CHAPTER 1

INTRODUCTION TO BIOSTATISTICS

PRIOR TO the twentieth century, medical research was primarily based on trial and error and empirical evidence. Diseases and the risk factors associated with a disease were not well understood. Drugs and treatments for treating diseases were generally untested. The rapid scientific breakthroughs and technological advances that took place in the latter half of the twentieth century have provided the modern tools and methods that are now being used in the study of the causes of diseases, the development and testing of new drugs and treatments, and the study of human genetics and have been instrumental in eradicating some infectious diseases.

Modern biomedical research is evidence-based research that relies on the scientific method, and in many biomedical studies it is the scientific method that guides the formulation of well-defined research hypotheses, the collection of data through experiments and observation, and the honest analysis whether the observed data support the research hypotheses. When the data in a biomedical study support a research hypothesis, the research hypothesis becomes a theory; however, when data do not support a research hypothesis, new hypotheses are generally developed and tested. Furthermore, because statistics is the science of collecting, analyzing, and interpreting data, statistics plays a very important role in medical research today. In fact, one of the fastest growing areas of statistical research is the development of specialized data collection and analysis methods for biomedical and healthcare data. The science of collecting, analyzing, and interpreting biomedical and healthcare data is called *biostatistics*.

COPYRIGHTED MATERIAL

1.1 WHAT IS BIOSTATISTICS?

Biostatistics is the area of statistics that covers and provides the specialized methodology for collecting and analyzing biomedical and healthcare data. In general, the purpose of using biostatistics is to gather data that can be used to provide honest information about unanswered biomedical questions. In particular, biostatistics is used to differentiate between chance occurrences and possible causal associations, for identifying and estimating the effects of risk factors, for identifying the causes or predispositions related to diseases, for estimating the incidence and prevalence of diseases, for testing and evaluating the efficacy of new drugs or treatments, and for exploring and describing the well being of the general public.

Applied Biostatistics for the Health Sciences. By Richard J. Rossi
Copyright © 2010 by John Wiley & Sons, Inc.

A biostatistician is a scientist trained in statistics who also works in disciplines related to medical research and public health, who designs data collection procedures, analyzes data, interprets data analyses, and helps summarize the results of the studies. Biostatisticians may also develop and apply new statistical methodology required for analyzing biomedical data. Generally, a biostatistician works with a team of medical researchers and is responsible for designing the statistical protocol to be used in a study.

Biostatisticians commonly participate in research in the biomedical fields such as epidemiology, toxicology, nutrition, and genetics, and also often work for pharmaceutical companies. In fact, biostatisticians are widely employed in government agencies such as the National Institutes of Health (NIH), the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), and the Environmental Protection Agency (EPA). Biostatisticians are also employed by pharmaceutical companies, medical research units such as the MAYO Clinic and Fred Hutchison Cancer Research Center, Sloan-Kettering Institute, and many research universities. Furthermore, some biostatisticians serve on the editorial boards of medical journals and many serve as referees for biomedical journal articles in an effort to ensure the quality and integrity of data-based biomedical results that are published.

1.2 POPULATIONS, SAMPLES, AND STATISTICS

In every biomedical study there will be research questions to define the particular population that is being studied. The population that is being studied is called the *target population*. The target population must be a well-defined population so that it is possible to collect representative data that can be used to provide information about the answers to the research questions. Finding the actual answer to a research question requires that the entire target population be observed, which is usually impractical or impossible. Thus, because it is generally impractical to observe the entire target population, biomedical researchers will use only a subset of the population units in their research study. A subset of the population is called a *sample*, and a sample may provide information about the answer to a research question but cannot definitively answer the question itself. That is, complete information on the target population is required to answer the research question, and because a sample is only a subset of the target population, it can only provide information about the answer. For this reason, statistics is often referred to as “the science of describing populations in the presence of uncertainty.”

The first thing a biostatistician generally must do is to take the research question and determine a particular set of characteristics of the target population that are related to the research question being studied. A biostatistician then must determine the relevant statistical questions about these population characteristics that will provide answers or the best information about the research questions. A characteristic of the target population that can be summarized numerically is called a *parameter*. For example, in a study of the body mass index (BMI) of teenagers, the average BMI value for the target population is a parameter, as is the percentage of teenagers having a BMI value less than 25. The parameters of the target population are based on the information about the entire population, and hence, their values will be unknown to the researcher.

To have a meaningful statistical analysis, a researcher must have well-defined research questions, a well-defined target population, a well-designed sampling plan, and an observed sample that is representative of the target population. When the sample is representative of the target population, the resulting statistical analysis will provide useful information about the research questions; however, when the observed sample is not representative of the target population the resulting statistical analysis will often lead to misleading or incorrect inferences being drawn about the target population, and hence, about the research questions, also. Thus, one of the goals of a biostatistician is to obtain a sample that is representative of the target population for estimating or testing the unknown parameters.

Once a representative sample is obtained, any quantity computed from the information in the sample and known values is called *statistic*. Thus, because any estimate of the unknown parameters will be based only on the information in the sample, the estimates are also statistics. Statements made by extrapolating from the sample information (i.e., statistics) about the parameters of the population are called *statistical inferences*, and good statistical inferences will be based on sound statistical and scientific reasoning. Thus, the statistical methods used by a biostatistician for making inferences need to be based on sound statistical and scientific reasoning. Furthermore, statistical inferences are meaningful only when they are based on data that are truly representative of the target population. Statistics that are computed from a sample are often used for estimating the unknown values of the parameters of interest, for testing claims about the unknown parameters, and for modeling the unknown parameters.

1.2.1 The Basic Biostatistical Terminology

In developing the statistical protocol to be used in a research study, biostatisticians use the following basic terminology:

- The *target population* is the population that is being studied in the research project.
- The *units* of a target population are the objects on which the measurements will be taken. When the units of the population are human beings, they are referred to as *subjects* or *individuals*.
- A *subpopulation* of the target population is a well-defined subset of the population units.
- A *parameter* is a numerical measure of a characteristic of the target population.
- A *sample* is a subset of the target population units. A *census* is sample consisting of the entire set of population units.
- The *sample size* is the number of units observed in the sample.
- A *random sample* is a sample that is chosen according to a sampling plan where the probability of each possible sample that can be drawn from the target population is known.
- A *statistic* is any value that is computed using only the sample observations and known values.
- A *cohort* is a group of subjects having similar characteristics.
- A *variable* is a characteristic that will be recorded or measured on a unit in the target population.
- A *response variable* or *outcome variable* is the variable in a research study that is of primary interest or the variable that is being modeled. The response variable is also sometimes called the *dependent variable*.
- An *explanatory variable* is a variable that is used to explain or is believed to cause changes in the response variable. The explanatory variables are also called *independent variables* or *predictor variables*.

- A *treatment* is any experimental condition that is applied to the units.
- A *placebo* is an inert or inactive treatment that is applied to the units.
- A *statistical inference* is an estimate, conclusion, or generalization made about the target population from the information contained in an observed sample.
- A *statistical model* is a mathematical formula that relates the response variable to the explanatory variables.

One of the most misunderstood and abused concepts in statistics is the difference between a parameter and a statistic, and researchers who do not have a basic understanding of statistics often use these terms interchangeably, which is incorrect. Whether a number is a parameter or a statistic is determined by asking whether or not the number was computed from the entire set of units in the target population (parameter) or from a sample of the units in the target population (statistic). It is important to distinguish whether a number is a parameter or a statistic because a parameter will provide the answer to a statistical research question, while a statistic can provide information only regarding the answer, and there is a degree of uncertainty associated with the information contained in a statistic.

Example 1.1

In a study designed to determine the percentage of obese adults in the United States, the BMI of 500 adults was measured at several hospitals across the country. The resulting percentage of the 500 adults classified as obese was 24%.

In this study, the target population was adults in the United States, 500 adults constitute a sample of the adults in the United States, the parameter of interest is the percentage of obese adults in the United States, and 24% is a statistic since it was computed from the sample, not the target population.

In designing a biomedical research study, the statistical protocol used in the study is usually determined by the research team in conjunction with the biostatistician. The statistical protocol should include the identification of the target population, the units in the population, the response variable and explanatory variables, the parameters of interest, the treatments or subpopulations being studied, the sample size, and models that will be fit to the observed data.

Example 1.2

In a study investigating the average survival time for stage IV melanoma patients receiving two different doses of interferon, $n = 150$ patients will be monitored. The age, sex, race, and tumor thickness of each patient will be recorded along with the time they survived after being diagnosed with stage IV melanoma. For this study, determine the following components of the statistical protocol:

a. The target population.

b. The units of target population.

c. The response variable.

d. The explanatory variables.

e. The parameter of interest.

f. The treatments.

g. The sample size.

Solutions

a. The target population in this study is individuals diagnosed with stage IV melanoma.

b. Units of the target population are the individuals diagnosed with stage IV melanoma.

c. The response variable in this study is the survival time after diagnosis with stage IV melanoma.

d. Explanatory variables in this study are age, sex, race, and tumor thickness.

e. The parameter of interest in this study is the average survival time after diagnosis with stage IV melanoma.

f. Treatments are the two different doses of interferon.

g. The sample size is $n = 150$.

1.2.2 Biomedical Studies

There are many different research protocols that are used in biomedical studies. Some protocols are forward looking studying what will happen in the future, some look at what has already occurred, and some are based on a cohort of subjects having similar characteristics. For example, the Framingham Heart Study is a large study conducted by the National Heart, Lung, and Blood Institute (NHLBI) that began in 1948 and continues today. The original goal of the Framingham Heart Study was to study the general causes of heart disease and stroke, and the three cohorts that have or are currently being studied in the Framingham Heart Study are

1. the original cohort that consists of a group of 5209 men and women between the ages of 30 and 62 recruited from Framingham, Massachusetts.
2. The second cohort, called the Offspring Cohort, consists of 5124 of the original participants' adult children and their spouses.
3. the third cohort that consists of children of the Offspring Cohort. The third cohort is recruited with a planned target study size of 3500 grandchildren from members of the original cohort.

Two other large ongoing biomedical studies are the Women's Health Initiative (WHI), which is a research study focusing on the health of women, and the National Health and Nutrition Examination Survey (NHANES), which is designed to assess the health and nutritional status of adults and children in the United States.

Several of the commonly used biomedical research protocols are described below.

- A *cohort study* is a research study carried out on a cohort of subjects. Cohort studies often involve studying the patients over a specified time period.
- A *prospective study* is a research study where the subjects are enrolled in the study and then followed forward over a period of time. In a prospective study, the outcome of interest has not yet occurred when the subjects are enrolled in the study.
- A *retrospective study* is a research study that looks backward in time. In a retrospective study, the outcome of interest has already occurred when the subjects are enrolled in the study.
- A *case–control study* is a research study in which subjects having a certain disease (cases) are compared with subjects who do not have the disease (controls).
- A *longitudinal study* is a research study where the same subjects are observed over an extended period of time.
- A *cross-sectional study* is a study to investigate the relationship between a response variable and the explanatory variables in a target population at a particular point in time.

- A *blinded study* is a research study where the subjects in the study are not told which treatment they are receiving. A research study is a *double-blind study* when neither the subject nor the staff administering the treatment know which treatment a subject is receiving.
- A *clinical trial* is a research study performed on humans and designed to evaluate a new treatment or drug or to investigate a specific health condition.
- A *randomized controlled study* is a research study in which the subjects are randomly assigned to the treatments with one of the treatments being a control treatment; a control treatment may be a standard treatment, a placebo, or no treatment at all.

It is important to note that a research study may actually involve more than one of these protocols. For example, a longitudinal study is often a cohort study, a case–control study is a retrospective study, a longitudinal study is a prospective study, and a clinical trial may be run as a double-blind randomized controlled study. Also, the nature of a particular research study will dictate the research protocol that is used. Finally, of all of the study protocols, the randomized controlled study is the gold standard in biomedical research because it provides more control over the external factors that can bias the results of a study.

Most of the medical journals that publish biomedical research require the authors of an article to describe the research protocol that was used in their study. In fact, during the peer-review process a journal article undergoes prior to publication, the research protocol will be carefully scrutinized and research based on poor research protocols will not be published. Several examples of the different research protocols used in published biomedical research articles are given in Examples 1.3–1.7.

Example 1.3

In the article "A prospective study of coffee consumption and the risk of symptomatic gallstone disease in men" published in the *Journal of the American Medical Association* (Leitzmann et al., 1999), the authors reported the results of a prospective cohort study designed to investigate whether coffee consumption helps prevent symptomatic gallstone disease. This study consisted of $n = 46,008$ men, aged 40–75 years in 1986, without any history of gallstone disease, and the subjects were monitored for a 10-year period from 1986 to 1996.

Example 1.4

In the article "Hospitalization before and after gastric bypass surgery" published in the *Journal of the American Medical Association* (Zingmond et al., 2005), the authors reported the results of a retrospective research study designed to investigate the amount of time spent in a hospital 1–3 years after an individual receives a Roux-en-Y gastric bypass (RYGB). This study consisted of $n = 60,077$ patients who underwent RYGB from 1995 to 2004 in California.

Example 1.5

In the article "Pesticides and risk of Parkinson disease: a population-based case–control study" published in the *Archives of Neurology* (Firestone et al., 2005), the authors reported the results of a case–control research study designed to investigate association between occupational and home-based pesticide exposure and idiopathic Parkinson disease. This study consisted of 250 subjects with Parkinson's disease and 388 healthy control subjects.

Example 1.6

In the article "Randomized, double-blind, placebo-controlled trial of 2 dosages of sustained-release bupropion for adolescent smoking cessation" published in the *Archives of Pediatric and Adolescent Medicine* (Muramoto et al., 2007), the authors reported the results of a randomized controlled double-blind research study designed to investigate the efficacy of sustained release of bupropion hydrochloride for adolescent smoking cessation. This study consisted of $n = 312$ subjects recruited through media and community venues from March 1, 1999 to December 31, 2002, who were aged 14–17 years, smoked at least six cigarettes per day, had an exhaled carbon monoxide level of 10 ppm or greater, had at least two previous quit attempts, and had no other current major psychiatric diagnosis.

Example 1.7

In the article "Antidepressant efficacy of the antimuscarinic drug scopolamine: a randomized, placebo-controlled clinical trial" published in *Archives of General Psychiatry* (Furey et al., 2006), the authors reported the results of a double-blind, placebo-controlled, dose finding clinical trial designed to investigate the antidepressant efficacy of scopolamine. This study consisted of $n = 19$ currently depressed outpatients aged 18–50 years with recurrent major depressive disorder or bipolar disorder.

1.2.3 Observational Studies Versus Experiments

When two or more subpopulations or treatments are to be compared in a biomedical research study, one of the most important aspects of the research protocol is whether the researchers can assign the units to the subpopulations or treatment groups that are being compared. When the researchers control the assignment of the units to the different treatments that are being compared, the study is called an *experiment*, and when units come to the researchers already assigned to the subpopulations or treatment groups, the study is called an *observational study*. Thus, in an experiment the researcher has the ability to assign the units to the groups that are being compared, while in an observational study the units come to the researcher already assigned to the groups.

One of the main reasons an observational study is used instead of an experiment in a biomedical research study is that it would be unethical to assign some subjects to a treatment that is known to be harmful and the remaining subjects to a treatment that is not harmful. For example, in a prospective 30-year study of the effects of smoking cigarettes, it would be unethical to assign some subjects to be smokers and others to be non-smokers.

For ethical reasons, observational studies are often used in epidemiological studies designed to investigate the risk factors associated with a disease. Also, a retrospective study is always an observational study because it looks backward in time and the units have already been assigned to the groups being compared. On the other hand, a prospective study and a clinical trial can be run as either experiments or observational studies depending on whether it is possible for the researcher to assign the units to the groups.

Example 1.8

Determine whether it would be possible to perform an experiment in each of the scenarios given below.

a. A nutritionist is interested in comparing several different diets in a prospective study. The treatments that will be compared are 10% fat in the diet, 15% fat in the diet, and 25% fat in the diet.

b. A pediatrician is interested in studying the effects of a mother's use of tobacco on the birth weight of her baby. The two treatments that are to be compared are smoking during pregnancy and not smoking during pregnancy.

c. A medical researcher is studying the efficacy of vitamin C as a preventive measure against the common cold. The two treatments that are being compared are 1000 mg vitamin C and 1000 mg placebo.

Solutions

a. Because the researcher can assign the subjects to each of the three diets in this study, it could be performed as an experiment.

b. Because smoking is known to have harmful effects on a fetus, it would be unethical for a pediatrician to assign some mothers smoke during pregnancy and others to not smoke during pregnancy. This study would have to be performed as an observational study by comparing the weights of babies born to mothers who chose to smoke during pregnancy with babies born to mothers who did not smoke during pregnancy.

c. Because the medical researcher could assign the subjects to these two treatments, it could be performed as an experiment.

An important advantage experiments have over observational studies is that it is possible in an experiment to control for external factors that might cause differences between the units of the target population. By controlling for the external factors in an experiment, it is possible to make the groups of units assigned to different treatments (i.e., treatment groups) as alike as possible before the treatments are applied. Moreover, in a well-designed experiment when the value of an explanatory variable is changed while no other changes take place in the experimental conditions, any differences in the responses are most likely due to the change in the value of this explanatory variable.

On the other hand, it is much harder to control external factors in an observational study because the units come to the researcher already assigned to the treatments, and thus, in an observational study there is no guarantee that the treatment groups were alike before the treatments were assigned to the units. Because experiments can be designed to control external factors, they can be used to establish evidence of causal relationships; an observational study generally cannot provide strong evidence of a causal relationship because uncontrolled external factors cannot be ruled out as the potential cause of the results.

Example 1.9

To study whether echinacea is effective in shortening the duration of the common cold, a random sample of 200 volunteers is taken. The 200 subjects are divided into two groups of size 100. Each group gets a supply of 300 mg pills and is instructed to take a 300 mg pill as soon as they recognize cold symptoms and to continue taking a pill each day until their cold symptoms go away. One group will receive 300 mg echinacea pills and the other group 300 mg placebo pills. The subjects are asked to record the duration of each cold they have in the following year.

a. Is this study an experiment or an observational study?

b. What is the target population in this study?

c. What is the response variable in this study?

d. What are the treatments in this study?

e. Is this a prospective or retrospective study?

Solutions

a. This study is an experiment because the researcher assigned the subjects to the treatments.

b. The target population in this study is people having the common cold.

c. The response variable in this study is the duration of the common cold.

d. The treatments in this study are 300 mg echinacea and 300 mg placebo.

e. This is a prospective study because the subjects are being followed forward in time.

Example 1.10

To study whether or not there is a relationship between childhood obesity and parental obesity in the United States, a random sample of 500 families was selected. The parents and children in each family were then classified as normal weight, overweight, or obese. The goal was to compare the percentage in each of the weight classifications of the children with normal weight parents against the percentages of the children with overweight and obese parents.

- **a.** Is this study an experiment or an observational study?
- **b.** What is the target population in this study?
- **c.** What is the response variable in this study?
- **d.** What are the treatments in this study?

Solutions

- **a.** This study is an observational study because the subjects came to the researcher assigned to their respective weight classifications.
- **b.** The target population in this study is parents with children living in the United States.
- **c.** The response variable in this study is weight classification of a child. The weight classification of the parent is an explanatory variable.
- **d.** The treatments in this study consist of weight classifications of the parents (normal, overweight, or obese).

1.3 CLINICAL TRIALS

Clinical trials are generally associated with biomedical research studies that are carried out on people for testing how well a new medical approach works, for testing the efficacy and safety of new drugs, for evaluating new biomedical procedures or technological advances, and for diagnosing, treating, managing, or preventing a disease. In the United States, a clinical trial is often highly regulated to ensure that it follows a well-designed research protocol that is ethical and preserves the safety of the participants.

For example, in the development of a new drug, a pharmaceutical company often begins by testing the drug on human cells and animals in a laboratory setting. If the initial laboratory research indicates that the drug may be beneficial to humans, the next step is to submit a new drug application (NDA) to the FDA. The NDA will contain information on the drug, the results of all prior test data on the drug, and descriptions of the manufacturing process used to make the drug. The FDA will then determine whether the drug is safe and effective for its proposed use(s), whether the benefits of the drug outweigh its risks, whether the drug's proposed labeling is appropriate, and, if not, what the drug's appropriate labeling is, and whether the methods used in manufacturing the drug and the controls used to maintain the drug's quality are adequate to preserve the drug's identity, strength, quality, and purity. Supervised clinical trials represent the final testing ground for a new drug, and the results of the clinical trials will be used in the final approval or disapproval of a new drug.

1.3.1 Safety and Ethical Considerations in a Clinical Trial

Every well-designed clinical trial will have a predetermined research protocol that outlines exactly how the clinical trial will be conducted. The clinical trial protocol will describe what will be done in the trial, the rules for determining who can participate, the specific research questions being investigated, the schedule of tests, procedures, medications, and dosages used in the trial, and the length of the trial. During the clinical trial, the participants are closely monitored by the research staff to determine the safety and effectiveness of their treatment. In fact, the ethical treatment and safety of the participants are carefully controlled in clinical trials performed in the United States.

In general, a clinical trial run in the United States must be preapproved by an independent committee of physicians, biostatisticians, and members of the community, which makes sure that the risks to the participants in the study are small and are worth the potential benefits of the new drug or treatment. Many, if not most, externally funded or university-based clinical trials must be reviewed and approved by an Institutional Review Board (IRB) associated with the funding agency. The IRB has the power to decide how often to review the clinical trial, and once started whether the clinical trial should continue as initially planned or modifications need to be made to the research protocol. Furthermore, the IRB may end a clinical trial when a researcher is not following the prescribed protocol, the trial is unsafe, or there is clear and strong evidence that the new drug or treatment is effective.

1.3.2 Types of Clinical Trials

Clinical trials can generally be classified as one of the following types of trials:

- Treatment trials that are clinical trials designed to test experimental treatments, new drugs, or new medical approaches or technology.
- Prevention trials that are clinical trials designed to investigate ways to prevent diseases or prevent the recurrence of a disease.
- Screening trials that are clinical trials designed to determine the best way to detect certain diseases or health conditions early on.
- Diagnostic trials that are clinical trials designed to determine tests or procedures that can be used for diagnosing a particular disease or condition.
- Quality-of-life trials that are clinical trials designed to explore ways to improve the comfort and quality of life for individuals with a chronic or terminal disease or condition.
- Genetic trials that are clinical trials designed to investigate the role genetics plays in the detection, diagnosis, or response to a drug or treatment.

Pharmaceutical companies commonly use treatment trials in the development and evaluation of new drugs, epidemiologists generally use prevention, screening, and diagnostic trials in their studies of diseases, public health officials often use quality-of-life trials, and geneticists often use genetic trials for studying tissue or blood samples from families or large groups of people to understand the role of genes in the development of a disease.

The results of a clinical trial are generally published in peer-reviewed scientific or medical journals. The peer-review process is carried out by experts who critically review a research report before it is published. In particular, the peer reviewers are charged with examining the research protocol, analysis, and conclusions drawn in a research report to ensure the integrity and quality of the research that is published. Following the publication of the results of a clinical trial or biomedical research study, further information is generally obtained as new studies are carried out independently by other researchers. The follow-up research is generally designed to validate or expand the previously published results.

1.3.3 The Phases of a Clinical Trial

Clinical research is often conducted in a series of steps, called phases. Because a new drug, medicine, or treatment must be safe, effective, and manufactured at a consistent quality, a series of rigorous clinical trials are usually required before the drug, medicine, or treatment can be made available to the general public. In the United States the FDA regulates and oversees the testing and approval of new drugs as well as dietary supplements, cosmetics, medical devices, blood products, and the content of health claims on food labels. The approval of a new drug by the FDA requires extensive testing and evaluation of the drug through a series of four clinical trials, which are referred to as *phase I*, *II*, *III*, and *IV* trials.

Each of the four phases is designed with a different purpose and to provide the necessary information to help biomedical researchers answer several different questions about a new drug, treatment, or biomedical procedure. After a clinical trial is completed, the researchers use biostatistical methods to analyze the data collected during the trial and make decisions and draw conclusions about the meaning of their findings and whether further studies are needed. After each phase in the study of a new drug or treatment, the research team must decide whether to proceed to the next phase or stop the investigation of the drug/treatment. Formal approval of a new drug or biomedical procedure generally cannot be made until a phase III trial is completed and there is strong evidence that the drug/treatment is safe and effective.

The purpose of a *phase I* clinical trial is to investigate the safety, efficacy, and side effects of a new drug or treatment. Phase I trials usually involve a small number of subjects and take place at a single or only a few different locations. In a drug trial, the goal of a phase I trial is often to investigate the metabolic and pharmacologic actions of the drug, the efficacy of the drug, and the side effects associated with different dosages of the drug. Phase I drug trials are also referred to as *dose finding trials*.

When the results of a phase I trial suggest that a treatment or drug appears to have promise, the treatment or drug is generally next studied in a *phase II* trial. In phase II clinical trials, the drug or treatment being studied is evaluated on a larger group of subjects to further investigate its effectiveness and safety. In general, the goal of a phase II trial is to study the feasibility and level of activity of the drug or treatment. Thus, phase II trials are designed to provide more information about the effective dosage of a drug, the severity of the side effects, and how to manage the side effects. Phase II trials are also referred to as *safety and efficacy trials* and usually involve more subjects than phase I trials.

When the preliminary results of a new drug or treatment from a phase II trial suggest the drug or treatment will be effective and safe, a *phase III* trial is designed to gather additional information that can be used in evaluating the overall benefit–risk relationship of the drug. Phase III trials are usually designed to compare the new drug/treatment with standard or commonly used drugs/treatments, to confirm its effectiveness, to further monitor side effects, and to determine how the new drug or treatment can be safely used. Phase III trials generally are large trials and may enroll subjects at a wide variety of locations. Phase III trials are also referred to as *comparative treatment efficacy trials*.

Finally, when a new drug or treatment has been examined in phase I, II, and III trials and has been approved for the general public, a *phase IV* trial is usually initiated. A phase IV trial is a postmarketing study designed to obtain additional information on the risks associated with the drug/treatment, its benefits, and its optimal use. The primary aim of a phase IV trial is to evaluate the long-term safety and effectiveness of a drug/treatment. Phase IV trials sometimes result in a drug being taken completely off the market or new restrictions being placed on the use of the drug. Phase IV trials are also referred to as *expanded safety trials* and usually involve a very large number of subjects.

Note that the number of subjects in a trial usually increases as the phases of the study progress. That is, a phase I trial usually involves fewer subjects than a phase II trial, a phase II trial usually involves fewer subjects than a phase III trial, and a phase III trial usually involves fewer subjects than a phase IV trial. Also, some research studies involving human subjects will have less than four phases. For example, it is not unusual for screening, prevention, diagnostic, genetic, and quality-of-life studies to be conducted in only phase I or II trials. However, new drugs and biomedical procedures almost always require phase I, II, and III clinical trials for approval and a phase IV trial to track the safety of the drug after its approval. The development of a new drug may take many years to proceed through the first three phases of the approval process, and following approval, the phase IV trial usually extends over a period of many years.

1.4 DATA SET DESCRIPTIONS

Throughout this book several data sets will be used in the examples and exercises. These data sets are given in Appendix B and are also available at http://www.mtech.edu/clsps/math/Faculty/rossi_book.htm as Excel files, text files, and MINITAB worksheets. Permission to use the Birth Weight, Intensive Care Unit, Coronary Heart Disease, UMASS Aids Research Unit, and Prostate Cancer data sets has been granted by John Wiley & Sons, Inc. These data sets were first published in *Applied Logistic Regression* (Hosmer, 2000). Permission to use the Body Fat data set has been provided by Roger W. Johnson, Department of Mathematics & Computer Science, South Dakota School of Mines & Technology and *Journal of Statistics Education*.

1.4.1 Birth Weight Data Set

The Birth Weight data set consists of data collected on 189 women to identify the risk factors associated with the birth of a low birth weight baby. The data set was collected at the Baystate Medical Center in Springfield, Massachusetts. The variables included in this data set are summarized in Table 1.1.

1.4.2 Body Fat Data Set

The Body Fat data set consists of data collected on 252 adult males. The data were originally collected to build a model relating body density and percentage of body fat in adult males to several body measurement variables. These data were originally used in the article "Generalized body composition prediction equation for men using simple measurement techniques," published in *Medicine and Science in Sports and Exercise* (Penrose et al., 1985). The variables included in this data set are summarized in Table 1.2.

1.4.3 Coronary Heart Disease Data Set

The Coronary Heart Disease data set consists of 100 observations on patients who were selected in a study on the relationship between the age and the presence of coronary heart disease. The variables included in this data set are summarized in Table 1.3.

TABLE 1.1 A Description of the Variables in the Birth Weight Data Set

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Low birth weight	1 = BWT≤2500 g 0 = BWT>2500 g	LOW
3	Age of mother	Years	AGE
4	Weight of mother at last menstrual period	Pounds	LWT
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Smoking status during pregnancy	0 = No 1 = Yes	SMOKE
7	History of premature labor	0 = None 1 = One 2 = Two, etc.	PTL
8	History of hypertension	0 = No 1 = Yes	HT
9	Presence of uterine irritability	0 = No 1 = Yes	UI
10	Number of physician visits during the first trimester	0 = None 1 = One 2 = Two, etc.	FTV
11	Birth weight	Grams	BWT

1.4.4 Prostate Cancer Study Data Set

The Prostate Cancer Study data set consists of 380 patients in a study to determine whether the variables measured at a baseline medical examination can be used to predict whether the prostatic tumor has penetrated a prostatic capsule. The data were collected by Dr. Donn

TABLE 1.2 A Description of the Variables in the Body Fat Data Set

Variable	Description	Codes/Values	Name
1	Density determined from underwater weighing		Density
2	Percent body fat from Siri's (1956) equation	Percent	PCTBF
3	Age	Years	Age
4	Weight	Pounds	Weight
5	Height	Inches	Height
6	Neck circumference	Centimeters	Neck
7	Chest circumference	Centimeters	Chest
8	Abdomen circumference	Centimeters	Abdomen
9	Hip circumference	Centimeters	Hip
10	Thigh circumference	Centimeters	Thigh
11	Knee circumference	Centimeters	Knee
12	Ankle circumference	Centimeters	Ankle
13	Biceps extended circumference	Centimeters	Biceps
14	Forearm circumference	Centimeters	Forearm
15	Wrist circumference	Centimeters	Wrist

TABLE 1.3 A Description of the Variables in the Coronary Heart Disease Data Set

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Age in years	Years	Age
3	Coronary heart disease	0 = Absent 1 = Present	CHD

Young at the Ohio State University Comprehensive Cancer Center and the data have been modified to protect subject confidentiality. Variables included in this data set are summarized in Table 1.4.

1.4.5 Intensive Care Unit Data Set

The Intensive Care Unit data set consists of 200 observations on subjects involved in a study on the survival of patients following admission to an adult intensive care unit (ICU). The data set was collected at the Baystate Medical Center in Springfield, Massachusetts, and the variables included in this data set are summarized in Table 1.5.

1.4.6 Mammography Experience Study Data Set

The Mammography Experience Study data set consists of 412 observations on subjects from a study designed to investigate the factors associated with a woman's knowledge, attitude, and behavior toward mammography exams. The data were collected by Dr. J. Zapka and Ms. D. Spotts of the University of Massachusetts, Division of Public Health. The variables included in this data set are summarized in Table 1.6.

TABLE 1.4 A Description of the Variables in the Prostate Cancer Study Data Set

Variable	Description	Codes/values	Name
1	Identification code	ID number	ID
2	Tumor penetration of prostatic capsule	0 = No penetration 1 = Penetration	CAPSULE
3	Age	Years	AGE
4	Race	1= White 2 = Black	RACE
5	Results of the digital rectal exam	1 = No nodule 2 = Unilobar nodule (left) 3 = Unilobar nodule (right) 4 = Bilobar nodule	DPROS
6	Detection of capsular involvement in rectal exam	1 = No 2 = Yes	DCAPS
7	Prostatic-specific antigen value	mg/ml	PSA
8	Tumor volume obtained from ultrasound	cm^3	VOL
9	Total Gleason score	0–10	GLEASON

TABLE 1.5 A Description of the Variables in the Intensive Care Unit Data Set

Variable	Description	Codes/values	Name
1	Identification code	ID number	ID
2	Vital status	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Sex	0 = Male 1 = Female	SEX
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Service at ICU admission	0 = Medical 1 = Surgical	SER
7	Cancer part of present problem	0 = No 1 = Yes	CAN
8	History of chronic renal failure	0 = No 1 = Yes	CRN
9	Infection probable at ICU admission	0 = No 1 = Yes	INF
10	CPR prior to ICU admission	0 = No 1 = Yes	CPR
11	Systolic blood pressure at ICU admission	mmHg	SYS
12	Heart rate at ICU admission	Beats/min	HRA
13	Previous admission to an ICU within 6 months	0 = No 1 = Yes	PRE
14	Type of admission	0 = Elective 1 = Emergency	TYP
15	Long bone, multiple, neck single area, or hip fracture	0 = No 1 = Yes	FRA
16	pO_2 from initial blood gases	0 = > 60 1 = < 60	PO2
17	pH from initial blood gases	0 = > 7.25 1 = < 7.25	PH
18	pCO_2 from initial blood gases	0 = < 45 1 = > 45	PCO
19	Bicarbonate from Initial blood gases	0 = > 18 1 = < 18	BIC
20	Creatinine from initial blood gases	0 = < 2.0 1 = > 2.0	CRE
21	Level of consciousness at ICU admission	0 = No coma or stupor 1 = Deep stupor 2 = Coma	LOC

1.4.7 Benign Breast Disease Study

The Benign Breast Disease Study data set consists of data collected from 200 women in a case–control study designed to investigate the risk factors associated with benign breast disease at two hospitals in New Haven, Connecticut. The variables included in this data set are summarized in Table 1.7.

TABLE 1.6 A Description of the Variables in the Mammography Experience Data Set

Variable	Description	Codes/Values	Name
1	Identification code	ID number	OBS
2	Mammograph experience	0 = Never 1 = Within 1 year 2 = Over 1 year ago	ME
3	"You do not need a mammogram unless you develop symptoms"	1 = Strongly agree 2 = Agree 3 = Disagree 4 = Strongly disagree	SYMPT
4	Perceived benefit of mammography	5–20	PB
5	Mother or sister with a history of breast cancer	0 = No 1 = Yes	HIST
6	"Has anyone taught you how to examine your own breasts: that is BSE"	0 = No 1 = Yes	BSE
7	"How likely is it that a mammogram could find a new case of breast cancer"	1 = Not likely 2 = Somewhat likely 3 = Very likely	DETC

TABLE 1.7 A Description of the Variables in the Benign Breast Disease Study Data Set

Variable	Description	Codes/Values	Name
1	Stratum	1–50	STR
2	Observation within a stratum	1 = Case 2–4 = Control	OBS
3	Age of the subject at the interview	Years	AGMT
4	Final diagnosis	1 = Case 0 = Control	FNDX
5	Highest grade in school		HIGD
6	Degree	0 = None 1 = High school 2 = Jr. college 3 = College 4 = Masters 5 = Doctoral	DEG
7	Regular medical checkups	1 = Yes 2 = No	CHK
8	Age at first pregnancy	Years	AGP1
9	Age at menarche	Years	AGMN
10	No. of stillbirths, miscarriages, etc.		NLV
11	Number of live births		LIV
12	Weight of the subject	Pounds	WT
13	Age at last menstrual period	Years	AGLP
14	Marital status	1 = Married 2 = Divorced 3 = Separated 4 = Widowed 5 = Never married	MST

GLOSSARY

Biostatistics Biostatistics is the science of collecting, an alyzing, and interpreting biomedical and healthcare data.

Blinded Study A blinded study is a research study where the subjects in the study are not told which treatment they are receiving. A research study is a double-blind study when neither the subject nor the staff administering the treatment know which treatment a subject receives.

Case–Control Study A case–control study is a retrospective study in which subjects having a certain disease or condition are compared with subjects who do not have the disease.

Census A census is a sample consisting of the entire set of population units.

Clinical Trial A clinical trial is a research study performed on humans and designed to evaluate a new treatment or drug or to investigate a specific health condition that follows a predefined protocol.

Cohort A cohort is a group of subjects having similar characteristics.

Cross–Sectional Study A cross-sectional study is a study to investigate the relationship between a response variable and the explanatory variables in a target population at a particular point in time.

Experiment An experiment is a study where the researcher controls the assignment of the units to the treatments.

Explanatory Variable An explanatory variable is a variable that is used to explain or is believed to cause changes in the response variable. The explanatory variables are also called independent variables or predictor variables.

Longitudinal Study A longitudinal study is a study where the same subjects are observed over a specific period of time. A longitudinal study could be either a prospective or a retrospective study.

Observational Study An observational study is any study where the units of the study come to the researchers already assigned to the subpopulations or treatment groups.

Parameter A parameter is a numerical measure of a characteristic of the population.

Phase I Clinical Trial A phase I clinical trial is designed for investigating the safety, efficacy, and side effects of a new drug or treatment. Phase I drug trials are also referred to as dose finding trials.

Phase II Clinical Trial A phase II clinical trial follows a phase I trial and is used to further investigate the effectiveness, feasibility, and safety of a drug or treatment. Phase II trials are also referred to as safety and efficacy trials and usually have a larger sample size than a phase I trial.

Phase III Clinical Trial A phase III clinical trial follows a phase II trial and is designed to gather additional information that will be used in evaluating the overall benefit–risk relationship of the drug. Phase III trials are generally large trials and are referred to as comparative treatment efficacy trials.

Phase IV Clinical Trial A phase IV clinical trial is a postmarketing study designed to obtain additional information on the risks associated with the drug/treatment, its benefits, and its optimal use. The primary use of a phase IV trials is to evaluate the long-term safety and effectiveness of a drug/treatment. Phase IV trials are referred to as expanded safety trials and usually involve a large number of subjects.

Population Units The units of a population are the objects on which measurements will be taken. When the units of the population are human beings, they are referred to as subjects or individuals.

Prospective Study A prospective study is a study that monitors the units over a period of time and analyzes what happens to the units in the study.

Randomized Controlled Study A randomized controlled study is a research study where the subjects are randomly assigned to the treatments with one of the treatments being a control treatment; a control treatment may be a standard treatment, a placebo, or no treatment at all.

Response Variable A response variable is an observed variable or outcome variable in an experiment or study that is believed to depend on other variables in the study. The response variable is also called a dependent variable.

Retrospective Study A retrospective study is a study that looks backward in time and analyzes what has happened to the units in the study.

Sample A sample is a subset of the population units. A random sample is a sample that is chosen according to a sampling plan where the probability of each possible sample that can be drawn from the target population is known, where the probability of sampling each unit in the population is known.

Statistic A statistic is any value that is computed from only the sample observations and known values.

Statistical Inferences Statistical inferences are estimates, conclusions, or generalizations made about the target population from the information contained in an observed sample.

Statistical Model A statistical model is a mathematical formula that relates the response variable to the explanatory variables.

Target Population The target population is the population of units that is being studied.

Treatment A treatment is any experimental condition that is applied to the units. A placebo treatment is an inert or inactive treatment that is applied to the units.

Variable A variable is a characteristic that will be recorded or measured on a unit in the target population.

EXERCISES

1.1 What is biostatistics?

1.2 What does a biostatistician do?

1.3 What are three federal agencies that employ biostatisticians?

1.4 What is a
(a) target population? **(b)** sample?
(c) census?

1.5 How is the target population different from a sample?

1.6 What is a
(a) parameter? **(b)** statistic?

1.7 How is a statistic different from a parameter?

1.8 How can the value of an unknown parameter be
(a) found exactly? **(b)** estimated?

1.9 What is a numerical value that is computed from only the information contained in a sample called?

1.10 What is a numerical value that is computed from a census called?

1.11 What is a statistical inference?

1.12 What is a
(a) random sample? **(b)** cohort?
(c) variable? **(d)** treatment?
(e) placebo? **(f)** statistical model?

1.13 What is the difference between a response variable and an explanatory variable.

1.14 In a study designed to determine the percentage of understaffed hospitals in the United States, 250 of roughly 7500 hundred hospitals in the United States were surveyed. The resulting percentage of the understaffed hospitals was 41%.
(a) What is the target population in this study?
(b) Is 41% a statistic or a parameter? Explain.

1.15 In a study designed to determine the percentage of doctors belonging to the American Medical Association (AMA) who perform pro bono work, the AMA found from a census of its membership that 63% performed pro bono work. The AMA also found from a sample of 1000 members that the average number of pro bono hours worked by a doctor in a year was 223.
(a) What is the target population in this study?
(b) Is 63% a statistic or a parameter? Explain.
(c) Is 223 a statistic or a parameter? Explain.

1.16 In a Red Cross sponsored laboratory study designed to investigate the average length of time blood can be stored safely in a blood bank in the United States, 20 freshly sampled 500 ml blood bags were monitored over a 6-month period. The results of this laboratory study showed that blood may be stored safely on average for 17 days.
(a) What is the target population in this study?
(b) Is 17 a statistic or a parameter? Explain.

1.17 In a study designed to investigate the effects of omega-3 fatty acids for lowering the risk of Alzheimer's disease, 300 participants aged 50 and older with mild to moderate Alzheimer's disease were selected for the study. The 300 participants were randomly assigned to two groups with one group receiving placebo pills and the other group receiving omega-3 supplement pills. Doctors and nurses will monitor the participants throughout the study, and neither the researchers conducting the trial nor the participants will know who is getting the omega-3 pills or the placebo pills.
(a) What is the target population in this study?.
(b) What are units of target population?
(c) What are the treatments being used in this study?
(d) What is the sample size in this study?
(e) Is this a blinded study? Explain.
(f) Is this a randomized controlled study? Explain.

1.18 To study whether vitamin C is effective in shortening the duration of a common cold, a random sample of 400 volunteers is taken. The 400 subjects are divided into two groups of size 200. Each group gets a supply of 1000 mg pills and is instructed to take a pill as soon as they recognize cold symptoms and to continue taking a pill each day until their cold symptoms go away. One group will receive 1000 mg vitamin C pills and the other group 1000 mg placebo pills. The subjects are asked to record the duration of each cold they have in the following year.

(a) What is the target population in this study?

(b) What is the response variable in this study?

(c) What are the treatments in this study?

(d) Could this be run as a double-blind study? Explain.

1.19 In a study designed to investigate the efficacy of three treatments for prostate cancer, $n = 1088$ patients were selected from Los Angeles public hospitals for the study. The treatments studied were surgery, hormone therapy, and radiation therapy. The goal of the study is to estimate the percentage of prostate cancer patients surviving at least 5 years for each of the treatments. The variables age, sex, and race were also recorded for each patient since they are believed to influence the survival time of a prostate cancer patient. Determine

(a) the target population in this study?.

(b) the unit of the target population?

(c) the parameters of interest in this study?

(d) the explanatory variables used in this study?

(e) the treatments being studied?

1.20 What is a

(a) retrospective study?

(b) prospective study?

(c) longitudinal study?

(d) case–control study?

(e) cross-sectional study?

(f) clinical trial?

(g) blinded study?

(h) double-blind study?

1.21 How do prospective and retrospective studies differ?

1.22 What is a randomized controlled study?

1.23 Use the Internet to find an article published in a biomedical journal where a prospective study was used and identify the target population, the units of the population, the response variable, the treatments used in the study, and the explanatory variables measured in the study.

1.24 Use the Internet to find an article published in a biomedical journal where a retrospective study was used and identify the target population, the units of the population, the response variable, the treatments used in the case–control study, and the explanatory variables measured in the study.

1.25 Use the Internet to find an article published in a biomedical journal where a cohort study was used and identify the target population, the units of the population, the response variable, the treatments used in the study, and the explanatory variables measured in the study.

1.26 Use the Internet to find an article published in a biomedical journal where a double-blind study was used and identify the target population, the units of the population, the response variable, the treatments used in the study, and the explanatory variables measured in the study.

1.27 Use the Internet to find an article published in a biomedical journal where a cross-sectional study was used and identify the target population, the units of the population, the response variable, the treatments used in the study, and the explanatory variables measured in the study.

1.28 Use the Internet to find an article published in a biomedical journal where a prospective case-control study was used and identify the target population, the units of the population, the response variable, the treatments used in the study, and the explanatory variables measured in the study.

1.29 What is an

(a) experiment? **(b)** observational study?

1.30 How is an experiment different from an observational study?

1.31 Why is an experiment preferred over an observation study?

1.32 Why are retrospective studies and case–control studies observational studies?

1.33 Explain why it would or would not be ethical to perform an experiment in each of the scenarios below.

(a) A researcher is interested in the effects of smoking cigarettes on human health. The researcher would like to assign subjects to the treatments smoke cigarettes for 25 years and do not smoke at all in a 25-year prospective study.

(b) A researcher is interested in determining the efficacy of a new HIV/AIDS drug. The researcher would like to assign subjects to standard treatment and the new drug in a randomized controlled study.

(c) A researcher is interested in identifying the risk factors associated with Alzheimer's disease. The researcher would like to assign subjects to several different risk factors in a 20-year prospective study.

(d) A researcher is interested in identifying the relationship between hormone therapy and breast cancer. The researcher would like to assign subjects to the treatments hormone therapy, time-reduced hormone therapy, and no hormone therapy at all in a 20-year prospective study.

1.34 What are the six different types of clinical trials that were discussed earlier in this chapter?

1.35 What is a

(a) prevention trial? **(b)** quality-of-life trial?
(c) screening trial? **(d)** treatment trial?

1.36 What are the four phases of clinical trials?

1.37 What is a

(a) dose finding trial?
(b) safety and efficacy trial?
(c) comparative treatment efficacy trial?
(d) expanded safety trial?

1.38 What phases must a new drug, treatment, or biomedical procedure go through to receive approval for widespread use in the United States?

1.39 What is the purpose of running a phase IV trial after a new drug, treatment, or biomedical procedure has been approved?

1.40 What reasons might be used for prematurely stopping a clinical trial?

1.41 Are all research studies based on clinical trials required to be studied in all four phases? Explain.

1.42 Use the Internet to find

(a) the FDA regulations for the approval of a new drug.

(b) the regulations used in the United Kingdom for the approval of a new drug.

(c) the regulations used in Japan for the approval of a new drug.

(d) two drugs that have been taken off the market for safety reasons after their approval.

(e) out what the CDER agency does with regard to drugs developed and marketed in the United States.