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*.