Study Designs

6

Chapter Outline

6.1 Design: Putting Chance to Work
6.2 Sample Surveys and Experiments
6.3 Sampling and Sample Designs
6.4 Designed Experiments
6.5 Variations in Study Designs

In meeting a set of data we must first check the credentials of the data: what the data represent and how the data were collected. In Chapter 2 we discussed the linkage between concepts and numbers — that is, what the data represent. As far as data collection is concerned, there are two basic methods used to obtain data: the sample survey and the designed experiment. In this chapter we examine these two basic methods and some variations of them.

6.1 Design: Putting Chance to Work

In collecting sample data we should try to avoid all potential causes of bias. When conducting an experiment, we should try to eliminate the effect of potential confounding factors. Strangely, adhering to these ideas involves the use of a chance mechanism. Let us explore why and how a chance mechanism plays a role in designing surveys and experiments.

A smart shopper is conscious of the possible variability in the quality of fruit between the top and the bottom of the fruit basket. The smart shopper looks at pieces of fruit throughout the basket, even though it is more convenient to look only at the pieces on top, before making a purchase. In the same way, a researcher is aware of the possible variability among observational units in the population. A good researcher takes steps to ensure that the process for selecting units from the population deals with this possible variability. The failure to take adequate steps would introduce a selection bias. Selecting a sample of units because of convenience also poses a problem for a researcher just as it did for the shopper. For example, the opinions of people interviewed during lunchtime on downtown street corners, although convenient to obtain, usually are not representative of the residents of the city. Those who never go to the center of the city during lunchtime are not represented in the sample, and they may have different opinions from those who go to the city center.

We are familiar with the use of a chance mechanism to remove possible biases. For example, to start a football game, a coin toss — a chance mechanism — is used in deciding which team receives the opening kickoff. The use of a chance mechanism is also involved in selecting a sample in an attempt to avoid biases. One method of drawing a fair sample is to place numbered slips of paper in a bowl, mix them up thoroughly, and then have a neutral party pick out the slips. This chance mechanism sounds fair but may not be satisfactory, as shown the following example.

Example 6.1

In 1970, Selective Service officials used a chance mechanism, a lottery, to determine who would be drafted for military service. Officials put slips of paper representing birthdates into cylindrical capsules, one birthdate per capsule, and then placed the capsules into a box. The January birthdates were put into the box first and pushed to one side, and then the February capsules were placed in the box and pushed to the side of the box with the January capsules, and then so on with March. The box was then closed, shaken several times, carried up three flights of stairs, and carried back down to the room, where the capsules were poured into a bowl. A public figure then selected the capsules to determine the order of drafting men. Figure 6.1 shows the lottery results (Fienberg 1971). It appears that the process did not work as intended, since the months at the end of the year, which were put into the container last and were not mixed thoroughly, have much smaller lottery numbers than the earlier months.

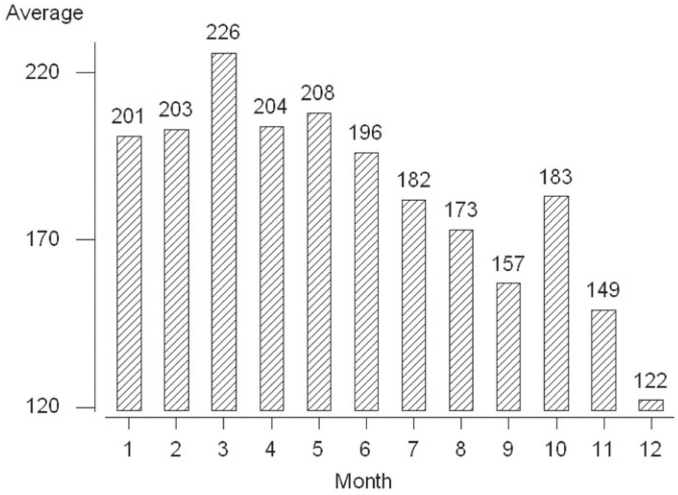

Figure 6.1 Average lottery number by month from the 1970 draft lottery.

A better way of selecting a fair sample is using random numbers that were used to estimate probabilities in Chapter 4. The random numbers can be described as the sequence of numbers we get when we draw balls numbered 0, 1, . . . 9 from an urn, replacing the ball drawn, thoroughly remixing the balls, and then drawing another ball. This process is repeated several times.

The first random numbers were produced by Tippett in 1927. It is said that Tippett obtained the numbers from the figures of areas of parishes given in the British census returns, and omitted the first two and last two digits in each figure of area. The truncated numbers were arranged in sets of four in eight columns. This 26-page book containing

Table 6.1 The first 10 rows from page 14 of Tippett's random numbers.

7816	6572	0802	6314	0702	4369	9728	0198
3204	9243	4935	8200	3623	4869	6938	7481
2976	3413	2841	4241	2424	1985	9313	2322
8303	9822	5888	2410	1158	2729	6443	2943
5556	8526	6166	8231	2438	8455	4618	4445
2635	7900	3370	9160	1620	3882	7757	4950
3211	4919	7306	4916	7677	8733	9974	6732
2748	6198	7164	4148	7086	2888	8519	1620
7477	0111	1630	2404	2979	7991	9683	5125
5379	7076	2694	2927	4399	5519	8106	8501

41,600 digits became the best-seller among technical books. Table 6.1 shows the first 10 rows of page 14 of his random number table and contains 320 digits. The appearance of each digit is random in the sense that we cannot predict the appearance of a particular digit based on the previous sequence of digits. Despite this uncertainty, we can expect that each digit is equally likely to appear. From Table 6.1, the frequencies of each digit from 0 to 9 are 27, 33, 39, 32, 36, 22, 31, 31, 35, and 34, which slightly deviates from the expect frequency (10 percent of 320).

Now random numbers are generated by computer algorithms, and most statistical software packages contain a random number generator. Table B1 in Appendix B show 1000 random digits generated from MINITAB. Considerable research is still devoted to random number generation. No definition of random numbers exists except for a vague description that they do not follow any particular pattern. The use of random numbers helps reduce the possibility of selection bias in surveys and also helps reduce the possible effect of confounders when designing experiments.

6.2 Sample Surveys and Experiments

There are many similarities as well as some differences between sample surveys and experiments. We learn the characteristics of some population from sample surveys. The sample survey focuses on the selection of individuals from the population. We discover the effect of applying a stimulus to subjects from experiments. The experimental design focuses on the formation of comparison groups that allow conclusions about the effect of the stimulus to be drawn.

As emphasized in Chapter 4, a probability sample is a carefully drawn blueprint or design as is an experiment. The blueprint or design of a survey or an experiment is based on both statistical and substantive considerations. An experiment is different from a sample survey in that the experimenter actively intervenes with the experimental subjects through the assignment of the subjects to groups, whereas the survey researcher passively observes or records responses of the survey subjects. Experiments and surveys often have different goals as well.

In a survey, the primary goal is to describe the population, and a secondary goal is to investigate the association between variables. In a survey, variables are usually not referred to as independent or dependent because all the variables can be viewed as being response variables. The survey researcher usually has not manipulated the levels of any of the variables as the experimenter does.

The goal in an experiment is to determine whether or not there is an association between the independent or predictor variables with the dependent or response variable. The different groups to which the subjects are assigned usually represent the levels of the independent variable(s). *Independent* and *dependent* were chosen as names for the variable types because it was thought that the response variable depended on the levels of the predictor variables. To determine whether or not there is an association, the experimenter assigns subjects to different levels of one or more variables — for example, to different doses of some medication. The effects of the different levels — the different doses — are found by measuring the values of an outcome variable — for example, change in blood pressure. An association exists if there is relationship between the change in blood pressure values and the dosage levels. Let us examine first how surveys are designed and then consider the basic principles of experimental design.

6.3 Sampling and Sample Designs

Sampling means selecting a few units from all the possible observational units in the population. For practical purposes, any data set is a sample. Even if a complete census is attempted, there are missing observations. This means that we must pay attention to the intended as well as the unintended sampling when evaluating a sample. This also suggests that we cannot evaluate a sample by looking at the sample itself, but we need to know what sampling method was used and how well it was executed. We are interested in the process of selection as well as the sample obtained.

Sampling is used extensively today for many reasons. In many situations a sample produces more accurate information about the population than that provided by a census. Two reasons for obtaining more accurate information from a sample are the following. As was just mentioned, a census often turns out to be incomplete, and the impact of the missing information is most often unknown. Additionally, in obtaining a sample, fewer interviewers are required, and it is likely that they will be better trained than the huge team of interviewers required when conducting a census.

Even more pragmatically, collecting data from a sample is cheaper and faster than attempting a complete census. In addition, in many situations a census is impractical or even impossible. The following three examples will illustrate situations in which sampling was used and reasons for the use of samples.

Example 6.2

Even in the U.S. Population Census, many data items are collected from a sample of households. In the 2000 Census, for example, only a few basic demographic data items — gender, age, race, and marital status — were asked from each individual in all households in the short form of the questionnaire. Many questions about socioeconomic characteristics such as education, income, and occupation are included in the long form of the questionnaire that was distributed to about 17 percent of U.S. households. In small towns, a larger proportion of households received the long form to ensure reliable estimates. Conversely, in large cities, proportionately fewer households received the long form. Use of sampling not only reduced the cost of the census, but also shortened the data collection burden and time.

Example 6.3

Pharmaceutical companies routinely sample a small fraction of their products to examine the quality and the chemical contents. On the basis of this examination, a decision is made whether to accept the entire lot and ship them or reject the lot and change the manufacturing process. In this case the sample is destroyed to check the quality and, therefore, a company cannot afford to inspect the entire lot.

Example 6.4

Health departments of large urban areas monitor ambient air quality. Since the health department cannot afford to monitor the air everywhere in its coverage area, sample sites are selected and the values of several different pollutants are continuously recorded.

6.3.1 Sampling Frame

Before performing any sampling, it is important to define clearly the population of interest. Similarly, when we are given a set of data, we need to know what group the sample represents — that is, to know from what population the data were collected. The definition of population is often implicit and assumed to be known, but we should ask what the population was before using the data or accepting the information. When we read an election poll, we should know whether the population was all adults or all registered voters to interpret the results appropriately. In practice, the population is defined by specifying the *sampling frame*, the list of units from which the sample was selected. Ideally, the sampling frame must include all units of the defined population. But as we shall see, it is often difficult to obtain the sampling frame and we need to rely on a variety of alternative approaches.

The failure to include all units contained in the defined population in the sampling frame leads to selecting a biased sample. A biased sample is not representative of the population. The average of a variable obtained from a biased sample is likely to be consistently different from the corresponding value in the population. *Selection bias* is the consistent divergence of a sample value (*statistic*) from the corresponding population value (*parameter*) due to an improper selection process. Even with a complete sampling frame, selection bias can occur if proper selection rules were not followed. Two basic sources of selection bias are the use of an incomplete sampling frame and the use of improper selection procedures. The following example illustrates the importance of the sampling frame.

Example 6.5

The *Report of the Second Task Force on Blood Pressure Control in Children* (1987) provides an example of the possibility of selection bias in data. This Task Force used existing data from several studies, only one of which could be considered representative of the U.S. noninstitutionalized population. In this convenience sample, over 70 percent of the data came from Texas, Louisiana, and South Carolina, with little data from the Northeast or the West. Data from England were also used for newborns and children up to three years of age. The representativeness of these data for use in the creation of blood pressure standards for U.S. children is questionable. Unlike the *Literary Digest* survey in which the errors in the sampling were shown to lead to a wrong conclusion, it is not clear that the blood pressure standards are wrong. All we can point to is the use of convenience sampling, and with it, the likely introduction of selection bias by the Second Task Force.

Example 6.6

Telephone surveys may provide another example of the sampling frame failing to include all the members of the target population. If the target population is all the resident households in a geographical area, a survey conducted using the telephone will miss a portion of the resident households. Even though more than 90 percent of the households in the U.S. have telephones, the percentage varies with race and socioeconomic status. The telephone directory was used frequently in the past as the sampling frame, but it excluded households without telephones as well as households with unlisted numbers. A technique called *random digit dialing* (RDD) has been developed to deal with the unlisted number problem in an efficient manner (Waksberg 1978). As the name implies, telephone numbers are basically selected at random from the prefixes — the first 3 digits — thought to contain residential numbers, instead of being selected from a telephone directory. But the concern about the possible selection bias due to missing households without telephones and people who do not have a stable place of residence remains.

In order to avoid or minimize selection bias, every sample needs to be selected based on a carefully drawn sample design. The design defines the population the sample is supposed to represent, identifies the sampling frame from which the sample is to be selected, and specifies the procedural rules for selecting units. The sample data are then evaluated based on the sample design and the way the design was actually executed.

6.3.2 Importance of Probability Sampling

Any sample selected using a random mechanism that results in known chances of selection of the observational units is called a random or a probability sample. This definition requires only that the chances of selection are known. It does not require that the chances of the observational units being selected into the sample are equal. Knowledge of the chance of selection is the basis for the statistical inference from the sample to the population. A sample selected with unknown chances of selection cannot be linked appropriately to the population from which the sample was drawn. This point was explained in Chapter 4. Various sampling designs are discussed in the following sections starting with simple random sampling.

6.3.3 Simple Random Sampling

The simplest probability sample is a *simple random sample* (SRS). In an SRS, each unit in the sampling frame has the same chance of being included in the sample as any other unit. The use of an SRS removes the possibility of any bias, conscious or unconscious, on the part of the researcher in selecting the sample from the sampling frame. An SRS is drawn by the use of a random number table or random numbers generated by a computer. If the population is relatively small, we can number all units sequentially. Next we locate a starting point in the random number table, Table B1 in Appendix B. We then begin reading random numbers in some systematic fashion — for example, across a row or down a column or diagonal — but the direction of reading should be decided ahead of time. The units in the sampling frame whose unique numbers match the random numbers that have been read are selected into the sample.

Example 6.7

Suppose that we have 50 students in a classroom and they are sequentially labeled from 00 to 49 by row starting at the left end of the first row. We wish to select an SRS of 10 students. We decide to use the left-hand corner of line 1 of Table B1 as our starting point, and we will go across the row. By reading the two-digit numbers from the first row of the random digit table, the following 10 numbers are obtained:

$$17,\ \underline{17},\ \underline{47},\ 59,\ \underline{08},\ \underline{43},\ \underline{30},\ 67,\ 70,\ 61$$

Since four numbers are greater than 49, they cannot be used, and we must draw additional numbers until we have 10 random numbers smaller than 50. In addition, the number 17 occurred twice. Since there is no good practical reason for including the same element twice in the sample, we should draw another number that has not been selected previously. We usually sample without replacement, as mentioned in Chapter 4. The next five valid numbers are 07, 44, 48, 36, and 47. Since the number 47 is already used, the next valid number 24 is drawn. The students whose labels match the 10 valid numbers drawn are selected as the sample.

Example 6.8

One way of dealing with the problem of drawing invalid numbers is to subtract 50 from values greater than or equal to 50 in the first set of 10 random numbers. For example, 59, 67, 70, and 61 become 09, 17, 20, and 11. We now select the students with labels 09, 17, 20, and 11. This procedure is based on the premise that each student is represented by two numbers differing by 50 in value. For example, the first student will be selected if either 00 or 50 were read, the second would be selected if either 01 or 51 were read, and so on until the last student would be selected if 49 or 99 were read. Note that even with the subtraction of 50, we again have another 17. We would still have to draw two more valid random numbers: 25 and 02 (obtained by subtracting 50 from 75 and 52) to have 10 distinct values.

In using the procedure in Example 6.8, each unit (student) in the sampling frame had the same number (two) of labels associated with it. If there are 30 students in a class, we can label them in three cycles, 1 through 30, 31 through 60, and 61 through 90, but we cannot assign 91 through 99 and 00 to any student. If we assigned these last 10 values to some of the students, some students would have three labels associated with them, whereas other students would have four labels. The students would have unequal chances of being selected. By not using the last 10 values, each student has three labels (numbers). The first student is assigned the numbers 01, 31, and 61, and the second student is assigned the numbers 02, 32, and 62, and so on for the other students.

In Examples 6.7 and 6.8, we used two-digit random numbers because we could not provide distinct labels for all 50 students with only a single digit. The number of digits to be used is dependent on the size of the population under consideration. For example, when we have 570 units in the population, we need to use three digits. A population that contains 7870 units would require four-digit random numbers.

The SRS design is modified to accommodate other theoretical and practical considerations. The common practical methods for selecting a sample include systematic sampling, stratified random sampling, single-stage cluster sampling, multistage cluster sampling, PPS (probability proportional to size) sampling, and other controlled selection procedures. These more practical designs deviate from SRS in two important ways. First, the inclusion probabilities for the elements (also the joint inclusion probabilities for sets for the elements) may be unequal. Second, the sampling unit(s) can be different from the population element of interest. These departures complicate the usual methods of estimation and variance calculation and, if no adjustments are made, can lead to a bias in estimation and statistical tests. We will consider these departures in detail, using several specific sampling designs, and examine their implications for survey analysis.

Computer packages can be used to draw random samples (see **Program Note 6.1** on the website).

6.3.4 Systematic Sampling

Systematic sampling is commonly used as an alternative to SRS because of its simplicity. It selects every kth element after a random start. Its procedural tasks are simple, and the process can easily be checked, whereas it is difficult to verify SRS by examining the results. It is often used in the final stage of multistage sampling when the field worker is instructed to select a predetermined proportion of units from the listing of dwellings in a street block. The systematic sampling procedure assigns each element in a population the same probability of being selected. This assures that the sample mean will be an unbiased estimate of the population mean when the number of elements in the population (N) is equal to k times the number of elements in the sample (n). If N is not exactly nk, then the equal probability is not guaranteed, although this problem can be ignored when N is large. When N is not exactly nk, we can use the *circular systematic sampling scheme*. In this scheme, the random starting point is selected between 1 and N (any element can be the starting point) and every kth element is selected assuming that the frame is circular (i.e., the end of list is connected to the beginning of the list).

Example 6.9

Suppose that we are taking a 1-in-4 systematic sample from a population of 11 elements: A, B, C, D, E, F, G, H, I, J, and K. Four possible samples can be drawn using the ordinary systematic sampling scheme and 11 possible samples using the circular systematic sampling. The possible samples and their selection probabilities using the ordinary systematic sampling and circular systematic sampling are shown in Table 6.2.

Ordinary systematic sampling does not guarantee equal probability sampling. For example, here the fourth sample has a different selection probability. Under the circular systematic sampling, each element can be a starting point and equal probability sampling is guaranteed in this scheme.

Table 6.2 Possible samples and selection probabilities taking 1-in-4 systematic samples from *N* = 11, using two different selection schemes.

Ordinary Systematic Sampling					Circular Systematic Sampling				
Samples				Selection Probability	Samples				Selection Probability
1.	A	E	I	3/11	1.	A	E	I	3/11
2.	B	F	J	3/11	2.	B	F	J	3/11
3.	C	G	K	3/11	3.	C	G	K	3/11
4.	D	H		2/11	4	D	H	A	3/11
					5.	E	I	B	3/11
					6.	F	J	C	3/11
					7.	G	K	D	3/11
					8.	H	A	E	3/11
					9.	I	B	F	3/11
					10.	J	C	G	3/11
					11.	K	D	H	3/11

Systematic sampling is convenient to use, but it can give an unrealistic estimate when the elements in the frame are listed in a cyclical manner with respect to a survey variable and the selection interval coincides with the listing cycle. For example, if one selects every 40th patient coming to a clinic and the average daily patient load is about 40, then the resulting sample would contain only those who came to the clinic at a certain time of the day. Such a sample may not be representative of the clinic patients. Moreover, even when the listing is randomly ordered, unlike SRS, different sets of elements may have unequal inclusion probabilities. For example, the probability of including both the ith and $(i + k)$th element is $1/k$ in a systematic sample, whereas the probability of including both the ith and $(i + k + 1)$th element is zero. This situation complicates the variance calculation.

Another way of viewing systematic sampling is that it is equivalent to selecting one cluster from k systematically formed clusters of n elements each. The sampling variance (between clusters) cannot be estimated from the one cluster selected. Thus, variance estimation from a systematic sample requires special strategies.

A modification to overcome these problems with systematic sampling is the so-called *repeated systematic sampling*. Instead of taking a systematic sample in one pass through the list, several smaller systematic samples are selected going down the list several times with a new starting point in each pass. This procedure not only guards against possible periodicity in the frame but also allows variance estimation directly from the data. The variance of an estimate from all subsamples can be estimated from the variability of the separate estimates from each subsample.

6.3.5 Stratified Random Sampling

Stratification is often used in complex sample designs. In a *stratified random sample* design, the units in the sampling frame are first divided into groups, called *strata*, and a separate SRS is taken in each stratum to form the total sample. The strata are formed to keep similar units together — for example, a female stratum and a male stratum. In this design, units need not have equal chances of being selected and some strata may be deliberately oversampled. For example, in the first National Health and Nutrition Examination Survey (NHANES I), the elderly, persons in poverty areas, and women of childbearing age were oversampled to provide sufficient numbers of these groups for in-depth analysis (NCHS 1973). If an SRS had been used, it is likely that too few people in these groups would have been selected to allow any in-depth analysis of these groups.

Another advantage of stratification is that it can reduce the variability of sample statistics over that of an SRS, thus reducing the sample size required for analysis. This reduction in variability occurs when the units in a stratum are similar, but there is variation across strata. Another way of saying this is that the reduction occurs when the variable used to form the strata is related to the variable being measured. Let us consider a small example that illustrates this point.

In this example, we wish to estimate the average weight of persons in the population. The population contains six persons: three females and three males. The weights of the females in the population are 110, 120, and 130 pounds, and the weights of the males are 160, 170, and 180 pounds. We shall form our estimate of the population average weight by taking a sample of size two without replacement.

If we use an SRS, the smallest possible estimate is 115 pounds (= [110 + 120]/2), and the largest possible estimate is 175 (= [170 + 180]/2). As an alternative, we could use a stratified random sample where the strata are formed based on gender. If one person is randomly selected from each stratum, the smallest estimate is 135 pounds (= [110 + 160]/2), and the largest estimate is 155 pounds (= [130 + 180]/2). The estimates from the stratified sample approach have less variation — that is, have greater precision than the SRS approach in this case.

The formulation of the strata requires that information on the stratification variables be available for the elements in the sampling frame. When such information is not available, stratification is not possible, but we still can take advantage of stratification by using the poststratification method. For example, stratification by race is usually desirable in social surveys but the racial identification is often not available in the sampling frame. In this case we can attempt to take race into account in the analysis after the sample is selected. Chapter 15 will provide further discussion on this topic.

6.3.6 Cluster Sampling

Most of the methods of statistical analysis assume that the data were collected using an SRS. However, when we attempt to use an SRS in the collection of data, we often encounter difficulties. Suppose we wanted an SRS of 500 adults from a large city. First, a sampling frame is not readily available. Developing a list of all adults in the city is very costly and should be considered impractical. Even though we are able to select an SRS of 500 adults from a reasonably complete list, it would be expensive to send interviewers to sample persons scattered all over the city. A solution to these practical difficulties is to sample people based on geographical areas — for example, census tracts. Most survey agencies and researchers use a *multistage cluster sample design* in this situation. First, a random sample of census tracts is selected and then neighborhood blocks within each selected tract are randomly selected. Within the selected neighborhood blocks a list of households can be prepared and a sample of households can be selected systematically from the list — say, every third household. Finally, within each of the selected households, an adult may be randomly chosen. In this example, the census tracks, neighborhood blocks, and households are the clusters used as the sampling units.

Cluster sampling is widely used but it complicates statistical estimation and analysis, since the sampling method deviates from SRS. For example, an SRS of unequal-sized clusters leads to the elements in the smaller clusters being more likely to be in the sample than those in the larger clusters. Such complications are handled either by using a special selection method or by a special analytical method. We will discuss these methods in Chapter 15.

6.3.7 Problems Due to Unintended Sampling

In analyzing data it is imperative to understand the sample design, as well as how the design was actually executed in the field. Deviations from the intended sample design are reflected in the data. Even in a well-designed survey, it is usually not possible to collect data from all the units sampled because there is almost always some nonresponse. Hence, the respondents, a subset of the sampled persons, are self-selected from the sampled persons through some procedure which is usually unknown to the designer of the study. Since the respondents are no longer a random sample of the study population, there is concern that the data may be unusable because of nonresponse bias.

If the percentage of nonresponse is small — say, less than 5 to 10 percent — there is usually little concern because the bias, if any, is also likely to be small. If the nonresponse is on the order of 20 to 30 percent, the possibility of a substantial bias exists. For example, assume that we wish to estimate the proportion of people without health insurance in our community. We select an SRS and find that 20 percent of the respondents were without health insurance. However, 1/4 of those selected to be in the sample did not respond. If we knew the proportion of those without health insurance among the nonrespondents, it would be easy to combine this value with that of the respondents to obtain the total sample estimate. The proportions of those without health insurance among the respondents and nonrespondents would be weighted by the corresponding proportion of respondents and nonrespondents in the sample.

For example, if none of these nonrespondents had health insurance, the total sample estimate would be 40 percent (= {20% × 0.75} + {100% × 0.25}), twice as large as the rate for the respondents only. If all of the nonrespondents had health insurance, then the total sample estimate becomes 15 percent (= {20% × 0.75} + {0% × 0.25}). Hence, although 20 percent of the respondents were without health insurance, the total sample estimate can range from 15 to 40 percent when 1/4 of the sample are nonrespondents.

For nonresponse bias to occur, the nonrespondents must differ from the respondents with regard to the variable of interest. In the preceding example, it may be that many of the nonrespondents were unemployed homeless whereas few of the respondents were unemployed or homeless. In this case, the respondents and nonrespondents would likely differ with regard to health insurance coverage. If they do differ, there would be a large nonresponse bias. With larger percentages of nonresponse, the likelihood of a substantial nonresponse bias is very high, and this makes the use of the data questionable. Unfortunately, many large surveys have a high percentage of nonresponse or do not mention the level of nonresponse. Data from these surveys are problematic.

Example 6.10

An example of a survey with poor response is the Nationwide Food Consumption Survey conducted in 1987–1988 for the U.S. Department of Agriculture. This survey, conducted once per decade, was to be the basis for policy decisions regarding food assistance programs. However, only about one-third of the persons who were selected for the sample participated, and, hence, the sample may not be representative of the U.S. population. An independent expert panel and the Government Accounting Office of the U.S. Congress have concluded that information from this survey may be unusable (Government Accounting Office 1991).

There is no easy solution to the nonresponse problem. The best approach is a preventive one — that is, to exert every effort to obtain a high response rate. Even if you are unable to contact the sample person, perhaps a neighbor or family member can provide some basic demographic data about the person. If a sample person refuses to participate, again try to obtain some basic data about the person. If possible, try to obtain some information about the main topic of interest in the survey. The basic demographic data can be used to compare the respondents and nonrespondents. Even if there are no differences between the two groups on the demographic variables, that does not necessarily guarantee the absence of nonresponse bias. However, it does eliminate the demographic variables as a cause of the potential nonresponse bias. If there is a difference, it may be possible to take those differences into account and create an adjusted estimator. The following calculations show one of many possible adjustment methods.

Suppose we found that there was a difference in the gender distribution between the respondents and nonrespondents. Sixty percent of the respondents were females and 40 percent were males, whereas 30 percent of the nonrespondents were females and 70 percent were males. If there were no difference in the proportions of females and males with health insurance, this difference in the gender distribution between the respondents and nonrespondents would be no problem. However, for this example, assume there was a difference. In the respondent group, 30 percent of the females were without health insurance compared to only 5 percent of the males. Figure 6.2 is a display of these percentages and of the calculations involved in creating an adjusted rate.

The corresponding percentages with health insurance are unknown for the nonrespondent group. However, if we assume that the female and male respondents' percentages with health insurance hold in the nonrespondent group, we can obtain an

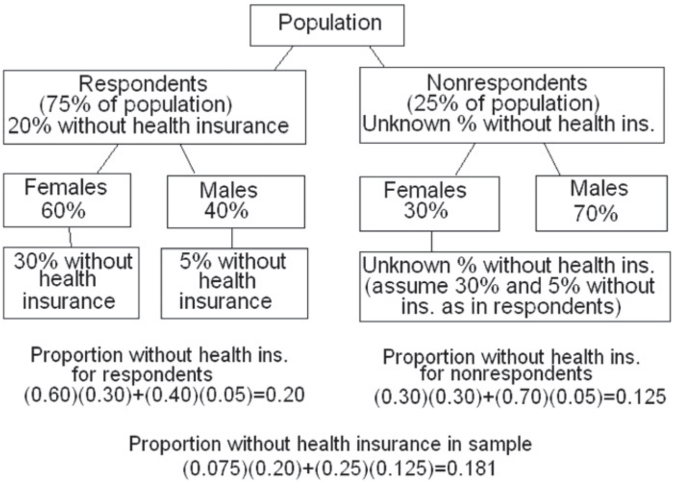
Population
Respondents (75% of population) 20% without health insurance
Nonrespondents (25% of population) Unknown % without health ins.
Females 60%
Males 40%
Females 30%
Males 70%
30% without health insurance
5% without health insurance
Unknown % without health ins. (assume 30% and 5% without ins. as in respondents)
Proportion without health ins. for respondents
(0.60)(0.30)+(0.40)(0.05)=0.20
Proportion without health ins. for nonrespondents
(0.30)(0.30)+(0.70)(0.05)=0.125
Proportion without health insurance in sample
(0.075)(0.20)+(0.25)(0.125)=0.181

Figure 6.2 Display of the percentages for the health insurance example and calculation of the adjusted rate.

adjusted rate. The percentage of those without health insurance in the nonrespondent group under this assumption is found by weighting the proportions of females and males without health insurance by their proportions in the nonrespondent group — that is, {30% × 0.3} + {70% × 0.05}, which is 12.5 percent. We then use this value for the proportion of nonrespondents without health insurance and combine it with the proportion of respondents without health insurance to obtain a sex-adjusted estimate of the proportion of our community without health insurance. This adjusted estimated is 18.1 percent (= {75% × 0.20} + {25% × 0.125}).

The adjusted rate does not differ much from the rate for the respondents only. However, this adjusted rate was based on the assumption that the proportions of females and males without health insurance were the same for respondents and nonrespondents. If this assumption is false, which we cannot easily check, this adjusted estimate then is incorrect. Whatever method of adjustment is employed, an assumption similar to the above must be made at some stage in the adjustment process (Kalton 1983). Our message is to prevent nonresponse from occurring or to keep its rate of occurrence small.

The discussion so far has focused on *unit nonresponse* — that is, the observational unit did not participate in the survey. There is also *item nonresponse*, in which the sample person did not provide the requested information for some of the items in the survey. Just as there are no easy answers to unit nonresponse, item nonresponse or missing data also is a source of difficulty for the data analyst. Again, if the percentage of item nonresponse is small — say, less than 5 to 10 percent — it probably will not have much of an effect on the data analysis. In this case, the observations with the missing values may be deleted from the analysis. As the percentage of missing data increases, there is increasing concern about the representativeness of the sample persons remaining in the analysis. Because of this concern, statisticians have developed methods for *imputing* or creating values for the missing data (Kalton 1983). By imputing values, it is no longer necessary to delete the sample persons with the missing data from the analysis. The imputation methods range from the very simple to the complex, depending on the amount of auxiliary data available.

As an example, suppose that in a survey to estimate the per capita expenditure for health care, we decided to substitute the respondents' sample average for those with a missing value on this variable. That is a reasonable imputation. However, since age is highly related to health care expenditures, if we know the age of the sample persons, a better imputation would be to use the average expenditure from respondents in the same age group. There are other variables that could be used with age that would be even better than using age alone — for example, the combination of age and health insurance status. The sample mean from the respondents in the same age and health insurance group should be an even better estimate of the missing value than the mean from the age group or the overall mean. In using any imputation method, we must remember that the number of observations is really the number of sample persons with no missing data for the analysis performed, not the number of sample persons. We must also realize that we are assuming that the mean of the group is a reasonable value to substitute for the missing value. Using the mean smoothes the data and likely reduces variability.

Other more complicated procedures are also available. However, none of these procedures guarantee that the value substituted for the missing data is correct. It is possible that the use of imputation procedures can lead to wrong conclusions being drawn from the data. Again, the best procedure for dealing with missing data is preventive — that is, make every effort to avoid missing data in the data collection process.

6.4 Designed Experiments

Designed experiments have been used in biostatistics in the evaluation of (1) the efficacy and safety of drugs or medical procedures, (2) the effectiveness and cost of different health care delivery systems, and (3) the effect of exposure to possible carcinogens. In the following, we present the principles underlying such experiments. Limitations of experiments and ethical issues related to experiments, especially when applied to humans, are also raised. Let us consider a couple of examples to illustrate the essential points in the experimental design.

Example 6.11

The Hypertension Detection and Follow-up Program (HDFP 1979) was a community-based, clinical trial conducted in the early 1970s by the National Heart, Lung and Blood Institute (NHLBI) with cooperation of 14 clinical centers and other supporting groups. The purpose of the trial was to assess the effectiveness of treating hypertension, a major risk factor for several different forms of heart disease. For this trial, it was decided that the major outcome variable would be total mortality.

At the time of designing the HDFP trial, results of a Veterans Administration (VA) Cooperative Study were known. This study had already demonstrated the effectiveness of antihypertensive drugs in reducing morbidity and mortality due to hypertension among middle-aged men with sustained elevated blood pressure. However, the VA study included only a subset of the entire community. Applicability of its findings to those with undetected hypertension in the community, to women and to minority persons was uncertain. Therefore, it was decided to perform a study, the HDFP study, in the general community. Instead of including only people who knew that they had high blood pressure, subjects were recruited by screening people in the community.

In this clinical trial, antihypertensive therapy was the *independent or predictor variable* and the mortality rate was the *dependent or response variable*. To determine the effectiveness of the antihypertensive therapy, a comparison group was required. Thus, the study was intended to have a *treatment group* — those who received the therapy — and a *control group* — those who did not receive the therapy. However, this classic experimental design could not be used. Since the antihypertensive therapy was already known to be effective, it could not ethically be withheld from the control group. Recognizing this, the HDFP investigators decided to compare a systematic antihypertensive therapy given to those in the treatment group (Stepped Care) to the therapy received from their usual sources of care for those in the control group (Regular Care). As a result, no one was denied treatment.

Example 6.12

In Chapter 3 we introduced a data set from the Digitalis Investigation Group trial. The primary objective of the DIG trial was to determine the effect of digoxin as the cause of mortality in patients with clinical heart failure who were in sinus rhythm and whose ejection fraction was ≤0.45. A total of 302 clinical centers in the United States and Canada enrolled 7788 patients between February 1991 and September 1993 and follow-up continued until December 1995 (DIG 1995).

Eligible patients were recruited and randomized to either digoxin or placebo (dummy pill) using a random block size method (to be explained later) within each clinical center; 3889 to digoxin and 3899 to placebo. This large sample size was required to detect a 12 percent reduction in mortality by treatment and to take non-compliance into account. The trial was double blinded (both investigators and patients were not informed about the group assignment). We discuss these essential feathers of an experimental design in this section.

6.4.1 Comparison Groups and Randomization

A simple experiment may be conducted without any comparison group. For example, a newly developed AIDS education course was taught to a class of ninth graders in a high school for a semester. The level of knowledge regarding AIDS was tested before and after the course to assess the effect of the course on students' knowledge. The difference in test scores between the pre- and posttests would be taken as the effect of the instructional program. However, it may be inappropriate to attribute the change in scores to the instructional program. The change may be entirely or partially due to some influence outside the AIDS course — for example, mass media coverage of AIDS-related information. Therefore, we have to realize that when this simple experimental design is used, the outside influence, if any, is mixed with the effect of the course and it is not possible to separate them.

Thus, in studying the effect of an independent variable on a dependent variable, we have to be aware of the possible influence of an extraneous variable(s) on the dependent variable. When the effects of the independent variable and the extraneous variable cannot be separated, the variables are said to be *confounded*. In observational studies such as sample surveys, all variables are confounded with one another and the analytical task is to untangle the comingled influence of many variables that are measured at the same time. In experimental studies, the effects of extraneous variables are separated from the effect of the independent variable by adopting an appropriate design.

The basic tool for separating the influence of extraneous variables from that of the independent variable is the use of comparison groups. For example, giving the treatment to one of two equivalent groups of subjects and withholding it from the other group means that the observed difference in the outcome variable between the two groups can be attributed to the effect of the treatment. In this design, any extraneous variables would presumably influence both groups equally, and, thus, the difference between the two groups would not be influenced by the extraneous variables. The key to the successful use of this design is that the groups being compared are really equivalent before the experiment begins.

Matching is one method that is used in an attempt to make groups equivalent. For example, subjects are often matched on age, gender, race, and other characteristics, and then one member of each matched pair receives the treatment and the other does not. However, it is difficult to match subjects on many variables, and also, the researcher may not know all the important variables that should be used in the matching process. A method for dealing with these difficulties with matching is the use of *randomization*.

Randomization is the random assignment of subjects to groups. By using randomization, the researcher is attempting to (1) eliminate intentional or nonintentional selection bias — for example, the assignment of healthier subjects to the treatment group and sicker subjects to the control group; and (2) remove the effect of any extraneous variables. With large samples, the random assignment of subjects to groups should cause the distributions of the extraneous variables to be equivalent in each group, thus removing their effects.

6.4.2 Random Assignment

One way of randomly assigning subjects to groups is the use of the random sampling without replacement procedure discussed in the earlier section.

Example 6.13

Consider the case of randomly assigning 50 subjects to two groups. An SRS (without replacement) of 25 from the 50 sequentially numbered subjects is selected using a computer package (see **Program Note 6.1** on the website):

```
 2    4    5    6    11    12    16    17    18    20
21   25   16   27    30    31    32    33    35    36
40   41   44   47    48
```

These subjects are assigned to the treatment group and the remaining 25 subjects form the control group. In many randomized experiments, subjects are assigned to the groups sequentially as soon as subjects are identified, as in the HDFP trial. In that case, the preceding results can be put into the following sequence of letters T (treatment group) and C (control group) that can be used to show the assignment:

C	T	C	T	T	T	C	C	C	C
T	T	C	C	C	T	T	T	C	T
T	C	C	C	T	T	T	C	C	T
T	T	T	C	T	T	C	C	C	T
T	C	C	T	C	C	T	T	C	C

If one were to assign 60 subjects to three groups, the first random sample of 20 will be assigned to the first group, the second random sample of 20 to the second group, and the remaining 20 subjects to the third group.

The method of random allocation illustrated in Example 6.13 poses some problem when the subjects are to be randomized in sequence as they are recruited in a clinical trial because the total number of eligible patients is not known in advance. As a result, it is difficult to balance the size of comparison groups. For example, the sequence of letters T and C in Example 6.13 works fine if 50 eligible patients can be recruited in a clinical center. But if only 10 eligible patients are available, then there are 4 Ts and 6 Cs in the first 10 letters in the sequence, making the size of comparison groups unbalanced.

An alternative to the preceding method is a *random block size method* (or random permuted blocks method). This is the randomization method used in the DIG trial described in Example 6.12. We illustrate an example for the blocks of size 4. We can list all the different possible sequences of allocations of four successive patients containing two Ts and two Cs as follows:

1. T T C C
2. T C T C
3. T C C T
4. C T T C
5. C T C T
6. C C T T

Blocks are then chosen at random by selecting random numbers between 1 and 6. This could be done, for example, with a fair die or by using Table B1, ignoring the digits 7, 8, 9, and 0. The first five eligible random digits from the first row of Table B1 are 1, 1, 4, 5, and 4. By choosing the previous corresponding blocks, we have the following sequence of allocations:

T, T, C, C, T, T, C, C, C, T, T, C, C, T, C, T, C, T, T, C.

It may be possible for an investigator to discover the pattern when the block size is small. To alleviate this problem, the block size is often changed in as allocation proceeds. It will be difficult to discover a pattern of sequence when the block size is 10 or more.

6.4.3 Sample Size

The random assignment of subjects to groups does not guarantee the equivalence of the distributions of the extraneous variables in the groups. There must be a sufficiently large number of subjects in each group for randomization to have a high probability of causing the distributions of the extraneous variables to be similar across groups. As discussed earlier, use of larger random samples decreases the sample-to-sample variability and increases our confidence that the sample estimates are closer to the population parameters. In the same way, a greater number of subjects in the treatment and control groups increase our confidence that the two groups are equivalent with respect to all extraneous factors.

To make this point clearer, consider the following example. A sample of 10 adults is taken from the Second National Health and Nutrition Examination Survey (NHANES II) data file, and 5 of the 10 persons are randomly assigned to the treatment group, and the other 5 are assigned to the control group. The two groups are compared with respect to five characteristics. The same procedure is repeated for sample sizes of 40, 60, and 100, and the results are shown in Table 6.3.

Table 6.3 Comparison of treatment and control groups for different group sizes.

Characteristics[a]	Treatment	Control	Treatment	Control
	($n_1 = 5$)	($n_2 = 5$)	($n_1 = 20$)	($n_2 = 20$)
Percent male	60	20	60	35
Percent black	0	20	5	20
Mean years of education	12.6	11.2	12.9	13.0
Mean age	38.8	41.6	40.7	34.0
Percent smokers	60	40	27	23
	($n_1 = 30$)	($n_2 = 30$)	($n_1 = 50$)	($n_2 = 50$)
Percent male	43	50	42	44
Percent black	17	10	16	16
Mean years of education	12.7	12.9	11.7	12.5
Mean age	39.7	40.2	42.1	42.5
Percent smokers	32	35	34	34

[a] Observations are weighted using the NHANES II sampling weights.

The treatment and control groups are not very similar when *n* is 10. As the sample size increases, the treatment and control groups become more similar. When *n* is 100, the two groups are very similar. It appears that at least 30 to 50 persons are needed in each of the treatment and control groups for them to be reasonably similar. The sample size considerations will be discussed further in Chapters 7 and 9.

In the HDFP clinical trial shown in Example 6.11, over 10,000 hypertensive persons were screened through community surveys and included in the study. These subjects were randomly assigned to either the Stepped Care or Regular Care groups. Because of this random assignment and the large number of subjects included in the trial, the Stepped Care and the Regular Care groups were very similar with respect to many important characteristics at the beginning of the trial. Table 6.4 is a demonstration of the similarities. The randomization and the sufficiently large sample size also give us confidence that these two groups were equivalent with respect to other characteristics that are not listed in Table 6.4.

Table 6.4 Comparison of Stepped Care and Regular Care participants by selected characteristics at entry to the hypertension detection and follow-up program.

Characteristics (Number of Participants)	Stepped Care (5485)	Regular Care (5455)
Mean age in years	50.8	50.8
Percent black men	19.4	19.9
Percent black women	24.5	24.8
Mean systolic blood pressure, mmHg	159.0	158.5
Mean diastolic blood pressure, mmHg	101.1	101.1
Mean pulse beats/minute	81.7	82.2
Mean serum cholesterol, mg/dL	235.0	235.4
Mean plasma glucose, mg/dL	178.5	178.9
Percent smoking >10 cigarettes/day	25.6	26.2
Percent with history of stroke	2.5	2.5
Percent with history of myocardial infarction	5.1	5.2
Percent with history diabetes	6.6	7.5
Percent taking antihypertension medication	26.3	25.7

Source: HDFP, 1979

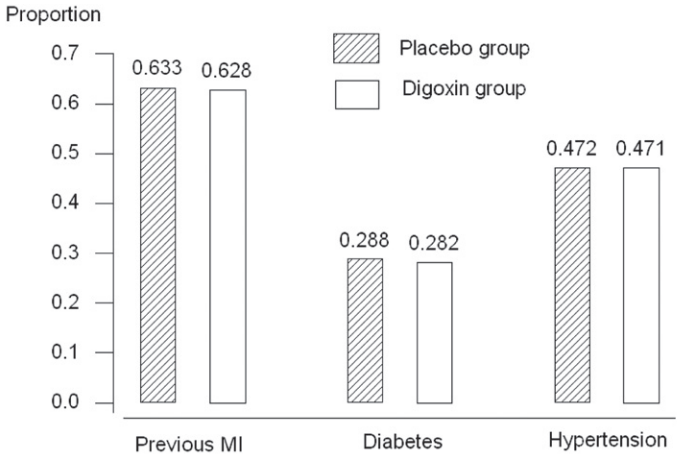

Figure 6.3 Proportions of patients with previous myocardial infarction, diabetes, and hypertension for placebo and digoxin groups in the DIG trial. *Source:* Digitalis Investigation Group, 1995

The DIG trial shown in Example 6.12 also used a large sample size recruited by clinical centers in the United States and Canada. The recruited patients were assigned to either placebo or digoxin treatment by a random block size method just described. As shown in Figure 6.3, medical history of the subjects with respect to myocardial infarction, diabetes, and hypertension is about the same, providing assurance that these and other clinical conditions would not pose as confounding factors for the comparison of two experimental groups.

The sample size required for an experiment depends on three factors: (1) the amount of variation among the experimental subjects, (2) the magnitude of the effect to be detected, and (3) the level of confidence associated with the study. When the experimental subjects are similar, a smaller sample size can be used than when the subjects differ. For example, a laboratory experiment using genetically engineered mice does not require as large a sample size as the same experiment using mice trapped in the wild. There is less likelihood of extraneous variables existing in the study using the genetically engineered mice. Hence, a smaller sample should be acceptable, since there is less need to control for extraneous variables. The fact that the sample size for the experiment depends on the size of the effect to be detected is not surprising. Since it should be more difficult to detect a small effect of the independent variable than a large effect, the sample size must reflect this. This is one of the reasons that the HDFP trial and DIG trial used a large sample size. Both trials attempt to detect a relatively small difference between the treatment and control groups. The relation between the sample size and the confidence associated with the study will be explored further in Chapters 7 and 8.

6.4.4 Single- and Double-Blind Experiments

So far we have been concerned with the statistical aspects of the design of an experiment. This means the use of comparison groups, the random assignment of subjects to the groups, and the need for an adequate number of subjects in the groups. An additional concern is the possible bias that can be introduced in an experiment. Let us consider some possible sources of bias and possible ways to avoid them.

In drug trials, particularly in those involving a placebo, the subjects are often *blinded* — that is, they are not informed whether they have received the active medication or a placebo. This is done because knowledge of which treatment has been provided may affect the subject's response. For example, those assigned to the control group may lose interest, whereas those receiving the active medication, because of expectations of a positive result, may react more positively. Studies in which the treatment providers know but the subjects are unaware of the group assignment are called *single-blind* experiments.

In most drug trials, both the subjects and the treatment providers are unaware of the group assignment. The treatment providers are blinded because they also have expectations about the reaction to the treatment. These expectations may affect how the experimenter measures or interprets the results of the experiment. When both the subjects and the experimenters are unaware of the group assignment, it is called a *double-blind* experiment.

Example 6.14

Let us examine one double-blind, randomized experiment conducted by a Veterans Administration research team (Goldman et al. 1988). They used the experimental design shown in Figure 6.4 to determine whether antiplatelet therapies improve saphenous vein graft patency after coronary artery bypass grafting.

In this experimental design, there are four treatment groups (four regimens of drug therapy) and a control group (placebo). Both the patients and the doctors were blinded, and only the designers of the trial, who were not directly involved in patient treatment, knew the group assignment. A total of 772 consenting patients were randomized, and postoperative treatment was started six hours after surgery and continued for one year.

As was to be expected, this experiment encountered problems in retaining subjects during the course of the experiment. The final analysis was based on 502 patients who underwent the late catheterization. These patients had a total of 1618 grafts. Of the 270 patients not included in the final analysis, 154 refused to undergo catheterization, 32 were lost to follow-up, 31 died during treatment, 42 had medical

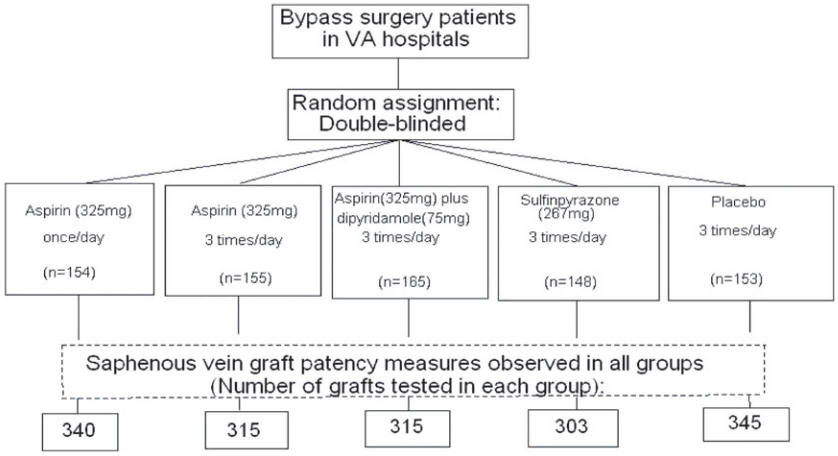

Figure 6.4 Experimental design for Veterans Administration Cooperative Study on Effect of Antiplatelet Therapy.
Source: Goldman et al., 1988

complications, and data on 11 patients were not available in the central laboratory (Goldman et al. 1989). Although we may expect that these problems are fairly evenly distributed among the groups because of the random assignment of subjects, the sample size was reduced considerably. This suggests that we needed to increase the initial sample size in anticipation of the loss of some subjects during the experiment.

There are other types of precautions that must be taken to avoid potential biases. In addition to statistical aspects, the experiment designer must provide detailed procedures for handling experimental subjects, monitoring compliance of all participants, and collecting data. For this purpose a study protocol must be developed, and the experimenter is responsible for adherence to the protocol by all participants. Similar to the problem of nonresponse in sample surveys, the integrity of experiments are often threatened by unexpected happenings such as the loss of subjects during the experiment and changes in the experimental environment. Steps must be taken to minimize such threats.

6.4.5 Blocking and Extraneous Variables

Thus far we have considered the simplest randomization, the random assignment of subjects to groups without any restriction. This design is known as a *completely randomized design*. The role of this design in experimental design is the same as that of the simple random sample design in survey sampling. As was mentioned earlier, in completely randomized designs, we attempt to remove the effects of extraneous variables by randomization. However, a reasonably large sample size is required before we can have confidence in the randomization process.

Another experimental design for eliminating the effects of extraneous variables known or thought to be related to the dependent variable uses *blocking*. Blocking means directly taking these extraneous variables into account in the design. For example, in a study of the effects of different diets on weight loss, subjects are often blocked or grouped into different initial weight categories. Within each block, the subjects are then randomly assigned to the different diets. We block based on initial weight because it is thought that weight loss may be related to the initial weight. Designs using blocking do not rely entirely on randomization to remove the effects of these important extraneous variables. Blocking guarantees that each diet has subjects with the same distribution of initial weights; randomization cannot guarantee this. Blocking in experiments is similar to stratification in sample surveys. The experimental design that uses blocks to control the effect of one extraneous variable is called a *randomized block design*. This name indicates that randomization is performed separately within each block.

Blocking is also used for administrative convenience. The VA Cooperative Study discussed in the previous section had 11 participating hospitals located throughout the United States. Since the subjects were randomized separately at each site, each participating hospital was a block. In this case, the blocking was done for administrative convenience while also controlling for the variation among hospitals.

In the previous section, we saw that the SRS design can be modified and extended as required to meet the demands of a wide variety of sampling situations. The completely randomized experimental design can similarly be expanded to accommodate many different needs in experimentation. Only one factor is considered in a completely randomized design. When two or more factors are considered, a *factorial design* can be used. For example, a clinical trial testing two different drugs simultaneously can be conducted in a 2 by 2 or 2^2 factorial design. Two levels of drug A and two levels of drug B will form four experimental groups: both drug A and B, A only, B only, and control (no drug). Study subjects are randomly assigned to the four groups. From this design we can examine the main effects of A and B as well as the interaction of two drugs.

The randomized block design just examined can also be expanded to accommodate more than one independent variable or block on more than one extraneous variable. We will discuss further these more complex experimental designs in Chapter 12.

6.4.6 Limitations of Experiments

The results of an experiment apply to the population from which the experimental subjects were selected. Sometimes this population may be very limited — for example, patients may be selected from only one hospital or from one clinic within the hospital. In situations like these, does this mean that we must perform similar experiments in many more hospitals to determine if the results can be generalized to a larger population — for example, to all patients with the condition being studied? From a statistical perspective, the answer is yes. However, if based on substantive reasons, we can argue that there is nothing unique about this hospital or clinic that should affect the experiment, then it may be possible to generalize the results to the larger population of all patients with the condition. This generalization is based on substantive reasoning, not on statistical principles.

For example, the results of the VA Cooperative Study may be valid only for male veterans. It certainly would be difficult to generalize the results to females without more information. It may be possible to generalize the results to all males who are known to have hypertension, but this requires careful scrutiny. We must know whether or not the

VA medical treatment of hypertension is comparable to that received by males in the general population. Does the fact that the men served in the military cause any difference, compared to those who were not in the military, in the effect of the medical intervention? If differences are suspected, then we should not generalize beyond the VA system.

On the other hand, the results of the HDFP should apply more widely, since the subjects were screened from random samples of residents in 14 different communities and then randomly assigned to the comparison groups. This use of accepted statistical principles of random sampling from the target population and randomizing these subjects to comparison groups makes it reasonable to generalize the results.

Another limitation of an experiment stems from its dependency on the experimental conditions (Deming 1975). Often experiments take place in a highly controlled, artificial environment, and the observed results may be confounded with these factors. Dr. Lewis Thomas's experience (Thomas 1984) is a case in point. While he was waiting to return home from Guam at the end of World War II, he conducted an experiment on several dozen rabbits left in a medical science animal house. He tested a mixed vaccine consisting of heat-killed streptococci and a homogenate of normal rabbit heart tissue, and the test produced spectacular and unequivocal results. All the rabbits that received the mixture of streptococci and heart tissue became ill and died within two weeks. The histologic sections of their hearts showed the most violent and diffuse myocarditis he had ever seen. The control rabbits injected with streptococci alone or with heart tissue alone remained healthy and showed no cardiac lesions. Upon returning to the Rockefeller Institute, he replicated the experiment using the Rockefeller stock of rabbits. He repeated the experiment over and over, but he never saw a single sick rabbit. One explanation for the spectacular results of the Guam experiment is that there may have been some type of a latent virus in the Guam rabbit colony. As Dr. Thomas said, "I had all the controls I needed. I wasn't bright enough to realize that Guam itself might be a control."

As Dr. Thomas's experience shows, we have to be careful not to deceive ourselves and extrapolate beyond our data. The experimental data consist of not only the observed difference between the treatment and control groups, but also the conditions and circumstances under which the experiment was conducted. These include the method of investigation, the time and place, and the duration of the test and other conditional factors. For example, in interpreting the results of drug trials, there is no statistical method by which to extrapolate the safety record of a drug beyond the period of the experiment, nor to a higher level of dosage, nor to other types of patients. The toxic effect of the medication may manifest itself only after a longer exposure, at higher levels of dosage or for other types of patients. Therefore, extrapolation of experimental results must be done with great care, if at all. Better than extrapolation is a replication of the study for different types of subjects under different conditions.

Implicit in the naming of experimental variables as being dependent and independent is the idea of cause and effect — that is, changes in the levels of the independent variables cause corresponding changes in the dependent variable. However, it is difficult to demonstrate a cause-and-effect relationship. It is sometimes possible to demonstrate this in very carefully designed experiments. However, in most situations in which statistics are used, positive results do not mean a cause-and-effect relationship but only the existence of an association between the dependent and independent variables.

Finally, statistical principles of experimentation can sometimes be in conflict with our cultural values and ethical standards. Experimenting, especially on human beings, can lead to many problems. If the experiment can potentially harm the subjects or impinge upon their privacy or individual rights, then serious ethical questions arise. The harm can be direct physical, psychological, or mental damage, or it may be the withholding of potential benefits. As was seen in the HDFP study, to avoid withholding the benefits of antihypertensive therapy, the study designers used the Regular Care group instead of a placebo group as the control group. When the potential direct harm is obvious, we cannot subject human beings to an experiment.

To protect human subjects from potential harm or from an invasion of privacy, an *informed consent* is required for experiments and even for interviews in sample surveys. This consent has to be voluntary. However, it is not difficult to recognize the possibility for pressuring patients to participate in a clinical trial. To prevent undue pressure being applied to patients or other potential study participants, all organizations receiving funds from the Federal government are required to have an institutional review committee (OSTP 1991). It is this committee's task to evaluate the study protocol to see if it provides adequate safeguards for the rights of the study participants.

6.5 Variations in Study Designs

As just seen, the essential characteristics of an experiment are that the investigators initially randomly assign the study subjects to the treatment and control groups (parallel comparison groups), administer the treatment, and observe what happens prospectively. Such an experimental design is hard to use in practice due to ethical and other practical reasons. The following quasi-experimental designs attempt to emulate an experimental situation, accommodating certain practical constraints.

6.5.1 The Crossover Design

The *crossover design* uses one group of experimental subjects, and each level of treatment is given at different times to each subject. The simplest crossover uses two periods and two levels of treatments. To control the effect of the order of applying two treatments, subjects are often randomly assigned group A and group B. Subjects in group A receive treatment 1 and subjects in group B receive treatment 2 in the first period. In the second period the treatments are switched. It is possible that a treatment effect in the first period can carry residual effects to the second period. In order to minimize the carryover effect, a *washout period* is often established between the two treatment periods. Some studies have a *run-in period* before the first treatment period begins, so as to wash out any residual effects of any previous medications. An obvious advantage of a crossover design is the cost saved by using a fewer subjects. This design, however, would require a long period of experimentation. The lack of a parallel comparison group and random assignment would make it difficult to distinguish the within-subject variation from the between-subject variation and to control the confounding effects.

Example 6.15

A 2 × 2 crossover design was used to compare ibuprofen (usual prescribed treatment) and lysine acetyl salicylate (Aspergesic, over-the-counter medicine) in the treatment of rheumatoid arthritis (Hill et al. 1990). Thirty-six patients were randomly assigned to the two equal treatment order groups at entry. After two weeks on their first treatment, patients crossed over to the other treatment, and two weeks later, the trial ended. There was no run-in or wash-out period, but the trial was double-blinded. They used the *double-dummy* procedure where each patient always received a combination of two pills, the appropriate active treatment plus a placebo that was indistinguishable from the other active treatment. The treatment periods were of equal length and relatively short. At baseline a general medical examination was carried out, and the recorded baseline values of the two groups were found to be similar. At the end of each treatment period a clinic visit was made to assess grip strength, blood pressure, and so forth. During the treatment periods, patients recorded a pain assessment score on a 1–5 scale (1 = no pain, 2 = mild pain, 3 = moderate pain, 4 = severe pain, 5 = unbearable pain). Five patients withdrew from the trial, one patient was considered noncompliant, and one patient failed to report his scores for the second period. The average pain scores for the 29 patents were analyzed.

6.5.2 The Case-Control Design

The *case-control design* identifies a set of subjects with a disease or certain condition (the cases) and another set without the disease (the controls). These two groups are compared with respect to a risk factor (the treatment). In this study the outcome is specified first, and the risk factor is assessed retrospectively. Since the subjects are not assigned to the case and control groups, the effects of confounding factors are not controlled. To overcome this problem, the two groups are matched with respect to some confounders such as age and gender. Although the two groups may be comparable in terms of age and gender, the effect of other confounders still remain. Another drawback of the case-control design is that it cannot be used to measure the incidence or prevalence of the disease because of the retrospective assessment of the risk factor.

Example 6.16

In January 1984 six cases of Legionnaires' disease were reported to the health authority in Reading, United Kingdom (Anderson 1985). All of them became ill between December 15 and 19, 1983. After a thorough investigation, the authority discovered seven unreported cases. The cases had no obvious factor in common except for visiting the Reading town center just before their illness. A case-control study was conducted to compare exposure between the 13 cases and a selected set of 36 people without the disease (the controls). Frequencies of visiting the town center just before Christmas were high for both groups, but most of the cases and fewer controls visited the Butts Center shopping mall. This suggested that the Butts Center might be a source of the legionella bacterium. In their analysis of data, they matched the cases and controls with respect to age, sex, neighborhood, and mobility status. One case had one control, another case had two controls, and the remaining

11 cases had three controls. In this investigation the case-control approach appears to be the only one possible. The investigation needed to be conducted quickly in fear of further infections. Other study designs may be impractical in this type of situation.

6.5.3 The Cohort Study Design

In the *cohort study design*, study subjects are followed up through time to record incidences of disease. The simplest approach is to select two groups of subjects at the baseline. One group consists of subjects who possess some special attribute that is thought to be a possible risk factor for a disease of interest, while the other group does not. For instance, a group of coal miners who are free of lung cancer and a group of lung cancer–free farmers are followed up several years to record the lung cancer incidences. This is a prospective study design, mimicking conditions of an experiment, and thus we can measure incidence. However, in the cohort study design the study subjects are not assigned to the groups and confounding is not controlled. Similar to the case-control study, the groups can be matched with respect to selected confounders, but the matching cannot provide protection against all possible confounders.

Example 6.17

In a cohort study of 34,387 menopausal women in Iowa, intakes of vitamin A, C, and E were assessed in 1986 (Kushi 1996). In the period up to the end of 1992, 879 of these women were newly diagnosed with breast cancer. The investigators examined the effect of vitamin use and level of intake for each vitamin on breast cancer incidence. Since women were not randomly assigned to vitamin use and level of intake groups, confounding factors were not controlled effectively. However, this study provided valuable information for further investigation. Randomization might not be possible for ethical and practical reasons.

These alternative study designs are widely used in epidemiological studies because of ethical and practical reasons. But the analysis of data from these studies requires the use of special methods and the analytical results need to be interpreted recognizing the limitations of the study design used. The data from matched studies need to be analyzed taking into account the matching. These methods will be discussed in the subsequent chapters. Data from these studies will be used to illustrate the methods of analysis in subsequent chapters.

Conclusion

In this chapter we saw how to collect data using sample surveys and designed experiments. We examined the use of a chance mechanism in drawing samples and assigning experimental subjects to comparison groups. We also presented some practical issues that cause more complicated sample designs to be used and experimental designs to be modified. Regardless of the complexity of the sample design, as long as we know the selection probability, we can infer from the sample to the population. A requirement of a good experimental design is that it reduces the chance of extraneous variables being confounded with the experimental variables. Randomization and blocking are basic tools for preventing this confounding. When these tools are used appropriately, it is possible to analyze the data to determine whether or not it is likely that an association exists between the dependent variable and the independent variables. Analysis of data from complex surveys would require special considerations, which we will discuss in Chapter 15. Experimental data are analyzed using the designs described here in Chapters 8 and 12. Even after performing the experiment appropriately, care must be used in interpreting the experimental results. We must not unduly extrapolate the findings from our experiment, but recognize that replication may be necessary for the appropriate generalization to the target population.

EXERCISES

6.1 Choose the most appropriate response from the choices listed after each question.

a. To determine whether a given set of data is a random sample from a defined population, one must ___________.
__ analyze the data.
__ know the procedure used to select the sample.
__ use a mathematical proof.

b. A simple random sample is a sample chosen in such a way that every unit in the population has a(n) ___________ chance of being selected into the sample.
__ equal
__ unequal
__ known

c. In the random number table, Appendix Table B1, approximately what percent of numbers are 9 or 2?
__ 20
__ 10
__ unknown

d. Sampling with replacement from a large population gives virtually the same result as sampling without replacement.
__ true
__ false

e. In a stratified random sample, the selection probability for each element within a stratum is _____.
__ equal.
__ unequal.
__ unknown.

f. A probability sample is a sample chosen in such a way that each possible sample has a(n) _________ chance of being selected.
__ equal
__ unequal
__ known
__ unknown

6.2 If a population has 2000 members in it, how would you use Table B-1, the table of random numbers, to select a simple random sample of size 25? Assume that the 2000 members in the population have been assigned numbers from 0 to 1999. Beginning with the first row in Table B1, select the 25 subjects for the sample.

6.3 In the following situations, do you consider the selected sample to be a simple random sample? Provide your reasoning for your answer.

a. A college administrator wishes to investigate students' attitudes concerning the college's health services program. A 10 percent random sample is to be selected by distributing questionnaires to students whose student ID number ends with a 5.

b. A medical researcher randomly selected five letters from the alphabet and abstracted data from the charts of patients whose surnames start with any of those five letters.

6.4 In the NHANES II, 27 percent of the target sample did not undergo the health examination. In the examined sample, the weighted estimate of the percent overweight was 25.7 percent (NCHS 1992).

a. Assuming that these data were collected via an SRS, what is the range for the percent overweight in the target sample?

b. Should any portion of the population be excluded in the measurement of overweight?

6.5 Discuss how sampling can be used in the following situations by defining (1) the population, (2) the unit from which data will be obtained, (3) the unit to be used in sampling, and (4) the sample selection procedure:

a. A student is interested in estimating the total number of words in this book.

b. A city planner is interested in estimating the proportion of passenger cars that have only one occupant during rush hours.

c. A county public health officer is interested in estimating the proportion of dogs that have been vaccinated against rabies.

6.6 For each of the following situations discuss whether or not random sampling is used appropriately and why the use of random sampling is important:

a. A doctor selected every 20th file from medical charts arranged alphabetically to estimate the percent of patients who have not had any clinic visits during the past 24 months.

b. A city public health veterinarian randomly selected 50 out of 500 street corners and designated a resident at each corner to count the number of stray dogs for one week. He multiplied the number of stray dogs counted at the 50 corners by 10 as an estimate of the number of stray dogs in the city.

c. A hospital administrator reported to the board of directors that his extensive conversations with two randomly selected technicians revealed no evidence of support for a walkout by hospital technicians this year.

6.7 An epidemiologist wishes to estimate the average length of hospitalization for cancer patients discharged from the hospitals in her region of the country. There are 500 hospitals with the number of beds ranging from 30 to 1200 in the region.

a. Discuss what difficulties the researcher might encounter in drawing a simple random sample.

b. Offer suggestions for drawing a random sample.

6.8 Discuss the advantages and disadvantages of the following sampling frames for a survey of the immunization levels of preschool children:

a. Telephone directory

b. The list of children in kindergarten

c. The list of registered voters

6.9 Discuss the interpretation of the following surveys:

a. A mail survey was conducted of 1000 U.S. executives and plant managers. After a month, 112 responses had been received. The report of the survey results stated that Japan, Germany, and South Korea were viewed as being better competitors than the U.S. in the world economy. Also one-third of the managers did not believe their own operations were making competitive improvements.

b. A weekly magazine reported that most American workers are satisfied with the amount of paid vacation they are allowed to take. This conclusion was based on the results of a telephone poll of 522 full-time employees (margin of error is plus or minus 4%; "Not sure" omitted). The question asked was "Should you have more time off or is the amount of vacation you have fair?"

More time off	33%
Current amount fair	62%

6.10 Choose the most appropriate response from the choices listed under each question:

a. Which of the following is not required in an experiment?

__ designation of independent and dependent variables

__ random selection of the subjects from the population

__ use of a control group

__ random assignment of the subjects to groups

b. The main purpose of randomization is to balance between experimental groups the effects of extraneous variables that are ____________________.

__ known to the researcher.

__ not known to the researcher.

__ both known and unknown to the researcher.

c. The experimental groups obtained by randomization may fail to be equivalent to each other, especially when ______________________________.

__ the sample size is very small.

__ blocking is not used.

__ matching is not used.

d. Which, if any, of the following is an inappropriate analogy between random sampling and randomized experiments?

__ simple random sampling–completely randomized experiment

__ stratified random sampling–randomized complete block design

__ random selection–random assignment

e. A randomized experiment is intended to eliminate the effect of ____________________.

__ independent variable.

__ confounded extraneous variables.

__ dependent variable.

f. If the number of subjects randomly assigned to experimental groups increases, then the treatment and control groups are likely to be ____________________.

__ more similar to each other.

__ less similar to each other.

__ neither of the above.

6.11 A middle school principal wants to implement a newly developed health education curriculum for 30 classes of 7th graders that are taught by 6 teachers. However, the available budget for teacher training and resource material is sufficient for implementing the new course in only half of the classes. A teacher suggests that an experiment can be designed to compare the effectiveness of the new and old curricula.

a. Design an experiment to make this comparison, explaining how you would carry out the random assignment of classes and what precautions you would take to minimize hidden bias.

b. How would you select teachers for the new curriculum?

6.12 To examine the effect of the seat belt laws on traffic accident casualties, the National Highway Traffic Safety Administration compared fatalities among those jurisdictions that were covered by seat belt laws (the Covered Group) with those jurisdictions that were not covered by seat belt laws (the Other Group). They found that among the Covered Group, 24 belt law jurisdictions, fatalities were 6.6 percent lower than the number forecasted from past trends. In the Other Group, observed fatalities were 2 percent above the forecasted level (Campbell and Campbell 1988).

a. Explain whether or not you attribute the difference between these two groups to seat belt laws.

b. Provide some possible extraneous variables that might have influenced the effect difference and explain why these variables may have had an effect.

6.13 A large-scale experiment was carried out in 1954 to test the effectiveness of the Salk poliomyelitis vaccine (Francis et al. 1955). After a considerable debate, the randomized placebo (double-blind) design was used in approximately half of the participating areas and the "observed control" design was used in the remaining areas. In the latter areas, children in the second grade were vaccinated and children in the first and third grades were considered as controls (no random assignment was used). In both areas, volunteers participated in the study, but polio cases were monitored among all children in participating areas. The following results were announced on April 12, 1955, at the University of Michigan:

Study Type and Group	Study Subjects	Polio Case Rate[a] (per 100,000)			
		Total	Paralytic	Nonparalytic	Fatal
Placebo Control Areas:					
Vaccinated	200,745	28	16	12	0
Placebo	201,229	71	57	13	2
Not inoculated[b]	338,778	46	36	11	0
Observed Control Areas:					
Vaccinated	221,998	25	17	8	0
Controls	725,173	54	46	8	2
Not inoculated[c]	123,605	44	35	9	0

[a] Based on confirmed cases

[b] Nonvolunteers in the participating areas

[c] Second graders not inoculated

Source: Reference 19, Tables 2 and 3

a. Why was it necessary to use so many subjects in this trial?

b. What extraneous variables could have been confounded with the vaccination in the observed control areas?

6.14 To test whether or not oat bran cereal diet lowers serum lipid concentrations (as compared with a corn flakes diet), an experiment was conducted (Anderson et al. 1990). In this experiment 12 men with undesirably high serum total-cholesterol concentrations were randomly assigned to one of the two diets upon admission to the metabolic ward. After completing the first diet for two weeks, the subjects were switched to the other diet for another two weeks. This is a crossover design in which each subject received both diets in sequence. Eight subjects were hospitalized in the metabolic ward for a continuous four-week period, and the remaining subjects were allowed a short leave of absence, ranging from 3 to 14 days, between diet regiments for family emergencies or holidays. The results indicated that the oat bran cereal diet compared with the corn flakes diet lowered serum total-cholesterol and serum LDL-cholesterol concentrations significantly by 5.4 percent and 8.5 percent, respectively.

a. Discuss how this crossover design is different from the two-group comparison design studied in this chapter. What are the advantages of a crossover design?

b. The nutritional effects of the first diet may persist during the administration of the second diet. Is the carryover effect effectively controlled in this experiment?

c. Discuss any other factors that may have been confounded with the type of cereal.

6.15 To determine the efficacy of six different antihypertensive drugs in lowering blood pressure, a large experiment was conducted at 15 clinics (Materson 1993). After a washout phase lasting four to eight weeks (using a placebo without informing the subjects), a total of 1292 male veterans whose diastolic blood pressure was between 95 and 109 mmHg were randomly assigned in a double-blind manner to one of the six drugs or a placebo. Each medication was prepared in three dose levels (low, medium, and high). The average age of the subjects was 59; 48 percent were black, and 71 percent were already on antihypertensive treatment at screening. All medications were started at the lowest dose, and the dose was increased every two weeks, as required, until a diastolic blood pressure of less than 90 mmHg was reached without intolerance to the drug on two consecutive visits or until the maximal drug dose was reached.

The blood pressure measurement during treatment was taken as the mean of the blood pressures recorded during the last two visits. The following table shows the number of subjects assigned, the number that withdrew during the treatment and the results on reduction in diastolic blood pressure:

Experimental Group	Number Assigned	Number Withdrawn	Reduction in diastolic BP: Mean	Std	% Success*
1. Hydrochlorothiazide	188	15	10	6	57
2. Atenlol	178	16	12	6	65
3. Captopril	188	23	10	7	54
4. Clonidine	178	13	12	6	65
5. Diltiazem	185	12	14	5	75
6. Prazosin	188	29	11	7	56
7. Placebo	187	29	5	7	33
Total	1292	137			

*Proportion of patients reaching the target blood pressure (diastolic blood pressure <90 mmHg)

Source: Reference 22, Tables 2 and 3, Figure 1

a. Discuss why the patients were not informed about the use of a placebo during the initial washout period.

b. More than 10 percent of the subjects withdrew from the study during the treatment, and there were more withdrawals in some groups than in other groups. Discuss how the withdrawals may affect the experimental results.

c. Discuss how widely you can generalize the results of this experiment.

6.16 A randomized trial was conducted to test the effects of an educational program to reduce the use of psychoactive drugs in nursing homes. Six matched pairs of nursing homes were selected for this trial. The matching was based on the size of nursing home, type of ownership, and level of drug use. Professional staff and aides participated in an educational program at one randomly selected nursing home in each pair. At baseline, the drug use status was determined for all residents of the nursing homes ($n = 823$), and a blinded observer performed standardized clinical assessments of the residents who were taking psychoactive medications. After the five-month program, drug use and patient clinical status were reassessed and the educational program was found to have reduced the use of psychoactive drugs in the nursing homes (Avorn et al. 1992).

a. How would you characterize the experimental design used in this study?

b. If the effectiveness of the educational program is related to the organizational and leadership types of the nursing home staff, is the effect of this confounder effectively controlled in this study? If not, how would you modify the experimental design?

c. Obviously not all the nursing homes that could be matched were included in this study. How might this limitation affect the study findings?

d. Discuss to what extent the study findings can be extrapolated to nursing homes in other states.

REFERENCES

Anderson, J. W., D. B. Spencer, C. C. Hamilton, et al. "Oat-Bran Cereal Lowers Serum Total and LDL Cholesterol in Hypercholesterolemic Men." *American Journal of Clinical Nutrition* 52:495–499, 1990.

Anderson, P., C. Bartlett, G. Cook, and M. Woodward. "Legionnaires Disease in Reading — Possible Association with a Cooling Tower." *Community Medicine* 7:202–207, 1985.

Avorn, J., S. B. Soumerai, D. E. Everitt, et al. "A Randomized Trial of a Program to Reduce the Use of Psychoactive Drugs in Nursing Homes." *The New England Journal of Medicine* 327:168–173, 1992.

Campbell, B. J., and F. A. Campbell. “Injury Reduction and Belt Use Associated with Occupant Restraint Laws.” In Graham, J. D. (ed.) *Preventing Automobile Injury: New Findings from Evaluation Research*, Chapter 2. Dover, MA: Auburn House Publishing Co., 1988.

Deming, W. E. “On Probability As a Basis For Action.” *The American Statistician* 29:146–152, 1975.

Digitalis Investigation Group. “Rationale, Design, Implementation, and Baseline Characteristics of Patients in the DIG Trial: A Large, Simple, Long-Term Trial to Evaluate the Effect of Digitalis on Mortality in Heart Failure.” *Controlled Clinical Trials* 17:77–97, 1995.

Fienberg, S. E. “Randomization and Social Affairs: The 1970 Draft Lottery.” *Science* 171:255–261, 1971.

Francis, T., Jr., R. F. Korns, R. B. Voight, et al. “An Evaluation of the 1954 Poliomyelitis Vaccine Trials: Summary Report.” *American Journal of Public Health* 45, Supplement:1–63, 1955.

Goldman, S., J. Copeland, T. Moritz, et al. “Improvement in Early Saphenous Vein Graft Patency After Coronary Artery Bypass Surgery with Antiplatelet Therapy: Results of a Veterans Administration Cooperative Study.” *Circulation* 77:1324–1332, 1988.

Goldman, S., J. Copeland, T. Moritz, et al. “Saphenous Vein Graft Patency 1 Year After Coronary Artery Bypass Surgery and Effects of Antiplatelet Therapy: Results of a Veterans Administration Cooperative Study.” *Circulation* 80:1190–1197, 1989.

Government Accounting Office. “Nutrition Monitoring: Mismanagement of Nutrition Survey Has Resulted in Questionable Data.” GAO/RCED-91-117, 1991.

Hill, J., H. A. Bird, G. C. Fenn, C. E. Lee, M. Woodward, and V. Wright. “A Double Blind Crossover Study to Compare Lysine Acetyl Salicylate (Aspergesic) with Ibuprofen in the Treatment of Rheumatoid Arthritis.” *Journal of Clinical Pharmacologic Therapeutics*, 15:205–211, 1990.

Hypertension Detection and Follow-up Program (HDFP) Cooperative Group. “Five-Year Findings of the Hypertension Detection and Follow-up Program,” *Journal of the American Medical Association* 242:2562–2571, 1979.

Kalton, G. *Compensating for Missing Survey Data*. Research Report Series, Institute for Social Research, the University of Michigan, 1983.

Kushi, L. H., R. M. Fee, T. A. Sellers, W. Zheng, and A. R. Folsom. “Intake of Vitamins A, C, and E and Postmenopausal Breast Cancer. The Iowa Women’s Health Study.” *American Journal of Epidemiology* 144:165–174, 1996.

Materson, B. J., D. J. Reda, W. C. Cushman, et al. “Single-Drug Therapy for Hypertension in Men: A Comparison of Six Antihypertensive Agents with Placebo.” *The New England Journal of Medicine* 328:914–921, 1993.

National Center for Health Statistics: Plan and Operation of the Health and Nutrition Examination Survey, United States, 1971–73. *Vital and Health Statistics*. Series 1, No. 10a. DHEW Pub. No. (HSM) 73-1310, 1973.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile*. Hyattsville, MD: Public Health Service. DHHS Pub. No. 92-1232, 1992.

The NHLBI Task Force on Blood Pressure Control in Children. “The Report of the Second Task Force on Blood Pressure Control in Children, 1987.” *Pediatrics* 79:1–25, 1987.

Office of Science and Technology Policy (OSTP). “Federal Policy for the Protection of Human Subjects: Notices and Rules.” *Federal Register* 56:28003–28032, 1991.

Thomas, L. *The Youngest Science: Notes of a Medicine Watcher*. New York: Bantam Books, 1984.

Tippett, L. H. C. “Random Sampling Numbers”. *Tracks of Computers.* No. 15. Ed. E. S. Pearson, Cambridge University Press, 1927.

Waksberg, J. “Sampling Methods for Random Digit Dialing.” *Journal of the American Statistical Association* 73:40–46, 1978.