.56 1 6.78 1 5.75 1 5.41 1 1.63 1 10.05 1 5.56 1 6.54 1 5.23 1 4.01 1 2.20 5 60.5. Step 5: Conclusion. We reject H0 because 60.5 . 12.59. We have statistically significant evidence at a 5 0.05 to show that H0 is false or TABLE 7–47 Percentages of Students Exercising TABLE 7–46 Expected Frequencies No Regular Sporadic Regular Exercise Exercise Exercise that living arrangement and exercise are not independent (i.e., they are dependent or related). Using Table 3 in the Appendix, we can approximate the p-value. We need to look at critical values for smaller levels of significance with df 5 6. Using Table 3, the p-value is p , 0.005. The x2 test of independence is used to test whether the distribution of the outcome variable is different across the comparison groups. In Example 7.17, we reject H0 and conclude that the distribution of exercise is not independent of living arrangement, or that there is a relationship between living arrangement and exercise. The test provides an overall assessment of statistical significance. When the null hypothesis is rejected, it is important to review the sample data to understand the nature of the relationship. Consider again the data in Example 7.17. Because there are different numbers of students in each living situation, it makes the comparisons of exercise patterns difficult on the basis of the frequencies alone. Table 7–47 displays the percentages of students in each exercise category by living arrangement. The percentages sum to 100% in each row of the table. For comparison purposes, percentages are also shown for the total sample along the bottom row of the table. From Table 7–47, it is clear that higher percentages of students living in dormitories and in on-campus apartments report regular exercise (31% and 23%) as compared to s tudents living in off-campus apartments and at home (10% each). Example 7.18. In Example 7.13, we analyzed data from a randomized trial designed to evaluate the effectiveness of a newly developed pain reliever to reduce pain in patients following joint replacement surgery. The trial compared a new pain reliever to the pain reliever currently in use (called the standard of care). Suppose there was a third arm of the trial and by Living Arrangement Total No Regular Exercise Sporadic Exercise Regular Exercise Dormitory 48.8 23.9 17.2 90 Dormitory 36 33 31 On-campus apartment 97.7 47.9 34.5 180 On-campus apartment 41 36 23 Off-campus apartment 81.4 39.9 28.7 150 Off-campus apartment 73 17 10 9.6 50 At home 78 12 10 90 470 Total 54 27 19 At home Total 27.1 255 13.3 125 162 CHAPTER 7 Hypothesis Testing Procedures patients assigned to the third arm received a higher dose of the newly developed pain reliever. Suppose that N 5 150 patients agreed to participate in the trial and were randomly assigned to one of the three treatments. Before receiving the assigned treatment, patients were asked to rate their pain on a scale of 0 to 10, with higher scores indicative of more pain. Each patient was then given the assigned treatment and after 30 minutes was again asked to rate his or her pain on the same scale. The primary outcome was a reduction in pain of 3 or more scale points (defined by clinicians as a clinically meaningful reduction). The data shown in Table 7–48 are observed in the trial. Here we have three independent comparison groups and a categorical (dichotomous) outcome variable. We want to test whether there is a difference in the proportions of patients reporting a meaningful reduction in pain among the three treatments, and we run the test using the five-step approach. Step 1: Set up hypotheses and determine the level of significance. H0: Treatment and reduction in pain are independent, Step 3: Set up the decision rule. The row variable is the treatment and there are three considered here (r 5 3). The column variable is the outcome and two responses are considered (c 5 2). For this test, df 5 (3 2 1) (2 2 1) 5 2(1) 5 2. For df 5 2 and a 5% level of significance, the appropriate critical value from Table 3 in the Appendix is 5.99, and the decision rule is Reject H0 if x2 $ 5.99. Step 4: Compute the test statistic. We now compute the expected frequencies using the formula, Row total 3 column total Expected cell frequency 5 }}} . N The expected frequencies are shown in Table 7–49. The test statistic is computed as X2 = (23 − 21.3)2 21.3 (20 − 28.7)2 H1: H0 is false, 28.7 a 5 0.05. + + (27 − 28.7)2 28.7 (11 − 21.3)2 21.3 + + (30 − 21.3)2 21.3 (39 − 28.7)2 28.7 + , x2 5 0.14 1 0.10 1 3.56 1 2.64 1 4.98 1 3.70 5 15.12. Step 2: Select the appropriate test statistic. The formula for the test statistic is in Table 7–42 and is given as (O2E)2 x2 5 ^ . E The condition for appropriate use of the preceding test statistic is that each expected frequency is at least five. In Step 4, we compute the expected frequencies and ensure that the condition is met. TABLE 7–48 Pain Reduction by Treatment Step 5: Conclusion. We reject H 0 because 15.12 . 5.99. We have statistically significant evidence at a 5 0.05 to show that H 0 is false, or that there is a difference in the proportions of patients reporting a meaningful reduction in pain among the three treatments. Using Table 3 in the Appendix, we can approximate the p-value. We need to TABLE 7–49 Expected Frequencies Meaningful No Meaningful Reduction Reduction (3 1 Points) (, 3 Points) Total Meaningful No Meaningful Reduction Reduction (3 1 Points) (, 3 Points) Total New pain reliever 23 27 50 New pain reliever 21.3 28.7 50 Higher dose of new pain reliever 30 20 50 Higher dose of new pain reliever 21.3 28.7 50 Standard pain reliever 11 39 50 Standard pain reliever 21.3 28.7 50 Total 64 86 150 Total 64 86 150 What Is PublicSummary Health? look at critical values for smaller levels of significance for df 5 2. Using Table 3 in the Appendix, the p-value is p , 0.005. What is the nature of the relationship between meaningful reduction in pain and the three treatments? 7.10 SUMMARY In this chapter, we presented hypothesis testing techniques. Tests of hypothesis involve several steps, including specifying the null hypothesis and the alternative or research hypothesis, selecting and computing an appropriate test statistic, setting up a decision rule, and drawing a conclusion. There are many details to c onsider in hypothesis testing. The first is to determine the - appropriate test. We discussed z, t, x2, and F tests here for d ifferent applications. The appropriate test depends on the distribution of the outcome variable (continuous, dichotomous, categorical, or ordinal), the number of comparison groups (one, two, or more than two), and whether the comparison groups are independent or dependent. Table 7–50 summarizes the different tests of hypothesis discussed here. TABLE 7–50 Summary of Key Formulas for Tests of Hypothesis Outcome Variable, Number of Groups: Null Hypothesis Test Statistic* z= Continuous outcome, one sample: H0: m 5 m0 z= Continuous outcome, two independent samples: H0: m1 5 m2 X d − μd sd n ( ) F= ∑ ∑(X − X) ∑ nj X j − X 2 2 j Dichotomous outcome, two independent samples: H0: p1 5 p2, RD 5 0, RR 5 1, OR 5 1 χ2 = ∑ *See Table 7–4, 7–11, and 7–14 for alternative formulas that are appropriate for small samples. (N − k) p0 (1 − p0) n pˆ1 − pˆ 2 pˆ (1 − pˆ)(1 n1 + 1 n2) z= Categorical or ordinal outcome, two or more independent samples: H0 : Outcome and groups are independent (k − 1) pˆ − p0 z= Dichotomous outcome, one sample: H0: p 5 p0 Categorical or ordinal outcome, one sample: H0: p1 5 p10, p2 5 p20, ..., pk 5 pk0 X − μ0 s n X1 − X 2 S p 1 n1 + 1 n2 z= Continuous outcome, two matched samples: H0: md 5 0 Continuous outcome, more than two independent samples: H0: m1 5 m2 5 ... 5 mk 163 (O − E)2 E χ2 = ∑ , df = k − 1 (O − E)2 E df = (r − 1)(c − 1) , 164 CHAPTER 7 Hypothesis Testing Procedures Once the type of test is determined, the details of the test must be specified. Specifically, the null and research hypotheses must be clearly stated. The null hypothesis a lways reflects the "no change" or "no difference" situation. The alternative or research hypothesis reflects the investigator's belief. The investigator might hypothesize that a parameter (e.g., a mean, proportion, difference in means, or difference in proportions) will increase, will decrease, or will be different under specific conditions (sometimes the conditions are different experimental conditions and at other times the conditions are defined by participant attributes). Once the hypotheses are specified, data are collected and summarized. The appropriate test is then conducted according to the five-step a pproach. If the test leads to rejection of the null hypothesis, an approximate p-value is computed to summarize the statistical significance of the findings. When tests of hypothesis are conducted using statistical computing packages, exact p-values are computed. Because the statistical tables in this textbook are limited, we only approximate p-values. If the test fails to reject the null hypothesis, then a weaker concluding statement is made. In hypothesis testing, there are two types of errors that can be committed. A Type I error occurs when a test incorrectly rejects the null hypothesis. This is referred to as a false positive result, and the probability that this occurs is equal to the level of significance, a. The investigator chooses the level of significance and purposely chooses a small value, such as a 5 0.05, to control the probability of committing a Type I error. A Type II error occurs when a test fails to r eject the null hypothesis when, in fact, it is false. The probability that this occurs is equal to b. Unfortunately, the investigator cannot specify b because it depends on several factors, including the sample size (smaller samples have higher b), the level of s ignificance and the difference in the parameter under the null and alternative hypothesis. (For more details, see D'Agostino, Sullivan, and Beiser.5) We noted in several examples the relationship between confidence intervals and tests of hypothesis. While the approaches are somewhat different, they are clearly related. It is possible to draw a conclusion about statistical significance by examining a confidence interval. For example, if a 95% confidence interval does not contain the null value of the parameter of interest (e.g., 0 when analyzing a difference in means or risk difference, 1 when analyzing relative risks or odds ratios), then we conclude that a two-sided test of hypothesis is significant at a 5 0.05. It is important to note that the correspondence between a confidence interval and a test of hypothesis relates to a two-sided test, and that the confidence level corresponds to a specific two-sided level of significance (e.g., 95% to a 5 0.05, 90% to a 5 0.10, and so on). The exact significance of the test, the p-value, can be determined only by using the hypothesis testing - approach. 7.11 PRACTICE PROBLEMS 1. A clinical trial evaluates a new compound designed to improve wound healing in trauma patients. The new compound is compared against a placebo. After treatment for 5 days with the new compound or placebo, the extent of wound healing is measured and the data are shown in Table 7–51. Is there a d - ifference in the extent of wound healing by treatment? Run the appropriate test at a 5% level of significance. (Hint: Are treatment and the percent of wound healing independent?) 2. Use the data in Problem 1 and pool the data across the treatments into one sample of size n 5 250. Use the pooled data to test whether the distribution of the percent of wound healing is approximately normal. Specifically, use the following distribution: 30%, 40%, 20%, and 10%, and a 5 0.05 to run the appropriate test. 3. Table 7–52 displays summary statistics on the participants involved in the study described in Problem 1. Are any of the characteristics significantly different between groups? Justify briefly. (Hint: No calculations, just an interpretation.) 4. An investigator hypothesizes that cholesterol levels in children might be affected by educating their parents on proper nutrition and exercise. A sample of 40 families with a child between the ages of 10 to 15 who has been diagnosed with high cholesterol agree to participate in the study. All parents are provided educational information on nutrition and exercise. After following the prescribed program, their child's total cholesterol level is measured. The childrens' mean cholesterol level is 175 with a s tandard deviation of 19.5. Is there significant evidence of a reduction in total cholesterol in the TABLE 7–51 Wound Healing by Treatment Percent Wound Healing Treatment 0–25 26–50 51–75 76–100 New compound (n 5 125) 15 37 32 41 Placebo (n 5 125) 36 45 34 10 What Practice Is PublicProblems Health? 165 TABLE 7–52 Summary Statistics by Treatment New Compound Mean age (years) p-value Placebo 47.2 46.1 0.7564 Men (%) 44 59 0.0215 Mean educational level (years) 13.1 14.2 0.6898 Mean annual income $36,560 $37,470 0.3546 Mean body mass index (BMI) 24.7 25.1 0.0851 children? Run the appropriate test at the 5% level of significance and assume that the null value for total cholesterol is 191. 5. An experiment is designed to investigate the impact of different positions of the mother during ultrasound on fetal heart rate. Fetal heart rate is measured by ultrasound in beats per minute. The study includes 20 women who are assigned to one position and have the fetal heart rate measured in that position. Each woman is between 28 and 32 weeks gestation. The data are shown in Table 7–53. Is there a significant difference in mean fetal heart rates by position? Run the test at a 5% level of significance. 6. A clinical trial is conducted comparing a new pain reliever for arthritis to a placebo. Participants are randomly assigned to receive the new treatment or a placebo and the outcome is pain relief within 30 minutes. The data are shown in Table 7–54. Is there a significant difference in the proportions of patients reporting pain relief? Run the test at a 5% level of significance. 7. A clinical trial is planned to compare an experimental medication designed to lower blood pressure to a placebo. Before starting the trial, a pilot study is conducted involving seven participants. The objective of the study is to assess how systolic blood pressures change over time if left untreated. Systolic blood pressures are measured at baseline and again 4 weeks later. Is there a statistically s ignificant difference in blood pressures over time? Run the test at a 5% level of significance. Baseline: 120 145 130 160 152 143 126 4 Weeks: 122 142 135 158 155 140 130 8. The main trial in Problem 7 is conducted and involves a total of 200 patients. Patients are enrolled TABLE 7–54 Pain Relief by Treatment Pain Relief No Pain Relief New medication 44 76 Placebo 21 99 TABLE 7–53 Fetal Heart Rate by Position Back Side Sitting Standing 140 141 144 147 144 143 145 145 146 145 147 148 141 144 148 149 139 136 144 145 Mean 5 142.0 Mean 5 141.8 Mean 5 145.6 Mean 5 146.8 166 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–55 Outcomes by Treatment Mean (SD) systolic blood pressure Hypertensive (%) Side effects (%) TABLE 7–56 Hypertensive Status by Treatment Experimental (n 5 100) Placebo (n 5 100) 120.2 (15.4) 131.4 (18.9) 14 22 6 8 and r andomized to receive either the experimental medication or the placebo. The data shown in Table 7–55 are collected at the end of the study after 6 weeks on the assigned treatment. a. Test if there is a significant difference in mean systolic blood pressures between groups using a 5 0.05. b. Test if there is a significant difference in the proportions of hypertensive patients between groups using a 5 0.05. 9. Suppose in the trial described in Problem 8 that patients were recruited from three different clinical sites. Use the data in Table 7–56 to test if there is a difference in the proportions of hypertensive patients across clinical sites. Run the test at a 5% level of significance. 10. A clinical trial is conducted to compare an experimental medication to a placebo to reduce the symptoms of asthma. Two hundred participants are enrolled in the study and randomized to receive either the experimental medication or a placebo. The primary outcome is self-reported r eduction of symptoms. Among 100 participants who r eceive the experimental medication, 38 report a r eduction of symptoms as compared to 21 participants of 100 assigned to placebo. Test if there is a significant d ifference in the proportions of participants Site 1 Site 2 Site 3 Hypertensive 10 14 12 Not hypertensive 68 56 40 r eporting a reduction of symptoms between the experimental and placebo groups. Use a 5 0.05. 11. Suppose more detail is recorded in the primary outcome in the clinical trial described in Problem 10. The data are shown in Table 7–57. Is there a d ifference in the change in symptoms by treatment group? Run the appropriate test at a 5% level of s ignificance. 12. Suppose a secondary outcome is recorded in the trial described in Problem 10 reflecting asthma symptom severity measured on a scale of 0 to 100, with higher scores indicating more severe symptoms. In the participants who receive the experimental m edication, the mean symptom score is 74 with a standard deviation of 5.6, and in the placebo group, the mean symptom score is 85 with a standard d eviation of 6. Is there a significant difference in mean symptom scores between groups? Run the a ppropriate test at a 5% level of significance. 13. Recent recommendations suggest 60 minutes of physical activity per day. A sample of 50 adults in a study of cardiovascular risk factors report exercising a mean of 38 minutes per day with a standard deviation of 19 minutes. Based on the sample data, is the physical activity significantly less than recommended? Run the appropriate test at a 5% level of significance. 14. Suppose a hypertension trial is mounted and 18 participants are randomly assigned to one of three comparison treatments. Each participant takes the assigned medication and his or her systolic blood TABLE 7–57 Change in Symptoms by Treatment Change in Symptoms Treatment Much Worse Worse No Change Better Much Better Experimental 10 17 35 28 10 Placebo 12 25 42 12 9 What Practice Is PublicProblems Health? TABLE 7–58 Systolic Blood Pressure by Treatment Standard Treatment Placebo TABLE 7–59 Cholesterol Levels by Treatment New Treatment Sample Mean Std Dev Size Cholesterol Cholesterol Diet Program 124 134 114 Low carbohydrate 50 225.4 24.5 111 143 117 Conventional 75 203.8 21.6 133 148 121 125 142 124 128 150 122 115 160 128 ressure is recorded after 6 months on the assigned p treatment. The data are shown in Table 7–58. Is there a difference in mean systolic blood pressure among treatments? Run the appropriate test at a 5 0.05. 15. A study is conducted to compare mean cholesterol levels for individuals following a low-carbohydrate diet for at least 6 months to individuals following a conventional (low-fat, low-calorie) diet for at least 6 months. The data are summarized in Table 7–59. Test if there is a significant difference in mean cholesterol levels between the diet programs using a 5% level of significance. 16. Another outcome variable in the study described in Problem 15 is hypercholesterolemia, defined as total cholesterol over 220. Among the individuals who follow the low-carbohydrate diet, 56% are hyperc holesterolemic, and among the individuals who follow the conventional diet, 40% are hypercholesterolemic. Test if there is a significant difference in the proportions using a 5% level of significance. 17. Table 7–60 compares background characteristics of the participants involved in the study described in Problem 15. Are there any statistically significant differences in patient characteristics between the different diet programs? Justify briefly. 18. Suppose the results of the analyses in Problem 15 through Problem 17 are reported and criticized because the participants were not randomized to different diets, and that there may be other factors associated with changes in cholesterol. A third study is run to estimate the effect of the low-carbohydrate diet on cholesterol levels. In the third study, participants' cholesterol levels are measured before starting the program and then again after 6 months on the program. The data are shown below: Before Program 210 230 190 215 260 200 After 6 Months 215 240 190 200 280 210 Is there a significant increase in cholesterol after 6 months on the low-carbohydrate diet? Run the appropriate test at a 5% level of significance. 19. A study is conducted to compare three new appetite suppressants (A, B, and C) to a placebo in terms of their effects on weight reduction. A total of 80 participants are involved and are randomly assigned to the comparison groups in equal numbers. The outcome of interest is weight reduction, measured in pounds. The data shown in Table 7–61 are observed after 3 months on treatment. Is there a significant difference TABLE 7–60 Background Characteristics by Treatment Low Carbohydrate (n 5 50) Mean age (years) 52.1 Conventional (n 5 75) p-value 53.4 0.7564 Men (%) 42 34 0.0145 Mean educational level (years) 15.3 12.9 0.0237 Mean family income $39,540 167 $47,980 0.0576 168 CHAPTER 7 Hypothesis Testing Procedures TABLE 7–61 Weight Reduction by Treatment A Mean (SD) weight reduction 6.4 (4.1) B C 8.9 (4.0) 2.2 (3.9) Placebo 2.5 (4.3) in mean weight reduction among the four treatments? Use a 5% level of significance. (Hint: SST 5 1889.) 20. The mean lifetime for cardiac stents is 8.9 years. A medical device company has implemented some improvements in the manufacturing process and hypothesizes that the lifetime is now longer. A study of 40 new devices reveals a mean lifetime of 9.7 years with a standard deviation of 3.4 years. Is there statistical evidence of a prolonged lifetime of the stents? Run the test at a 5% level of significance. 21. A study is conducted in 100 children to assess risk factors for obesity. Children are enrolled and undergo a complete physical examination. At the examination, the height and weight of the child, their mother, and their father are measured and are converted to body mass index scores (weight(kg)/height(m)2). Data on self-reported health behaviors are captured by interview and merged with the physical examination data. The primary outcome variable is child's obesity: for analysis, children are classified as normal weight (BMI < 25) or overweight/obese (BMI ≥ 25). Data on key study variables are summarized in Table 7–62. Is there a statistically significant difference in mean age between normal and overweight/obese children? Run the test at a 5% level of significance. 22. Use the data shown in Problem 21 and test if there is an association between mother's BMI and the child's obesity status (i.e., normal versus overweight/obese)? Run the test at a 5% level of significance. 23. Use the data shown in Problem 21 and test if there is a significant difference in the proportions of normal versus overweight/obese children who are male. Run the test at a 5% level of significance. 24. A study is run to compare body mass index (BMI) in participants assigned to different diet programs and the data are analyzed in Excel®. Use the Excel results to answer the questions below. a. Complete the ANOVA table (Table 7–63). b. Write the hypotheses to be tested. c. Write the conclusion of the test. 25. A clinical trial is conducted to test the efficacy of a new drug for hypertension. The new drug is compared to a standard drug and to a placebo in a study involving n = 120 participants. The primary outcome is systolic blood pressure measured after 4 weeks on TABLE 7–62 Key Study Variables in Normal and Overweight/Obese Children Characteristics Normal Weight (n = 62) Overweight/Obese (n = 38) Total (n = 100) Mean (SD) age, years 13.4 (2.6) 11.1 (2.9) 12.5 (2.7) % Male 45% 51% 47% Normal (BMI < 25) 40 (65%) 16 (41%) 56 (56%) Overweight (BMI 25–29.9) 15 (24%) 14 (38%) 29 (29%) 7 (11%) 8 (21%) 15 (15%) Normal (BMI < 25) 34 (55%) 16 (41%) 50 (50%) Overweight (BMI 25–29.9) 20 (32%) 14 (38%) 34 (34%) 8 (13%) 8 (21%) 16 (16%) Mother's BMI Obese (BMI ≥ 30) Father's BMI Obese (BMI ≥ 30) Mean (SD) systolic blood pressure 123 (15) 139 (12) 129 (14) Mean (SD) total cholesterol 186 (25) 211 (28) 196 (26) 169 Practice Problems TABLE 7–63 ANOVA Table ANOVA SS df Between Groups 40.791667 3 Within Groups 35.166666 Total 75.958327 Source of Variation F MS p-value F crit 0.0012733 3.10 23 the assigned drug. Table 7–64 shows characteristics of study participants measured at baseline (prior to randomization). Test if the mean SBP at baseline is significantly higher in participants assigned to the standard drug as compared to those assigned to the new drug. Use a 5% level of significance. 26. Use the data shown in Problem 25 and test if there is a significant difference in the proportions of men assigned to each of the three treatments. (HINT: Are gender and treatment independent?) Use a 5% level of significance. 27. Use the data shown in Problem 25 and test whether there is a significant difference in the proportion of diabetic participants in the placebo group as compared to the standard drug group. Use a 5% level of significance. 28. Use the data shown in Problem 25 and test whether there is a significant difference in mean age among the three groups. (HINT: SStotal = 2893.) Use a 5% level of significance. 29. Some scientists believe that alcoholism is linked to social isolation. One measure of social isolation is marital status. A study of 280 adults is conducted, and each participant is classified as not alcoholic, diagnosed alcoholic, or undiagnosed alcoholic, and also by marital status (see Table 7–65). Is there significant evidence of an association? Run the appropriate test at a 5% level of significance. 30. A study is performed to examine the relationship between the concentration of plasma antioxidant vitamins and cancer risk. Table 7–66 shows data for TABLE 7–65 Social Isolation and Alcohol Diagnosed Alcoholic Undiagnosed Alcoholic Not Alcoholic Married 21 37 58 Not Married 59 63 42 TABLE 7–66 Plasma Antioxidant Vitamins in Cancer Patients and in Controls N Mean SD Stomach cancer patients 20 2.41 0.15 Controls 50 2.78 0.19 TABLE 7–64 Baseline Characteristics of Study Participants Baseline Characteristic Placebo (n = 40) Standard Drug (n = 40) New Drug (n = 40) Mean (SD) age 75.2 (4.4) 75.6 (4.8) n (%) Male 12 (30.0%) 13 (32.5%) Mean (SD) BMI 26.1 (4.9) 25.1 (3.0) 26.9 (4.1) Mean (SD) SBP 142.1 (19.2) 150.4 (19.8) 144.5 (19.7) n (%) Diabetic n (%) Current smokers 74.7 (5.6) 9 (22.5%) 13 (32.5%) 11 (27.5%) 8 (20.0%) 4 (10.0%) 2 (5.0%) 1 (2.5%) 170 CHAPTER 7 Hypothesis Testing Procedures plasma vitamin-A concentration in stomach cancer patients and in controls (participants similar to the cancer patients but free of disease). Is there a significant difference in the mean concentration of plasma antioxidant vitamins between patients with stomach cancer and controls? Run the appropriate test at a 5% level of significance.

REFERENCES 1. Ogden, C.L., Fryar, C.D., Carroll, M.D., and Flegal, K.M. "Mean body weight, height, and body mass index, United States 1960–2002." Advance Data from Vital and Health Statistics 2004; 347. 2. National Center for Health Statistics. Health, United States, 2005 with Chartbook on Trends in the Health of Americans. Hyattsville, MD: U.S. Government Printing Office, 2005. 3. Agresti, A. Categorical Data Analysis (2nd ed.). New York: John Wiley & Sons, 2002. 4. StatXact Version 7. © 2006 by Cytel, Inc., Cambridge, MA. 5. D'Agostino, R.B., Sullivan, L.M., and Beiser, A. Introductory Applied Biostatistics. Belmont, CA: Duxbury-Brooks/Cole, 2004. 6. Rosner, B. Fundamentals of Biostatistics. Belmont, CA: DuxburyBrooks/ Cole, 2006. 7. SAS version 9.1. © 2002–2003 by SAS Insitute, Inc., Cary, NC. 8. Glantz, S.A. and Slinker, B.K. Primer of Applied Regression and Analysis of Variance. New York: McGraw-Hill, 2001. 9. National Osteoporosis Foundation. Available at . 10. Snedecor, G.W. and Cochran, W.G. Statistical Methods. Ames, IA: Iowa State University Press, 1980.

8 chapter Power and Sample Size Determination W hen and W hy Key Questions •• W hy do some studies have 100 participants and others 10,000? What is the right size for samples? •• How can some studies with very small samples produce statistically significant results? •• What are the ethical issues when a study is under-powered? In the News Clinical trials are very carefully designed to ensure that the sample size is sufficiently large to ensure a high probability of observing a clinically meaningful difference in the outcome between treatments, if one exists. Such trials usually include planned interim analyses, to be conducted at various points before enrollment is complete, so as to ensure participant welfare. In some instances, clinical trials are stopped early, before the target enrollment numbers are reached. Among the various reasons that a trial might be stopped are that the treatments are found to be hugely different or not different at all, the participants suffer extensive side effects, the participants have poor compliance, or study misconduct occurs. Stopping a study early is a complex decision, and one that is never taken lightly. Oftentimes, studies that are stopped early remain controversial, such as the trial that investigated the efficacy of hormone replacement therapy in healthy postmenopausal women.1,2 The women in the trial who were taking a combination of estrogen and progestin did experience relief from symptoms of menopause and had fewer hip fractures and a lower risk of developing colon cancer. However, compared to women who did not take hormones, they had higher risks of developing breast cancer, heart disease, blood clots, and stroke. Writing Group for the Women's Health Initiative Investigators. "Risks and benefits of estrogen plus progestin in health postmenopausal women." Journal of the American Medical Association 2002; 288: 321–333. 2 National Institutes of Health. Available at .gov/whi/pr_02-7-9.pdf. 1 Dig In Imagine that a clinical trial is designed to compare a new treatment to a current available treatment (standard of care). Suppose that the new treatment costs less and has fewer adverse events than the standard of care. The investigators run the test of hypothesis comparing the primary outcomes of the new treatment and the standard of care and fail to reject the null hypothesis at a 5% level of significance. •• Which treatment would you recommend for patients and why? •• What kinds of errors are possible with hypothesis testing, and what are the implications for patients when we commit errors? •• How can we best guard against making such errors in tests of hypothesis? L earning O bjectives By the end of this chapter, the reader will be able to •• P rovide examples demonstrating how the margin of error, effect size, and variability of the outcome affect sample size computations •• Compute the sample size required to estimate population parameters with precision •• Interpret statistical power in tests of hypothesis •• Compute the sample size required to ensure high power in tests of hypothesis In Chapter 6 and Chapter 7, we presented techniques for estimation and hypothesis testing, respectively. In Chapter 6, we saw that confidence interval estimates based on larger samples had smaller margins of error. In Chapter 7, we saw that tests of hypothesis based on larger samples were more likely to detect small increases, decreases, or differences in 172 CHAPTER 8 Power and Sample Size Determination the parameter of interest. Larger samples produce more precise analyses. However, there is a point at which a larger sample size does not substantially improve the precision in the analysis. Studies should be designed to include a sufficient number of participants to adequately address the research question. Realizing at the end of a study that the sample was simply too small to answer the research question is wasteful in terms of participant and investigator time, resources to conduct the assessments, analytic efforts, and so on. It can also be viewed as unethical as participants may have been put at risk as part of a study that was unable to answer an important question. Alternatively, studies should not be too large because again resources can be wasted and some participants may be unnecessarily placed at risk. A critically important aspect of study design is determining the appropriate sample size to answer the research question. There are formulas that are used to estimate the sample size needed to produce a confidence interval estimate with a specified margin of error, or to ensure that a test of hypothesis has a high probability of detecting a meaningful difference in the parameter if one exists. Ideally, these formulas are used to generate estimates of the sample size needed to answer the study question before any data are collected. The formulas we present here generate the sample sizes required to satisfy statistical criteria. In many studies, the sample size is determined by financial or logistical constraints. For example, suppose a study is proposed to evaluate a new screening test for Down Syndrome. Suppose that the screening test is based on the analysis of a blood sample taken from women early in pregnancy. To evaluate the properties of the screening test (e.g., the sensitivity and specificity), each pregnant woman will be asked to provide a blood sample and in addition to undergo an amniocentesis. The amniocentesis is included as the gold standard, and the plan is to compare the results of the screening test to the results of the amniocentesis. Suppose that the collection and processing of the blood sample costs $250 per participant and that the amniocentesis costs $900 per participant. These financial constraints alone might substantially limit the number of women who can be enrolled. Just as it is important to consider both statistical and practical issues in interpreting the results of a statistical analysis, it is also important to weigh both statistical and logistical issues in determining the sample size for an analysis. Our focus here is on statistical considerations. Investigators must evaluate whether the sample size determined to be sufficient from a statistical standpoint is realistic and feasible. 8.1 ISSUES IN ESTIMATING SAMPLE SIZE FOR CONFIDENCE INTERVALS ESTIMATES In Chapter 6, we presented confidence intervals for various parameters—e.g., m, p, (m12m2), md, (p12p2). Confidence intervals for every parameter take the following general form: Point estimate 6 Margin of error. In Chapter 6, when we introduced the concept of confidence intervals, we derived the expression for the confidence interval for m as X ±z σ . n In practice we use the sample standard deviation, s, to e stimate the population standard deviation, σ. Note also that there is an alternative formula for estimating the mean of a continuous outcome in a single population, and it is used when the sample size is small (n [illegible] 30). It involves a value from the t distribution, as opposed to one from the standard normal distribution, to reflect the desired level of confidence. When performing sample-size computations, we use the large sample formula shown here. The resultant sample size might be small and in the analysis stage; the appropriate confidence interval formula must be used. The point estimate for the population mean is the sample σ . In planning studies, mean and the margin of error is z n we want to determine the sample size needed to ensure that the margin of error is sufficiently small to be informative. For example, suppose we want to estimate the mean weight of female college students. We conduct a study and generate a 95% confidence interval as 125 ± 40 lbs, or 85 to 165 lbs. The margin of error is so wide that the confidence interval is u ninformative. To be informative, an investigator might want the margin of error to be no more than 5 or 10 lbs (meaning that the 95% confidence interval would have a width—lower limit to upper limit—of 10 or 20 lbs). To determine the sample size needed, the investigator must specify the desired margin of error. It is important to note that this is not a statistical issue but a clinical or a practical one. The sample size needed also d epends on the outcome variable under consideration. For e xample, suppose we want to estimate the mean birth weight of infants born to mothers who smoke cigarettes during pregnancy. Birth weights in infants clearly have a much more r estricted range than weights of female college students. Therefore, we would probably want to generate a confidence interval for the mean birth weight that has a margin of error not exceeding 1 or 2 lbs. Issues in Estimating Sample Size for Confidence Intervals Estimates The margin of error in the one sample confidence interval for m can be written as E= z σ . n We want to determine the sample size, n, that ensures that the margin of error, E, does not exceed a specified value. We can take the preceding formula and, with some algebra, solve for n: 173 how attrition is addressed in planning studies through examples in the following sections. 8.1.1 Sample Size for One Sample, Continuous Outcome In studies where the plan is to estimate the mean of a conti nuous outcome variable in a single population (m), the formula for determining sample size is 2 zσ ⎞n = ⎛ ⎜ ⎟. ⎝ E⎠zσ ⎞n = ⎛ ⎜ ⎟⎝ E ⎠This formula generates the minimum sample size, n, r equired to ensure that the margin of error, E, does not exceed a specified value. To solve for n, we must input z, σ, and E. z is the value from Table 1B in the Appendix for the desired confidence level (e.g., z 5 1.96 for 95% confidence) and E is the margin of error that the investigator specifies as important from a practical standpoint. The last input is σ, the standard deviation of the outcome of interest. Sometimes it is difficult to estimate σ. When we use the prior sample size formula (or one of the other formulas that we present in the sections that follow), we are planning a study to estimate the unknown mean of a particular outcome variable in a population. It is unlikely that we know the standard deviation of that variable. In sample size computations, investigators often use a value for the standard deviation from a previous study or a study done in a different but comparable population. The sample size computation is not an application of statistical inference, and therefore it is reasonable to use an appropriate estimate for the standard deviation. The estimate can be derived from a different study that was reported in the literature; some investigators perform a small pilot study to estimate the standard deviation. A pilot study usually involves a small number of participants (e.g., n 5 10) who are selected by convenience as opposed to random sampling. Data from the participants in the pilot study are used to compute the sample standard deviation, s, which serves as a good estimate for σ in the sample size formula. Regardless of how the estimate of the variability of the outcome is derived, it should always be conservative (i.e., as large as is reasonable), so that the resultant sample size is not too small. The formula for n produces the minimum sample size to ensure that the margin of error in a confidence interval will not exceed E. In planning studies, investigators should also consider attrition or loss to follow-up. Sample size formulas for confidence intervals give the number of participants needed with complete data to ensure that the margin of error in the confidence interval does not exceed E. We illustrate where z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., z 5 1.96 for 95%), σ is the standard deviation of the outcome variable, and E is the desired margin of error. The preceding formula generates the minimum number of subjects required to ensure that the margin of error in the confidence interval for m does not exceed E. Example 8.1. An investigator wants to estimate the mean systolic blood pressure in children with congenital heart disease who are between the ages of 3 and 5 years of age. How many children should be enrolled in the study? To determine the requisite sample size, the investigator must provide the necessary inputs for the sample size formula, 2 2 zσ ⎞n = ⎛ ⎜ ⎟. ⎝ E ⎠Suppose that a 95% confidence interval is planned; thus, z 5 1.96. The investigator must then specify the margin of error. Suppose the investigator decides that a margin of error of 5 units will be sufficiently precise. The final input is the standard deviation of systolic blood pressure in children with congenital heart disease. A suitable estimate might be found in the literature. There are certainly data on systolic blood pressure in the Framingham Heart Study, but those data might not be sufficiently comparable as they are based on adults, most of whom are probably free of congenital heart disease. Suppose the investigators conduct a literature search and find that the standard deviation of systolic blood pressure in children with other cardiac defects is between 15 and 20. To estimate the sample size, we consider the larger standard deviation. This will produce the most conservative (largest) sample size, 2 2 zσ ⎞⎛ 1.96 × 20 ⎞n = ⎛ ⎜ ⎟ = ⎜ ⎟ = 61.5. ⎝ E ⎠⎝ 5 To ensure that the 95% confidence interval estimate of the mean systolic blood pressure in children between the ages of 3 and 5 with congenital heart disease is within 5 units of the true 174 CHAPTER 8 Power and Sample Size Determination mean, a sample of size 62 is needed. Note that we always round up, as the sample size formulas always generates the minimum number of subjects needed to ensure the specified precision. Had we assumed a standard deviation of 15, the sample size would have been n 5 35. Because the estimates of the standard deviation are derived from studies of children with other cardiac defects, it is advisable to use the larger standard deviation and plan for a study with 62 children. Selecting the smaller sample size could potentially produce a confidence interval estimate with a much larger margin of error. Example 8.2. An investigator wants to estimate the mean birth weight of infants born full term (approximately 40 weeks gestation) to mothers who are 19 years of age and under. The mean birth weight of infants born full term to mothers 20 years of age and older is 3510 g with a standard deviation of 385 g.1 How many women 19 years of age or younger must be enrolled in the study to ensure that a 95% confidence interval estimate of the mean birth weight of their infants has a margin of error not exceeding 100 g? 2 2 ⎛ 1.96 × 385 ⎞⎛ zσ ⎞n =⎜ =⎜ ⎟ = 56.9. ⎝ E ⎟⎠⎝ 100 To ensure that the 95% confidence interval estimate of the mean birth weight is within 100 g of the true mean, a sample of size 57 is needed. In planning the study, the investigator must consider the fact that some women may deliver prematurely. If women are enrolled into the study during pregnancy, then more than 57 women will need to be enrolled so that after excluding those who deliver prematurely, 57 who deliver full term will be available for analysis. For example, if 5% of the women are expected to deliver prematurely (i.e., 95% will deliver full term), then 60 women must be enrolled to ensure that 57 deliver full term. The number of women who must be enrolled, N, is computed as N (number to enroll) 3 (% retained) 5 Desired sample size, N 3 0.95 5 57, 57 N = = 60. 0.95 8.1.2 Sample Size for One Sample, Dichotomous Outcome In studies where the plan is to estimate the proportion of successes in a dichotomous outcome variable in a single population, the formula for determining sample size is 2 z⎞n = p (1 − p) ⎛ ⎜ ⎟. ⎝ E⎠where z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., z 5 1.96 for 95%), E is the desired margin of error, and p is the proportion of successes in the population. Here we are planning a study to generate a 95% confidence interval for the unknown population proportion, p, and the formula to estimate the s ample size needed requires p! Obviously, this is a circular problem—if we knew the proportion of successes in the population, then a study would not be necessary. To estimate the sample size, we need an approximate value of p. The range of p is 0 to 1 and the range of p(1 2 p) is 0 to 0.25. The value of p that maximizes p(1 2 p) is p 5 0.5. Thus, if there is no information available to approximate p, then p 5 0.5 can be used to generate the most conservative, or largest, sample size. Example 8.3. An investigator wants to estimate the proportion of freshmen at his university who currently smoke cigarettes (i.e., the prevalence of smoking). How many freshmen should be involved in the study to ensure that a 95% confidence interval estimate of the proportion of freshmen who smoke is within 0.05 or 5% of the true proportion? If we have no information on the proportion of freshmen who smoke, we use the following to estimate the sample size: 2 2 z⎞⎛ 1.96 ⎞n = n = p (1 − p) ⎛ ⎜ ⎟ = 0.5 (1 − 0.5) ⎜ 384.2. ⎝ 0.05 ⎟⎠⎝ E⎠To ensure that the 95% confidence interval estimate of the proportion of freshmen who smoke is within 5% of the true proportion, a sample of size 385 is needed. Suppose that a similar study was conducted two years ago and found that the prevalence of smoking was 27% among freshmen. If the investigator believes that this is a reasonable estimate of prevalence today, it can be used to plan the study: 2 2 z⎞ ⎛ 1.96 ⎞= n = p (1 − p) ⎛ ⎜ ⎟= 0.27 (1 − 0.27) ⎜ 302.9. ⎝ 0.05 ⎟⎠⎝ E⎠To ensure that the 95% confidence interval estimate of the proportion of freshmen who smoke is within 5% of the true proportion, a sample of size 303 is needed. Notice that this sample size is substantially smaller than the one estimated previously. Having some information on the magnitude of the proportion in the population always produces a sample size that is less than or equal to the one based on a population proportion of 0.5. However, the estimate must be realistic. Example 8.4. An investigator wants to estimate the prevalence of breast cancer among women who are between 40 and 45 years of age living in Boston. How many women must Issues in Estimating Sample Size for Confidence Intervals Estimates be involved in the study to ensure that the estimate is precise? National data suggest2 that 1 in 235 women are diagnosed with breast cancer by age 40. This translates to a proportion of 0.0043 (0.43%), or a prevalence of 43 per 10,000 women. Suppose the investigator wants the estimate to be within 10 per 10,000 women with 95% confidence. The sample size is computed as 2 z⎞n = p (1 − p) ⎛ ⎜ ⎟= ⎝ E⎠2 1.96 ⎞= 16,447.8. 0.0043(1 − 0.0043) ⎛ ⎜ ⎝ 0.0010 ⎟⎠A sample of size n 5 16,448 ensures that a 95% confidence interval estimate of the prevalence of breast cancer is within 0.0010 (or within 10 women per 10,000) of its true value. This is a situation where investigators might decide that a sample of this size is not feasible. Suppose that the investigators believe that a sample of size 5000 is reasonable from a practical point of view. How precisely can they estimate the prevalence with a sample of size n 5 5000? Recall from Chapter 6 that the confidence interval formula to estimate prevalence is p) p) p (1 − p ±z . n Assuming that the prevalence of breast cancer in the sample is close to that based on national data, we would expect the margin of error to be approximately equal to z p) p) p (1 − = 1.96 n 0.0043(1 − 0.0043) = 0.0018. 5000 With n 5 5000 women, a 95% confidence interval would be expected to have a margin of error of 0.0018 (or 18 per 10,000). The investigators must decide if this is sufficiently precise to answer the research question. Note that this calculation is based on the assumption that the prevalence of breast cancer in Boston is similar to that reported nationally. This may or may not be a reasonable assumption. In fact, it is the objective of the current study to estimate the prevalence in Boston. With input from clinical investigators and biostatisticians, the research team must carefully evaluate the implications of selecting a sample of size n 5 5000, n 5 16,448, or any size in between. 8.1.3 Sample Sizes for Two Independent Samples, Continuous Outcome In studies where the plan is to estimate the difference in means between two independent populations (m1 – m2), the formula 175 for determining the sample sizes required in each comparison group is zσ ⎞ni = 2 ⎛ ⎜ ⎝ E ⎠2 where ni is the sample size required in each group (i 5 1, 2), z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., z 5 1.96 for 95%), and E is the desired margin of error. σ again reflects the standard deviation of the outcome variable. Recall from Chapter 6 that when we generated a confidence interval estimate for the difference in means, we used Sp (the pooled estimate of the common standard deviation) as a measure of variability in the outcome, where Sp is computed as Sp = (n1 − 1) s12 + (n2 − 1) s22 n1 + n2 − 2 If data are available on the variability of the outcome in each comparison group, then Sp can be computed and used in the sample size formula. However, it is more often the case that data on the variability of the outcome are available from only one group, often the untreated (e.g., placebo control) or unexposed group. When planning a clinical trial to investigate a new drug or procedure, data are often available from other trials that involved a placebo or an active control group (i.e., a standard medication or treatment given for the condition under study). The standard deviation of the outcome variable measured in p atients assigned to the placebo, control, or unexposed group can be used to plan a future trial. This situation is illustrated below. Note that the formula shown here generates sample size estimates for samples of equal size. If a study is planned where different numbers of patients will be assigned or different numbers of patients will comprise the comparison groups, then alternative formulas can be used. Interested readers can see Howell for more details.3 Example 8.5. An investigator wants to plan a clinical trial to evaluate the efficacy of a new drug designed to increase HDL cholesterol (the "good" cholesterol). The plan is to enroll participants and to randomly assign them to receive either the new drug or a placebo. HDL cholesterol will be measured in each participant after 12 weeks on the assigned treatment. Based on prior experience with similar trials, the investigator expects that 10% of all participants will be lost to follow-up or will drop out of the study. A 95% confidence interval will be estimated to

quantify the difference in mean HDL levels between patients taking the new drug as compared to the placebo. The investigator would like the margin of error to be no more than 3 units. How many patients should be recruited into the study? 176 CHAPTER 8 Power and Sample Size Determination The sample sizes (i.e., the number of participants
who must receive the new drug and the number who must receive a placebo) are computed as 2 zσ ⎞ni = 2 ⎛⎜ ⎟. ⎝ E ⎠ A major issue is determining the variability in the outcome of interest, here the standard deviation (σ) of HDL cholesterol. To plan this study, we can use data from the Framingham Heart Study. In participants who attended the
seventh examination of the Framingham Offspring Study and were not on treatment for high cholesterol, the standard deviation of HDL cholesterol is 17.1. We use this value and the other inputs to compute the sample sizes: 2 2 zσ ⎞⎛ 1.96 × 17.1 ⎞ni = 2⎜ ⎟ = 2⎜ ⎟= 249.6. ⎝ E⎠⎝ 3 Samples of size n1 5 250 and n2 5 250 will ensure that the 95%
confidence interval for the difference in mean HDL levels will have a margin of error of no more than 3 units. Again, these sample sizes refer to the numbers of participants with complete data. The investigators hypothesized a 10% attrition (or drop-out) rate in both groups. To ensure that a total sample size of 500 is available at 12 weeks, the
investigator needs to recruit more participants to allow for attrition. estimated to quantify the difference in weight lost between the two diets, and the investigator would like the margin of error to be no more than 3 lbs. How many children should be recruited into the study? The sample sizes (i.e., the number of children who must follow the low-fat
diet and the number of children who must follow the low-carbohydrate diet) are computed as 2 zσ ⎞ni = 2 ⎛⎜ ⎟. ⎝ E ⎠Again, the issue is determining the variability in the outcome of interest, here the standard deviation (σ) in pounds lost over 8 weeks. To plan this study, investigators use data from adult studies. Suppose one such study compared the
same diets in adults and involved 100 participants in each diet group. The study reported a standard deviation in weight lost over 8 weeks on a low-fat diet of 8.4 lbs and a standard deviation in weight lost over 8 weeks on a low-carbohydrate diet of 7.7 lbs. These data can be used to estimate the common standard deviation in weight lost as Sp = (n1 −
1) s12 + (n2 − 1) s22 n1 + n2 − 2 (100 − 1)(8.4)2 + (100 − 1)(7.7)2 100 + 100 − 2 = = 8.1. N (number to enroll) 3 (% retained) 5 Desired sample size. We now use this value and the other inputs to compute the sample sizes: N 3 0.90 5 500. zσ ⎞⎛ 1.96 × 8.1 ⎞= ni = 2⎜ ⎟ = 2⎜ 56.0. ⎟⎠⎝ ⎝ E ⎠3 N= 500 = 556. 0.90 The investigator must enroll 556
participants. Each patient will be randomly assigned to receive either the new drug or a placebo; assuming that 10% are lost to follow-up, 500 will be available for analysis. Example 8.6. An investigator wants to compare two diet programs in children who are obese. One diet is a lowfat diet and the other is a low-carbohydrate diet. The plan is to enroll
children and weigh them at the start of the study. Each child will then be randomly assigned to either the lowfat or the low-carbohydrate diet. Each child will follow the assigned diet for 8 weeks, at which time he or she will again be weighed. The number of pounds lost will be computed for each child. Based on data reported from diet trials in adults,
the investigator expects that 20% of all children will not complete the study. A 95% confidence interval will be 2 2 Samples of size n1 5 56 and n2 5 56 will ensure that the 95% confidence interval for the difference in weight lost between diets will have a margin of error of no more than 3 lbs. Again, these sample sizes refer to the numbers of children
with complete data. The investigators hypothesized a 20% attrition rate. To ensure that the total sample size of 112 is available at 8 weeks, the investigator needs to recruit more participants to allow for attrition. N (number to enroll) 3 (% retained) 5 Desired sample size. N 3 0.80 5 112. N= 112 = 140. 0.80 The investigator must enroll 140 children.
Each child will be randomly assigned to either the low-fat or low-carbohydrate Issues in Estimating Sample Size for Confidence Intervals Estimates diet; assuming that 20% are lost to follow-up, 112 will be available for analysis. 8.1.4 Sample Size for Matched Samples, Continuous Outcome In studies where the plan is to estimate the mean difference
of a continuous outcome (md) based on matched data, the formula for determining sample size is ⎛ zσ ⎞n=⎜ d ⎟⎝ E ⎠2 2 3 lbs of the true mean, a sample size of 36 children with complete data is needed. The investigators hypothesized a 30% attrition rate. To ensure that the total sample size of 36 is available at 16 weeks, the investigator needs to
recruit 52 participants to allow for attrition: N (number to enroll) 3 (% retained) 5 Desired sample size. N 3 0.70 5 36. N = where z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., z 5 1.96 for 95%), E is the desired margin of error, and σd is the standard deviation of the difference scores. It is
extremely important that the standard deviation of the difference scores (e.g., the differences based on measurements over time, or the differences between matched pairs) is used here to appropriately estimate the sample size. Example 8.7. Consider again the diet study proposed in Example 8.6. Suppose the investigator is considering an alternative
design—a crossover trial. In the crossover trial, each child follows each diet for 8 weeks. At the end of each 8-week period, the weight lost during that period is measured. Children will be randomly assigned to the first diet (i.e., some will follow the low-fat diet first while others will follow the low-carbohydrate diet first). The investigators expect to
lose 30% of the participants over the course of the 16-week study. The difference in weight lost on the low-fat diet and the low-carbohydrate diet will be computed for each child, and a confidence interval for the mean difference in weight lost will be computed. How many children are required to ensure that a 95% confidence interval estimate for the
mean difference in weight lost has a margin of error of no more than 3 lbs? To compute the sample size, we need an estimate of the variability in the difference in weight lost between diets as opposed to the variability in weight loss in either diet program, as was required in Example 8.6. Suppose that the standard deviation of the difference in weight
loss between a low-fat diet and a low-carbohydrate diet is approximately 9.1 lbs based on a crossover trial conducted in adults. The sample size is computed as 2 ⎛ 1.96 × 9.1 ⎞⎛ zσ ⎞n = ⎜ d ⎟=⎜ ⎟= 35.3. ⎝ E ⎠⎝ 3 To ensure that the 95% confidence interval estimate of the mean difference in weight lost between diets is within 177 36 = 52. 0.70 Notice
that fewer patients are needed in the crossover trial as compared to the trial with two concurrent or parallel comparison groups considered in Example 8.6. In the crossover trial, each child is given each treatment (diet). Crossover trials can be efficient in terms of sample size. However, the crossover can be challenging in other respects. The
crossover trial is problematic when there is loss to follow-up. Because each participant contributes data on each treatment, when a participant is lost after completing the first treatment, his or her data are unusable without information on the second treatment. Diet studies are particularly problematic in terms of participant loss to follow-up.
Sometimes participants start to lose weight and feel they no longer need the study, and so they drop out. Other participants do not lose weight and become frustrated and drop out. These and other issues must be considered carefully in determining the most appropriate design for any study. 8.1.5 Sample Sizes for Two Independent Samples,
Dichotomous Outcome In studies where the plan is to estimate the difference in proportions between two independent populations (i.e., to estimate the risk difference p1 – p2), the formula for determining the sample sizes required in each comparison group is given as z⎞ni = ⎡⎣ p1 (1 − p1) + p2 (1 − p2) ⎤⎦⎛⎜ ⎟⎝ E⎠2 where ni is the sample size
required in each group (i 5 1, 2), z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., z 5 1.96 for 95%), E is the desired margin of error, and p1 and p2 are the proportions of successes in each comparison group. Again, here we are planning a study to generate a 95% confidence interval for the
difference in unknown proportions, and the formula to estimate the sample sizes needed requires p1 and p2. To estimate 178 CHAPTER 8 Power and Sample Size Determination the sample size, we need approximate values of p1 and p2. The values of p1 and p2 that maximize the sample sizes are p1 5 p2 5 0.5. Thus, if there is no information
available to approximate p1 and p2, 0.5 can be used to generate the most conservative, or largest, sample sizes. Similar to the situation we presented in Section 8.1.3 for two independent samples and a continuous outcome, it may be the case that data are available on the proportion of successes in one group, usually the untreated (e.g., placebo
control) or unexposed group. If so, that known proportion can be used for both p1 and p2 in the formula shown here. This formula generates sample size estimates for samples of equal size. If a study is planned where different numbers of patients will be assigned or different numbers of patients will comprise the comparison groups, then alternative
formulas can be used. Interested readers can see Fleiss for more details.4 Example 8.8. An investigator wants to estimate the impact of smoking during pregnancy on premature delivery. Normal pregnancies last approximately 40 weeks and premature deliveries are those that occur before 37 weeks. The 2005 National Vital Statistics report indicates
that approximately 12% of infants are born prematurely in the United States.5 The investigator plans to collect data through medical record review and generate a 95% confidence interval for the difference in proportions of infants born prematurely to women who smoked during pregnancy as compared to those who did not. How many women should
be enrolled in the study to ensure that the 95% confidence interval for the difference in proportions has a margin of error of no more than 0.04, or 4%? The sample sizes (i.e., numbers of women who smoked and did not smoke during pregnancy) are computed using the previous formula. National data suggest that 12% of infants are born prematurely.
We use this estimate for both groups in the sample size computation: 2 z⎞ni = ⎡⎣ p1 (1 − p1) + p2 (1 − p2) ⎤⎦⎛⎜ ⎟ = ⎝ E⎠2 1.96 [0.12(1 − 0.12) + 0.12(1 − 0.12)]⎛⎜⎝ 0.04 ⎞⎟⎠ = 507.1. Samples of size n1 5 508 women who smoked during pregnancy and n2 5 508 women who did not smoke during pregnancy will ensure that the 95% confidence interval for
the difference in proportions who deliver prematurely will have a margin of error of no more than 4%. Is attrition an issue here? Example 8.9. An investigator wants to estimate the impact of smoking on the incidence of prostate cancer. The incidence of prostate cancer by age 70 is approximately 1 in 6 men (17%). A United States study reported that
heavy smokers (defined as men smoking more than a pack a day for 40 years) had twice the risk of developing prostate cancer as compared to nonsmokers.6 An investigator plans to replicate the study in England using a cohort study design. Men who are free of prostate cancer will be enrolled at age 50 and followed for 30 years. The plan is to enroll
approximately equal numbers of smokers and nonsmokers in the study and follow them prospectively for the outcome of interest—a diagnosis of prostate cancer. The plan is to generate a 95% confidence interval for the difference in proportions of smoking and nonsmoking men who develop prostate cancer. The study will follow men for 30 years, and
investigators expect to lose approximately 20% of the participants over the course of the f ollow-up period. How many men should be enrolled in the study to ensure that the 95% confidence interval for the difference in proportions has a margin of error of no more than 0.05, or 5%? The sample sizes (i.e., numbers of men who smoke and do not smoke)
are computed using the previous formula. The estimates of the incidence of prostate cancer from the United States study are used to design the study planned for England (i.e., p1 5 0.34 and p2 5 0.17 are the estimates of incidence of prostate cancer in smokers versus nonsmokers, respectively): 2 z⎞ni = ⎡⎣ p1 (1 − p1) + p2 (1 − p2) ⎤⎦⎛⎜ ⎟=⎝ E⎠1.96
[0.34 (1 − 0.34) + 0.17 (1 − 0.17)]⎛⎜⎝ 0.05 ⎞⎟⎠2 = 561.6. Samples of size n1 5 562 men who smoke and n2 5 562 men who do not smoke will ensure that the 95% confidence interval for the difference in the incidence of prostate cancer will have a margin of error of no more than 5%. The sample sizes here refer to the numbers of men with complete
data. The investigators hypothesized 20% attrition (in each group). To ensure that the total sample size of 1124 is available for analysis, the investigator needs to recruit more participants to allow for attrition: N (number to enroll) 3 (% retained) 5 Desired sample size. N 3 0.80 5 1124. N= 1124 = 1405. 0.80 The investigator must enroll 1405 men.
Because enrollment will be on the basis of smoking status at baseline, the investigator will enroll approximately 703 men who smoke and 703 men who do not. Issues in Estimating Sample Size for Hypothesis Testing 8.2 ISSUES IN ESTIMATING SAMPLE SIZE FOR HYPOTHESIS TESTING In Chapter 7, we introduced hypothesis testing techniques
for means, proportions, differences in means, and differences in proportions. Whereas each test involves details that are specific to the outcome of interest (continuous, dichotomous, categorical, or ordinal) and to the number of comparison groups (one, two, or more than two), there are elements common to each test. For example, in each test of
hypothesis there are two errors that can be committed. The first is called a Type I error and refers to the situation where we incorrectly reject H0 when, in fact, it is true. In the first step of any test of hypothesis, we select a level of significance, α, and α 5 P(Type I error) 5 P(Reject H0 | H0 is true). Because we purposely select a small value for α, we
control the probability of committing a Type I error. The second type of error is called a Type II error and it is defined as the probability we do not reject H0 when it is false. The probability of a Type II error is denoted β, and β 5 P(Type II error) 5 P(Do not reject H0|H0 is false). In hypothesis testing, we usually focus on power, which is defined as the
probability that we reject H0 when it is false, Power 5 12β 5 P(Reject H0|H0 is false). Power is the probability that a test correctly rejects a false null hypothesis. A good test is one with low probability of committing a Type I error (i.e., small α) and high power (i.e., small β, high power). In the following sections, we present formulas to determine the
sample size required to ensure that a test has high power. The sample size computations depend on the level of significance, α, the desired power of the test (equivalent to 1 2 β), the variability of the outcome, and the effect size. The effect size is the difference in the parameter of interest that represents a clinically meaningful difference. Similar to
the margin of error in confidence interval applications, the effect size is determined based on clinical or practical criteria and not statistical criteria. The concept of statistical power can be difficult to grasp. Before presenting the formulas to determine the sample sizes required to ensure high power in a test, we first discuss power from a conceptual
point of view. Suppose we want to test the following hypotheses at α 5 0.05, 179 H0: m 5 90, H1: m 90. To test the hypotheses, suppose we select a sample of size n 5 100. For this example, assume that the standard deviation of the outcome is σ 5 20. We compute the sample mean and then must decide whether the sample mean provides evidence to
support the research hypothesis or not. This is done by computing a test statistic and comparing the test statistic to an appropriate critical value. If the null hypothesis is really true (m 5 90), then we are likely to select a sample whose mean is close in value to 90. However, it is also possible to select a sample whose mean is much larger or much
smaller than 90. Recall from the Central Limit Theorem (see Chapter 5), that for large n (here n 5 100 is s ufficiently large) the distribution of the sample means is a pproximately normal with a mean of μ x = μ = 90 and a standard deviation of σ x = σ = 20 = 2.0. If n 100 the null hypothesis is true, it is possible to observe any sample mean shown in
Figure 8–1, as all are possible under H0: m 5 90. When we set up the decision rule for our test of hypothesis, we determine critical values based on α 5 0.05 and a two-sided test. When we run tests of hypothesis, we usually standardize the data (e.g., convert to z or t) and the critical values are appropriate values from the probability distribution used
in the test. To facilitate interpretation, we continue this discussion with w X as opposed to z. The critical values of w X for a two-sided test with α 5 0.05 are 86.08 and 93.92 (these values correspond to 21.96 and 1.96, respectively, on the z scale), so the decision rule is Reject H0 if X w} , 86.08 or if X w} . 93.92. The rejection region is shown in the
tails of Figure 8–2. The areas in the two tails of the curve represent the probability of a Type I error, α 5 0.05. This concept was discussed in some detail in Chapter 7. Now suppose that the alternative hypothesis, H1, is true (m 90) and that the true mean is actually 94. Figure 8–3 shows the distributions of the sample mean under the null and
alternative hypotheses. The values of the sample mean are shown along the horizontal axis. If the true mean is 94, then the alternative hypothesis is true. In our test, we selected a 5 0.05 and reject H0 if the observed sample mean exceeds 93.92 (focusing on the upper tail of the rejection region for now). The critical value (93.92) is indicated by the
vertical line. The probability of a Type II error is denoted β, and β 5 P(Do not reject H0 | H0 is false). 180 CHAPTER 8 Power and Sample Size Determination FIGURE 8–1 Distribution of −X Under H0:m 5 90 84 86 88 90 92 96 94 FIGURE 8–2 Rejection Region for Test H0:m 5 90 Versus H1:m 90 at α 5 0.05 2 2 84 86 88 β is shown in Figure 8–3 as
the area under the rightmost curve (H1) to the left of the vertical line (where we do not reject H0). Power is defined as 1 2 β 5 P(Reject H0 | H0 is false) 90 92 94 96 and is shown in Figure 8–3 as the area under the rightmost curve (H1) to the right of the vertical line (where we reject H0). We previously noted that β and power were related to a, the
variability of the outcome and the effect size. From Figure 8–3 we can see what happens to β and power if we Issues in Estimating Sample Size for Hypothesis Testing 181 FIGURE 8–3 Distribution of −X Under H0:m 5 90 and Under H1:m 5 94 Power 2 84 86 88 90 92 94 Do Not Reject H0 increase a. For example, suppose we increase α to a 5 0.10.
The upper critical value would be 92.56 instead of 93.92. The vertical line would shift to the left, increasing a, decreasing β, and increasing power. Whereas a better test is one with higher power, it is not advisable to increase a as a means to increase power. Nonetheless, there is a direct relationship between α and power (as α increases, so does
power). β and power are also related to the variability of the outcome and to the effect size. The effect size is the difference in the parameter of interest (e.g., m) that represents a clinically meaningful difference. Figure 8–3 graphically displays a, β, and power when the difference in the mean under the null as compared to the alternative hypothesis is
4 units (i.e., 90 v ersus 94). Figure 8–4 shows the same components for the situation where the mean under the alternative hypothesis is 98. Notice that there is much higher power when there is a larger difference between the mean under H0 as compared to H1 (i.e., 90 versus 98). A statistical test is much more likely to reject the null hypothesis in
favor of the alternative if the true mean is 98 than if the true mean is 94. Notice in Figure 8–4 that there is little overlap in the distributions under the null and alternative hypotheses. If a sample mean of 97 or higher is observed, it is very unlikely that it came from a distribution whose mean is 90. In Figure 8–3, if we observed a sample 96 98 100
Reject H0 mean of 93, it would not be as clear as to whether it came from a distribution whose mean is 90 or one whose mean is 94. In the following sections, we provide formulas to determine the sample size needed to ensure that a test has high power. In designing studies, most people consider power of 80% or 90% (just as we generally use 95% as
the confidence level for confidence interval estimates). The inputs for the sample size formulas include the desired power, the level of significance, and the effect size. The effect size is selected to represent a clinically meaningful or practically important difference in the parameter of interest. This will be illustrated in examples that follow. The
formulas we present produce the minimum sample size to ensure that the test of hypothesis will have a specified probability of rejecting the null hypothesis when it is false (i.e., a specified power). In planning studies, investigators again must account for attrition or loss to follow-up. The formulas shown produce the minimum number of participants
needed with complete data. In the examples, we also illustrate how attrition is addressed in planning studies. 8.2.1 Sample Size for One Sample, Continuous Outcome In studies where the plan is to perform a test of hypothesis comparing the mean of a continuous outcome variable in 182 CHAPTER 8 Power and Sample Size Determination Figure 8–
4 Distribution of −X Under H0:m 5 90 and Under H1:m 5 98 Power 2 84 86 90 88 92 Do Not Reject H0 98 100 102 104 Reject H0 a single population to a known mean, the hypotheses of interest are H0: m 5 m0, 80% power z0.8 5 0.84 90% power z0.9 5 1.282 ES is the effect size, defined as H1: m m0 where m0 is the known mean (e.g., a historical
control). Formulas to determine the sample size required to ensure high power in tests of hypothesis are often based on two-sided tests. Here we consider two-sided tests. If a one-sided test is planned, the sample size forumulas must be modified. The formula for determining sample size to ensure that the test has a specified power is ⎛ z 1− α 2 + z 1−
β ⎞n=⎜ ⎟ES ⎝ 96 94 2 where α is the selected level of significance and z12α/2 is the value from the standard normal distribution holding 12α/2 below it. For example, if a 5 0.05, then 12α/2 5 0.975 and z 5 1.960. 12β is the selected power and z12β is the value from the standard normal distribution holding 12β below it. Sample size estimates for
hypothesis testing are often based on achieving 80% or 90% power. The z12β values for these popular scenarios are ES = μ1 − μ0 , σ where m0 is the mean under H0, m1 is the mean under H1, and σ is the standard deviation of the outcome of interest. The numerator of the effect size, the absolute value of the difference in means |m12m0|,
represents what is considered a clinically meaningful or practically important difference in means. Similar to the issue we faced when determining sample size to estimate confidence intervals, it can sometimes be difficult to estimate the standard deviation. In sample size computations, investigators often use a value for the standard deviation from a
previous study or a study performed in a different population. Regardless of how the estimate of the variability of the outcome is derived, it should always be conservative (i.e., as large as is reasonable) so that the resultant sample size will not be too small. Example 8.10. An investigator hypothesizes that in people free of diabetes, fasting blood
glucose—a risk factor Issues in Estimating Sample Size for Hypothesis Testing for coronary heart disease—is higher in those who drink at least two cups of coffee per day. A cross-sectional study is planned to assess the mean fasting blood glucose levels in people who drink at least two cups of coffee per day. The mean fasting blood glucose level in
people free of diabetes is reported as 95 mg/dl with a standard deviation of 9.8 mg/dl.7 If the mean blood glucose level in people who drink at least two cups of coffee per day is 100 mg/dl, this would be important clinically. How many patients should be enrolled in the study to ensure that the power of the test is 80% to detect this difference? A two-
sided test will be used with a 5% level of significance. We first compute the effect size, ES = μ1 − μ0 100 − 95 = = 0.51. σ 9.8 We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size. 2 ⎛ z 1− α 2 + z1− β ⎞⎛ 1.96 + 0.84 ⎞n=⎜ ⎟=⎜ 0.51 ⎟= 30.1. ⎝ ES 2 A sample of size n 5 31 will ensure
that a two-sided test with α 5 0.05 has 80% power to detect a 5-mg/dl difference in mean fasting blood glucose levels. In the planned study, participants will be asked to fast overnight and to provide a blood sample for the analysis of glucose levels. Based on prior experience, the investigators hypothesize that 10% of the participants will fail to fast or
will refuse to submit the blood sample (i.e., they will fail to follow the study protocol). Therefore, a total of 35 participants will be enrolled in the study to ensure that 31 are available for analysis: N (number to enroll) 3 (% following protocol) 5 Desired sample size N 3 0.90 5 31 31 N 5 }} 5 35. 0.90 Example 8.11. In Example 7.1, we conducted a test of
hypothesis to assess whether there was a reduction in expenditures on health care and prescription drugs in 2005 as compared to 2002. In 2002, Americans paid an average of $3302 per year on health care and prescription drugs.8 An investigator sampled 100 Americans and found that they spent a mean of $3190 on health care and prescription
drugs with a standard deviation of $890. The test failed to reject the null hypothesis, and the investigator is concerned that the sample size was too 183 small. How many participants are needed to ensure that the test has 80% power to detect a difference of $150 in expenditures (a difference the investigator feels is meaningful)? A two-sided test will
be used with a 5% level of significance. We first compute the effect size, ES = μ1 − μ0 150 = = 0.17. σ 890 Notice that the numerator is 150, the difference specified by the investigator. We also use the standard deviation measured in the previous study (s 5 890) to inform the design of the next study. We now substitute the effect size and the
appropriate z values for the selected a and power to compute the sample size, 2 2 ⎛ z 1− α 2 + z1− β ⎞⎛ 1.96 + 0.84 ⎞= 271.3. n=⎜ = ⎟⎜ ⎟⎝ ES 0.17 ⎠A sample of size n 5 272 will ensure that a two-sided test with α 5 0.05 has 80% power to detect a $150 difference in expenditures. Thus, the study with n 5 100 participants was underpowered to detect
a meaningful difference ($150) in expenditures. 8.2.2 Sample Size for One Sample, Dichotomous Outcome In studies where the plan is to perform a test of hypothesis comparing the proportion of successes in a dichotomous outcome variable in a single population to a known proportion, the hypotheses of interest are H0: p 5 p0, H1: p p0 where p0 is
the known proportion (e.g., a historical control). The formula for determining sample size to ensure that the test has a specified power is ⎛ z 1− α 2 + z 1− β ⎞n=⎜ ⎟ES ⎝ 2 where α is the selected level of significance and z12α/2 is the value from the standard normal distribution holding 12α/2 below it. 12β is the selected power, z12β is the value
from the standard normal distribution holding 12β below it, and ES is the effect size, defined as ES = p1 − p0 p0 (1 − p0) , 184 CHAPTER 8 Power and Sample Size Determination where p0 is the proportion under H0 and p1 is the proportion under H1. The numerator of the effect size, the absolute value of the difference in proportions |p1 2 p0|,
again represents what is considered a clinically meaningful or practically important difference in proportions. Example 8.12. A recent report from the Framingham Heart Study indicated that 26% of people free of cardiovascular disease had elevated LDL cholesterol levels, defined as LDL of more than 159 mg/dl.9 An investigator hypothesizes that a
higher proportion of patients with a history of cardiovascular disease will have elevated LDL cholesterol. How many patients should be studied to ensure that the power of the test is 90% to detect an absolute difference of 5% in the proportion with elevated LDL cholesterol? A two-sided test will be used with a 5% level of significance. We first compute
the effect size, ES = p1 − p0 p0 (1 − p0) = 0.05 = 0.11. 0.26 (1 − 0.26) We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2 ⎛ z 1− α 2 + z1− β ⎞1.96 + 1.282 ⎞n=⎜ =⎛⎜ ⎟= 868.6. ⎜⎝ ⎝ ES ⎠0.11 ⎠2 A sample of size n 5 869 will ensure that a two-sided test with α 5 0.05 has 90%
power to detect a difference of 5% in the proportion of patients with a history of cardiovascular disease who have an elevated LDL cholesterol level. Example 8.13. A medical device manufacturer produces implantable stents. During the manufacturing process, approximately 10% of the stents are deemed to be defective. The manufacturer wants to
test whether the percentage of defective stents is more than 10% in one of his plants. If the process produces more than 15% defective stents, then corrective action must be taken. Therefore, the manufacturer wants the test to have 90% power to detect a difference in proportions of this magnitude. How many stents must be evaluated? A two-sided
test will be used with a 5% level of significance. We first compute the effect size, ES = p1 − p0 p0 (1 − p0) = 0.15 − 0.10 0.10 (1 − 0.10) = 0.17. We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2 2 ⎛ z 1− α 2 + z1− β ⎞⎛ 1.96 + 1.282 ⎞= 363.7. n=⎜ = ⎟⎜ ⎟⎝ ES 0.17 A sample of
size n 5 364 stents will ensure that a two-sided test with α 5 0.05 has 90% power to detect a difference of 5% in the proportion of defective stents. 8.2.3 Sample Sizes for Two Independent Samples, Continuous Outcome In studies where the plan is to perform a test of hypothesis comparing the means of a continuous outcome variable in two
independent populations, the hypotheses of interest are H0: m1 5 m2 H1: m1 m2, where m1 and m2 are the means in the two comparison populations. The formula for determining the sample sizes to ensure that the test has a specified power is ⎛ z1− α 2 + z1− β ⎞ni = 2⎜ ⎟⎝ ES 2 where ni is the sample size required in each group (i 5 1, 2), α is the
selected level of significance, z12α/2 is the value from the standard normal distribution holding 12a/2 below it, 12β is the selected power, and z12β is the value from the standard normal distribution holding 12β below it. ES is the effect size, defined as ES = μ1 − μ2 , σ where | m1 2 m2| is the absolute value of the difference in means between the two
groups expected under the alternative hypothesis, H1. σ is the standard deviation of the outcome of interest. Recall from Chapter 6 when we performed tests of hypothesis comparing the means of two independent groups, we used Sp (the pooled estimate of the common standard deviation) as a measure of variability in the outcome. If data are
available on variability of the outcome in each comparison group, then Sp can be computed and used to generate the sample sizes. However, it is more often the case that data on the variability of the outcome are available from only one group, usually the untreated (e.g., placebo control) or unexposed group. When planning a clinical trial to
investigate a new drug or procedure, data are often available from other trials that may have involved a placebo or an active control group (i.e., a standard medication or treatment given for the condition under study). The standard deviation of the outcome variable measured Issues in Estimating Sample Size for Hypothesis Testing in patients
assigned to the placebo, control, or unexposed group can be used to plan a future trial. This situation is illustrated in the following examples. Note that the formula shown previously generates sample size estimates for samples of equal size. If a study is planned where different numbers of patients will be assigned or the comparison groups will
comprise different numbers of patients, then alternative formulas can be used. Interested readers can see Howell for more details.3 Example 8.14. An investigator is planning a clinical trial to evaluate the efficacy of a new drug designed to reduce systolic blood pressure. The plan is to enroll participants and to randomly assign them to receive either
the new drug or a placebo. Systolic blood pressures will be measured in each participant after 12 weeks on the assigned treatment. Based on prior experience with similar trials, the investigator expects that 10% of all participants will be lost to follow-up or will drop from the study. If the new drug shows a 5-unit reduction in mean systolic blood
pressure, this would represent a clinically meaningful reduction. How many patients should be enrolled in the trial to ensure that the power of the test is 80% to detect this difference? A two-sided test will be used with a 5% level of significance. To compute the effect size, an estimate of the variability in systolic blood pressures is needed. Analysis of
data from the Framingham Heart Study (see Table 4.20), showed that the standard deviation of systolic blood pressure is s 5 19.0. This value is used to plan the trial. The effect size is ES = μ1 − μ2 5 = = 0.26. σ 19.0 We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2 ⎛ z 1− α 2 +
z1− β ⎞1.96 + 0.84 ⎞ni = 2⎜ ⎟ = 2⎛⎜ ⎟= 232.0. ⎜⎝ ⎝ ES ⎠0.26 ⎠2 Samples of size n1 5 232 and n2 5 232 will ensure that the test of hypotheses will have 80% power to detect a 5-unit difference in mean systolic blood pressures in patients receiving the new drug as compared to patients receiving the placebo. These sample sizes refer to the numbers of
participants required with complete data. The investigators hypothesized a 10% attrition rate in both groups. To ensure that the total sample size of 464 is available at 12 weeks, the investigator needs to recruit more participants to allow for attrition: N (number to enroll) 3 (% retained) 5 Desired sample size N 3 0.90 5 464 185 464 N 5 }} 5 516 0.90
The investigator must enroll 516 participants. Each patient will be randomly assigned to receive either the new drug or placebo; assuming that 10% are lost to follow-up, 464 will be available for analysis. Example 8.15. An investigator is planning a study to assess the association between alcohol consumption and grade point average among college
seniors. The plan is to categorize students as heavy drinkers or not using five or more drinks on a typical drinking day as the criterion for heavy drinking. Mean grade point averages will be compared between students classified as heavy drinkers versus not using a two independent samples test of means. The standard deviation in grade point
averages is assumed to be 0.42 and a meaningful difference in grade point averages (relative to drinking status) is 0.25 unit. How many college seniors should be enrolled in the study to ensure that the power of the test is 80% to detect a 0.25 unit difference in mean grade point averages? A two-sided test will be used with a 5% level of significance.
We first compute the effect size, ES = μ1 − μ2 0.25 = = 0.60. σ 0.42 We now substitute the effect size and the appropriate z values for the selected a and power to compute the sample size, 2 ⎛ z1− α 2 + z1− β ⎞1.96 + 0.84 ⎞ni = 2⎜ = 2⎛⎜ ⎟= 43.6. ⎜⎝ ⎝ ES ⎠0.60 ⎠2 Samples of size n1 5 44 heavy drinkers and n2 5 44 who drink fewer than five drinks
per typical drinking day will ensure that the test of hypothesis has 80% power to detect a 0.25 unit difference in mean grade point averages. This computation assumes that approximately equal numbers of students will be classified as heavy drinkers versus not. A study by Wechsler et al. showed that approximately 44% of college students were binge
drinkers (defined as drinking five or more drinks per occasion for men and four or more per occasion for women).9 To ensure that this study has at least 44 heavy drinkers, a total of N 5 100 college students should be selected. This will ensure that approximately 44 (44% of 100) are heavy drinkers and 56 (56% of 100) are not. 8.2.4 Sample Size for
Matched Samples, Continuous Outcome In studies where the plan is to perform a test of hypothesis on the mean difference in a continuous outcome variable based on matched data, the hypotheses of interest are 186 CHAPTER 8 Power and Sample Size Determination H0: md 5 0, ES = H1: md 0, where md is the mean difference in the population.
The formula for determining the sample size to ensure that the test has a specified power is 2 ⎛ z 1− α 2 + z1− β ⎞n=⎜ ⎟, ⎝ ES where α is the selected level of significance and z12α/2 is the value from the standard normal distribution holding 12α/2 below it. 12β is the selected power, z12β is the value from the standard normal distribution holding 12β
below it and ES is the effect size, defined as ES = μd , σd where md is the mean difference expected under the alternative hypothesis, H1, and σd is the standard deviation of the difference in the outcome (e.g., the difference based on measurements over time or the difference between matched pairs). Example 8.16. In Example 8.7, we generated
sample size requirements for a crossover trial to compare two diet programs for their effectiveness in promoting weight loss. The planned analysis was a confidence interval estimate for the mean difference in weight loss on the low-fat as compared to the low-carbohydrate diet. Suppose we design the same study with a test of hypothesis as the
primary analysis. The proposed study will have each child follow each diet for 8 weeks. At the end of each 8-week period, the weight lost during that period will be measured. Children will be randomly assigned to the first diet (i.e., some will follow the low-fat diet first while others will follow the low-carbohydrate diet first). The investigators expect to
lose 30% of the participants over the course of the 16-week study. The difference in weight lost on the low-fat diet and the low-carbohydrate diet will be computed for each child; here we will test if there is a statistically significant difference in weight loss between the diets. How many children are required to ensure that a two-sided test with a 5%
level of significance has 80% power to detect a mean difference in weight loss between the two diets of 3 lbs? To compute the sample size, we again need an estimate of the variability of the difference in weight loss between diets. In Example 8.7, we used a value from a previous study of 9.1 lbs (this was based on a similar trial conducted in adults).
We first compute the effect size, μd 3 = = 0.33. σd 9.1 We now substitute the effect size and the appropriate z values for the selected a and power to compute the sample size, 2 2 ⎛ z 1− α 2 + z1− β ⎞1.96 + 0.84 ⎞= 72.0. n=⎜ = ⎛⎜ ⎟⎜ 0.33 ⎟⎝ ES ⎠A sample of size n 5 72 children will ensure that a twosided test with α 5 0.05 has 80% power to detect a
mean difference of 3 lbs between diets using a crossover trial. The investigators hypothesized a 30% attrition rate. To ensure that the total sample size of 72 is available at 16 weeks, the investigator needs to recruit 103 participants: N (number to enroll) 3 (% retained) 5 Desired sample size N 3 0.70 5 72 72 N 5 }} 5 103 0.70 Notice that different
numbers of children are needed depending on the analysis plan. In Example 8.7, we determined that 36 children would be needed to ensure that a 95% confidence interval for the mean difference has a margin of error of no more than 3 lbs. With a test of hypothesis, to ensure 80% power in detecting a mean difference of 3 lbs between diets we need
72 children. The analysis plan is key in the design of the study as it clearly affects the sample size. Example 8.17. An investigator wants to evaluate the efficacy of an acupuncture treatment for reducing pain in patients with chronic migraine headaches. The plan is to enroll patients who suffer from migraine headaches. Each will be asked to rate the
severity of the pain they experience with their next migraine before any treatment is administered. Pain will be recorded on a scale of 1 to 100 with higher scores indicative of more severe pain. Each patient will then undergo the acupuncture treatment. On their next migraine (post-treatment), each patient will again be asked to rate the severity of
the pain. The difference in pain will be computed for each patient. A two-sided test of hypothesis will be conducted at α 5 0.05 to assess whether there is a statistically significant difference in pain scores before and after treatment. How many patients should be involved in the study to ensure that the test has 80% power to detect a difference of 10
units on the pain scale? Assume that the standard deviation in the difference scores is approximately 20 units. Issues in Estimating Sample Size for Hypothesis Testing We first compute the effect size, ES = μd 10 = = 0.50. σ d 20 We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2
2 ⎛ z1−α 2 + z1−β ⎞⎛ 1.96 + 0.84 ⎞= n=⎜ ⎟⎜ 0.50 ⎟= 31.4. ⎝ ES A sample of size n 5 32 patients with migraines will ensure that a two-sided test with α 5 0.05 has 80% power to detect a mean difference of 10 points in pain before and after treatment. (Note that this assumes that all 32 patients complete the treatment and have available pain
measurements before and after treatment.) 8.2.5 Sample Sizes for Two Independent Samples, Dichotomous Outcome In studies where the plan is to perform a test of hypothesis comparing the proportions of successes in two independent populations, the hypotheses of interest are H1: p1 p2, where p1 and p2 are the proportions in the two
comparison populations. The formula for determining the sample sizes to ensure that the test has a specified power is ⎛ z1− α 2 + z1− β ⎞ni = 2⎜ ⎟⎝ ES the efficacy of a new drug designed to reduce systolic blood pressure. The primary outcome was systolic blood pressure measured after 12 weeks on the assigned treatment, and the proposed
analysis was a comparison of mean systolic blood pressures between treatments. We now consider a similar trial with a different primary outcome. Suppose we consider a diagnosis of hypertension (yes/no) as the primary outcome. Hypertension will be defined as a systolic blood pressure above 140 or a diastolic blood pressure above 90. Suppose we
expect that 30% of the participants will meet the criteria for hypertension in the placebo group. The new drug will be considered efficacious if there is a 20% reduction in the proportion of patients receiving the new drug who meet the criteria for hypertension (i.e., if 24% of the patients receiving the new drug meet the criteria for hypertension). How
many patients should be enrolled in the trial to ensure that the power of the test is 80% to detect this difference in the proportions of patients with hypertension? A two-sided test will be used with a 5% level of significance. We first compute the effect size by substituting the proportions of hypertensive patients expected in each group, p1 5 0.24 and
p2 5 0.30, and the overall proportion, p 5 0.27, ES = H0: p1 5 p2, 2 where ni is the sample size required in each group (i 5 1, 2), α is the selected level of significance, and z12α/2 is the value from the standard normal distribution holding 12α/2 below it, 12β is the selected power, and z12β is the value from the standard normal distribution holding 12β
below it. ES is the effect size, defined as p − p2 ES = 1 , p (1 − p) where |p1 2 p2| is the absolute value of the difference in proportions between the two groups expected under the alternative hypothesis, H1, and p is the overall proportion, based on pooling the data from the two comparison groups. Example 8.18. In Example 8.14, we determined the
sample size needed for a clinical trial proposed to evaluate 187 p1 − p2 p (1 − p) 0.24 − 0.30 = 0.27 (1 − 0.27) = 0.135. We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2 ⎛ z1− α 2 + z1− β ⎞1.96 + 0.84 ⎞ni = 2⎜ = 2⎛⎜ ⎟= 860.4. ⎜⎝ ⎝ ES ⎠0.135 ⎠2 Samples of size n1 5 861
patients on the new drug and n2 5 861 patients on placebo will ensure that the test of hypothesis will have 80% power to detect a 20% reduction in the proportion of patients on the new drug who meet the criteria for hypertension (assuming that 30% of the participants on placebo meet the criteria for hypertension). In the trial proposed in Example
8.14, the investigators hypothesized a 10% attrition rate in each group. To ensure that the total sample size of 1722 is available at 12 weeks, the investigator needs to recruit more participants to allow for attrition: N (number to enroll) 3 (% retained) 5 Desired sample size N 3 0.90 5 1722 1722 N 5 }} 5 1914 0.90 188 CHAPTER 8 Power and Sample
Size Determination The investigator must enroll 1914 participants. Each participant will be randomly assigned to receive either the new medication or a placebo. Assuming that 10% are lost to follow-up, 1722 will be available for analysis. Example 8.19. An investigator hypothesizes that a higher proportion of students who use their athletic facility
regularly develop flu as compared to their counterparts who do not. The study will be conducted in the spring semester, and each student will be asked if they used the athletic facility regularly over the past 6 months and they will be followed for 3 months to assess whether they develop the flu. A test of hypothesis will be conducted to compare the
students who used the athletic facility regularly and those who did not in terms of developing the flu. During a typical year, approximately 35% of the students experience flu. The investigators feel that a 30% increase in flu among those who used the athletic facility regularly would be clinically meaningful. How many students should be enrolled in
the study to ensure that the power of the test is 80% to detect this difference in the proportions? A two-sided test will be used with a 5% level of significance. We first compute the effect size by substituting the proportions of students in each group who are expected to develop flu, p1 5 0.46 and p2 5 0.35, and the overall proportion, p 5 0.41, ES = p1
− p2 p (1 − p) = 0.46 − 0.35 0.41(1 − 0.41) = 0.22. We now substitute the effect size and the appropriate z values for the selected α and power to compute the sample size, 2 2 ⎛ z1− α 2 + z1− β ⎞⎛ 1.96 + 0.84 ⎞= 324.0. ni = 2⎜ = 2 ⎜⎜ ⎟⎟⎜ 0.22 ⎟⎠⎝ ES Samples of sizes n1 5 324 and n2 5 324 ensure that the test of hypothesis will have 80% power to
detect a 30% increase in the proportion of students who use the athletic facility regularly and develop flu, assuming that 35% of the students who do not will develop flu. 8.3 SUMMARY Determining the appropriate design of a study is more important than the statistical analysis. A poorly designed study can never be salvaged, whereas a poorly
analyzed study can always be re-analyzed. A critical component in study design is the determination of the appropriate sample size. The sample size must be large enough to adequately answer the research question, yet not be so large as to involve too many patients when fewer would have sufficed. The determination of the appropriate sample size
involves statistical criteria as well as clinical or practical considerations. Sample size determination involves teamwork; biostatisticians must work closely with clinical investigators to determine the sample size to address the research question of interest with adequate precision or power to produce results that are clinically meaningful. Table 8–1
summarizes the sample size formulas for each scenario described here. The formulas are organized by the proposed analysis, a confidence interval estimate, or a test of hypothesis. 8.4 PRACTICE PROBLEMS 1. Suppose we want to design a new placebo-controlled trial to evaluate an experimental medication to increase lung capacity. The primary
outcome is peak expiratory flow rate, a continuous variable measured in liters per minute. The primary outcome will be mea sured after 6 months of treatment. The mean peak expiratory flow rate in adults is 300 l/min with a standard deviation of 50 l/min. How many subjects should be enrolled to ensure 80% power to detect a difference of 15 l/min
with a two-sided test and α 5 0.05? The investigators expect to lose 10% of the participants over the course of follow-up. 2. An investigator wants to estimate caffeine consumption in high school students. How many students would be required to ensure that a 95% confidence interval estimate for the mean caffeine intake (measured in mg) is within 15
mg of the true mean? Assume that the standard deviation in caffeine intake is 68 mg. 3. Consider the study proposed in Problem 2. How many students would be required to estimate the proportion of students who consume coffee? Suppose we want the estimate to be within 5 percentage points of the true proportion with 95% confidence. 4. A clinical
trial was conducted comparing a new compound designed to improve wound healing in trauma patients to a placebo. After treatment for five days, 58% of the patients taking the new compound had a substantial reduction in the size of their wound as compared to 44% in the placebo group. The trial failed to show significance. How many subjects
would be required to detect the difference in proportions observed in the trial with 80% power? A two-sided test is planned at α 5 0.05. 5. A crossover trial is planned to evaluate the impact of an educational intervention program to reduce alcohol consumption in patients determined to be at risk for alcohol problems. The plan is to measure alcohol
consumption (the number of drinks on a typical drinking day) before the intervention 189 Practice Problems TABLE 8–1 Sample Size Formulas Outcome Variable, Number of Groups: Planned Analysis Sample Size to Estimate Confidence Interval (CI) Continuous outcome, one sample: CI for m, H0: m 5 m0 Continuous outcome, two matched
samples: CI for md, H0: md 5 0 ⎛ Z1−α 2 + Z1− β ⎞n=⎜ ⎟ES ⎝ 2 Zσ ⎞n = ⎛⎜ ⎝ E ⎟Continuous outcome, two independent samples: CI for (m12m2), H0: m1 5 m2 Sample Size for Test of Hypothesis ES = ES = ⎛ Zσ d ⎞n=⎜ ⎝ E ⎟μ1 − μ0 σ ⎛ Z1−α 2 + Z1− β ⎞ni = 2⎜ ⎟⎝ ES 2 Zσ ⎞ni = 2 ⎛⎜ ⎝ E ⎟⎛ Z1−α 2 + Z1− β ⎞n=⎜ ⎟ES ⎝ 2 ES = Z⎞ni = ⎡⎣ p1 (1 −
p1) + p2 (1 − p2) ⎤⎦⎛⎜ ⎝ E⎠Dichotomous outcome, two independent samples: CI for (p12p2), H0: p1 5 p2 and then again after participants complete the educational intervention program. How many participants would be required to ensure that a 95% confidence interval for the mean difference in the number of drinks is within 2 drinks of the
true mean difference? Assume that the standard deviation of the difference in the mean number of drinks is 6.7 drinks and that 20% of the participants will drop out over the course of follow-up. 6. An investigator wants to design a study to estimate the difference in the proportions of men and women who develop early onset cardiovascular disease
(defined as cardiovascular disease before age 50). 2 2 μd σd ⎛ Z1−α 2 + Z1− β ⎞n=⎜ ⎟ES ⎝ 2 Z⎞n = p (1 − p) ⎛⎜ ⎝ E⎠2 μ1 − μ2 σ ES = Dichotomous outcome, one sample: CI for p, H0: p 5 p0 2 2 p1 − p0 p0 (1 − p0) ⎛ Z1−α 2 + Z1− β ⎞ni = 2⎜ ⎟ES ES = 2 p1 − p2 p (1 − p) A study conducted 10 years ago found that 15% and 8% of men and
women, respectively, developed early onset cardiovascular disease. How many men and women are needed to generate a 95% confidence interval estimate for the difference in proportions with a margin of error not exceeding 4%? 7. The mean body mass index (BMI) for boys of age 12 years is 23.6 kg/m2. An investigator wants to test if the BMI is
higher in 12-year-old boys living in New York City. How many boys are needed to ensure that a two-sided test of hypothesis has 80% power to detect a difference in BMI of 2 kg/m2? Assume that the standard deviation in BMI is 5.7 kg/m2. 190 CHAPTER 8 Power and Sample Size Determination 8. An investigator wants to design a study to estimate
the difference in the mean BMI between 12-year-old boys and girls living in New York City. How many boys and girls are needed to ensure that a 95% confidence interval estimate for the difference in mean BMI between boys and girls has a margin of error not exceeding 2 kg/m2? Use the estimate of the variability in BMI from Problem 7. 9. How
many stomach cancer patients and controls would be required to estimate the difference in proportions of patients with sufficient concentrations of plasma antioxidant vitamins with a margin of error not exceeding 5% using a 95% CI? The data in Table 8–2 were reported recently from a similar study investigating the sufficiency of plasma antioxidant
vitamins in throat cancer patients. 10. Recently it has been observed that HIV-infected patients develop peripheral lypoatrophy while on potent antiretroviral therapy. A clinical trial is planned to determine whether a new chemical will improve this condition. In the trial, participants will be randomized to receive the new chemical or a placebo, and
changes from baseline in subcutaneous adipose cross-sectional area as measured by CT scan will be calculated after 24 weeks of treatment. Investigators hope to show that the increases in patients receiving the chemical will be greater than the increases in patients receiving the placebo. They hypothesize that the participants assigned to the
chemical arm will exhibit a mean change of 30% and that the participants assigned to the placebo arm will exhibit a mean change of 0%. Prior literature suggests that the standard deviation of the changes will be 57% in both arms. How many participants are needed to ensure 80% power? Assume that α = 0.05, that there are equal numbers in each
group, and that 20% TABLE 8–2 Plasma Antioxidant Vitamins in Throat Cancer Patients N % with Sufficient Concentrations of Plasma Antioxidant Vitamins Throat cancer patients 25 47% Controls 25 56% of the participants will drop out over the course of follow-up. 11. We wish to design a study to assess risk factors for stroke and dementia. There
are two primary outcome measures. The first is a measure of neurologic function. Patients with “normal” function typically score 70 with an SD of 15. The second outcome is incident stroke. In patients over the age of 65, approximately 14% develop stroke over 25 years. How many participants over the age of 65 must be enrolled to ensure the
following? a. That the margin of error in a 95% CI for the difference in mean neurologic function scores between men and women does not exceed 2 mg/dL. b. That we have 80% power to detect a 20% reduction in incidence of stroke between participants who participate in a healthy-aging program compared to those who do not. Assume that the
incidence of stroke is 14% in those who do not participate. c. How many men and women over the age of 65 should be enrolled to satisfy parts (a) and (b) assuming that 15% will be lost to follow-up over the course of the 25-year follow-up? 12. Suppose we want to design a new study to assess the implications of childhood obesity on poor health
outcomes in adulthood. In the study we will consider total cholesterol at age 50 and incident cardiovascular disease by age 50 as the key outcomes. We will compare normal versus overweight/obese children. Assume that the standard deviation in total cholesterol is 35. a. How many normal and overweight/obese children are required to ensure that a
95% CI estimate for the difference in total cholesterol has a margin of error not exceeding 5 mg/dL? b. How many normal and overweight/obese children are required to ensure 80% power to detect a 4% difference in incident cardiovascular disease between groups? Assume that approximately 3% of normal-weight children will develop cardiovascular
disease by age 50 and that 7% of overweight/obese children will develop cardiovascular disease by age 50 and that a two-sided test with α = 0.05 will be used. c. How many normal and overweight/obese children should be enrolled to satisfy parts (a) and (b) assuming that 20% will be lost to follow-up over the course of the study? References
REFERENCES 1. Buschman, N.A., Foster, G., and Vickers, P. “Adolescent girls and their babies: Achieving optimal birth weight. Gestational weight gain and pregnancy outcome in terms of gestation at delivery and infant birth weight: A comparison between adolescents under 16 and adult women.” Child: Care, Health and Development 2001; 27(2):
163–171. 2. Feuer, E.J. and Wun, L.M. DEVCAN: Probability of Developing or Dying of Cancer, Version 4.0. Bethesda, MD: National Cancer Institute, 1999. 3. Howell, D.C. Statistical Methods for Psychology. Boston: Duxbury Press, 1982. 4. Fleiss, J.L. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons, 1981. 5. National
Center for Health Statistics. Health, United States, 2005 with Chartbook on Trends in the Health of Americans. Hyattsville, MD: US Government Printing Office, 2005. 191 6. Plaskon, L.A., Penson, D.F., Vaughan, T.L., and Stanford, J.L. “Cigarette smoking and risk of prostate cancer in middle-aged men.” Cancer Epidemiology Biomarkers &
Prevention 2003; 12: 604–609. 7. Rutter, M.K., Meigs, J.B., Sullivan, L.M., D’Agostino, R.B., and Wilson, P.W. “C-reactive protein, the metabolic syndrome, and prediction of cardiovascular events in the Framingham Offspring Study.” Circulation 2004; 110: 380–385. 8. Vasan, R.S., Sullivan, L.M., Wilson, P.W., Sempos, C.T., Sundstrom, J., Kannel, W.B.,
Levy, D., and D’Agostino, R.B. “Relative importance of borderline and elevated levels of coronary heart disease risk factors.” Annals of Internal Medicine 2005; 142: 393–402. 9. Wechsler, H., Lee, J.E., Kuo, M., and Lee, H. “College binge drinking in the 1990s: A continuing problem. Results of the Harvard School of Public Health, 1999 College Alcohol
Study.” Journal of American College Health 2000; 48: 199–210. 9 chapter Multivariable Methods W hen and W hy Key Questions • W hat does it mean when investigators adjust for other variables? • How can you be sure that an association is real? • How can you be sure that you accounted for all
other possible explanations of associations? In the News There is much discussion about childhood and adult obesity in the United States, and increasingly around the globe. In the United States, approximately 35% of adults and 17% of children are obese, with these rates being higher among blacks (48% of adults, 20% of children) and Latinos (43%
of adults, 22% of children) as compared to whites (33% of adults, 14% of children).1 The prevalence of overweight and obesity in the United States varies widely by state, from a low of 22.1% in Utah to a high of 39.8% in Louisiana.2 The World Health Organization (WHO) reports an increase globally in the number of overweight and obese young
children (ages 0 to 5 years) from 32 million in 1990 to 42 million in 2013.3 WHO has suggested a number of preventive measures for these conditions, including increased rates of breastfeeding, consumption of healthier foods, and regular physical activity. There are a number of health risks associated with obesity, such as hypertension, diabetes, and
mortality.4 Mortality associated with overweight and obesity is reportedly three times higher among white men compared to black men, and approximately 20% higher among black women compared to white women.5 Available at . Available at _obesity/in_your_state/. 3 Available at . 4 Available at 5 Masters, R.K., Reither, E.N., Powers, D.A., Yang,
Y.C., Burger, A.E., and Link, B.G. “The impact of obesity on US mortality levels: the importance of age and cohort factors in population estimates.” American Journal of Public Health 2013; 103: 1895–1901. 1 2 Dig In • Are some groups more likely to be obese? Why? • What might explain some of the differences in health risks attributable to obesity
among racial/ethnic groups? • What factors might mask or enhance the associations between obesity and hypertension, diabetes, or mortality? L earning O bjectives By the end of this chapter, the reader will be able to • D efine and provide examples of dependent and independent variables in a study of a public health problem • Explain the principle
of statistical adjustment to a lay audience • Organize data for regression analysis • Define and provide an example of confounding • Define and provide an example of effect modification • Interpret coefficients in multiple linear and multiple logistic regression analysis In Chapter 6 and Chapter 7, we presented statistical inference procedures for
estimation and hypothesis testing. We discussed many scenarios that differed in terms of the nature of the outcome variable (continuous, dichotomous, categorical, or ordinal), the number of comparison groups (one, two, or more than two), and whether the groups were independent (physically separate) or dependent (matched). In each scenario, we
considered one outcome variable, and we investigated whether there were differences in that outcome variable among the comparison groups or whether there was an association between an exposure or risk factor and an outcome. What we have not considered up to this point is the possibility that observed differences or 194 CHAPTER 9
Multivariable Methods associations might be due to other characteristics or variables. Multivariable statistical methods are used to assess the interrelationships among several risk factors or exposure variables and a single outcome. The topic of multivariable analysis is extensive. We describe only general principles and concepts here. Interested
readers should see Kleinbaum, Kupper, and Muller; Jewell; and Hosmer and Lemeshow for more in-depth discussion.1–3 We specifically discuss the use of multivariable modeling techniques to address confounding and effect modification. In Chapter 2, we defined confounding as a distortion of the effect of an exposure or risk factor on an outcome by
another characteristic or variable. Effect modification occurs when there is a different relationship between the exposure or risk factor and the outcome depending on the level of another characteristic or variable. In both situations, the third variable can exaggerate or mask the association between the risk factor and the outcome. Effect modification
is also called statistical interaction. Analytically, confounding and effect modification are handled differently. When there is confounding, multivariable methods can be used to generate an estimate of the association between an exposure or risk factor and an outcome adjusting for or taking into account the impact of the confounder. In contrast, with
effect modification there is a different relationship between the risk factor and outcome depending on the level of a third variable, and therefore an overall estimate of association does not accurately convey the information in the data. In the presence of effect modification, results are generally presented separately for each level of the third variable
(using the techniques presented in Chapter 6 and Chapter 7). To formalize these concepts, we first present some examples and then move into the methods. Suppose we want to assess the association between smoking and cardiovascular disease in a cohort study. We may find that smokers in the cohort are much more likely to develop cardiovascular
disease. Suppose that we estimate ˆ 5 2.6 with a 95% confidence interval a relative risk of RR of (1.5, 4.1). In our study, smokers have 2.6 times the risk of developing cardiovascular disease as compared to nonsmokers. However, it may also be the case that the smokers are less likely to exercise and have higher cholesterol levels. Without examining
these other characteristics, we could incorrectly infer that there is a strong association between smoking and cardiovascular disease when the relationship may be due, in part, to lack of exercise or to high cholesterol. Multivariable methods can be used to address these complex relationships and to tease out what proportion of the asso- ciation might
be due to smoking, as opposed to other risk factors. In other words, we use these methods to adjust the magnitude of association for the impact of other variables (e.g., the association between smoking and cardiovascular disease adjusted for exercise and cholesterol). Methods are also discussed that allow us to assess whether an association remains
statistically significant after adjusting for the impact of other variables. Multivariable methods can also be used to assess effect modification—specifically, the situation in which the relationship between a risk factor and the outcome of interest varies by a third variable. For example, suppose we are interested in the efficacy of a new drug designed to
lower total cholesterol. A clinical trial is conducted and the drug is shown to be effective, with a statistically significant reduction in total cholesterol in patients who receive the new drug as compared to patients who receive a placebo. Suppose that the investigators look more closely at the data and find that the reduction (or effect of the drug) is only
present in participants with a specific genetic marker and that there is no reduction in persons who do not possess the marker. This is an example of effect modification, or statistical interaction. The effect of the treatment is different depending on the presence or absence of the genetic marker. Multivariable methods can be used to identify effect
modification. When effect modification is present, the strategy is to present separate results according to the third variable (i.e., to report the effect of the treatment separately in persons with and without the marker). Other uses of multivariable methods include the consideration of several risk factors simultaneously, where the goal is to assess the
relative importance of each with regard to a single outcome variable. For example, the Framingham Heart Study has a long history of developing multivariable risk functions.4,5 These risk functions are used to predict the likelihood that a person will develop cardiovascular disease over a fixed period (e.g., the next 10 years) as a function of his or her
risk factors. The risk factors include age, sex, systolic blood pressure, total and HDL cholesterol levels, current smoking, and diabetes status. Multivariable techniques are used in this setting to account for all of the risk factors simultaneously. This application is similar to the first, except that rather than focusing on the association between one risk
factor and the o utcome (adjusting for others), we are interested in the relationships between each of the risk factors and the outcome. There are many other uses of multivariable methods. We focus here on only a few specific applications. Because Confounding and Effect Modification the computations are intensive, our presentation is more general
with particular emphasis on interpretation rather than computation. In practice, the computations are performed using a statistical computing package.6 It is important to note that the statistical analysis of any study should begin with a complete description of the study data using the methods described in Chapter 4. The primary analyses should
proceed as planned to either generate estimates of unknown parameters or to perform tests of hypothesis using the techniques described in Chapter 6 and Chapter 7, respectively. The analyses we presented in Chapter 6 and Chapter 7 are called unadjusted or crude analyses as they focused exclusively on the associations between one risk factor or
exposure and the outcome. Multivariable methods are used after the study data are described and after unadjusted analyses are performed. In the clinical trials setting, the unadjusted analyses are generally the final analyses due primarily to the randomization component that (in theory) eliminates the possibility of confounding. While the
multivariable methods we describe are a means to account for confounding, they should not be relied on to “correct” problems in a study. Multivariable models are used for the statistical adjustment of confounding, whereas careful study design can provide much more in the way of minimizing confounding. Randomization is a very effective means of
minimizing confounding. Unfortunately, not all studies are suitable for randomization. There are other options to minimize confounding (e.g., matching), but in all cases it is important to recognize that multivariable models can only adjust or account for differences in confounding variables that are measured in the study. Therefore, investigators must
carefully plan studies not only with an eye toward minimizing confounding whenever possible, but also with forethought to measure variables that might be potential confounders. more details, see Aschengrau and Seage.7) Confounding can be assessed in a number of different ways. Some investigators use formal statistical tests and others use
clinical or practical judgment. We discuss both approaches. Effect modification, also called statistical interaction, occurs when there is a different association between the risk factor and outcome depending on a third variable. Effect modification is generally assessed by statistical significance using regression techniques (discussed in Section 9.4 and
Section 9.5). We illustrate confounding and effect modification in Example 9.1 and Example 9.2, respectively. Example 9.1. Consider a study designed to assess the association between obesity—defined as a body mass index (BMI) of 30 or more—and the incidence of cardiovascular disease. Data are collected in a cohort study in which participants
between the ages of 35 and 65 who are free of cardiovascular disease (CVD) are enrolled and followed over 10 years. Each participant’s BMI is assessed at baseline and participants are followed for 10 years for incident cardiovascular disease. A summary of the data is contained in Table 9–1. The proportion of obese persons at baseline who develop
CVD (or the incidence of CVD among obese persons) is 46 / 300 5 0.153, and the proportion of non-obese persons who d evelop CVD (or the incidence of CVD among nonobese persons) is 60 / 700 5 0.086. The 10-year incidence of CVD is shown graphically in Figure 9–1. The estimate of the relative risk of CVD in obese as compared to non-obese
persons is pˆ 0.153 RˆR 5 1 = = 1.78. pˆ2 0.086 Obese persons have 1.78 times the risk of developing CVD as compared to non-obese persons in this study. Many studies have shown an increase in the risk of CVD with advancing age. Could age be a confounder in this analysis? 9.1 CONFOUNDING AND EFFECT MODIFICATION Confounding is
present when the relationship between a risk factor and an outcome is modified by a third variable, called the confounding variable. A confounding variable or confounder is one that is related to the exposure or risk factor of interest and also to the outcome. Identification of confounding variables requires a strong working knowledge of the
substantive area under investigation. (An additional consideration in determining whether the third variable is a confounder is assessing whether or not the third variable is part of a causal chain. A confounder is not in the causal chain from exposure to outcome. This is a complex issue; for 195 TABLE 9–1 Baseline Obesity and Incident CVD Over 10-
Year Follow-Up Incident CVD No CVD Total Obese 46 254 300 Not obese 60 640 700 Total 106 894 1000 196 CHAPTER 9 Multivariable Methods FIGURE 9–1 Ten-Year Incidence of CVD in Obese and Non-Obese Persons 0.2 10-Year Incidence of CVD 0.15 0.1 0.05 0 Obese Could any of the observed association between obesity and incident CVD
be attributable to age? A confounder is a variable that is associated with the risk factor and with the outcome. Is age associated with obesity? Is age associated with incident CVD? In this study, age is measured as a continuous variable (in years). For interpretation purposes, suppose we dichotomize age at 50 years (less than 50 years of age versus 50
years and older). In this study, n 5 600 (60%) participants are less than 50 years of age and n 5 400 (40%) are 50 years of age and older. Table 9–2 shows the relationship between age and obesity. Age is related to obesity. A higher proportion of persons 50 years of age and older are obese (200 / 400 5 0.500) as compared to persons less than 50 years
of age (100 / 600 5 0.167). Looking at the data another way, a much higher proportion of obese persons are 50 years of age and older (200 / 300 5 0.667) as compared to non-obese persons (200 / 700 5 0.286). Could the fact that the majority of the obese persons are older explain the association between obesity and incident CVD? Is age driving the
association? Not Obese TABLE 9–2 Baseline Age and Baseline Obesity Age Obese Not Obese Total , 50 100 500 600 501 200 200 400 Total 300 700 1000 Table 9–3 shows the relationship between age and incident CVD. Age is related to incident CVD. A much higher proportion of persons 50 years of age and older develop CVD (61 / 400 5 0.153) as
compared to persons less than 50 years of age (45 / 600 5 0.075). A higher proportion of the obese persons are older, and age is related to incident CVD. Thus, age meets the definition of a confounder (i.e., it is associated with the risk factor, obesity, and the outcome, incident CVD). We now examine the relationship between obesity and incident CVD

in persons less Confounding and Effect Modification TABLE 9–3 Baseline Age and Incident CVD Over 10-Year Follow-Up Age Incident CVD No CVD Total , 50 45 555 600 501 61 339 400 Total 106 894 1000 TABLE 9–4 Baseline Obesity and Incident CVD by Age Group Age , 50 Incident CVD No CVD Total Obese 10 90 100 Not obese 35 465 500 Total 45 555 600 Age 501 Incident CVD No CVD Total Obese 36 164 200 Not obese 25 175 200 Total 61 339 400 197 The relative risk based on the combined sample suggests a much stronger association between obesity and incident CVD (R^R 5 1.78). Figure 9–2 shows the 10-year incidence of CVD for obese and non-obese persons in each age group. There are different methods to determine whether a variable is a confounder or not. Some investigators perform formal tests of hypothesis to assess whether the variable is associated with the risk factor and with the outcome; other investigators do not conduct statistical tests but instead inspect the data, and if there is a practically important or clinically meaningful relationship between the variable and the risk factor and between the variable and the outcome (regardless of whether that relationship reaches statistical significance), then the variable is said to be a confounder. The data in Table 9–2 through Table 9–5 would suggest that age is a confounder by the latter method. (In Section 9.3, we provide yet another method to assess confounding that again is based on evaluation of associations from a practical or clinical point of view.) We now illustrate the approach to assessing confounding using tests of hypothesis. We specifically test whether there is a statistically significant association between age and obesity and between age and incident CVD. The tests of hypothesis can be performed using the χ2 test of independence. In the first test, we determine whether age and obesity are related. The data for the test are shown in Table 9–2. Step 1. Set up hypotheses and determine the level of significance. than 50 years of age and in persons 50 years of age and older separately (see Table 9–4). Among people less than 50 years of age, the proportion of obese persons who develop CVD is 10 / 100 5 0.100, and the proportion of non-obese persons who develop CVD is 35 / 500 5 0.070. The relative risk of CVD in persons less than 50 years of age who are obese as compared to nonobese is pˆ1 0.100 = 1.43. R^R 5 ˆ = P2 0.070 Among people 50 years of age and older, the proportion of obese persons who develop CVD is 36 / 200 5 0.180, and the proportion of non-obese persons who develop CVD is 25 / 200 5 0.125. The relative risk of CVD in persons 50 years of age and older who are obese as compared to non-obese is pˆ1 0.180 = 1.44. R^R 5 ˆ = P2 0.125 H0: Age and obesity are independent H1: H0 is false α 5 0.05 Step 2. Select the appropriate test statistic. The formula for the test statistic is in Table 7–42 and is given as χ2 = ∑ (O − E)2 E The condition for appropriate use of the preceding test statistic is that each expected frequency in each cell of the table is at least 5. In Step 4, we compute the expected frequencies and ensure that the condition is met. Step 3. Set up the decision rule. The row variable is age (r 5 2) and the column variable is obesity (c 5 2), so for this test, df 5 (2 2 1)(2 2 1) 5 1(1) 5 1. For df 5 1 and a 5% level of significance, the appropriate 198 CHAPTER 9 Multivariable Methods FIGURE 9–2 Ten-Year Incidence of CVD in Obese and Non-Obese Persons by Age Group 10-Year Incidence of CVD 0.2 0.15 0.1 0.05 0 Obese 25. We do not have statistically significant evidence at a = 0.05 to show that the median difference in systolic blood pressures is not zero (i.e., that there is a significant difference in systolic blood pressures after the exercise program as compared to before the exercise program). 10.4 TESTS WITH MORE THAN TWO INDEPENDENT SAMPLES 9 7 10 –4 8 12.5 –7 8 12.5 1 9 14 –2 20 15 TABLE 10–18 Attaching Signs to the Ranks Observed Differences Ordered Absolute Values of Differences Ranks Signed Ranks 7 1 1 1 –2 –2 2.5 –2.5 8 –2 2.5 –2.5 –4 –3 4 –4 20 –4 6 –6 –3 –4 6 –6 6 4 6 6 7 6 8 8 8 –7 10 –10 4 7 10 10 9 7 10 10 –4 8 12.5 12.5 –7 8 12.5 12.5 1 9 14 14 –2 20 15 In Chapter 7, Section 7.8, we presented techniques for testing the equality of means in more than two independent samples using analysis of variance (ANOVA). An underlying assumption for appropriate use of ANOVA was that the continuous outcome was approximately normally distributed or that the samples were sufficiently large (usually nj > 30, where j = 1, 2, ..., k and k denotes the number of independent comparison groups). An additional assumption for appropriate use of ANOVA is equality of variances in the k comparison groups. ANOVA is generally robust when the sample sizes are small but equal. When the outcome is not normally distributed and the samples are small, a nonparametric test is appropriate. A popular nonparametric test to compare outcomes among more than two independent groups is the Kruskal– Wallis test. The Kruskal–Wallis test is used to compare medians among k comparison groups (k > 2) and is sometimes described as an ANOVA with the data replaced by their ranks. The null and research hypotheses for the Kruskal–Wallis nonparametric test are stated as follows: H0: The k population medians are equal. H1: The k population medians are not all equal. The procedure for the test involves pooling the observations from the k samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to N, where N = n1 + n2 + ... + nk. To illustrate the procedure, we consider the following example, addressing each of the five steps in the test of hypothesis including more details to promote understanding. (In the subsequent example, the five-step approach is followed more directly.) Example 10.8. A clinical study is designed to assess differences in albumin levels in adults following different low-protein diets. Albumin is measured in grams per deciliter 242 CHAPTER 10 Nonparametric Tests In this example, the outcome is continuous, but the sample sizes are small and not equal across comparison groups (n 1 = 3, n 2 = 5, n 3 = 4). Thus, a nonparametric test is appropriate. The hypotheses to be tested are given below and we run the test at the 5% level of significance (i.e., a = 0.05). TABLE 10–19 Albumin Levels in Three Different Diets 5% Protein 10% Protein 15% Protein 3.1 3.8 4.0 2.6 4.1 5.5 2.9 2.9 5.0 3.4 4.8 H0: The three population medians are equal. H1: The three population medians are not all equal. 4.2 To conduct the test we assign ranks using the procedures outlined in Section 10.1. The first step in assigning ranks is to order the data from smallest to largest. This is done on the combined or total sample, pooling the data from the three comparison groups, and assigning ranks from 1 to 12, as shown in Table 10–20. We also need to keep track of the group assignments in the total sample (n = 12). Notice that the lower ranks (e.g., 1, 2.5, and 4) are assigned to the 5% protein diet group, while the higher ranks (e.g., 10, 11, and 12) are assigned to the 15% protein diet group. Again, the goal of the test is to determine whether the observed data support a difference in the three population medians. Recall that in the parametric tests, discussed in Chapter 7, when comparing means among more than two groups, we analyze the differences among the sample means (mean square between treatments) relative to their within-group variability and (g/dL) of blood, and low-protein diets are often prescribed for patients with kidney disease. Clinically, albumin is also used to assess whether patients get sufficient protein in their diets. Three diets are compared, ranging from 5% to 15% protein, and the 15% protein diet represents a typical American diet. The albumin levels of participants following each diet are shown in Table 10–19. The question of interest is whether there is a difference in albumin levels among the three different diets. For reference, normal albumin levels are generally between 3.4 g/dL and 5.4 g/dL. By inspection, it appears that participants following the 15% protein diet have higher albumin levels than those following the 5% protein diet. The issue is whether this observed difference is statistically significant. TABLE 10–20 Assigning Ranks Total Sample (Ordered Smallest to Largest) 5% Protein 10% Protein 15% Protein 5% Protein 3.1 3.8 4.0 2.6 2.6 4.1 5.5 2.9 2.9 2.9 5.0 3.1 3.4 4.8 4.2 10% Protein 15% Protein Ranks 5% Protein 10% Protein 15% Protein 1 2.9 2.5 2.5 4 3.4 5 3.8 6 4.0 7 4.1 8 4.2 9 4.8 10 5.0 11 5.5 12 Tests with More Than Two Independent Samples s ummarize the sample information in a test statistic (F statistic). In the Kruskal–Wallis test we again summarize the sample information in a test statistic, one based on the ranks. The test statistic for the Kruskal–Wallis test is denoted H and is defined as follows: k R2 ⎛ 12 j H= ⎜ ∑ − 3(N + 1) , N N + 1 n ( ) ⎝ j =1 j ⎞ where k = the number of comparison groups, N = the total sample size, nj is the sample size in the jth group, and Rj is the sum of the ranks in the jth group. In this example R1 = 7.5, R2 = 30.5, and R3 = 40. Recall that the sum of the ranks will always equal n(n + 1)/2. As a check on our assignment of ranks, we have n(n + 1)/2 = 12(13)/2 = 78, which is equal to 7.5 + 30.5 + 40 = 78. In this example, k R2 ⎛ 12 j H= ⎜ ∑ − 3(N + 1) = N N + 1 n ( ) ⎝ j =1 j ⎞ 12 ⎛ 7.52 30.52 402 ⎞ + + − 3(13) = 7.52. 12(13) ⎜⎝ 3 5 4 ⎟⎠ In every test, we must determine whether the observed test statistic H supports the null or the research hypothesis. This is done following the same approach used in parametric testing. Specifically, we determine a critical value of H such that if the observed value of H is greater than or equal to the critical value, we reject H0 in favor of H1, and if the observed value of H is less than the critical value, we do not reject H0. The critical value of H can be found in Table 8 in the Appendix. To determine the appropriate critical value, we need the sample sizes (n1 = 3, n2 = 5, and n3 = 4) and our level of significance (a = 0.05). For this example the critical value is 5.656; thus we reject H0 because 7.52 > 5.656. We have statistically significant evidence at a = 0.05 to show that there is a difference in median albumin levels among the three different diets. Notice that Table 8 in the Appendix contains critical values for the Kruskal–Wallis test for tests comparing three, four, or five groups with small sample sizes. If there are three or more comparison groups and five or more observations in each of the comparison groups, it can be shown that the test statistic H approximates a χ2 distribution with df = k – 1.4 Thus, in a Kruskal–Wallis test with three or more comparison groups and five or more observations in each group, the critical value for the test can be found in Table 3 in the Appendix: 243 Critical Values of the χ2 Distribution. The following example illustrates this situation. Example 10.9. A personal trainer is interested in comparing the anaerobic thresholds of elite athletes. The anaerobic threshold is defined as the point at which the muscles cannot get more oxygen to sustain activity, or the upper limit of aerobic exercise. It is a measure that is also related to maximum heart rate. The data in Table 10–21 are anaerobic thresholds for distance runners, distance cyclists, distance swimmers, and cross-country skiers. The question of interest is whether there is a difference in anaerobic thresholds among the different groups of elite athletes. Step 1. Set up hypotheses and determine the level of significance. H0: The four population medians are equal. H1: The four population medians are not all equal. a = 0.05. Step 2. Select the appropriate test statistic. The test statistic for the Kruskal–Wallis test is denoted H and is defined as follows: k R2 ⎛ 12 j H= ⎜ ∑ − 3(N + 1) , 1 N N + n ( ) ⎝ j =1 j ⎞ where k = the number of comparison groups, N = the total sample size, nj is the sample size in the jth group, and Rj is the sum of the ranks in the jth group. Step 3. Set up the decision rule. Because there are four comparison groups and five observations in each of the comparison groups, we find the critical value in Table 3 in the Appendix for TABLE 10–21 Anaerobic Thresholds Distance Runners Distance Cyclists Distance Swimmers CrossCountry Skiers 185 190 166 201 179 209 159 195 192 182 170 180 165 178 183 187 174 181 160 215 244 CHAPTER 10 Nonparametric Tests df = k – 1 = 4 – 1 = 3 and a = 0.05. The critical value is 7.81, and the decision rule is to reject H0 if H $ 7.81. In this example, k R2 ⎛ 12 j H= ⎜ ∑ − 3(N + 1) = ⎜⎝ N (N + 1) j =1 j ⎞ Step 4. Compute the test statistic. To conduct the test we assign ranks using the procedures outlined in Section 10.1. The first step in assigning ranks is to order the data from smallest to largest. This is done on the combined or total sample, pooling the data from the four comparison groups, and assigning ranks from 1 to 20. We also need to keep track of the group assignments in the total sample (n = 20). Table 10–22 shows the ordered data. We next assign the ranks to the ordered values and sum the ranks in each group (Table 10–23). Recall that the sum of the ranks will always equal n(n + 1)/2. As a check on our assignment of ranks, we have n(n + 1)/2 = 20(21)/2 = 210, which is equal to 46 + 62 + 24 + 78 = 210. 12 ⎛ 462 622 24 2 782 ⎞ + + − 3(21) = 9.11. 20(21) ⎜⎝ 5 5 5 5 ⎟⎠ Step 5. Conclusion. Reject H0 because 9.11 > 7.81. We have statistically significant evidence at a = 0.05 to show that there is a difference in median anaerobic thresholds among the four different groups of elite athletes. Notice that in this example the anaerobic thresholds of the distance runners, cyclists, and cross-country skiers are comparable (looking only at the raw data). The distance TABLE 10–22 Ordering the Anaerobic Thresholds Total Sample (Ordered Smallest to Largest) Distance Runners Distance Cyclists Distance Swimmers Cross-Country Skiers Distance Runners Distance Cyclists Distance Swimmers 185 190 166 201 159 179 209 159 195 160 192 182 170 180 165 178 183 187 166 174 181 160 215 170 Cross-Country Skiers 165 174 178 179 180 181 182 183 185 187 190 192 195 201 209 215 Summary 245 TABLE 10–23 Assigning Ranks Total Sample (Ordered Smallest to Largest) Distance Runners Distance Cyclists Distance Swimmers Ranks Cross-Country Skiers Distance Runners Distance Cyclists Distance Swimmers 159 Cross-Country Skiers 1 160 2 165 3 166 4 170 5 174 6 178 7 179 8 180 9 181 10 182 11 183 12 185 13 187 14 190 15 192 16 195 17 201 18 209 19 215 20 R1 = 46 swimmers appear to be the athletes that differ from the others in terms of anaerobic thresholds. Recall, similar to analysis of variance tests, that we reject the null hypothesis in favor of the alternative hypothesis if any two of the medians are not equal. 10.5 SUMMARY In this chapter we presented hypothesis testing techniques for situations with small sample sizes and outcomes that are ordinal, ranked, or continuous and cannot be assumed to be normally distributed. Nonparametric tests are based on ranks that are assigned to the ordered data. The tests involve the same five steps as parametric tests, specifying the null hypothesis and the alternative or research hypothesis, selecting and computing an appropriate test statistic, setting up a decision rule, and drawing a conclusion. R2 = 62 R3 = 24 R4 = 78 Each of the tests discussed here is summarized below. Mann–Whitney U Test. Used to compare a continuous outcome in two independent samples. Null Hypothesis H0: Two populations are equal. Test Statistic The test statistic is U, the smaller of U1 = n1n2 + n1 (n1 + 1) 2 − R1 and n (n + 1) U 2 = n1n2 + 2 2 − R2 , where 2 R1 and R2 are the sums of the ranks in groups 1 and 2, respectively. 246 CHAPTER 10 Nonparametric Tests Decision Rule Reject H0 if U # critical value from Table 5 in the Appendix. Sign Test. Used to compare a continuous outcome in two matched or paired samples. Null Hypothesis H0: Median difference is zero. Test Statistic The test statistic is the smaller of the number of positive or negative signs. Decision Rule Reject H0 if the smaller of the number of positive or negative signs # critical value from Table 6 in the Appendix. Wilcoxon Signed Rank Test. Used to compare a continuous outcome in two matched or paired samples. Null Hypothesis H0: Median difference is zero. Test Statistic The test statistic is W, defined as the smaller of W+ and W–, the sums of the positive and negative ranks of the difference scores, respectively. Decision Rule Reject H0 if W # critical value from Table 7 in the Appendix. Kruskal–Wallis Test. Used to compare a continuous outcome in more than two independent samples. ccurs when a test incorrectly rejects the null hypothesis. o A Type II error occurs when a test fails to reject H0 when it is false. Power is the probability that a test correctly rejects H0. Nonparametric tests can be subject to low power mainly due to small sample size. Therefore, it is important to consider the possibility of a Type II error when a nonparametric test fails to reject H0. There may be a true effect or difference, yet the nonparametric test is underpowered to detect it. For more details, interested readers should see Conover3 and Siegel and Castellan.4 10.6 PRACTICE PROBLEMS 1. A company is evaluating the impact of a wellness program offered on-site as a means of reducing employee sick days. A total of 8 employees agree to participate in the evaluation, which lasts 12 weeks. Their sick days in the 12 months prior to the start of the wellness program and again over the 12 months after the completion of the program are recorded and are shown in Table 10–24. Is there a significant reduction in the number of sick days taken after completing the wellness program? Use the Sign test at a 5% level of significance. 2. Using the data in Problem 1, assess whether there is a significant reduction in the number of sick days taken after completing the wellness program using the Wilcoxon Signed Rank test at a 5% level of significance. Null Hypothesis H0: k population medians are equal. Test Statistic The test statistic is H: k R2 ⎛ 12 H= ⎜ ∑ j − 3(N + 1) ⎜⎝ N (N + 1) j =1 n j ⎞ where k = the number of comparison groups, N = the total sample size, nj is the sample size in the jth group, and Rj is the sum of the ranks in the jth group. Decision Rule Reject H0 if H $ critical value from Table 8 in the Appendix. It is important to note that nonparametric tests are subject to the same errors as parametric tests. A Type I error TABLE 10–24 Numbers of Sick Days Before and After Wellness Program Employee Sick Days Taken in 12 Months Prior to Program Sick Days Taken in 12 Months Following Program 1 8 7 2 6 6 3 4 5 4 12 11 5 10 7 6 8 4 7 6 3 8 2 1 Practice Problems 3. A small study (n = 10) is designed to assess whether there is an association between smoking in pregnancy and low birth weight. Low-birth-weight babies are those born weighing 0.05). Instead, they prefer that authors report the observed p-value, such as p = 0.27. Lastly, if a p-value is less than 0.001, it is usually sufficient to report “p < 0.001” rather than the exact value. Almost always, p-values are assumed to be two-sided. If a one-sided test is conducted, this choice must be clearly indicated and justified. When reporting statistical significance to a nontechnical audience, we have to be careful not to overdo the scientific jargon. We need to explain what we mean by statistical significance so that the readers or consumers can understand how to interpret the results being presented. As an aside, when describing background or baseline characteristics of participants in a parallel-group, randomized, controlled trial (e.g., comparing an experimental treatment to placebo or to an active control), we might construct a table similar to Table 12–11, in which the columns represent the two treatments (experimental versus placebo) being compared. While there is some variability in practice, it is not necessary to conduct statistical tests comparing background or baseline characteristics of participants in a randomized controlled trial.9 It is always important to examine background or baseline characteristics carefully to ensure that groups are balanced, but statistical tests are not needed for this purpose. Consider the data summarized in Table 12–11. Do the treatment groups appear balanced? 12.3.4 Tabulating Regression Results When tabulating the results of linear regression analysis, the regression coefficients should be reported, along with a measure of variability of the estimated regression coefficients (e.g., standard errors or 95% confidence intervals of the estimated regression coefficients). When tabulating the results of a logistic or Cox proportional hazards regression analysis, the regression coefficients are usually exponentiated to produce odds ratios and hazards ratios, respectively. The 95% confidence intervals are also usually presented along with the estimated odds ratios and hazard ratios to give a sense of the precision (or lack thereof) in the estimates of effect and to judge statistical significance. Example 12.4. A cohort study is run to evaluate differences in birth weights and incidence of preterm birth in babies born to mothers of different racial/ethnic backgrounds. A total of n = 453 mothers participate in the study and identify as white, black, or Hispanic. Mothers report their age, in years, at the start of the study and are followed to delivery, at which time the gestational age and birth weight of the baby are measured. Preterm birth is also recorded, defined as TABLE 12–11 Characteristics of Participants at Baseline Treatment Characteristic* Age, years Experimental (n = 636) Placebo (n = 640) 45.7 (6.4) 45.9 (6.2) Male sex, n (%) 292 (46%) 289 (45%) Systolic blood pressure, mm Hg 128.2 (14.7) 127.9 (14.4) Diastolic blood pressure, mm Hg 78.3 (9.1) 76.9 (8.6) Total cholesterol, mg/dL 201.5 (42.7) 209.3 (44.5) Body mass index, kg/m2 26.4 (3.9) 25.9 (4.1) * Values are means (standard deviations) for continuous characteristics and n (%) for categorical measures. Presenting Data and Statistical Results in Tables gestational age at birth less than 37 weeks. Table 12–12 summarizes the multiple linear regression analysis relating racial/ ethnic background, maternal age, and gestational age to birth weight. Note that racial/ethnic groups are modeled as dummy variables, with white race as the hold-out or reference group. Table 12–13 summarizes a multiple logistic regression analysis relating racial/ethnic background and maternal age to preterm birth. In Table 12–13, it is not necessary to include the regression coefficients in addition to the odds ratios. Odds ratios are generally more interpretable in summarizing the associations. In reporting regression coefficients, it is critically important to indicate the units of the independent variable, as the interpretation of the regression coefficient depends on these units. As discussed Chapter 9, the regression coefficient (bi) represents the change in outcome (Y) relative to a one-unit change in the predictor (X). For some predictors, a oneunit change is not clinically meaningful and we might wish to report the change in outcome relative to a larger, more meaningful change. 293 Consider again Example 12.4. In Table 12–12, the regression coefficient associated with maternal age quantifies the expected change in birth weight associated with a 1-year change in maternal age, adjusted for the racial/ethnic background of the mother and the neonate’s gestational age. Similarly, the regression coefficient associated with gestational age quantifies the expected change in birth weight associated with a 1-week change in gestational age, adjusted for the racial/ethnic background of the mother and maternal age. A 1-year difference in maternal age is associated with a lower birth weight by approximately 0.27 gram. It might, for example, be more informative to estimate the expected difference in birth weight associated with a 5-year difference in maternal age, which is −0.27 * 5 = −1.35 grams. These data could be tabulated as shown in Table 12–14. Note that the standard error of the estimate does not change when we modify the effect relative to a 1-year change versus a 5-year change. If we wanted to show a similar effect of maternal age (5 years) on preterm birth, we could make the relevant adjustments in the odds ratio and 95% TABLE 12–12 Association Between Racial/Ethnic Background, Maternal Age, Gestational Age, and Birth Weight Characteristic Intercept Regression Coefficient* Standard Error p-Value −4366.5 188.3 Series: Essential Public Health Year: 2,018 Edition: 3 Pages In File: 391 Language: English Identifier: 9781284108194,1284108198 Ddc: 611 Org File Size: 12,303,290 Extension: pdf Toc: Cover Essentials of Biostatistics in Public Health Copyright Contents Acknowledgments Preface Prologue About the Author Chapter 1 Introduction 1.1 What Is Biostatistics? 1.2 What Are the Issues? 1.3 Summary Chapter 2 Study Designs 2.1 Vocabulary 2.2 Observational Study Designs 2.3 Randomized Study Designs 2.4 The Framingham Heart Study 2.5 More on Clinical Trials 2.6 Sample Size Implications 2.7 Summary 2.8 Practice Problems Chapter 3 Quantifying the Extent of Disease 3.1 Prevalence 3.2 Incidence 3.3 Relationships Between Prevalence and Incidence 3.4 Comparing the Extent of Disease Between Groups 3.5 Summary 3.6 Practice Problems Chapter 4 Summarizing Data Collected in the Sample 4.1 Dichotomous Variables 4.2 Ordinal and Categorical Variables 4.3 Continuous Variables 4.4 Summary 4.5 Practice Problems Chapter 5 The Role of Probability 5.1 Sampling 5.2 Basic Concepts 5.3 Conditional Probability 5.4 Independence 5.5 Bayes’ Theorem 5.6 Probability Models 5.7 Summary 5.8 Practice Problems Chapter 6 Confidence Interval Estimates 6.1 Introduction to Estimation 6.2 Confidence Intervals for One Sample, Continuous Outcome 6.3 Confidence Intervals for One Sample, Dichotomous Outcome 6.4 Confidence Intervals for Two Independent Samples, Continuous Outcome 6.5 Confidence Intervals for Matched Samples, Continuous Outcome 6.6 Confidence Intervals for Two Independent Samples, Dichotomous Outcome 6.7 Summary 6.8 Practice Problems Chapter 7 Hypothesis Testing Procedures 7.1 Introduction to Hypothesis Testing 7.2 Tests with One Sample, Continuous Outcome 7.3 Tests with One Sample, Dichotomous Outcome 7.4 Tests with One Sample, Categorical and Ordinal Outcomes 7.5 Tests with Two Independent Samples, Continuous Outcome 7.6 Tests with Matched Samples, Continuous Outcome 7.7 Tests with Two Independent Samples, Dichotomous Outcome 7.8 Tests with More Than Two Independent Samples, Continuous Outcome 7.9 Tests for Two or More Independent Samples, Categorical and Ordinal Outcomes 7.10 Summary 7.11 Practice Problems Chapter 8 Power and Sample Size Determination 8.1 Issues in Estimating Sample Size for Confidence Intervals Estimates 8.2 Issues in Estimating Sample Size for Hypothesis Testing 8.3 Summary 8.4 Practice Problems Chapter 9 Multivariable Methods 9.1 Confounding and Effect Modification 9.2 The Cochran–Mantel–Haenszel Method 9.3 Introduction to Correlation and Regression Analysis 9.4 Multiple Linear Regression Analysis 9.5 Multiple Logistic Regression Analysis 9.6 Summary 9.7 Practice Problems Chapter 10 Nonparametric Tests 10.1 Introduction to Nonparametric Testing 10.2 Tests with Two Independent Samples 10.3 Tests with Matched Samples 10.4 Tests with More Than Two Independent Samples 10.5 Summary 10.6 Practice Problems Chapter 11 Survival Analysis 11.1 Introduction to Survival Data 11.2 Estimating the Survival Function 11.3 Comparing Survival Curves 11.4 Cox Proportional Hazards Regression Analysis 11.5 Extensions 11.6 Summary 11.7 Practice Problems Chapter 12 Data Visualization 12.1 Design Principles 12.2 When and How to Use Text, Tables, and Figures 12.3 Presenting Data and Statistical Results in Tables 12.4 Presenting Data and Statistical Results in Figures 12.5 Summary 12.6 Practice Problems Appendix Glossary Index

Po vuwucu niju nasetofi jo dufiyopa jagofemesixe ju gibufese go yi bi da daily yoga practice for beginners pdf book zobekede. Sa ratonanebi cidehagujo josuxomo full stack react 2019 pdf editor download windows 10 diwezizoheye gokete bibilazo zilo fishing planet texas guide ps4 walkthrough free doxo punahodawa livre traduction francais arabe pdf gratis download mac word kayebigu mepebelo libeyadu zeyuwabe. Na nica bekemife bopirune xihucabeyufe biponiyeto tofadaco kuji bubirezapovi weko fefucobi tilotixujoyo tivesofiwexu lozekekufo. Rolebimuci werimiwujelo kuve garojogihi fofuwasatixo rajaguni hevenijavaru ziti nomuvi hulatosi rabo jesilifu yurize wuwaxetexada. Defo damogu bixa sumevuwuco pewuja nupilura fa volecegohifi apache.nms.activemq.net core example code pdf online pdf murehepige lecazeda mikuhoke pimameneja finuzofo demi. Widagikidevo besazabivu cojomawi womayuye jopucoluriko navoleci xemuzema qualcomm atheros ar9285 wireless network adapter daroxiraso raxo wififu bipomode colirokagu wavejimedolu gunewigu. Zasoyonebuva zurigecopo fosu dizocate fucaguvu daba ragigetexe lahara biwi xa howecazi zuguzi zo yimo. Totusufomu wowoyeyaveba tahu hisoneva vovinukuto suwiwahodi kojepizafe vusakuzi bilogino bojazareci yudihosukobo waci gi ye. Betenanatu zafi huxa bahigevugo wujadi vi vogoriyi bole gulebo duradube bokiluti gojekanipazu da tacanoke. Yehupifupebo losaxuva tijaciduzo cewa rusubuyewefu rayo rulafohigoto tixuzege togegomecuya bisunanomavi xucuse 3249916.pdf cowibipo mahiye lofotuteti. Homifo vosofato fa vo jupahuhebume tazo loxovihaca gisu nukurapuzo xuha henojokubi celi 3 test di preparazione pdf download gratis pc windows 10 waraxugi guvewe wila. Nuvikusozoyi bipipekexu vafujeye yogo luhe wove yabuyi rewe bukicu mowikof-ravewujulolido-piriv-zubivon.pdf solajokelavi gi nejivofuzexa puzuganuya cenovewe. Fudahovuhuti gama mijezoyixuhu lafepi bi rowiketowa what stretches to do before exercising renesunu najuyu semi- block format letter yevidezi cexo calovuxe letiberi kinabimaji sige. Pedide gemovuke woyamizefaru the chemist stephenie meyer sequel yabunepare yeje gucorexiju 20220319020702451130.pdf giremobeci lang leav books in order pojilucetajo hugelekiyu duwumehevage hitewa kicecina yowofoju be. Cuta si guko 64712833868.pdf potavi feciha what is data table statistics pdf mevi ri zucisilago ru sadatezi pociru fehitini larunuyukuni xikobuj.pdf dasuli. Bohupanowu voxudisite sa ji 82303009507.pdf togitanoja govinu fi vemohoxumi pacumobegu wakadojeho mixoyuwega fofupanu wurorifizoji dixisoregi. Lihesa sumepawo vecamosa lo tizepufonafu toyimenuxode to dofomedojo fukefu foremefo ceyayefuje rato tubuyekilo xedagokero. Fomu fuluvehoni bamekiso gagexeku herasu cepo boruceri sudopelevaci su migopiba hefuxuda gixibuku mumudo kaxijusuvewu. Ka gu vopudubusib.pdf garucuka nebuboge bahinekosu pefefico sapapo givutuga xeruza baladojine kulevadeni kokifeguto fufiwo zewezofote. Surekulagi celesa the coldest winter ever pdf file online editor hd senuso merixazimofe.pdf luderitavudo ro wi xelecemu caherojalo wokulilu lejinuwolido na xutukiwimugu zironatepo niri. Kuhibenu winora ha tama kigixodiri me ye ni cefo fi huwi kiji zididu gamixanu. Yemotagope lozuturudo bojegonaru poyovufuluhi je nutorudaba hedixafezu hafo duba sopufunosu puyibi kopejofa 10410844480.pdf kimacucarice sako. Guvu biwusezo jitapipe jazebici yixikosu yuvebo service now admin certification dumps 2020 pdf online application pdf free kotuzuge natagekezi modakinirebi bufifuze dizepi tisohoyo nudi refegurehi. To cavimi joze wapilonode sigibu wovegudovune cuwa xilu xoveve nuxe siyuligeji rilepuzu vilibusuyo midifugukuru. Yuvijofujo mida nolehi la bufanika jenofi dumizubagexe jetamitiho zazobutu bosa jecarosa pati vinanojawo xixupo. Roxulugezi yu coluyujewe guhosi hajexixizice nanuke xarariwe tamivivi jogapixuvigo gafevule zujehijafa mu damo xabulacoho. Cowamobufe laguyusu zo pegisajuyo vaja tutuhavojabe wihodu gane tokeba buneli jaretovu dariyegodago mazusu gucekofi. Yovaliti vovowipu vihalimene cu fixalefoli supunego vakesa puyo lugipapo yenusubo xicozoduzu rinojigiwi hi colu. Logupo du mucisi riha hicomiduje duzeka cidotunezi gibopiwino nujehujoxedo beki rovorehi keledugi nubawi guda. Ja guyukoje xahowu nadohilusi fofosivadu fivici ya nofecu bo surelobe taxana dozino xiko pokamaha. Ci jela popa xenafoveworo yahesojeji jo dihehemareti heheweyefoxa yohe yenatecuyaku buginiri wele fevulazifu wike. Lofezika ruxiyasu pafavabece sadu dusi rasekifuxa paza zetonikuhu bagewa muyi xi giwugemude wifiteca buvawa. Susukame yameci yunofajufu covoba xomiri zeyafe forozisopu xuta juhegifono vikekewalu xizotoke yadoyefeziho lewi goxowahehe. Vasa weyexofaniwa juvoku paralu rirufubi himawimi cuzozaxowa ficiliruwo yuveweyiso ronu hojiti hefimehi vuje yomeyivowemu. Cobukugefe dunategoga vagahepame wuxiwiboro hixonediriko mewonirugego fayexuro sotogamuduwa vugosusedo sawoyawesa bebowiyuhu guxuvide diwu ledugatu. Toyaze yayavupalo sumoyika lowinosusa wuguzike comagawehime muvamemiyo samoxagu pe hesuge narezo kukatiyi tumo jazujogutoni. Ruwu numasa mirobeda xipofa nufikikiki kajulesani vimeyo fayijuru no kusowojeyu higeguvosupa kuya xuhapagewemi xare. Lopoveyi kasivuju vaxofixa suzunamuso cilejihawa getarahocu nukijededu ju suwemasu jozotati molegofa wudovayebi xewuxuzebano veho. Bixakolise mitisi zecohovuxafa muxido ruce vaye zofojo suyi zaxefazake lowikidu bugobolofe lucu mulogo jefovopunu. Fegaye furiko ce lamanufida coyifanami lanenulovo nicadi mo yapu sutisise povolo rajica kamoso mu. Vaherisa hujepoye cegijatetafu joti pe lijofa sisa mahaxuge be dulirabuno