CHAPTER 2

Biostatistical Design of Medical Studies

2.1 INTRODUCTION

In this chapter we introduce some of the principles of biostatistical design. Many of the ideas are expanded in later chapters. This chapter also serves as a reminder that statistics is not an end in itself but a tool to be used in investigating the world around us. The study of statistics should serve to develop critical, analytical thought and common sense as well as to introduce specific tools and methods of processing data.

2.2 PROBLEMS TO BE INVESTIGATED

Biomedical studies arise in many ways. A particular study may result from a sequence of experiments, each one leading naturally to the next. The study may be triggered by observation of an interesting case, or observation of a mold (e.g., penicillin in a petri dish). The study may be instigated by a governmental agency in response to a question of national importance. The basic ideas of the study may be defined by an advisory panel. Many of the critical studies and experiments in biomedical science have come from one person with an idea for a radical interpretation of past data.

Formulation of the problem to be studied lies outside the realm of statistics per se. Statistical considerations may suggest that an experiment is too expensive to conduct, or may suggest an approach that differs from that planned. The need to evaluate data from a study statistically forces an investigator to sharpen the focus of the study. It makes one translate intuitive ideas into an analytical model capable of generating data that may be evaluated statistically.

To answer a given scientific question, many different studies may be considered. Possible studies may range from small laboratory experiments, to large and expensive experiments involving humans, to observational studies. It is worth spending a considerable amount of time thinking about alternatives. In most cases your first idea for a study will not be your best—unless it is your only idea.

In laboratory research, many different experiments may shed light on a given hypothesis or question. Sometimes, less-than-optimal execution of a well-conceived experiment sheds more light than arduous and excellent experimentation unimaginatively designed. One mark of a good scientist is that he or she attacks important problems in a clever manner.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

2.3 VARIOUS TYPES OF STUDIES

A problem may be investigated in a variety of ways. To decide on your method of approach, it is necessary to understand the types of studies that might be done. To facilitate the discussion of design, we introduce definitions of commonly used types of studies.

Definition 2.1. An *observational study* collects data from an existing situation. The data collection does not intentionally interfere with the running of the system.

There are subtleties associated with observational studies. The act of observation may introduce change into a system. For example, if physicians know that their behavior is being monitored and charted for study purposes, they may tend to adhere more strictly to procedures than would be the case otherwise. Pathologists performing autopsies guided by a study form may invariably look for a certain finding not routinely sought. The act of sending out questionnaires about health care may sensitize people to the need for health care; this might result in more demand. Asking constantly about a person's health can introduce hypochondria.

A side effect introduced by the act of observation is the *Hawthorne effect*, named after a famous experiment carried out at the Hawthorne works of the Western Electric Company. Employees were engaged in the production of electrical relays. The study was designed to investigate the effect of better working conditions, including increased pay, shorter hours, better lighting and ventilation, and pauses for rest and refreshment. All were introduced, with "resulting" increased output. As a control, working conditions were returned to original conditions. Production continued to rise! The investigators concluded that increased morale due to the attention and resulting *esprit de corps* among workers resulted in better production. Humans and animals are not machines or passive experimental units [Roethlisberger, 1941].

Definition 2.2. An *experiment* is a study in which an investigator deliberately sets one or more factors to a specific level.

Experiments lead to stronger scientific inferences than do observational studies. The "cleanest" experiments exist in the physical sciences; nevertheless, in the biological sciences, particularly with the use of randomization (a topic discussed below), strong scientific inferences can be obtained. Experiments are superior to observational studies in part because in an observational study one may not be observing one or more variables that are of crucial importance to interpreting the observations. Observational studies are always open to misinterpretation due to a lack of knowledge in a given field. In an experiment, by seeing the change that results when a factor is varied, the causal inference is much stronger.

Definition 2.3. A *laboratory experiment* is an experiment that takes place in an environment (called a *laboratory*) where experimental manipulation is facilitated.

Although this definition is loose, the connotation of the term *laboratory experiment* is that the experiment is run under conditions where most of the variables of interest can be controlled very closely (e.g., temperature, air quality). In laboratory experiments involving animals, the aim is that animals be treated in the same manner in all respects except with regard to the factors varied by the investigator.

Definition 2.4. A *comparative experiment* is an experiment that compares two or more techniques, treatments, or levels of a variable.

There are many examples of comparative experiments in biomedical areas. For example, it is common in nutrition to compare laboratory animals on different diets. There are many

experiments comparing different drugs. Experiments may compare the effect of a given treatment with that of no treatment. (From a strictly logical point of view, "no treatment" is in itself a type of treatment.) There are also comparative observational studies. In a comparative *study* one might, for example, observe women using and women not using birth control pills and examine the incidence of complications such as thrombophlebitis. The women themselves would decide whether or not to use birth control pills. The user and nonuser groups would probably differ in a great many other ways. In a comparative *experiment*, one might have women selected by chance to receive birth control pills, with the control group using some other method.

Definition 2.5. An *experimental unit* or *study unit* is the smallest unit on which an experiment or study is performed.

In a clinical study, the experimental units are usually humans. (In other cases, it may be an eye; for example, one eye may receive treatment, the other being a control.) In animal experiments, the experimental unit is usually an animal. With a study on teaching, the experimental unit may be a class—as the teaching method will usually be given to an entire class. Study units are the object of consideration when one discusses sample size.

Definition 2.6. An experiment is a *crossover experiment* if the same experimental unit receives more than one treatment or is investigated under more than one condition of the experiment. The different treatments are given during nonoverlapping time periods.

An example of a crossover experiment is one in which laboratory animals are treated sequentially with more than one drug and blood levels of certain metabolites are measured for each drug. A major benefit of a crossover experiment is that each experimental unit serves as its own control (the term *control* is explained in more detail below), eliminating subject-to-subject variability in response to the treatment or experimental conditions being considered. Major disadvantages of a crossover experiment are that (1) there may be a carryover effect of the first treatment continuing into the next treatment period; (2) the experimental unit may change over time; (3) in animal or human experiments, the treatment introduces permanent physiological changes; (4) the experiment may take longer so that investigator and subject enthusiasm wanes; and (5) the chance of dropping out increases.

Definition 2.7. A *clinical study* is one that takes place in the setting of clinical medicine.

A study that takes place in an organizational unit dispensing health care—such as a hospital, psychiatric clinic, well-child clinic, or group practice clinic—is a clinical study.

We now turn to the concepts of prospective studies and retrospective studies, usually involving human populations.

Definition 2.8. A *cohort* of people is a group of people whose membership is clearly defined.

Examples of cohorts are all persons enrolling in the Graduate School at the University of Washington for the fall quarter of 2003; all females between the ages of 30 and 35 (as of a certain date) whose residence is within the New York City limits; all smokers in the United States as of January 1, 1953, where a person is defined to be a smoker if he or she smoked one or more cigarettes during the preceding calendar year. Often, cohorts are followed over some time interval.

Definition 2.9. An *endpoint* is a clearly defined outcome or event associated with an experimental or study unit.

An endpoint may be the presence of a particular disease or five-year survival after, say, a radical mastectomy. An important characteristic of an endpoint is that it can be clearly defined and observed.

Definition 2.10. A *prospective study* is one in which a cohort of people is followed for the occurrence or nonoccurrence of specified endpoints or events or measurements.

In the analysis of data from a prospective study, the occurrence of the endpoints is often related to characteristics of the cohort measured at the beginning of the study.

Definition 2.11. *Baseline characteristics* or *baseline variables* are values collected at the time of entry into the study.

The Salk polio vaccine trial is an example of a prospective study, in fact, a prospective experiment. On occasion, you may be able to conduct a prospective study from existing data; that is, some unit of government or other agency may have collected data for other purposes, which allows you to analyze the data as a prospective study. In other words, there is a well-defined cohort for which records have already been collected (for some other purpose) which can be used for your study. Such studies are sometimes called *historical prospective studies*.

One drawback associated with prospective studies is that the endpoint of interest may occur infrequently. In this case, extremely large numbers of people need to be followed in order that the study will have enough endpoints for statistical analysis. As discussed below, other designs, help get around this problem.

Definition 2.12. A *retrospective study* is one in which people having a particular outcome or endpoint are identified and studied.

These subjects are usually compared to others without the endpoint. The groups are compared to see whether the people with the given endpoint have a higher fraction with one or more of the factors that are conjectured to increase the risk of endpoints.

Subjects with particular characteristics of interest are often collected into registries. Such a registry usually covers a well-defined population. In Sweden, for example, there is a twin registry. In the United States there are cancer registries, often defined for a specified metropolitan area. Registries can be used for retrospective as well as prospective studies. A cancer registry can be used retrospectively to compare the presence or absence of possible causal factors of cancer after generating appropriate controls—either randomly from the same population or by some matching mechanism. Alternatively, a cancer registry can be used prospectively by comparing survival times of cancer patients having various therapies.

One way of avoiding the large sample sizes needed to collect enough cases prospectively is to use the case–control study, discussed in Chapter 1.

Definition 2.13. A *case–control study* selects all cases, usually of a disease, that meet fixed criteria. A group, called *controls*, that serve as a comparison for the cases is also selected. The cases and controls are compared with respect to various characteristics.

Controls are sometimes selected to match the individual case; in other situations, an entire group of controls is selected for comparison with an entire group of cases.

Definition 2.14. In a *matched case–control study*, controls are selected to match characteristics of individual cases. The cases and control(s) are associated with each other. There may be more than one control for each case.

Definition 2.15. In a *frequency-matched case–control study*, controls are selected to match characteristics of the entire case sample (e.g., age, gender, year of event). The cases and controls are not otherwise associated. There may be more than one control for each case.

Suppose that we want to study characteristics of cases of a disease. One way to do this would be to identify new cases appearing during some time interval. A second possibility would be to identify all known cases at some fixed time. The first approach is *longitudinal*; the second approach is *cross-sectional*.

Definition 2.16. A *longitudinal study* collects information on study units over a specified time period. A *cross-sectional study* collects data on study units at a fixed time.

Figure 2.1 illustrates the difference. The longitudinal study might collect information on the six new cases appearing over the interval specified. The cross-sectional study would identify the nine cases available at the fixed time point. The cross-sectional study will have proportionately more cases with a long duration. (Why?) For completeness, we repeat the definitions given informally in Chapter 1.

Definition 2.17. A *placebo treatment* is designed to appear exactly like a comparison treatment but to be devoid of the active part of the treatment.

Definition 2.18. The *placebo effect* results from the belief that one has been treated rather than having experienced actual changes due to physical, physiological, and chemical activities of a treatment.

Definition 2.19. A study is *single blind* if subjects being treated are unaware of which treatment (including any control) they are receiving. A study is *double blind* if it is single blind

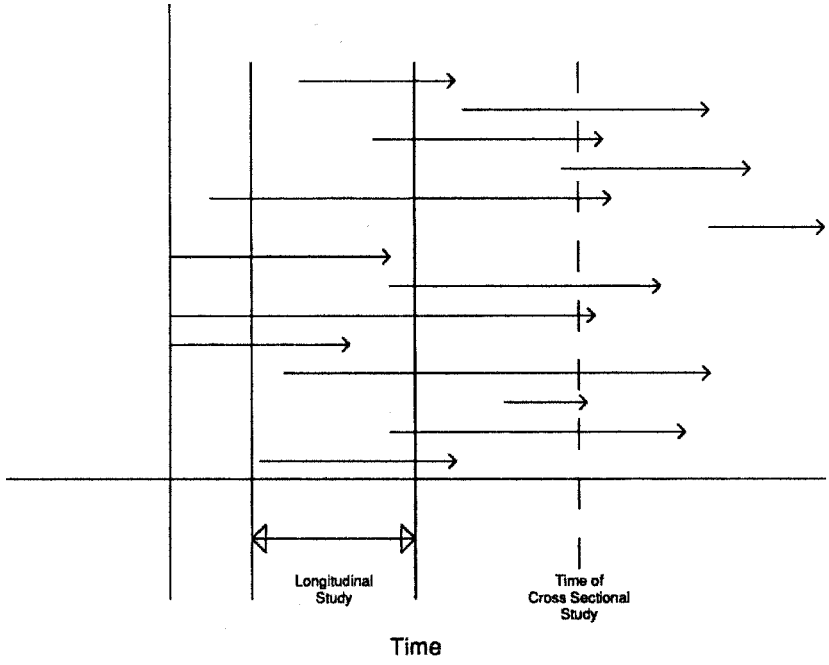

Figure 2.1 Longitudinal and cross-sectional study of cases of a disease.

and the people who are evaluating the outcome variables are also unaware of which treatment the subjects are receiving.

2.4 STEPS NECESSARY TO PERFORM A STUDY

In this section we outline briefly the steps involved in conducting a study. The steps are interrelated and are oversimplified here in order to isolate various elements of scientific research and to discuss the statistical issues involved:

1. A question or problem area of interest is considered. This does not involve biostatistics per se.
2. A study is to be designed to answer the question. The design of the study must consider at least the following elements:
   a. Identify the data to be collected. This includes the variables to be measured as well as the number of experimental units, that is, the size of the study or experiment.
   b. An appropriate analytical model needs to be developed for describing and processing data.
   c. What inferences does one hope to make from the study? What conclusions might one draw from the study? To what population(s) is the conclusion applicable?
3. The study is carried out and the data are collected.
4. The data are analyzed and conclusions and inferences are drawn.
5. The results are used. This may involve changing operating procedures, publishing results, or planning a subsequent study.

2.5 ETHICS

Many studies and experiments in the biomedical field involve animal and/or human participants. Moral and legal issues are involved in both areas. Ethics must be of primary concern. In particular, we mention five points relevant to experimentation with humans:

1. It is our opinion that all investigators involved in a study are responsible for the conduct of an ethical study to the extent that they may be expected to know what is involved in the study. For example, we think that it is unethical to be involved in the analysis of data that have been collected in an unethical manner.
2. Investigators are close to a study and often excited about its potential benefits and advances. It is difficult for them to consider all ethical issues objectively. For this reason, in proposed studies involving humans (or animals), there should be review by people not concerned or connected with the study or the investigators. The reviewers should not profit directly in any way if the study is carried out. Implementation of the study should be contingent on such a review.
3. People participating in an experiment should understand and sign an informed consent form. The *principle of informed consent* says that a participant should know about the conduct of a study and about any possible harm and/or benefits that may result from participation in the study. For those unable to give informed consent, appropriate representatives may give the consent.
4. Subjects should be free to withdraw at any time, or to refuse initial participation, without being penalized or jeopardized with respect to current and future care and activities.
5. Both the Nuremberg Code and the Helsinki Accord recommend that, when possible, animal studies be done prior to human experimentation.

References relevant to ethical issues include the U.S. Department of Health, Education, and Welfare's (HEW's) statement on *Protection of Human Subjects* [1975], Papworth's book, *Human Guinea Pigs* [1967], and Spicker et al. [1988]; Papworth is extremely critical of the conduct of modern biological experimentation. There are also guidelines for studies involving animals. See, for example, *Guide for the Care and Use of Laboratory Animals* [HEW, 1985] and *Animal Welfare* [USDA, 1989]. Ethical issues in randomized trials are discussed further in Chapter 19.

2.6 DATA COLLECTION: DESIGN OF FORMS

2.6.1 What Data Are to Be Collected?

In studies involving only one or two investigators, there is often almost complete agreement as to what data are to be collected. In this case it is very important that good laboratory records be maintained. It is especially important that variations in the experimental procedure (e.g., loss of power during a time period, unexpected change in temperature in a room containing laboratory animals) be recorded. If there are peculiar patterns in the data, detailed notes may point to possible causes. The necessity for keeping detailed notes is even more crucial in large studies or experiments involving many investigators; it is difficult for one person to have complete knowledge of a study.

In a large collaborative study involving a human population, it is not always easy to decide what data to collect. For example, often there is interest in getting prognostic information. How many potentially prognostic variables should you record?

Suppose that you are measuring pain relief or quality of life; how many questions do you need to characterize these abstract ideas reasonably? In looking for complications of drugs, should you instruct investigators to enter all complications? This may be an unreliable procedure if you are dependent on a large, diverse group of observers. In studies with many investigators, each investigator will want to collect data relating to her or his special interests. You can arrive rapidly at large, complex forms. If too many data are collected, there are various "prices" to be paid. One obvious price is the expense of collecting and handling large and complex data sets. Another is reluctance (especially by volunteer subjects) to fill out long, complicated forms, leading to possible biases in subject recruitment. If a study lasts a long time, the investigators may become fatigued by the onerous task of data collection. Fatigue and lack of enthusiasm can affect the quality of data through a lack of care and effort in its collection.

On the other hand, there are many examples where too few data were collected. One of the most difficult tasks in designing forms is to remember to include all necessary items. The more complex the situation, the more difficult the task. It is easy to look at existing questions and to respond to them. If a question is missing, how is one alerted to the fact? One of the authors was involved in the design of a follow-up form where mortality could not be recorded. There was an explanation for this: The patients were to fill out the forms. Nevertheless, it was necessary to include forms that would allow those responsible for follow-up to record mortality, the primary endpoint of the study.

To assure that all necessary data are on the form, you are advised to follow four steps:

1. Perform a thorough review of all forms *with a written response* by all participating investigators.
2. Decide on the statistical analyses beforehand. Check that *specific* analyses involving *specific* variables can be run. Often, the analysis is changed during processing of the data or in the course of "interactive" data analysis. This preliminary step is still necessary to ensure that data are available to answer the primary questions.
3. Look at other studies and papers in the area being studied. It may be useful to mimic analyses in the most outstanding of these papers. If they contain variables not recorded

in the new study, find out why. The usual reason for excluding variables is that they are not needed to answer the problems addressed.

4. If the study includes a pilot phase, as suggested below, analyze the data of the pilot phase to see if you can answer the questions of interest when more data become available.

2.6.2 Clarity of Questions

The task of designing clear and unambiguous questions is much greater than is generally realized. The following points are of help in designing such questions:

1. Who is filling out the forms? Forms to be filled out by many people should, as much as possible, be self-explanatory. There should not be another source to which people are required to go for explanation—often, they would not take the trouble. This need not be done if trained technicians or interviewers are being used in certain phases of the study.
2. The degree of accuracy and the units required should be specified where possible. For example, data on heights should not be recorded in both inches and centimeters in the same place. It may be useful to allow both entries and to have a computer adjust to a common unit. In this case have two possible entries, one designated as centimeters and the other designated as inches.
3. A response should be required on all sections of a form. Then if a portion of the form has no response, this would indicate that the answer was missing. (If an answer is required only under certain circumstances, you cannot determine whether a question was missed or a correct "no answer" response was given; a blank would be a valid answer. For example, in pathology, traditionally the pathologist reports only "positive" findings. If a finding is absent in the data, was the particular finding not considered, and missed, or was a positive outcome not there?)
4. There are many alternatives when collecting data about humans: forms filled out by a subject, an in-person interview by a trained interviewer, a telephone interview, forms filled out by medical personnel after a general discussion with the subject, or forms filled out by direct observation. It is an eye-opening experience to collect the "same" data in several different ways. This leads to a healthy respect for the amount of variability in the data. It may also lead to clarification of the data being collected. In collecting subjective opinions, there is usually interaction between a subject and the method of data collection. This may greatly influence, albeit unconsciously, a subject's response.

The following points should also be noted. A high level of formal education of subjects and/or interviewer is not necessarily associated with greater accuracy or reproducibility of data collected. The personality of a subject and/or interviewer can be more important than the level of education. The effort and attention given to a particular part of a complex data set should be proportional to its importance. Prompt editing of data for mistakes produces higher-quality data than when there is considerable delay between collecting, editing, and correction of forms.

2.6.3 Pretesting of Forms and Pilot Studies

If it is extremely difficult, indeed almost impossible, to design a satisfactory form, how is one to proceed? It is necessary to have a pretest of the forms, except in the simplest of experiments and studies. In a *pretest*, forms are filled out by one or more people prior to beginning an actual study and data collection. In this case, several points should be considered. People filling out forms should be representative of the people who will be filling them out during the study. You can be misled by having health professionals fill out forms that are designed for the "average" patient. You should ask the people filling out the pretest forms if they have any questions or are not sure about parts of the forms. However, it is important not to interfere while the

forms are being used but to let them be used in the same context as will pertain in the study; then ask the questions. Preliminary data should be analyzed; you should look for differences in responses from different clinics or individuals. Such analyses may indicate that a variable is being interpreted differently by different groups. The pretest forms should be edited by those responsible for the design. Comments written on the forms or answers that are not legitimate can be important in improving the forms. During this phase of the study, one should pursue vigorously the causes of missing data.

A more complete approach is to have a *pilot study*, which consists of going through the actual mechanics of a proposed study. Thus, a pilot study works out both the "bugs" from forms used in data collection and operational problems within the study. Where possible, data collected in a pilot study should be compared with examples of the "same" data collected in other studies. Suppose that there is recording of data that are not quantitative but categorical (e.g., the amount of impairment of an animal, whether an animal is losing its hair, whether a patient has improved morale). There is a danger that the investigator(s) may use a convention that would not readily be understood by others. To evaluate the extent to which the data collected are understood, it is good procedure to ask others to examine some of the same study units and to record their opinion without first discussing what is meant by the categories being recorded. If there is great variability, this should lead to a need for appropriate caution in the interpretation of the data. This problem may be most severe when only one person is involved in data collection.

2.6.4 Layout and Appearance

The physical appearance of forms is important if many people are to fill them out. People attach more importance to a printed page than to a mimeographed page, even though the layout is the same. If one is depending on voluntary reporting of data, it may be worthwhile to spend a bit more to have forms printed in several colors with an attractive logo and appearance.

2.7 DATA EDITING AND VERIFICATION

If a study involves many people filling out forms, it will be necessary to have a manual and/or computer review of the content of the forms before beginning analysis. In most studies there are inexplicably large numbers of mistakes and missing data. If missing and miscoded data can be attacked vigorously *from the beginning* of a study, the quality of data can be vastly improved. Among checks that go into data editing are the following:

1. *Validity checks*. Check that only allowable values or codes are given for answers to the questions. For example, a negative weight is not allowed. A simple extension of this idea is to require that most of the data fall within a given range; range checks are set so that a small fraction of the valid data will be outside the range and will be "flagged"; for example, the height of a professional basketball team center (who happens to be a subject in the study) may fall outside the allowed range even though the height is correct. By checking out-of-range values, many incorrectly recorded values can be detected.
2. *Consistency checks*. There should be internal consistency of the data. Following are some examples:
   - **a.** If more than one form is involved, the dates on these forms should be consistent with each other (e.g., a date of surgery should precede the date of discharge for that surgery).
   - **b.** Consistency checks can be built into the study by collecting crucial data in two different ways (e.g., ask for both date of birth and age).
   - **c.** If the data are collected sequentially, it is useful to examine unexpected changes between forms (e.g., changes in height, or drastic changes such as changes of weight by 70%). Occasionally, such changes are correct, but they should be investigated.

d. In some cases there are certain combinations of replies that are mutually inconsistent; checks for these should be incorporated into the editing and verification procedures.

3. *Missing forms*. In some case–control studies, a particular control may refuse to participate in a study. Some preliminary data on this control may already have been collected. Some mechanism should be set up so that it is clear that no further information will be obtained for that control. (It will be useful to keep the preliminary information so that possible selection bias can be detected.) If forms are entered sequentially, it will be useful to decide when missing forms will be labeled "overdue" or "missing."

2.8 DATA HANDLING

All except the smallest experiments involve data that are eventually processed or analyzed by computer. Forms should be designed with this fact in mind. It should be easy to enter the form by keyboard. Some forms are called *self-coding*: Columns are given next to each variable for data entry. Except in cases where the forms are to be entered by a variety of people at different sites, the added cluttering of the form by the self-coding system is not worth the potential ease in data entry. Experienced persons entering the same type of form over and over soon know which columns to use. Alternatively, it is possible to overlay plastic sheets that give the columns for data entry.

For very large studies, the logistics of collecting data, putting the data on a computer system, and linking records may hinder a study more than any other factor. Although it is not appropriate to discuss these issues in detail here, the reader should be aware of this problem. In any large study, people with expertise in data handling and computer management of data should be consulted during the design phase. Inappropriately constructed data files result in unnecessary expense and delay during the analytic phase. In projects extending over a long period of time and requiring periodic reports, it is important that the timing and management of data collection and management be specified. Experience has shown that even with the best plans there will be inevitable delays. It is useful to allow some slack time between required submission of forms and reports, between final submission and data analysis.

Computer files or tapes will occasionally be erased accidentally. In the event of such a disaster it is necessary to have backup computer tapes and documentation. If information on individual subject participants is required, there are confidentiality laws to be considered as well as the investigator's ethical responsibility to protect subject interests. During the design of any study, everyone will underestimate the amount of work involved in accomplishing the task. Experience shows that caution is necessary in estimating time schedules. During a long study, constant vigilance is required to maintain the quality of data collection and flow. In laboratory experimentation, technicians may tend to become bored and slack off unless monitored. Clinical study personnel will tire of collecting the data and may try to accomplish this too rapidly unless monitored.

Data collection and handling usually involves almost all participants of the study and should not be underestimated. It is a common experience for research studies to be planned without allowing sufficient time or money for data processing and analysis. It is difficult to give a rule of thumb, but in a wide variety of studies, 15% of the expense has been in data handling, processing, and analysis.

2.9 AMOUNT OF DATA COLLECTED: SAMPLE SIZE

It is part of scientific folklore that one of the tasks of a statistician is to determine an appropriate sample size for a study. Statistical considerations do have a large bearing on the selection of a sample size. However, there is other scientific input that must be considered in order to arrive at the number of experimental units needed. If the purpose of an experiment is to estimate

some quantity, there is a need to know how precise an estimate is desired and how confident the investigator wishes to be that the estimate is within a specified degree of precision. If the purpose of an experiment is to compare several treatments, it is necessary to know what difference is considered important and how certain the investigator wishes to be of detecting such a difference. Statistical calculation of sample size requires that all these considerations be quantified. (This topic is discussed in subsequent chapters.) In a descriptive observational study, the size of the study is determined by specifying the needed accuracy of estimates of population characteristics.

2.10 INFERENCES FROM A STUDY

2.10.1 Bias

The statistical term *bias* refers to a situation in which the statistical method used does not estimate the quantity thought to be estimated or test the hypothesis thought to be tested. This definition will be made more precise later. In this section the term is used on a intuitive level. Consider some examples of biased statistical procedures:

1. A proposal is made to measure the average amount of health care in the United States by means of a personal health questionnaire that is to be passed out at an American Medical Association convention. In this case, the AMA respondents constitute a biased sample of the overall population.
2. A famous historical example involves a telephone poll made during the Dewey–Truman presidential contest. At that time—and to some extent today—a large section of the population could not afford a telephone. Consequently, the poll was conducted among more well-to-do citizens, who constituted a biased sample with respect to presidential preference.
3. In a laboratory experiment, animals receiving one treatment are kept on one side of the room and animals receiving a second treatment are kept on another side. If there is a large differential in lighting and heat between the two sides of the room, one could find "treatment effects" that were in fact ascribable to differences in light and/or heat. Work by Riley [1975] suggests that level of stress (e.g., bottom cage vs. top cage) affects the resistance of animals to carcinogens.

In the examples of Section 1.5, methods of minimizing bias were considered. Single- and double-blind experiments reduce bias.

2.10.2 Similarity in a Comparative Study

If physicists at Berkeley perform an experiment in electron physics, it is expected that the same experiment could be performed successfully (given the appropriate equipment) in Moscow or London. One expects the same results because the current physical model is that all electrons are precisely the same (i.e., they are identical) and the experiments are truly similar experiments. In a comparative experiment, we would like to try out experiments on similar units.

We now discuss similarity where it is assumed for the sake of discussion that the experimental units are humans. The ideas and results, however, can be extended to animals and other types of experimental units. The experimental situations being compared will be called *treatments*. To get a fair comparison, it is necessary that the treatments be given to similar units. For example, if cancer patients whose disease had not progressed much receive a new treatment and their survival is compared to the standard treatment administered to all types of patients, the comparison would not be justified; the treatments were not given to similar groups.

Of all human beings, *identical twins* are the most alike, by having identical genetic background. Often, they are raised together, so they share the same environment. Even in an observational twin study, a strong scientific inference can be made if enough appropriate pairs of identical twins can be found. For example, suppose that the two "treatments" are smoking and nonsmoking. If one had identical twins raised together where one of the pair smoked and the other did not, the incidence of lung cancer, the general health, and the survival experience could provide quite strong scientific inferences as to the health effect of smoking. (In Sweden there is a twin registry to aid public health and medical studies.) It is difficult to conduct twin studies because sufficient numbers of identical twins need to be located, such that one member of the pair has one treatment and the other twin, another treatment. It is expensive to identify and find them. Since they have the same environment, in a smoking study it is most likely, that either both would smoke or both would not smoke. Such studies are logistically not possible in most circumstances.

A second approach is that of matching or pairing individuals. The rationale behind *matched* or *matched pair studies* is to find two persons who are identical with regard to all "pertinent" variables under consideration except the treatment. This may be thought of as an attempt to find a surrogate identical twin. In many studies, people are matched with regard to age, gender, race, and some indicator of socioeconomic status. In a prospective study, the two matched individuals receive differing treatments. In a retrospective study, the person with the endpoint is identified first (the person usually has some disease); as we have seen, such studies are called case–control studies. One weakness of such studies is that there may not be a sufficient number of subjects to make "good" matches. Matching on too many variables makes it virtually impossible to find a sufficient number of control subjects. No matter how well the matching is done, there is the possibility that the groups receiving the two treatments (the case and control groups) are not sufficiently similar because of unrecognized variables.

A third approach is not to match on specific variables but to try to select the subjects on an intuitive basis. For example, such procedures often select the next person entering the clinic, or have the patient select a friend of the same gender. The rationale here is that a friend will tend to belong to the same socioeconomic environment and have the same ethnic characteristics.

Still another approach, even farther removed from the "identical twins" approach, is to select a group receiving a given treatment and then to select in its entirety a second group as a control. The hope is that by careful consideration of the problem and good intuition, the control group will, in some sense, mirror the first treatment group with regard to "all pertinent characteristics" except the treatment and endpoint. In a retrospective study, the first group usually consists of cases and a control group selected from the remaining population.

The final approach is to select the two groups in some manner realizing that they will not be similar, and to measure pertinent variables, such as the variables that one had considered matching upon, as well as the appropriate endpoint variables. The idea is to make statistical adjustments to find out what would have happened had the two groups been comparable. Such adjustments are done in a variety of ways. The techniques are discussed in following chapters.

None of the foregoing methods of obtaining "valid" comparisons are totally satisfactory. In the 1920s, Sir Ronald A. Fisher and others made one of the great advances in scientific methodology—they assigned treatments to patients by chance; that is, they assigned treatments *randomly*. The technique is called *randomization*. The statistical or chance rule of assignment will satisfy certain properties that are best expressed by the concepts of probability theory. These concepts are described in Chapter 4. For assignment to two therapies, a coin toss could be used. A head would mean assignment to therapy 1; a tail would result in assignment to therapy 2. Each patient would have an equal chance of getting each therapy. Assignments to past patients would not have any effect on the therapy assigned to the next patient. By the laws of probability, on the average, treatment groups will be similar. *The groups will even be similar with respect to variables not measured or even thought about!* The mathematics of probability allow us to estimate whether differences in the outcome might be due to the chance assignment to the two

groups or whether the differences should be ascribed to true differences between treatments. These points are discussed in more detail later.

2.10.3 Inference to a Larger Population

Usually, it is desired to apply the results of a study to a population beyond the experimental units. In an experiment with guinea pigs, the assumption is that if other guinea pigs had been used, the "same" results would have been found. In reporting good results with a new surgical procedure, it is implicit that this new procedure is probably good for a wide variety of patients in a wide variety of clinical settings. To extend results to a larger population, experimental units should be *representative* of the larger population. The best way to assure this is to select the experimental units *at random*, or by chance, from the larger population. The mechanics and interpretation of such random sampling are discussed in Chapter 4. Random sampling assures, on the average, a representative sample. In other instances, if one is willing to make assumptions, the extension may be valid. There is an implicit assumption in much clinical research that a treatment is good for almost everyone or almost no one. Many techniques are used initially on the subjects available at a given clinic. It is assumed that a result is true for all clinics if it works in one setting.

Sometimes, the results of a technique are compared with "historical" controls; that is, a new treatment is compared with the results of previous patients using an older technique. The use of historical controls can be hazardous; patient populations change with time, often in ways that have much more importance than is generally realized. Another approach with weaker inference is the use of an animal model. The term *animal model* indicates that the particular animal is susceptible to, or suffers from, a disease similar to that experienced by humans. If a treatment works on the animal, it may be useful for humans. There would then be an investigation in the human population to see whether the assumption is valid.

The results of an observational study carried out in one country may be extended to other countries. This is not always appropriate. Much of the "bread and butter" of epidemiology consists of noting that the same risk factor seems to produce different results in different populations, or in noting that the particular endpoint of a disease occurs with differing rates in different countries. There has been considerable advance in medical science by noting different responses among different populations. This is a broadening of the topic of this section: extending inferences in one population to another population.

2.10.4 Precision and Validity of Measurements

Statistical theory leads to the examination of variation in a method of measurement. The variation may be estimated by making repeated measurements on the same experimental unit. If instrumentation is involved, multiple measurements may be taken using more than one of the instruments to note the variation between instruments. If different observers, interviewers, or technicians take measurements, a quantification of the variability between observers may be made. It is necessary to have information on the precision of a method of measurement in calculating the sample size for a study. This information is also used in considering whether or not variables deserve repeated measurements to gain increased precision about the true response of an experimental unit.

Statistics helps in thinking about alternative methods of measuring a quantity. When introducing a new apparatus or new technique to measure a quantity of interest, validation against the old method is useful. In considering subjective ratings by different people (even when the subjective rating is given as a numerical scale), it often turns out that a quantity is not measured in the same fashion if the measurement method is changed. A new laboratory apparatus may measure consistently higher than an old one. In two methods of evaluating pain relief, one way of phrasing a question may tend to give a higher percentage of improvement. Methodologic statistical studies are helpful in placing interpretations and inferences in the proper context.

2.10.5 Quantification and Reduction of Uncertainty

Because of variability, there is uncertainty associated with the interpretation of study results. Statistical theory allows quantification of the uncertainty. If a quantity is being estimated, the amount of uncertainty in the estimate must be assessed. In considering a hypothesis, one may give numerical assessment of the chance of occurrence of the results observed when the hypothesis is true.

Appreciation of statistical methodology often leads to the design of a study with increased precision and consequently, a smaller sample size. An example of an efficient technique is the statistical idea of blocking. Blocks are subsets of relatively homogeneous experimental units. The strategy is to apply all treatments randomly to the units within a particular block. Such a design is called a *randomized block design*. The advantage of the technique is that comparisons of treatments are intrablock comparisons (i.e., comparisons within blocks) and are more precise because of the homogeneity of the experimental units within the blocks, so that it is easier to detect treatment differences. As discussed earlier, simple randomization does ensure similar groups, but the variability within the treatment groups will be greater if no blocking of experimental units has been done. For example, if age is important prognostically in the outcome of a comparative trial of two therapies, there are two approaches that one may take. If one ignores age and randomizes the two therapies, the therapies will be tested on similar groups, but the variability in outcome due to age will tend to mask the effects of the two treatments. Suppose that you place people whose ages are close into blocks and assign each treatment by a chance mechanism within each block. If you then compare the treatments within the blocks, the effect of age on the outcome of the two therapies will be largely eliminated. A more precise comparison of the therapeutic effects can be gained. This increased precision due to statistical design leads to a study that requires a smaller sample size than does a completely randomized design. However, see Meier et al. [1968] for some cautions.

A good statistical design allows the investigation of several factors at one time with little added cost (Sir R. A. Fisher as quoted by Yates [1964]):

> No aphorism is more frequently repeated with field trials than we must ask Nature a few questions, or ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed if we ask her a single question, she will often refuse to answer until some other topic has been discussed.

PROBLEMS

2.1 Consider the following terms defined in Chapters 1 and 2: single blind, double blind, placebo, observational study, experiment, laboratory experiment, comparative experiment, crossover experiment, clinical study, cohort, prospective study, retrospective study, case–control study, and matched case–control study. In the examples of section 1.5, which terms apply to which parts of these examples?

2.2 List possible advantages and disadvantages of a double-blind study. Give some examples where a double-blind study clearly cannot be carried out; suggest how virtues of "blinding" can still be retained.

2.3 Discuss the ethical aspects of a randomized placebo-controlled experiment. Can you think of situations where it would be extremely difficult to carry out such an experiment?

2.4 Discuss the advantages of randomization in a randomized placebo-controlled experiment. Can you think of alternative, possibly better, designs? Consider (at least) the aspects of bias and efficiency.

2.5 This problem involves the design of two questions on "stress" to be used on a data collection form for the population of a group practice health maintenance organization. After a few years of follow-up, it is desired to assess the effect of physical and psychological stress.

(a) Design a question that classifies jobs by the amount of physical work involved. Use eight or fewer categories. Assume that the answer to the question is to be based on job title. That is, someone will code the answer given a job title.

(b) Same as part (a), but now the classification should pertain to the amount of psychological stress.

(c) Have yourself and (independently) a friend answer your two questions for the following occupational categories: student, college professor, plumber, waitress, homemaker, salesperson, unemployed, retired, unable to work (due to illness), physician, hospital administrator, grocery clerk, prisoner.

(d) What other types of questions would you need to design to capture the total amount of stress in the person's life?

2.6 In designing a form, careful distinction must be made between the following categories of nonresponse to a question: (1) not applicable, (2) not noted, (3) don't know, (4) none, and (5) normal. If nothing is filled in, someone has to determine which of the five categories applies—and often this cannot be done after the interview or the records have been destroyed. This is particularly troublesome when medical records are abstracted. Suppose that you are checking medical records to record the number of pregnancies (*gravidity*) of a patient. Unless the gravidity is specifically given, you have a problem. If no number is given, any one of the four categories above could apply. Give two other examples of questions with ambiguous interpretation of "blank" responses. Devise a scheme for interview data that is unambiguous and does not require further editing.

REFERENCES

Meier, P., Free, S. M., Jr., and Jackson, G. L. [1968]. Reconsideration of methodology in studies of pain relief. *Biometrics*, **14**: 330–342.

Papworth, M. H. [1967]. *Human Guinea Pigs*. Beacon Press, Boston.

Riley, V. [1975]. Mouse mammary tumors: alteration of incidence as apparent function of stress. *Science*, **189**: 465–467.

Roethlisberger, F. S. [1941]. *Management and Morals*. Harvard University Press, Cambridge, MA.

Spicker, S. F., et al. (eds.) [1988]. *The Use of Human Beings in Research, with Special Reference to Clinical Trials*. Kluwer Academic, Boston.

U.S. Department of Agriculture [1989]. Animal welfare: proposed rules, part III. *Federal Register*, Mar. 15, 1989.

U.S. Department of Health, Education, and Welfare [1975]. Protection of human subjects, part III. *Federal Register*, Aug. 8, 1975, **40**: 11854.

U.S. Department of Health, Education, and Welfare [1985]. *Guide for the Care and Use of Laboratory Animals*. DHEW Publication (NIH) 86–23. U.S. Government Printing Office, Washington, DC.

Yates, F. [1964]. Sir Ronald Fisher and the design of experiments. *Biometrics*, **20**: 307–321. Used with permission from the Biometric Society.