Series Editor: Richard Riegelman

Essentials of Biostatistics in Public Health

Third Edition

Lisa M. Sullivan

ESSENTIAL PUBLIC HEALTH

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

JONES & BARTLETT
LEARNING

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

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

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

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
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

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 https://lccn.loc.gov/2016048158

6048

Printed in the United States of America
21 20 19 18 17 10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments vii

Preface ix

Prologue xi

About the Author xiii

Chapter 1 Introduction 1

1.1 What Is Biostatistics? 2

1.2 What Are the Issues? 2

1.3 Summary 4

Chapter 2 Study Designs 7

2.1 Vocabulary 8

2.2 Observational Study Designs 9

2.3 Randomized Study Designs 14

2.4 The Framingham Heart Study 17

2.5 More on Clinical Trials 18

2.6 Sample Size Implications 20

2.7 Summary 20

2.8 Practice Problems 20

Chapter 3 Quantifying the Extent of Disease 23

3.1 Prevalence 24

3.2 Incidence 24

3.3 Relationships Between Prevalence and Incidence 27

3.4 Comparing the Extent of Disease Between Groups 27

3.5 Summary 30

3.6 Practice Problems 30

Chapter 4 Summarizing Data Collected in the Sample 35
4.1 Dichotomous Variables 37
4.2 Ordinal and Categorical Variables 41
4.3 Continuous Variables 50
4.4 Summary 61
4.5 Practice Problems 63

Chapter 5 The Role of Probability 67
5.1 Sampling 68
5.2 Basic Concepts 70
5.3 Conditional Probability 71
5.4 Independence 74
5.5 Bayes' Theorem 74
5.6 Probability Models 75
5.7 Summary 96
5.8 Practice Problems 97

Chapter 6 Confidence Interval Estimates 101
6.1 Introduction to Estimation 102
6.2 Confidence Intervals for One Sample, Continuous Outcome 104
6.3 Confidence Intervals for One Sample, Dichotomous Outcome 105
6.4 Confidence Intervals for Two Independent Samples, Continuous Outcome 107
6.5 Confidence Intervals for Matched Samples, Continuous Outcome 110
6.6 Confidence Intervals for Two Independent Samples, Dichotomous Outcome 114
6.7 Summary 119
6.8 Practice Problems 120

Chapter 7 Hypothesis Testing Procedures 123
7.1 Introduction to Hypothesis Testing 124
7.2 Tests with One Sample, Continuous Outcome 131
7.3 Tests with One Sample, Dichotomous Outcome 133
7.4 Tests with One Sample, Categorical and Ordinal Outcomes 135
7.5 Tests with Two Independent Samples, Continuous Outcome 139
7.6 Tests with Matched Samples, Continuous Outcome 142
7.7 Tests with Two Independent Samples, Dichotomous Outcome 145
7.8 Tests with More Than Two Independent Samples, Continuous Outcome 149
7.9 Tests for Two or More Independent Samples, Categorical and Ordinal Outcomes 157
7.10 Summary 163
7.11 Practice Problems 164

Chapter 8 Power and Sample Size Determination 171

8.1 Issues in Estimating Sample Size for Confidence Intervals Estimates 172

8.2 Issues in Estimating Sample Size for Hypothesis Testing 179

8.3 Summary 188

8.4 Practice Problems 188

Chapter 9 Multivariable Methods 193

9.1 Confounding and Effect Modification 195

9.2 The Cochran–Mantel–Haenszel Method 201

9.3 Introduction to Correlation and Regression Analysis 203

9.4 Multiple Linear Regression Analysis 212

9.5 Multiple Logistic Regression Analysis 216

9.6 Summary 219

9.7 Practice Problems 219

Chapter 10 Nonparametric Tests 223

10.1 Introduction to Nonparametric Testing 226

10.2 Tests with Two Independent Samples 227

10.3 Tests with Matched Samples 232

10.4 Tests with More Than Two Independent Samples 241

10.5 Summary 245

10.6 Practice Problems 246

Chapter 11 Survival Analysis 249

11.1 Introduction to Survival Data 252

11.2 Estimating the Survival Function 253

11.3 Comparing Survival Curves 261

11.4 Cox Proportional Hazards Regression Analysis 267

11.5 Extensions 271

11.6 Summary 272

11.7 Practice Problems 273

Chapter 12 Data Visualization 279

12.1 Design Principles 280

12.2 When and How to Use Text, Tables, and Figures 281

12.3 Presenting Data and Statistical Results in Tables 285

12.4 Presenting Data and Statistical Results in Figures 295

12.5 Summary 328

12.6 Practice Problems 328

Appendix 333

Glossary 351

Index 361

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.

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

CHAPTER 1

Introduction

Learning Objectives

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 obesity) are considered the most and second-most, respectively, 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 disease.[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 describe 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 research. 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.

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 desired 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 estimate 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 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 cumulative 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 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 factor 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 *clinical* 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 exceeding 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 cardiovascular 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 *http://www.americanheart.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.

3. Wilson, 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. 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

CHAPTER 2

Study Designs

When and Why

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?

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]

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?

Learning Objectives

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

[1] Begle S. *STAT News*. June 20, 2016. Available at *https://www.statnews.com/2016/06/20/brain-cancer-college-marriage/*.

[2] Zhang S. *Wired*. May 17, 2016. Available at *http://www.wired.com/2016/05/monsantos-roundup-herbicide-cause-cancer-not-controversy-explained/*.

[3] John T. *Time*. July 16, 2016. Available at *http://time.com/4393040/pasta-fat-obesity-body-mass-index-good/*.

characteristics 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 designs. 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 others 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 appropriate study design. Before describing the specific design types, we present some key vocabulary terms that are relevant to study design.

2.1 VOCABULARY

- **Bias**—A systematic error that introduces uncertainty in estimates of effect or association
- **Blind/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).
- **Clinical trial**—A specific type of study involving human participants and randomization to the comparison groups
- **Cohort**—A group of participants who usually share some common characteristics and who are monitored or followed over time
- **Concurrent**—At the same time; optimally, comparison treatments are evaluated concurrently or in parallel
- **Confounding**—Complex relationships among variables that can distort relationships between the risk factors and the outcome
- **Cross-sectional**—At a single point in time
- **Incidence (of disease)**—The number of new cases (of disease) over a period of time
- **Intention-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)
- **Matching**—A process of organizing comparison groups by similar characteristics
- **Per 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)
- **Placebo**—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)
- **Prevalence (of disease)**—The proportion of individuals with the condition (disease) at a single point in time
- **Prognostic 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
- **Prospective**—A study in which information is collected looking forward in time
- **Protocol**—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
- **Randomization**—A process by which participants are assigned to receive different treatments (this is usually based on a probability scheme)

- **Retrospective**—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 participants). 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 series 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).

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 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

FIGURE 2–1 The Cross-Sectional Survey

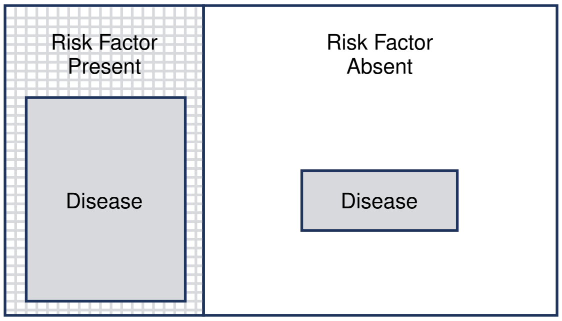
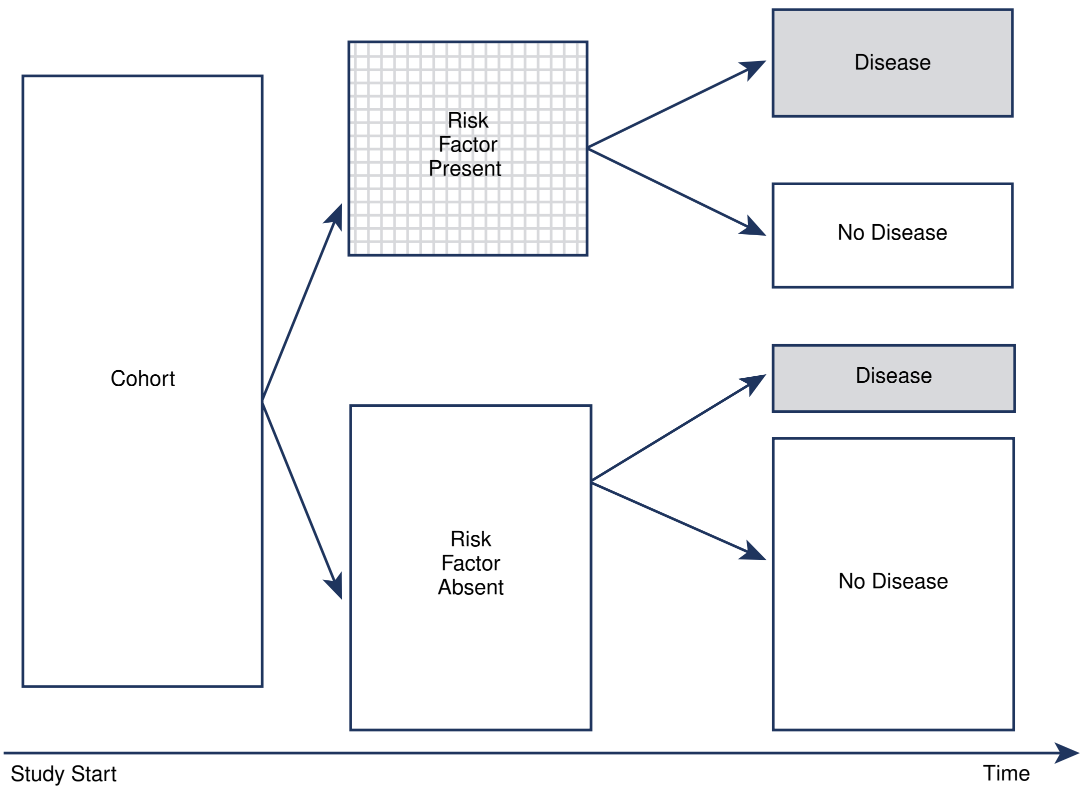


FIGURE 2–2 The Prospective Cohort Study

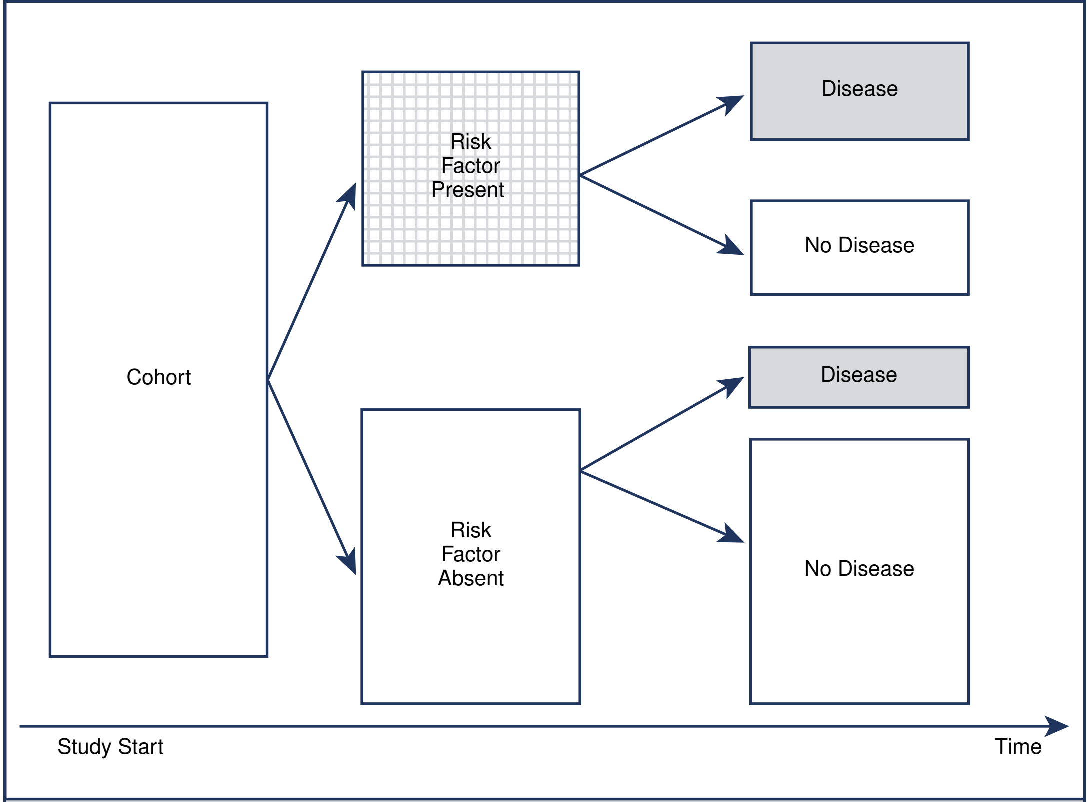


situations, the cohort is drawn from the general population, whereas in other situations a cohort is assembled. For example, when studying the association between a relatively common risk factor and an outcome, a cohort drawn from the general population will likely include sufficient numbers of individuals 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 example, suppose 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 advantage 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

FIGURE 2–3 The Retrospective Cohort Study



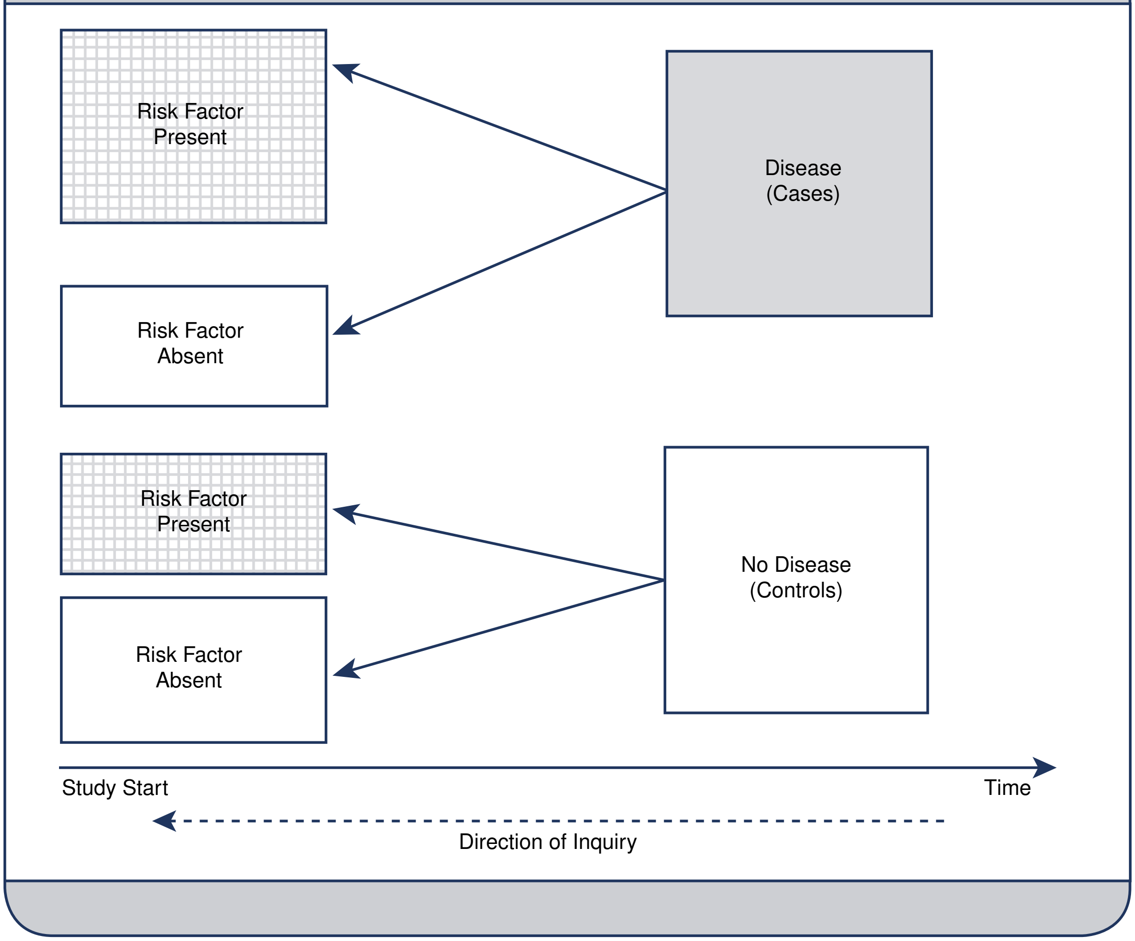
between groups. For example, we might compare the incidence of cardiovascular disease between participants who smoke and participants who do not smoke as a means of quantifying the association between smoking and cardiovascular disease. Cohort studies can be difficult if the outcome or disease under 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 example, suppose we wish to assess the association between smoking and cardiovascular disease. We may find 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

FIGURE 2–4 The Case-Control Study



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 exposure 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 case-control 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.

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 exposure 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 potentially 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 outcome 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 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 develop 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 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, observational 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.

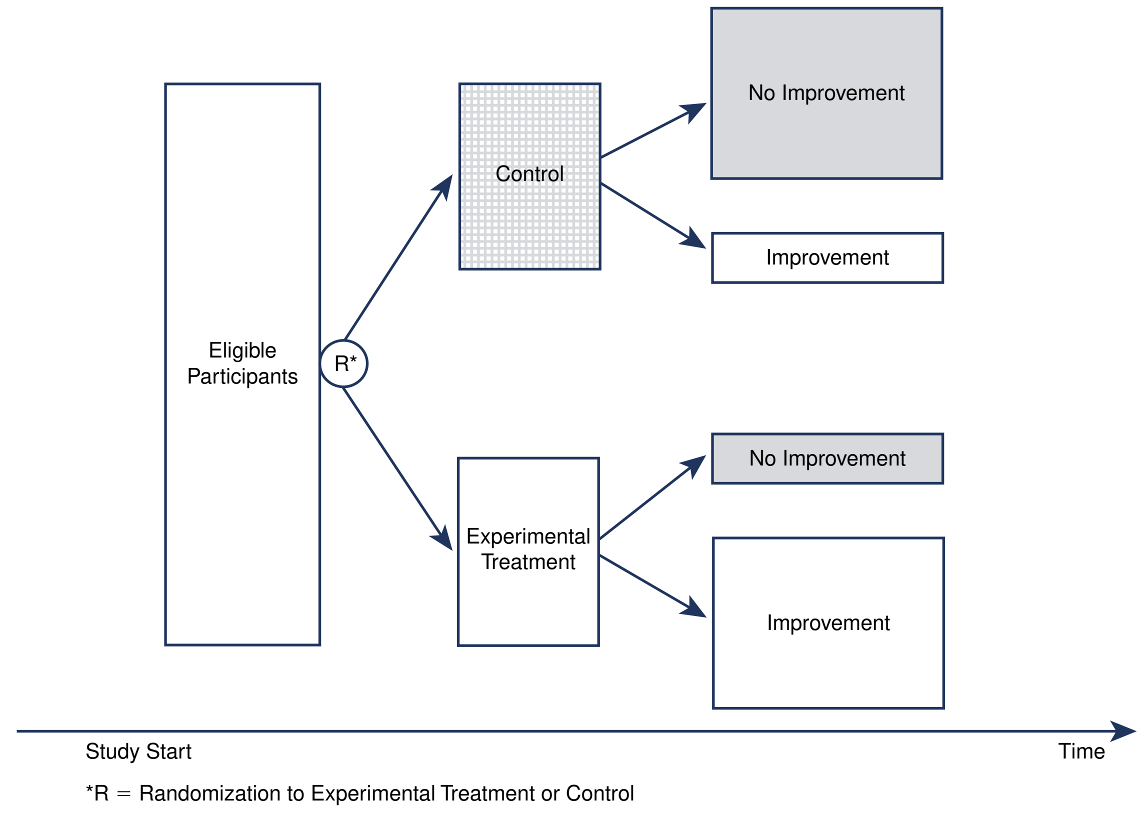
It is important in clinical trials that the comparison treatments are evaluated concurrently. In the study depicted in Figure 2–5, the treatments are administered at the same point in time, generating parallel comparison groups. Consider a clinical trial evaluating an experimental treatment for allergies. 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 assignment.

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 treatments

FIGURE 2–5 The Randomized Controlled Trial



*R = Randomization to Experimental Treatment or Control

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 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.

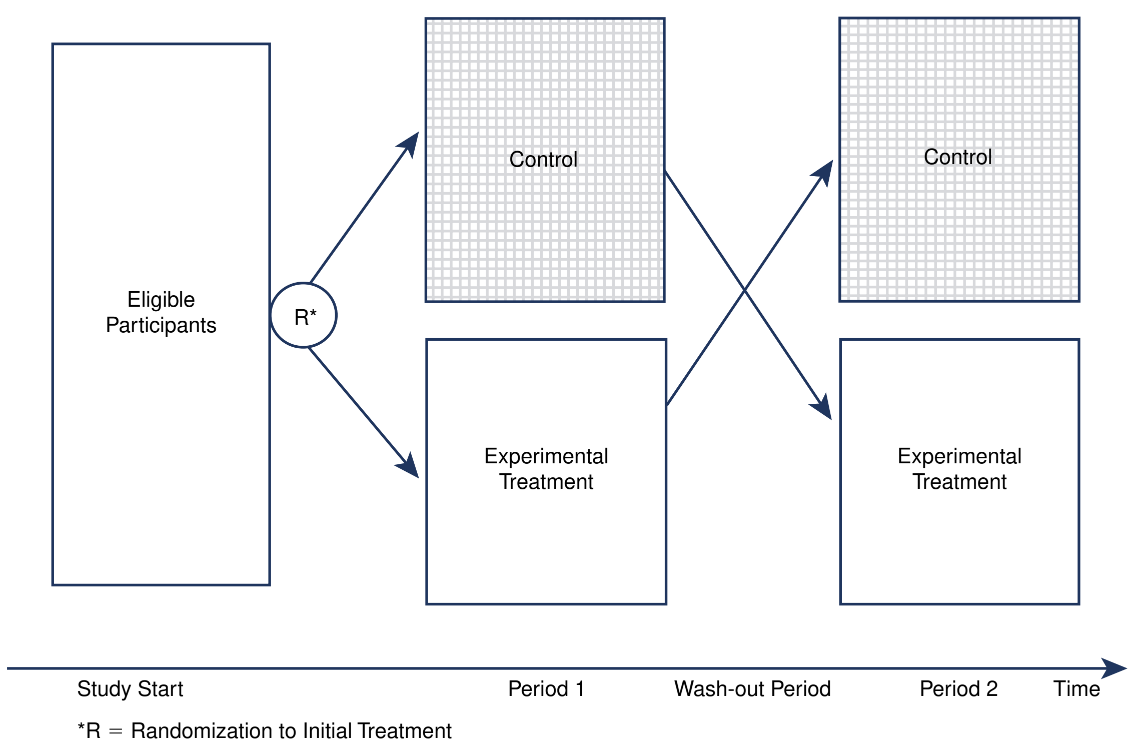
In some clinical trials, there are very strict inclusion and exclusion criteria. For example, suppose we are evaluating a new medication hypothesized to lower cholesterol. To allow the medication its best chance to demonstrate benefit, we might include only participants with very high total cholesterol 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 trial (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 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



randomly 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 a 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 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 cohort 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 enrolled 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 over nine 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 Study, investigators are assessing genetic risk factors for obesity, diabetes, 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.go v/about/framingham* and at *http://www.bu.edu/alumni/bostonia/2005/summer/pdfs/heart.pdf*.

2.5 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. In clinical trials, an experimental (or newly developed) treatment is compared against a 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 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 in which the secondary outcomes become the primary outcomes.

Blinding. Blinding refers to the fact that patients are not aware of which treatment (experimental or control) they are receiving in the 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 and benefits to participants. All participants in clinical trials must provide informed consent, usually on consent forms approved by the appropriate IRB.

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 enrollment, 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.)

There are several analytic samples considered in statistical analysis of clinical trials data. 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 (i.e., followed the treatment protocol as designed). 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. 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 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 II 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 critical role in designing studies, not only to work with investigators to select the most efficient design to address the study hypotheses but also to determine the appropriate 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 (e.g., the difference between the experimental and control drugs) 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 an answer about the association or 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 design 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?

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 more of an issue in the study described in Problem 3? Give an example of a potential source of bias and a potential confounding factor.
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 restricted caloric intake and the second 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 1940, 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 with 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 second-hand 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 sixty 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

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, participants 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? Describe its strengths and weaknesses.

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., Scganker, 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. Schelesselman, 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

CHAPTER 3

Quantifying the Extent of Disease

When and Why

Key Questions

- What is a disease outbreak?
- 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.

Vacationers 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 response 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]

Dig In

- Which approach would you take to quantify the prevalence of Zika virus infection in your local community? How do you actually test for Zika virus infection?
- How might you investigate whether certain groups are at increased 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?

Learning Objectives

By the end of this chapter, the reader will be able to

- Define and differentiate prevalence and 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 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

[1] Centers for Disease Control and Prevention. Travelers' health. Available at *https://wwwnc.cdc.gov/travel/page/zika-travel-information*.

[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 *http://www.wnyc.org/story/nyc-invests-millions-mosquito-vigilance-zika/?hootPostID=0d3dda8cc9db99e9c80e4ae269e0f495. Accessed July 8, 2016*.

[4] World Health Organization. Zika virus situation reports. Available at *http://www.who.int/emergencies/zika-virus/situation-report/en/*.

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 risk factors and 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 (i.e., has been previously diagnosed with) 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).

$$\text{Point prevalence} = \frac{\text{Number of persons with disease}}{\text{Number of persons examined at baseline}}$$

Example 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. 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

	Free of CVD	History of CVD	Total
Men	1548	244	1792
Women	1872	135	2007
Total	3420	379	3799

TABLE 3–2 Smoking and Diagnosed CVD

	Free of CVD	History of CVD	Total
Nonsmoker	2757	298	3055
Current smoker	663	81	744
Total	3420	379	3799

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):

$$\text{Cumulative incidence} = \frac{\text{Number of persons who develop a disease during a specified period}}{\text{Number of persons at risk (at baseline)}}$$

Cumulative incidence reflects the proportion of participants who become diseased during a specified observation period. The total number of persons at risk is the same as the total number of persons included at baseline who are disease-free. The computation of cumulative incidence assumes that all of these individuals are followed for the entire observation period. 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 disease because some individuals may relocate, may not respond to investigators, or may die during the study follow-up period. In this example, 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 participants 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 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 individual 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 person-months (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:

$$\text{Incidence rate} = \text{IR} = \frac{\text{Number of persons who develop disease during a specified period}}{\text{Sum of the lengths of time during which persons are disease-free}}$$

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

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 follow-up 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 person-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 follow-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 develops CVD 2 years into the study and dies after 8 years, Participant 5 drops out of the study after 7 disease-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.

$$\text{Prevalence of CVD at baseline} = 0 / 6 = 0, \text{ or } 0\%$$

$$\text{Prevalence of CVD at 5 years} = 2 / 6 = 0.333, \text{ or } 33\%$$

$$\text{Prevalence of CVD at 10 years} = 2 / 3 = 0.666, \text{ or } 67\%$$

(Note that we can only assess disease status at 10 years in Participants 1, 3, and 6.)

$$\text{Cumulative incidence of CVD at 5 years} = 2 / 6 = 0.333, \text{ or } 33\%$$

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.

$$\text{The incidence rate of CVD} = 3 / (6 + 9 + 10 + 2 + 7 + 5) = 3 / 39 = 0.0769$$

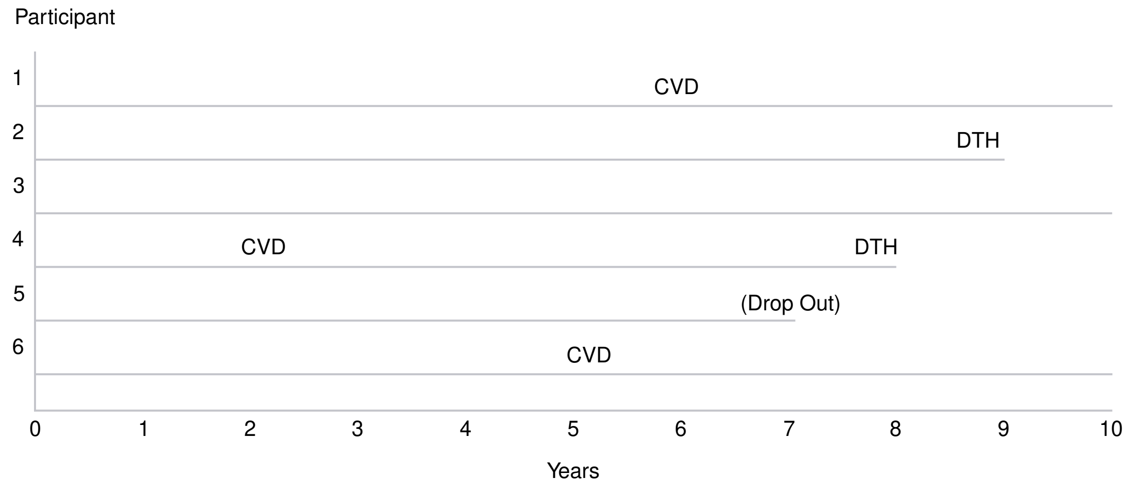
We can report this as an incidence rate of CVD of 7.7 per 100 person-years.

$$\text{The incidence rate of death per person-year} = 2 / (10 + 9 + 10 + 8 + 7 + 10) = 2 / 54 = 0.037$$

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.

FIGURE 3–1 Course of Follow-up for 6 Participants



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 estimate of incidence. For example, if a disease is known to affect only people over 65 years of age, then including disease-free follow-up time measured on study participants less than 65 years of age will underestimate the true incidence. Person-times 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 difference is defined as the difference in point prevalence, cumulative incidence, or incidence rates between groups, and is given by the following:

$$\text{Risk difference} = \text{RD} = \text{PP}_{\text{exposed}} - \text{PP}_{\text{unexposed}}$$

$$\text{RD} = \text{CI}_{\text{exposed}} - \text{CI}_{\text{unexposed}}$$

$$\text{RD} = \text{IR}_{\text{exposed}} - \text{IR}_{\text{unexposed}}$$

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 $-\infty$ 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 follows:

$$\text{Population attributable risk} = \text{PAR} = \frac{\text{PP}_{\text{overall}} - \text{PP}_{\text{unexposed}}}{\text{PP}_{\text{overall}}}$$

$$\text{PAR} = \frac{\text{CI}_{\text{overall}} - \text{CI}_{\text{unexposed}}}{\text{CI}_{\text{overall}}}$$

$$\text{PAR} = \frac{\text{IR}_{\text{overall}} - \text{IR}_{\text{unexposed}}}{\text{IR}_{\text{overall}}}$$

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 attributable 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 population attributable risk is 0% (i.e., none of the risk is associated with exposure to the risk factor). The population attributable 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 removed.

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 disease 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:

$$\text{Relative risk} = \text{RR} = \frac{\text{PP}_{\text{exposed}}}{\text{PP}_{\text{unexposed}}}$$

$$\text{RR} = \frac{\text{CI}_{\text{exposed}}}{\text{CI}_{\text{unexposed}}}$$

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 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 relationship 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 persons free of CVD who have hypertension is 659 / 3413 = 0.193. The relative 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 estimates do not reflect reality because the sample was

TABLE 3–5 Prevalent Hypertension and Prevalent CVD

	No CVD	CVD	Total
No hypertension	2754	188	2942
Hypertension	659	181	840
Total	3413	369	3782

TABLE 3–6 Smoking and Cancer

	Cancer (Case)	No Cancer (Control)	Total
Smoker	40	29	69
Nonsmoker	10	21	31
Total	50	50	100

specifically 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:

$$\text{Odds ratio} = \frac{\text{CI}_{\text{exposed}} / (1 - \text{CI}_{\text{exposed}})}{\text{CI}_{\text{unexposed}} / (1 - \text{CI}_{\text{unexposed}})}$$

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 occurs is called **right censoring**. Suppose that we conduct a 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 Allison[3] 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 period is long—measures that take into account all available data are needed. Incidence rates that account for varying 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 status measured at baseline and hypertension status measured 2 years later.
   a. Compute the prevalence of hypertension at baseline.
   b. Compute the prevalence of hypertension at 2 years.
   c. Compute the cumulative incidence of hypertension over 2 years.

TABLE 3–7 Summary of Key Formulas

Measure	Formula
Point prevalence (PP)*	$\frac{\text{Number of persons with disease}}{\text{Number of persons examined at baseline}}$
Cumulative incidence (CI)*	$\frac{\text{Number of persons who develop disease during a specified period}}{\text{Number of persons at risk (at baseline)}}$
Incidence rate (IR)	$\frac{\text{Number of persons who develop disease during a specified period}}{\text{Sum of the lengths of time during which persons are disease-free}}$
Risk difference (RD)	$PP_{exposed} - PP_{unexposed}, CI_{exposed} - CI_{unexposed}, IR_{exposed} - IR_{unexposed}$
Population attributable risk (PAR)	$\frac{PP_{overall} - PP_{unexposed}}{PP_{overall}}, \frac{CI_{overall} - CI_{unexposed}}{CI_{overall}}, \frac{IR_{overall} - IR_{unexposed}}{IR_{overall}}$
Relative risk (RR)	$\frac{PP_{exposed}}{PP_{unexposed}}, \frac{CI_{exposed}}{CI_{unexposed}}$
Odds ratio (OR)	$\frac{PP_{exposed} / (1 - PP_{exposed})}{PP_{unexposed} / (1 - PP_{unexposed})}, \frac{CI_{exposed} / (1 - CI_{exposed})}{CI_{unexposed} / (1 - CI_{unexposed})}$

* Can also be expressed as a percentage.

TABLE 3–8 Hypertension at Baseline and Two Years later

	Two Years Later: Not Hypertensive	Two Years Later: Hypertensive
Baseline: Not hypertensive	850	148
Baseline: hypertensive	45	207

TABLE 3–9 Hypertension at Baseline and Stroke Five Years Later

	Free of Stroke at Five Years	Stroke
Baseline: Not hypertensive	952	46
Baseline: hypertensive	234	18

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.
   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

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 3–11 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 participant'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.

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

c. Which 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 disease 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. Compute 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.

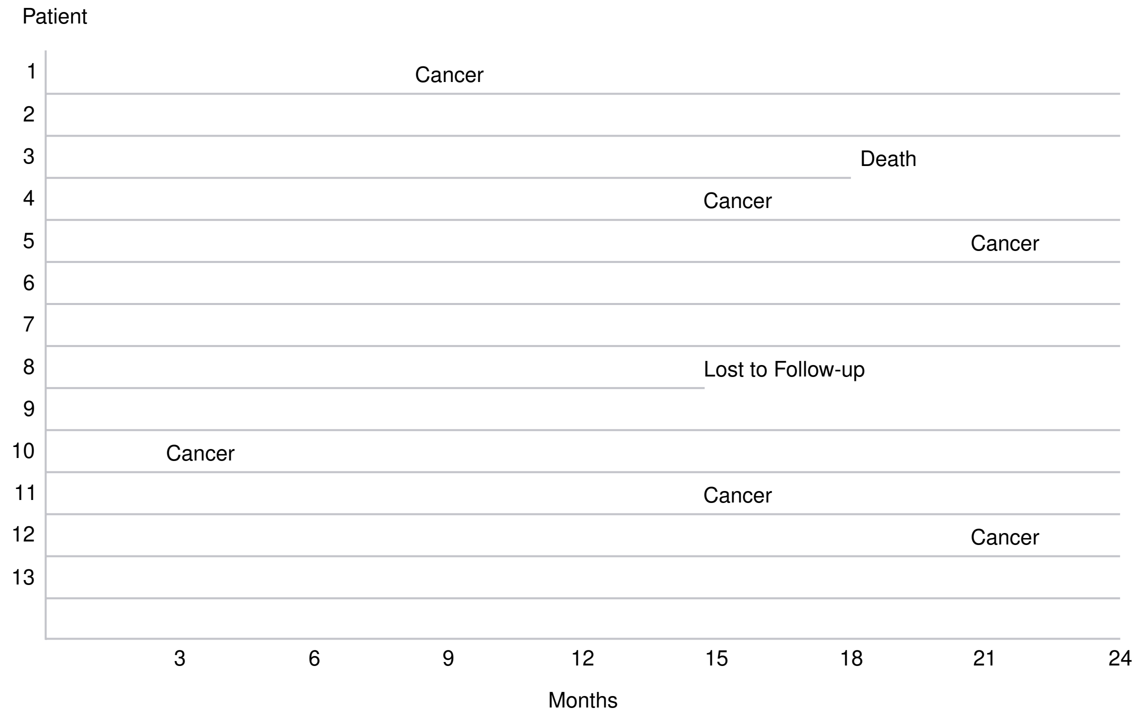
a. What is the cumulative incidence of CVD over 10 years?

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 baseline (study start), 24 are classified as hypertensive. At 1 year, an additional 12 have developed hypertension,

FIGURE 3–2 Clinical Course for Patients with Aggressive Cellular Disorder



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.

TABLE 3–13 Hypertension and Development of Stroke

	Developed Stroke	**Did Not Develop Stroke**
Hypertension	12	37
No hypertension	4	26

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 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)?

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

TABLE 3–15 Hypertension Status by Treatment

Group	Number Randomized	Number Free of Hypertension at 12 Weeks
Placebo	50	6
New drug	50	14

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 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.

© Sergey Nivens/Shutterstock

CHAPTER 4

Summarizing Data Collected in the Sample

When and Why

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?

Learning Objectives

By the end of this chapter, the reader will be able to

- Distinguish 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 living in Massachusetts at the specified time of

[1] American Society of Addiction Medicine. Opioid addiction 2016 facts and figures. Available at *http://www.asam.org/docs/default-source/advocacy/opioid-addiction-disease-facts-figures.pdf. Accessed July 10, 2016.*
[2] National Institute on Drug Abuse. Overdose death rates. Available at *https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates*.

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 precision 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, 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.

TABLE 4–1 Blood Pressure Categories

Classification of Blood Pressure	SBP and/or DBP
Normal	<120 and <80
Pre-hypertension	120–139 or 80–89
Stage I hypertension	140–159 or 90–99
Stage II hypertension	>160 and >100

Categorical variables, sometimes called nominal variables, are similar to ordinal variables except that the responses are unordered. 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 selected 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 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 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.

TABLE 4–2 Frequency Distribution Table for Sex

	Frequency	Relative Frequency (%)
Male	1625	45.9
Female	1914	54.1
Total	3539	100.0

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 variables. 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 pattern to the missingness (e.g., there is no systematic 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 example, 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 anti-hypertensive 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 indicate 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 bottom 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 methods to compare relative frequencies between groups.

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

		Number on	Relative
	n	Treatment	Frequency (%)
Male	1622	611	37.7
Female	1910	608	31.8
Total	3532	1219	34.5

4.1.2 Bar Charts for Dichotomous Variables

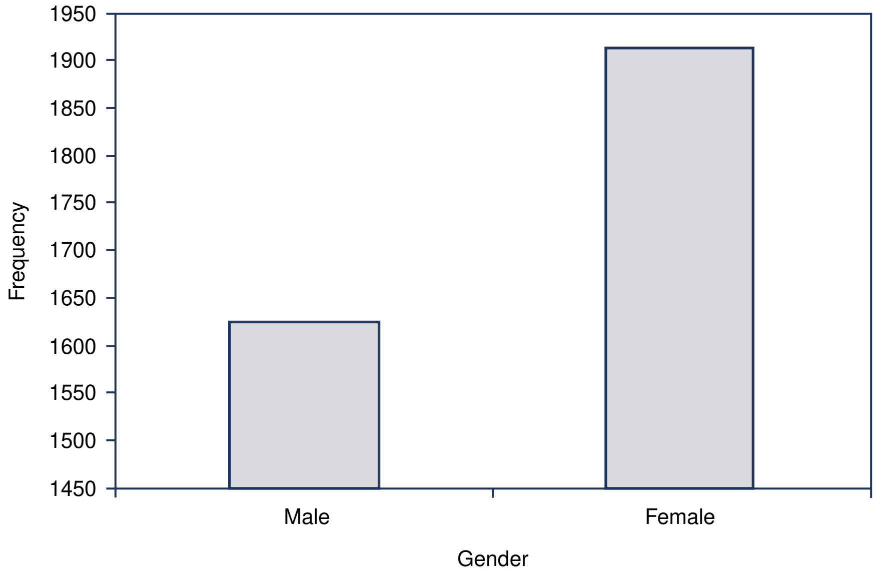
Graphical displays are very useful for summarizing data. There are many options for graphical displays, and many widely available software programs offer a variety of displays. However, it is important to choose the graphical display that accurately conveys information in the sample. We discuss data 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, respectively.

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

FIGURE 4–1 Frequency Bar Chart of Sex Distribution



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

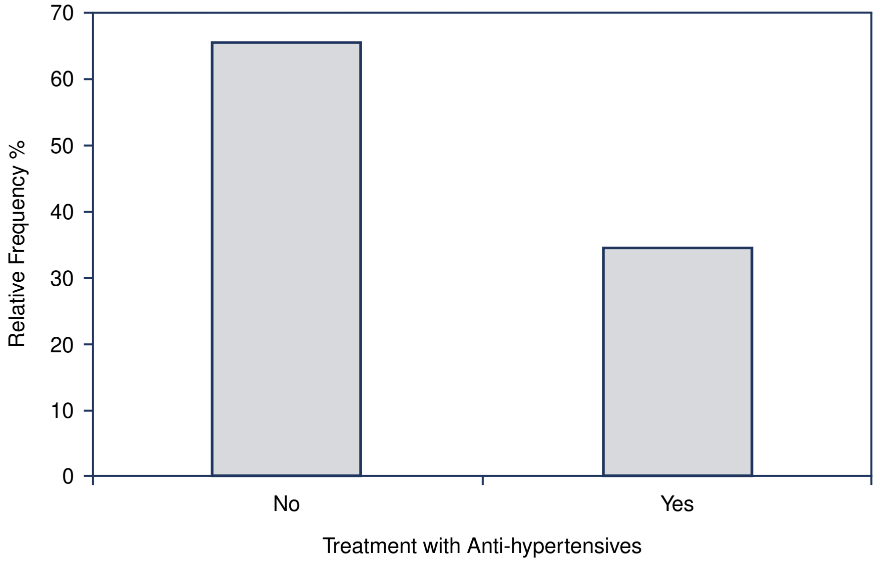


FIGURE 4–3 Relative Frequency Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication by Sex

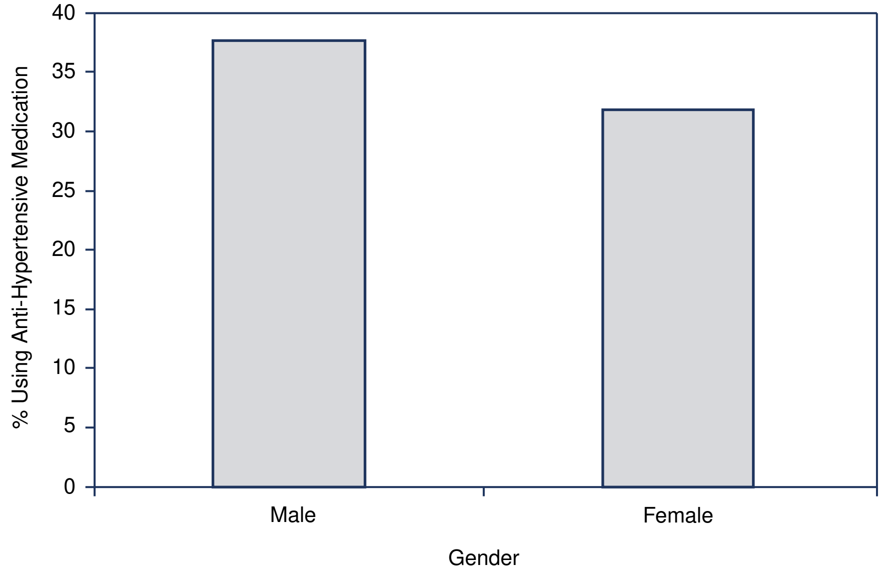
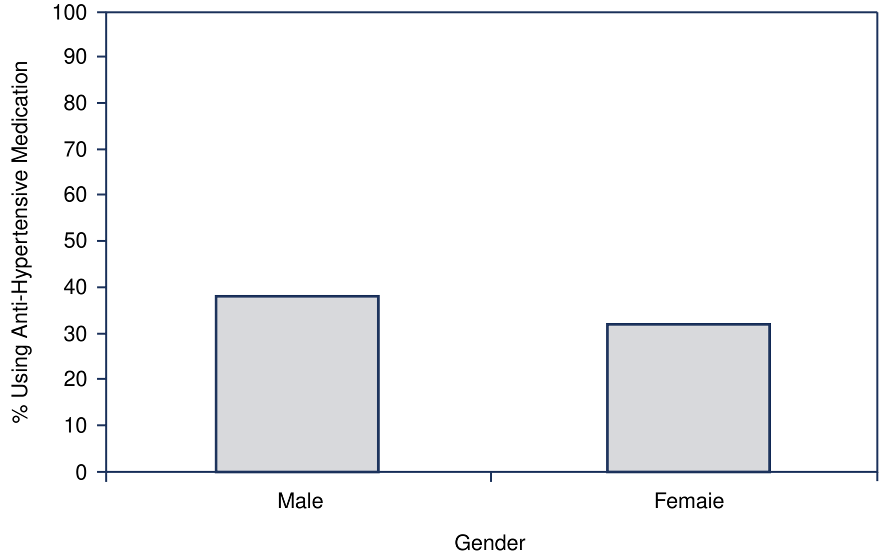


FIGURE 4–4 Relative Frequency Bar Chart of Distribution of Treatment with Anti-Hypertensive Medication by Sex


% Using Anti-Hypertensive Medication
Male
Femaie
Gender