CHAPTER 1

Variation, Control, and Bias

1.1 INTRODUCTION

This book deals with a class of studies, primarily in human populations, that have two characteristics:

1. The objective is to study the causal effects of certain agents, procedures, treatments, or programs.
2. For one reason or another, the investigator cannot use controlled experimentation, that is, the investigator cannot impose on a subject, or withhold from the subject, a procedure or treatment whose effects he desires to discover, or cannot assign subjects at random to different procedures.

In recent years the number of such studies has multiplied in government, medicine, public health, education, social science, and operations research. Examples include studies of the effects of habit-forming drugs, of contraceptive devices, of welfare or educational programs, of immunization programs, of air pollution, and so forth. The growing area of program evaluation has also caused more studies. Everywhere, administrative bodies —central, regional, and local—devote resources to new programs intended in some way to benefit part or all of the population, or to combat social evils. A business organization may institute changes in its operations in the hope of improving the way the business is run. The idea has spread that it is wise to plan, from the beginning of the program, to allocate some of the resources to try to measure both the intended and any major unintended effects of the program. This evaluation assists in judging whether the program should be expanded, continued at its current level, changed in some way, or discontinued.

Such studies will be called *observational*, because the investigator is restricted to taking selected observations or measurements that seem appropriate for the objectives, either by gathering new data or using those already collected by someone else.

The type of observational study described in this book is restricted in its objectives. The study concentrates on a small number of procedures, programs, or treatments, often only one, and takes one or more response measurements in order to estimate the effects. Examples include studies of the effect of wearing lap seat belts on the amount and type of injury sustained in automobile accidents; studies of the amount learned by people watching an educational television program; studies of the death rates and causes of death among people who smoke different amounts; and the National Halothane Study, which compared the death rates associated with the use of the five leading anesthetics in U.S. hospital operations. The objectives here are close to those in controlled experiments, leading some writers to call such studies either quasi experiments or the more disapproving term, pseudo experiments.

A basic difference between observational studies and controlled experiments is that the groups of people whom the investigator wishes to compare are already selected by some means not chosen by the investigator. They may, for example, be self-selected, as with smokers or wearers of seat belts; selected by other people, as in the choice of anesthetic used in an operation; or determined by various natural forces, as in the comparison of premature-birth children with normal-term children. In general, the investigator is limited to two choices. First, he may have a choice of different contrasting groups from which to draw his samples for comparison. For instance, for a comparison of people residing in heavily air-polluted areas with people living in relatively clear air, contrasting areas exist in different residential parts of many cities, and the investigator may select from several areas within the same city. Second, having selected contrasting groups, the investigator may have greater or less flexibility in the kinds of samples which can be drawn and measured for statistical analysis.

This class of observational studies may be distinguished from another class, sometimes called *analytical surveys*, that have broader and more exploratory objectives. The investigator takes a sample survey of a population of interest and conducts statistical analyses of the relations between variables of interest to him. A famous example is Kinsey's study (1948) of the relations between the frequencies of certain types of sexual behavior and the age, sex, social level, religious affiliation, rural–urban background, and social mobility (up or down) of the person involved. The Coleman Report (1966), based on a nationwide sample of schools, dealt with the question: to what extent do minority children in the Unites States (Blacks, Puerto Ricans, Mexican-Americans, American Indians, and Orientals) receive a poorer education in public schools than the majority whites? As part of the study, an extensive statistical analysis was made of the relation between school achievement and characteristics of the school (e.g., teachers, facilities), the child's aspirations and self-concept, and the home background. Much descriptive information was also obtained about the extent of racial segregation, the types of school facilities, and so forth.

Such analytical surveys vary in the extent to which the primary interest is in causal relations. The relations discovered in the statistical analyses often suggest possible causal hypotheses, later to be investigated more directly in the observational studies of the first class. Sometimes, causal hypotheses that appear plausible are adopted as a basis for action. The long-term Framingham Study [Dawber (1980)] took a sample of about 4500 middle-aged men, made numerous measurements on each man of variables that might be related to the development of heart disease, and followed the men for years to discover which men developed heart disease. Men who were obese, heavy cigarette smokers, and with high blood pressure were found to have the highest frequency of heart disease. Along with other data, these results were influential in leading to attempted control of these three variables as a standard preventive measure in medical practice.

1.2 STRATEGY IN CONTROLLED EXPERIMENTS—SAMPLED AND TARGET POPULATIONS

Although our concern is not with controlled experiments, it is worthwhile, for two reasons, to consider in succeeding sections the strategy that has been developed in controlled experimentation with variable material. First, the problems that face the experimenter are, in general, the same as those that face the investigator in an observational study. Second, the controlled experimenter has more power to study causal effects, and the techniques developed have been more fully worked out and described. Thus we may consider the two questions: What aspects of the approach in controlled experimentation can usefully be borrowed for observational studies? What are the most difficult problems?

The field of agriculture, in which the modern technique of experimentation with variable material was first developed, is convenient for illustration. Suppose that the objective is to compare the yield per acre of a new variety of a crop with a standard variety. If the new (N) and the standard (S) variety are each grown on a number of plots in the same field, the yield per plot will be found to vary from plot to plot. This variation immediately raises the problem: How far can we trust the mean difference $\bar{y}_N - \bar{y}_S$ over n plots of each variety as an estimate of the superiority of the new variety? The experimenter knows that if he increased n or decreased it, the quantity $\bar{y}_N - \bar{y}_S$ would change.

After some false starts, this problem was finally handled roughly as follows. At least conceptually, an experiment could be so large that the difference $\bar{y}_N - \bar{y}_S$ would finally assume a fixed value, say $\mu_N - \mu_S$. The observed $\bar{y}_N - \bar{y}_S$ from n repetitions is regarded not as an absolute quantity, but as an *estimate* of the value $\mu_N - \mu_S$ obtained for the population of repetitions of the trials. The theory of probability and a simple mathematical model were then invoked to prove that under certain assumptions the estimate $\bar{y}_N - \bar{y}_S$ is normally distributed about $\mu_N - \mu_S$ with standard error $\sqrt{2}\,\sigma/\sqrt{n}$, where σ^2 is the variance in yield from plot to plot. Student later removed the difficulty encountered when the investigator does not know σ, by showing that under the same assumptions, $(\bar{y}_N - \bar{y}_S)/(\sqrt{2}\,s/\sqrt{n})$ follows the t distribution, where s is the estimate of σ from the experiment. (The rules for calculating s and its number of degrees of freedom depend on the detailed structure of the experiment.) This theory led to tests of significance and confidence intervals for $\bar{y}_N - \bar{y}_S$ as tools in the interpretation of the results.

In the theory that led to these results, one basic assumption required is that the repetitions in the experiment are a random sample of the population of repetitions. To put it the other way round, statistical inferences about $\bar{y}_N - \bar{y}_S$ by these methods apply to the population of repetitions of which the experiment is a random sample. This immediately raises the question: Is this the population to which the experimenter would like the results to apply? The answer must usually be "no," particularly when the n repetitions completely fill the field, so that the population of indefinitely many repetitions is purely conceptual. In experiments and observational studies, the terms "the sampled population" (to denote this population of repetitions) and "the target population" (to denote the population for which the objective of the research is to make inferences) are useful. In experimental research the sampled population is nearly always much narrower and more restricted than the target population. Thus a medical experiment on the treatment of a disease may be done on the patients having the appropriate diagnosis, who turn up in a certain ward or clinic of a certain hospital in a certain six-month period. Experiments in behavioral psychology are often conducted using graduate students, and other volunteer students (paid or unpaid) in a university's psychology department. The target populations may be all patients in a certain age range with this diagnosis or all young persons in a certain age range.

Experimenters seldom have the resources or the interest (this is not their area of expertise) to conduct their experiments on a random sample of the target population. A partial exception occurs in certain problems in agriculture. For instance, the initial comparisons of the new (N) and the standard (S) varieties may be done on small plots at an agricultural experiment station. Small plots are used because with good design σ becomes smaller and n larger on a given area of land, making $\sqrt{2}\,\sigma/\sqrt{n}$, the standard error (SE), smaller. Uniform fields and good husbandry also lead to decreased σ. It is known, however, that comparisons $\bar{y}_N - \bar{y}_S$ for the sampled population of small plots at an experiment station (which often achieves higher than average yields) may not necessarily hold for the target population of farmers' fields in which the farmer may consider replacing S by N. Consequently, the purpose of the small-plot experiment is sometimes regarded as primarily to pick out promising new varieties. An N which beats S convincingly at the experiment station will then be compared with S on more nearly life-size plots at a sample of farmers' fields. This sample is seldom drawn strictly at random, since both willingness and some skill are required of the farmer; but the objective is to sample the range of conditions that occur in farmers' fields. These trials may be continued for several years to sample climatic variations.

In this example, moving from the sampled to the target population involves very substantial additional expenditure of resources and time after the original experiment or experiments have finished. Sometimes a step in this direction is taken by cooperative work between investigators of the same general problem. Cooperative experiments on the treatment of leprosy, for instance, were conducted simultaneously, with the same plan, treatments, and measurement of response, at leprosaria in Japan, The Philippines, and South Africa, while the same has been done on rheumatic fever in experiments conducted in England and the United States. With human subjects, a casual sampling of a broader population can also occur if the results of an initial experiment by an investigator excite interest. Other experimenters in different places with different subjects repeat the experiment, perhaps with slightly different techniques, to see if they get similar results. After a lapse of time, a more broadly based summary of such experiments may permit conclusions more nearly applicable to the target population.

It is sometimes stated that observational studies are often in a stronger position than experiments, with regard to the gap between sampled and target population. Analytical surveys may collect for analysis a random sample of the actual target population so that, apart from problems of nonresponse, there is no gap. In the restricted causally oriented studies the situation varies. We may have to take a group of persons subject to some program and a comparison group not subject to the program where we can find them. But sometimes, particularly in studies made from records, we can start with random samples of the immediate target population in constructing treated and nontreated groups.

More will be said on this problem later. At a minimum, the investigator should be aware of the nature of the target population when he selects comparison groups and should describe as clearly as possible the nature of the population that he believes he has sampled.

In this problem the experimenter has another weapon—factorial experimentation—that can be used to some extent in observational studies, provided that the composition and sizes of the samples are appropriate. In thinking whether to recommend N or S to the farmer, the experimenter knows that some farmers will apply one, two, or three of the common fertilizers, say sulphate of ammonia (S.A.) (supplying nitrogen), super phosphate [supplying phosphorus (P)], or potassium chloride [supplying potassium (K)]. Some sow the seed at heavier rates than others. Should the recommendations depend on the individual farmer's practices?

In dealing with this problem the experimenter might use what is called a 2^5 factorial experiment. There are now $2^5 = 32$ treatments, consisting of all combinations that can be made from the two levels of each factor, perhaps

$$\left\{\begin{matrix}\text{No S.A.}\\ \text{S.A.}\end{matrix}\right\}\left\{\begin{matrix}\text{No P}\\ \text{P}\end{matrix}\right\}\left\{\begin{matrix}\text{No K}\\ \text{K}\end{matrix}\right\}\left\{\begin{matrix}S_1\\ S_2\end{matrix}\right\}\left\{\begin{matrix}\text{Variety } N\\ \text{Variety } S\end{matrix}\right\}$$

A single replication of the experiment now requires 32 plots instead of 2. But within this replication N and S have been compared separately in each of the 16 combinations of levels of the other four factors. Thus for the *average* difference between N and S we obtain 16 comparisons from the 32 plots, just as if the experiment was nonfactorial but had 16 replications. The same is true of the average effects of S.A., P, and K and for the coverage difference between the two seeding rates S_1 and S_2.

We come to the question: Is the difference $\bar{y}_N - \bar{y}_S$ affected by the presence or absence of S.A.? The experimenter has 8 comparisons of N and S when S.A. is not applied and 8 comparisons when S.A. is applied, and their averages are comparable with respect to P, K, and the seeding rates. The experimenter can therefore estimate and test for significance the difference

$$(\bar{y}_N - \bar{y}_S)_{\text{S.A.}} - (\bar{y}_N - \bar{y}_S)_{\text{No S.A.}}$$

The same type of comparison can be made with respect to the effects of P, K, and different seeding rates on $\bar{y}_N - \bar{y}_S$, though the sample sizes per single

replication are now 8 instead of 16. This type of experiment and the resulting comparisons greatly help to broaden the basis for recommendations from experiments.

Consider an observational study in which a program is made available to certain subjects (P) but not to others (O). In trying to estimate the effect of the program on y, the investigator may wonder: Does the effect differ for men and women, for older and younger subjects, for persons with incomes above or below a certain level? In his statistical analysis he can compare $\bar{y}_P - \bar{y}_O$ in each of the eight subsamples formed by the combinations of the two levels of each of these other variables. The investigator then proceeds to estimate $\bar{y}_P - \bar{y}_O$ separately for men and women, older and younger persons, richer and poorer persons. Two difficulties arise: (1) The investigator will probably have unequal numbers of P and O subjects in each subsample, so that the analysis is more complex, involving multiple classifications with unequal numbers. (2) In some cells the numbers of P and O subjects may be so small that the comparisons of interest have poor precision and nothing very definite is learned. Nevertheless, it is useful to consider such variables as sex, age, and income, both in selecting the sample and in the analysis, and to try to determine the importance of attempting analyses of this kind, which may lead to sounder conclusions.

1.3 THE PRINCIPAL SOURCES OF VARIATION IN THE RESPONSES

To return to the discussion of strategy in experiments, the fact that the response y varies from plot to plot under both N and S implies, of course, that y is influenced by variables other than the treatment N or S. Commonly, there are numerous such sources of variation. Investigators, either in experiments or observational studies, who write down the sources that they know or suspect often end with an impressively long list. In considering such sources, the investigator may think of them as falling into one of three classes:

1. Sources whose effects the investigator tries to remove, wholly or partly, from the comparison $\bar{y}_N - \bar{y}_S$ by control during the course of the experiment or in the statistical analysis of the results.

2. Sources whose effects the investigator handles by randomization and replication. Randomization, unlike control, does not attempt to *remove* the effect of a source of variation, but instead makes this source act like a random variable, equally likely to favor N or S in any repetition. Consequently, if a given source of variation contributes an amount with standard

deviation, σ_1, to the variation in y, the contribution of this source to the SE of $\bar{y}_N - \bar{y}_S$ is $\sqrt{2}\,\sigma_1/\sqrt{n}$. By randomization and replication the contribution of any source can be made small if n is large enough.

3. Sources whose effects are neither controlled nor randomized. In an ideal experiment there should be no sources of this kind, and experimenters often forget the possible existence of these sources. If such sources happen to act like randomized variables in class 2, their effect is merely to increase σ. If, however, they are related to (confounded with) the treatment difference, they may give misleading results for which tests of significance are no protection. Well-known examples occur in medicine. If both the patient and the doctor measuring the response y know which treatment the patient has received, this may produce a biased overestimate $\bar{y}_N - \bar{y}_S$, especially if N is a new drug with an impressive name and S is simply bed rest. Whenever possible, medical experimenters go to considerable trouble to conduct "double blind" experiments, in which neither the patient nor the doctor measuring the response knows the treatment being measured for the patient. Another instance is the "novelty" effect. N may do well in the first experiment because it is a change from the usual routine, but later experiments (when N and S are both familiar) may show little difference. Sometimes a bias is introduced because of a wrong decision by the experimenter. Suppose that the measurement of y requires a complex laboratory analysis on a sample of the subject's blood, and that n is large enough so that two laboratories must be used. The experimenter sends all samples from N to lab 1 and all samples from S to lab 2. Any consistent difference between labs in the levels of y found when analyzing the same sample (and such differences are not uncommon) contributes a bias to $\bar{y}_N - \bar{y}_S$.

[Kish (1959) gives an excellent discussion of these sources of variation as they affect experiments and observational studies.]

1.4 METHODS OF CONTROL

In considering the variables whose effects on y should be removed by control, one might at first advise "so far as your knowledge of y permits, select for control those x variables that are the major contributors to the variation of y." Thus if y has a linear regression in the sampled population on each relevant x variable, this advice leads to selecting for control the x whose squared correlation ρ^2 with y is highest. If the regression model for some variable x is

$$y = \alpha + \beta(x - \mu) + e$$

where e is the residual term, it follows that

$$\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_e^2 = \rho^2\sigma_y^2 + (1 - \rho^2)\sigma_y^2$$

Successful removal of the effect of x by control reduces σ_y^2 to $\sigma_e^2 = (1 - \rho^2)\sigma_y^2$, the reduction being greatest when ρ^2 is greatest. But the decision to control or not to control with regard to x depends also on the inconvenience and expense that is required to control. With some variables it might be better in experiments to randomize with respect to x, thereby reducing its contribution to the SE of our comparison by extra replication.

The devices for control fall under three headings:

1. *Refinements of Technique.* Examples include use of intricate measuring instruments that reduce errors of measurement of y; of instruments that maintain constant temperature, humidity, and light throughout a laboratory; of experimental animals specially bred for uniformity (or of fields selected for test uniformity of yield). As has been noted, some of these refinements make the sampled population much more restricted than the target population. In general, such devices merely attempt to follow an old maxim for precise experimentation: "Keep everything constant except the difference in treatments." With variable material this maxim cannot be followed completely and the experimenter must decide to what extent he will try to follow it.

In observational studies there are some opportunities of this type also, particularly with standards of measurement and the choice between less- or more-variable groups in which to conduct the study. For instance, in the *Midtown Manhattan Study* (1962, 1975, 1977) interviews among male workers in New York City were used to conduct an analytical survey of the relationship between undiagnosed mental illness and various characteristics of the worker and his background. The investigators planned to use trained psychiatrists for the difficult measurement problems, but found that the scarcity of such specialists would limit their sample to a size much too small for their planned statistical analyses. Kinsey (1948) faced a similar conflict. Although his planned male sample size was 100,000, his high standards for the selection and training of interviewers restricted the interviewing force to a very small number.

2. *Blocking and Matching.* The idea is to arrange the experiment in separate individual replications and to try to keep the variables to be controlled constant *within each replication*. In this way, precise comparisons among the treatments can be made even if the controlled variables are far from constant throughout the available samples. In agricultural field experiments the replication is usually a compact block of land—hence the name

blocking—since it had been found that small plots close together tend to give similar yields. Additionally, operations like plowing, harvesting, and weighing the plot yields are applied in the same way at the same time within a block. Randomization is used in allotting the treatments to the individual plots in the block, with independent randomizations in each block.

The same blocking can be used to control several sources of variation simultaneously. For instance, an experiment was conducted to investigate whether injection of a certain chemical into rats would enable them to better withstand exposure to a poison gas. The two members of the same block were litter mates of the same sex, which made them of the same age and similar genetic constitution. They were put into a bell jar together into which poison gas was fed, the measurement being time to death, so that variations from trial to trial in the rate of feeding the gas affected each treatment equally. The two rats varied a little in weight, but random allocation of the treatment to the rats within a block made any advantage from selecting the heavier rat average out.

This technique is widely employed in observational studies, usually under the name *matching*. In comparing two groups of people under different programs, the investigator may judge that the response y will be affected by the age, sex, and educational level of the subjects. In matching, the investigator tries to find pairs of the same or nearly the same age, the same sex, and similar educational level. Matching may be performed with respect to a single variable or occasionally to as many as a dozen variables.

3. *Control During the Statistical Analysis.* This method is also widely used in observational studies. If the variables to be controlled are all classifications like sex, Republican, Democrat, or Independent, this control usually involves first calculating $\bar{y}_N - \bar{y}_S$ in each cell formed by these classifications as mentioned previously, and if appropriate, taking some weighted mean of the $\bar{y}_N - \bar{y}_S$ values. Since the controlled variables are constant within each cell, they do not affect this weighted mean.

If y and the x variables to be controlled are continuous, the investigator first constructs a regression model describing how the mean value of y depends on the x's. This method is often called the *analysis of covariance*. In the simplest case of a single x and a linear regression, the model for the sampled population in an experiment is (before any treatment effect is added)

$$y = \mu_y + \beta(x - \mu_x) + e$$

where e is a residual with mean 0 for any fixed x, representing the combined effects of uncontrolled randomized variables. If the treatments have effects τ_1 and τ_2, it follows that

$$\bar{y}_1 - \bar{y}_2 = \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2) + \bar{e}_1 - \bar{e}_2$$

The error in $\bar{y}_1 - \bar{y}_2$ as an estimate of $\tau_1 - \tau_2$ has a part $\beta(\bar{x}_1 - \bar{x}_2)$ due to this x variable and a part $\bar{e}_1 - \bar{e}_2$ due to other variables. Removal of the effect of x is done by computing a sample estimate b of β by standard regression methods, and replacing $\bar{y}_1 - \bar{y}_2$ by the adjusted estimate

$$(\bar{y}_1 - \bar{y}_2) - b(\bar{x}_1 - \bar{x}_2)$$

Since b will be subject to a sampling error and will not exactly equal β, the removal is not quite complete, but in large samples will be nearly so when all other assumptions hold. By use of multiple regression the method can adjust for more than one x variable and for curvilinear relations by including terms in x^2. In both experiments and surveys, regression adjustments for some variables can be combined with blocking or matching for others.

To summarize Sections 1.3 and 1.4, the experimenter tends to think of three types of variable, other than treatment differences deliberately introduced, that may affect the response y: (1) variables whose effects the experimenter tries to remove from $\bar{y}_1 - \bar{y}_2$ by control devices like blocking or adjustments in the analysis; (2) variables whose effects will be averaged out by randomization and replication, so that they create no systematic error or bias in $\bar{y}_1 - \bar{y}_2$ and that their effects are taken into account in the standard error associated with $\bar{y}_1 - \bar{y}_2$; and (3) variables neither controlled nor randomized. In some cases the experimenter believes that these variables act like randomized variables. Thus in some industrial experiments it is convenient and cost-effective to conduct and measure all replicates of a given treatment consecutively rather than randomizing this order among treatments. From the experimenter's knowledge of the chemical processes involved, the experimenter may argue that he sees no reason why this failure to randomize produces any systematic error in the comparison of the treatment means. In experiments in which numerous physical operations are required, the experimenter may argue that randomization at certain stages is a needless and perhaps troublesome step, although sometimes a single randomization, if planned from the beginning, can handle many variables simultaneously. In other cases devices like blindness may remove a bias likely to be related to the particular treatment, thus placing the variable in class (1) instead of (3).

In observational studies, randomization can sometimes be introduced at certain stages, with or without blocking. If several judges have to be employed to make a difficult subjective rating from the subjects' questionnaires, each judge might be assigned an equal-sized subsample of subjects from each treatment group, rated in random order to protect against systematic changes in the judge's levels of ratings, as has been claimed to occur in the marking of examinations. But such randomization usually handles only a few of the variables that affect y. Thus in observational studies there are normally only two classes—variables for which control is attempted and variables neither controlled nor randomized.

The techniques for control—matching and adjustment—therefore assume a more prominent role in observational studies than in experiments. Chapters 5 and 6 treat these topics systematically.

1.5 EFFECTS OF BIAS

For both theoretical and practical reasons, presented more fully in later chapters, the investigator may do well to adopt the attitude that, in general, estimates of the effect of a treatment or program from observational studies are likely to be biased. The number of variables affecting y on which control can be attempted is limited, and the controls may be only partially effective on these variables. One consequence of this vulnerability to bias is that the results of observational studies are open to dispute. Such disputes, often voluminous, may contribute little to understanding. One critic may believe that failure to adjust for x_4 made the results useless, while the investigator may believe that there is little risk of bias from x_4.

The investigator must use his judgment, assisted by any collateral evidence that he can find, in appraising the amount of bias that may be due to an x variable. This judgment is needed both in planning the variables to be controlled and in drawing conclusions. Often, the direction of a bias can be predicted. A television station may give a test to volunteer subscribers after an educational program on a certain topic, to estimate the amount learned from viewing the program. The volunteers included (1) some who viewed the program and (2) some who did not. It can usually be assumed that those who elected to see the program already knew more about the topic prior to the broadcast than those who did not see the program; therefore $\mu_{1y} > \mu_{2y}$. In studies of the possible inheritance of some forms of cancer, cancer patients may be better informed about cancer among relatives in the preceding generation than noncancer subjects.

In drawing conclusions, the investigator can sometimes reach a fairly firm judgment that the maximum bias is small relative to $\hat{\tau}_1 - \hat{\tau}_2$. This may have been the situation in studies of the relation between frequent cigarette smoking and the death rate from lung cancer. Cigarette smokers are self-selected, and numerous possible sources of bias in comparing them with nonsmokers have been mentioned in the literature, with supporting data from samples of the two populations. But the increase in the lung-cancer death rate for frequent cigarette smokers is so large that it is difficult to account for more than a small part of this increase by other differences in the two samples.

1.6 SUMMARY

Comparative observational studies in human populations have two distinguishing features: they address causal effects of certain treatments, and the data come from subjects in groups that have already been constituted by some means other than the investigator's choice.

Characteristically, the population from which samples are taken (the sampled population) is narrower than the population for which conclusions about treatment comparisons are desired (the target population). This commends that the investigator should give an account of the population sampled and how it may differ from the target population.

Issues are somewhat clarified by considering the process of arriving at (practical) conclusions in agricultural research, even though that research is based on experimental, not observational, studies. Experiments at an agricultural research station comparing two varieties, say new (N) and standard (S), result in estimates of treatment differences and statistical significance that relate to the particular fields and weather at the research station. Generalization from that sampled population to the more varied fields, weather, and diverse farming practices of the countryside is advanced by additional studies, over years, on farms in that target population. But generalization can also be aided during the experiment at the research station by using experiments that employ various levels of important factors that vary among farmers, such as the use of different seeding rates and various fertilizers. Such experiments lead to comparing outcomes in subsets of the data defined by combinations of these deliberately introduced factors. Similarly, it can be useful in observational studies to make comparisons between the treatments in subgroups that correspond to various combinations of important variables such as age, income level, and sex.

In experiments, variables that contribute to variation in outcome can be dealt with by randomization or control (or else ignored). In observational studies, since control usually offers the only alternative to ignoring influential variables, methods of control are quite important. There are three general methods of control: (1) refinement of techniques through devices such as training interviewers and improving questionnaires, (2) blocking and matching, and (3) statistical adjustments, such as regression.

Control can usually be attempted on only a few of the many variables that influence outcome; that control is likely to be only partially effective on these variables. Thus the investigator may do well to suppose that, in general, estimates from an observational study are likely to be biased. It is therefore worthwhile to think hard about what biases are most likely, and to think seriously about their sources, directions, and even their plausible magnitudes.

REFERENCES

Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. L. York (1966). *Equality of Educational Opportunity* (2 vols.). Office of Education, U.S. Department of Health, Education, and Welfare, U.S. Government Printing Office, Washington, D.C., No. OE-38001, Superintendent of Documents Catalog No. FS 5.238:38001.

Dawber, T. R. (1980). *The Framingham Study: The Epidemiology of Atherosclerotic Disease.* Harvard University Press, Cambridge, Massachusetts.

Kinsey, A. E., W. B. Pomeroy, and C. E. Martin (1948). *Sexual Behavior in the Human Male.* Saunders, Philadelphia.

Kish, L. (1959). Some statistical problems of research design. *Am. Sociological Rev.*, **24**, 328–338.

Srole, L., T. S. Langner, S. T. Michael, M. K. Opler, and Thomas A. C. Rennie, with Foreword by Alexander H. Leighton (1962). *Mental Health in the Metropolis: The Midtown Manhattan Study.* Thomas A. C. Rennie Series in Social Psychiatry, Vol. I. McGraw-Hill, New York.

Srole, L. and A. Kassen Fischer, Eds. (1975). *Mental Health in the Metropolis: The Midtown Manhattan Study.* Book One, Revised and Enlarged. Harper Torchbooks, Harper & Row, New York.

Srole, L. and A. Kassen Fischer, Eds. (1977). *Mental Health in the Metropolis: The Midtown Manhattan Study.* Book Two, Revised and Enlarged. Harper Torchbooks, Harper & Row, New York.