CHAPTER 3

Preliminary Aspects of Planning

3.1 INTRODUCTION

This chapter discusses some of the decisions that are faced in setting up an observational study designed to compare a limited number of groups of people. Since observational studies vary greatly, not all the points considered here will be relevant in a specific study; also, the issues that are most troublesome in some studies have been omitted in this chapter. The groups of people to be compared have been subjected to different experiences or agents, whose effects are the object of interest. We list below some examples of this type of study.

Experience or Agent	Effects or Response Variables
Wearing lap seat belts	Severity and type of injury in auto accidents
Head injury to child at birth	Performance in school
Viewing educational television program	Amount learned on the relevant topics
Urban-renewal program	Improvements by private landlords
Rise in taxes	Consumer spending and saving
Distribution of contraceptive device	Acceptability, birth rate
Town fluoridation of water	Status of children's teeth
Permissive or authoritarian kindergarten	Amount of quarreling among children
Smoking of cigarettes	Mortality and illness from specific causes

In controlled experiments the term *treatments* is often used for the agents which the experimenter applies in order to measure the agents' effects. Where appropriate we will continue to use this term to denote the different experiences or agents in different groups of people. The term *responses* will denote the measurements taken to throw light on the presumed effects of the treatments.

3.2 THE STATEMENT OF OBJECTIVES

Before planning actually begins, it is helpful to construct and have readily available as clear a statement as can be made about the objectives of the study. At first, the statement may have to be expressed in rather general language. As planning proceeds, the statement becomes helpful when decisions must be made about the treatments, responses, and other aspects of the conduct of the study, since one has an opportunity to check which choices seem most likely to aid in achieving the objectives of the study. In fact, without such a statement it is easy in a complex study to make later decisions that are costly and not particularly relevant to the original objectives; or worse still, the decisions may make the objectives actually harder to attain.

Some investigators like a statement in the form of a list of questions that the study is intended to answer; other investigators prefer a list of hypotheses about the expected causal effects of the agents. An example of the latter form occurs in a study by Buck et al. (1968) of the effects of the chewing of coca leaves by residents of four Peruvian Indian villages. This study lists three hypotheses:

> (1) Coca, by diminishing the sensation of hunger, has an unfavorable effect on the nutritional state of the habitual chewer. Malnutrition and conditions in which nutritional deficiencies are important disease determinants occur more frequently in coca chewers than among controls.
>
> (2) Coca chewing leads to a state of relative indifference which can result in inferior personal hygiene.
>
> (3) The work performance of coca chewers is lower than that of comparable non-chewers.

These three hypotheses are a good example of the general type of preliminary statement; the authors also explain that these particular effect

areas were selected because of a previous report on the consequences of coca chewing by a Commission of the United Nations. The statement does not specify either the treatments or the response variables, but provides a background against which alternative choices can be judged. Regarding the treatments, typical questions that arise in this type of study are as follows: Will this be a two-group study of coca chewers versus non chewers, or can any attempt be made to measure the level and duration of chewing per person, so as to give more-detailed knowledge of the dose–response relationship? Is it feasible to try to measure some effects separately for different kinds of people, for example, older and younger persons, or male and female? The choice of response variables obviously presents difficulties in this study, both in the number of possible variables that might be considered and in the question of the accuracy with which each variable can be measured.

This statement specifies only the direction of the effect of coca chewing on each of the three listed characteristics of chewers. From this statement, one might expect the statistical analysis of the results to consist mainly of tests of significance of differences between chewers and nonchewers, as indeed it does in the study cited. In initial studies a statement of directions of effects may be as far as the investigator feels able to go. But, reverting to the previous discussion of estimates and tests of significance (Sections 2.2 and 2.3), it is always worth considering how far the main interest should lie instead in estimating the sizes of the effects and forming some judgment about their importance.

The retort might be "Aren't the two objectives, estimation and testing significance, essentially similar? With a continuous response variable, the test criterion for two groups is $t = \bar{d}/s_{\bar{d}}$. This involves both the estimated size of difference $\bar{d}$ and its estimated standard error $s_{\bar{d}}$." This is so, but when the emphasis was concentrated on tests of significance, I have seen studies in which only the value of t, or only the value of χ^2 in the comparison of two percentages, was quoted in the presentation of results. The interested reader had to dig out the value of $\bar{d}$ from earlier parts of the studies, or found that $\bar{d}$ could not be calculated at all. There are also large-sample studies in which $\bar{d}$ was significant at the 1% level, and left the impression that the effect was large because of confusion between a highly significant difference and a large effect. Examination of $\bar{d}$ and $s_{\bar{d}}$, however, left the impression that the size of the effect was modest—perhaps of no great practical importance. Also, thinking in terms of the estimation of the sizes of effects tends to alert one more to possible sources of bias that must be handled than does thinking in terms of tests of significance.

This difference in point of view may also influence the approach to the statistical analysis. Suppose that there are four groups—the treatment

appearing at four different levels, $a_1, \ldots, a_4$. For a test of significance, the investigator may use the F test in a one-way analysis of variance, or the χ^2 test in a 2×4 table if the response is a $0 - 1$ variate. From the viewpoint of estimation, the objective becomes that of describing in its simplest terms the mathematical relationship, if any, between $\bar{y}_i$ or $\hat{p}_i$ and a_i. A simple approach would be to examine whether a straight-line relation appears to hold between $\bar{y}_i$ or $\hat{p}_i$ and a_i, or between simple transforms of these variables. This approach often produces a summary response curve that is more informative, and also provides a more powerful test of significance of the reality of the relationship.

As each decision is made in the later stages of planning, it is not a bad idea to record at the time the reason for the decision. This will be useful when the final report must be written, often a long time later, and may also be useful when facing related decisions. Moreover, in observational studies, the investigator sometimes knows so little about the merits of alternatives that when a decision is made he does not feel that the decision was well-informed; it may have been little better than tossing a coin. If this weak foundation for the decision is not recorded, the investigator is tempted at a much later date to invent a rationalization as to why this decision was a brilliant one.

When the statement of objectives has been constructed in as specific a form as possible, it leads to a number of questions about the actual conduct of the study: What locale is to be chosen? Where is the study to be done? (Often the locale has been tentatively selected before the investigator drafts his statement of objectives, but in exploratory studies on new possible causal agents the need for a study may be evident before any suitable locale has been found or even sought.) What aspects of the treatments or causal agents will it be necessary to measure? What measures of the responses will be taken? What groups of subjects are to be compared in order to appraise the effects of the treatments? What sample size is needed?

This chapter will consider questions relating to measurements of the treatments and the responses. Questions relating to comparisons and sample sizes will be considered in Chapter 4 and later chapters. Comments on the choice of a locale are best postponed until these other aspects of planning have been discussed, and are also given in Chapter 4.

3.3 THE TREATMENTS

In the simplest situations the treatment is a specific event, for example, wearing seat belts or viewing an educational television program, and it is

only a matter, in this case, of recording the subjects that wore seat belts or saw the program and those that did not. Often, however, the treatment is known to vary from subject to subject in amount or level on some scale. Regarding measurement of this level, at least three situations exist. (1) It may not be feasible to attempt any detailed measurement; the resources may permit only a comparison of a treated and an untreated group with no possibility of distinguishing between different levels. (2) The level of the treatment may be roughly the same for persons within the same subgroup of the study sample, but varies from one subgroup to another. Examples might be studies of the relationships between noise levels in different factory workshops (subgroups) and productivity, hearing losses, or irritability in the workers, or the relationship between air pollution and bronchitis in housewives living in different parts of a town. (3) It may be feasible to consider measurement of the treatment level separately for each person.

In situations (1) and (2), consider a two-group exploratory study (treated versus untreated) where the primary goal is to discover whether there appears to be an effect that might receive more careful investigation later. If the investigator is seeking a suitable treated group, or subgroup, it is advantageous to find one where the average level of treatment is high, so that a marked contrast is obtained. With two groups, the contrast is essentially the difference between the means. The power of a statistical test tells us the probability that the observed values will produce a statistically significant result.

Treatments often have approximately linear effects, either beneficial or deleterious, over the range of levels encountered in practice. Suppose that the true response to an amount τ of the treatment is $\beta\tau$. With treated and untreated groups each of size n, the value of $\delta/\sigma_{\bar{d}}$, the quantity that primarily determines the power of the t test, is $\sqrt{n}\,\beta\tau/\sqrt{2}\,\sigma_y$. Since n and τ enter into this formula in the forms $\sqrt{n}$ and τ, it follows that doubling τ (finding a treated group with a level twice as high) is equivalent to quadrupling n from the viewpoint of the power of the t test. If for a given level of τ the power of a two-sided t test at 5% significance level is 0.44, implying a less than 50–50 chance of revealing a true difference between treated and untreated groups, doubling τ increases this power to 0.91.

Of course, other factors also enter. For example, since people do not live in badly polluted air by choice, two areas offering the largest contrast in degree of pollution within a town may also show the largest difference in variables such as economic level and access to regular medical care. Thus areas with a large contrast in levels of treatment may also present a large contrast in potential sources of bias. But my judgment is that despite this difficulty a large contrast is a wise choice in exploratory studies.

What should the relative sizes of the two groups be? If the response y to an amount τ is roughly linear, $y = \alpha + \beta\tau$, and if the within-group variances σ_1^2 and σ_2^2 are roughly equal, experience in controlled experiments recommends two groups of equal size $N/2$, untreated and high-level. This plan minimizes the standard error of the estimate of β, the average change in y per unit increase in τ, for a given total number N of subjects. Incidentally, if there is evidence that σ_1^2 and σ_2^2 are substantially different in the two-group case, the optimum sample sizes are $n_1 = N\sigma_1/(\sigma_1 + \sigma_2)$ and $n_2 = N\sigma_2/(\sigma_1 + \sigma_2)$.

What about three subgroups—one untreated and the others having a high and an intermediate level. A third group, placed at one-half the high level, adds nothing to the estimate of β under these conditions. If the three groups are now of sizes $N/3$, the net effect of using three groups having equal variance instead of the two extremes is to increase the variance of the estimate of β by 50%. A third group whose level is somewhere between 30 and 70% of the high level adds little to the precision of the estimate of β. Even if we retained the end groups at size $N/2$ and added a third group of size $N/2$ whose level is somewhere between 30 and 70% of the high level, we would add little to the precision of the estimate of β.

The role of an intermediate level is to provide a test as to whether or not the line is straight. If the levels of the treatment are 0, ah, and h, the test is made by calculating the contrast $\bar{y}_2 - (1 - a)\bar{y}_1 - a\bar{y}_3$, which should be zero apart from sampling error if the line is straight. A t test for detecting a curve response is obtained by dividing this quantity by its estimated standard error.

The case for three levels rather than two is stronger in observational studies than in controlled experiments. An ever-present danger in observational studies is that some source of bias which affects the comparisons between groups has not been rendered unimportant by the plan and method of analysis. However, it is reassuring to find that $\bar{y}_2$ for the intermediate level lies between $\bar{y}_1$ and $\bar{y}_3$, though this result could still be produced by a bias which happened to operate in the expected direction of the treatment effect. If $\bar{y}_2$ was found to be less than $\bar{y}_1$, while $\bar{y}_3$ was greater, we would reexamine carefully both the suspected sources of bias and the arguments which led us to think that the response would be roughtly linear. For instance, in seeking comparison groups for radiologists (as a group exposed to a certain amount of radiation), Seltser and Sartwell (1965) deliberately chose two other groups: one group was ophthalmologists and otolaryngologists, who should have practically zero exposure, and the other group was physicians, who use x rays to some extent and have an intermediate level of exposure.

3.4 MEASUREMENT OF TREATMENT LEVELS FOR INDIVIDUAL PERSONS AND THE EFFECTS OF GROUPING

When the level of the treatment varies from person to person, it may be possible to measure the amount received by a given person more or less accurately. In an urban-renewal program in which sums were offered to repair rented houses, the amount given to each house might be known accurately. In studies of the aftereffects of the atom bomb dropped on Hiroshima, the dose of radiation received by each person has been estimated roughly from a person's memory of his location and the amount of nearby shielding when the bomb fell. In the smoking-health studies, the amount smoked by an individual may vary at different times, and both this and the number of years of smoking vary from person to person. The best single measure of the level of smoking is not clear even with full and accurate records. In the studies on smoking, the level used was a broad interval of number of cigarettes smoked per day (e.g., less than 10, 10–19, 20–39, 40 +), reported at the time when the study questionnaire was sent out.

Given an estimate a_i of the level of the treatment received by the ith person, one possibility is to analyze the relation between the response y_i and a_i using regression methods. An alternative (as in the studies on smoking) is to divide the range of a into a few classes, recording only the class into which each person falls. With two groups the sample is divided into upper and lower classes; with three groups, into low, middle, and high classes; and so forth.

When this method is under consideration, natural questions arise, such as: How many classes should be formed? Should they be made equal or unequal in numbers of persons? Is this method much inferior to the use of regression methods? Some answers can be given when there is a linear regression of y_i on a_i and the quantity of interest is the slope β of the line (average increase in y per unit increase in x). We assume that a total of N subjects are available so that with c classes the average class has N/c subjects. With the regression method, the estimate of β is $\hat{\beta} = \Sigma(y_i - \bar{y})(a_i - \bar{a})/\Sigma(a_i - \bar{a})^2$. We call the variance of this estimate based on the distributed values of $a, V(\hat{\beta})$. If, instead, two groups are formed, the estimate $\hat{\beta}_2 = (\bar{y}_2 - \bar{y}_1)/(\bar{a}_2 - \bar{a}_1)$. With three or more classes the regression of the $\bar{y}_c$ (class means of y) on the $\bar{a}_c$ (class means of a) is calculated to obtain the estimate $\hat{\beta}_3$ for this method. We call the variance of $\hat{\beta}_g$ based on the g groups $V(\hat{\beta}_g)$.

The values of $V(\hat{\beta}_g)/V(\hat{\beta})$ for $g = 2, 3, \ldots$ naturally depend on the shape of the distribution of the levels a_i. When this distribution is unimodal and roughly symmetrical, results for the normal distribution can be quoted,

since these results have been calculated by D. R. Cox (1957) for a similar purpose. In Table 3.4.1 the total number of persons, H, is constant and assumed large. For $g = 2$ to 5, the table shows the optimum class sizes, the corresponding values of $V(\hat{\beta}_g)/V(\hat{\beta})$, and the values of $V(\hat{\beta}_g)/V(\hat{\beta})$ for equal-sized classes. The table shows the increases in variance of the estimate when grouping is used. Thus grouping in two equal groups increases the variance by 57%. Another way of thinking about these numbers is that the efficiency of the grouped estimate compared with the ungrouped estimate is the reciprocal of the numbers given in the last two columns of the table. Thus for two classes, the relative efficiency of $\hat{\beta}_2$ compared with $\hat{\beta}$ is 100/1.57 or about 64%.

We conclude from Table 3.4.1 that (1) It pays to use more than two classes, although not much is gained by using more than four classes. The minimum variance of $\hat{\beta}_g$ with four classes is only 13% higher than the minimum variance attained by regression under these assumptions. (2) The class sizes that minimize $V(\hat{\beta}_g)$ are unequal; the central classes are substantially larger than the outside classes. (3) Nevertheless, the use of equal-sized classes (number of subjects) is only slightly less effective than the best set of classes; for example, 1.16 compared with 1.13 for $V(\hat{\beta}_g)/V(\hat{\beta})$ with four classes.

Investigations of several smooth nonnormal distributions of the a_i (Cochran, 1968) suggest that the results in Table 3.4.1 can also be used as a guide in such cases. However, there is a hint that with some nonnormal distributions, grouping into classes gives values of $V(\hat{\beta}_g)/V(\hat{\beta})$ that are slightly lower than with the normal distribution, thus losing less precision relative to regression methods than when the a_i are normal.

Thus far we have considered the loss in precision due to grouping into classes when the distribution of levels of a is smooth and unimodal. The loss

Table 3.4.1. Comparison of the Variance of $\hat{\beta}_g$ for g Optimally Grouped Classes by Level of Treatment into Classes with the Variance of $\hat{\beta}$ Based on the Normal Distribution for a Linear-Regression Situation

Number of Classes g	Optimum Class Sizes	$V(\hat{\beta}_g)/V(\hat{\beta})$	$V(\hat{\beta}_g)/V(\hat{\beta})$ for Equal-Sized Classes
2	$0.5N, 0.5N$	1.57	1.57
3	$0.27N, 0.46N, 0.27N$	1.23	1.26
4	$0.16N, 0.34N, 0.34N, 0.16N$	1.13	1.16
5	$0.11N, 0.24N, 0.30N, 0.24N, 0.11N$	1.09	1.11

is much less if the distribution has multiple modes, particularly if the classes can be centered around the modes. For instance, with cigarette smoking, about $\frac{1}{3}$ or more of the total samples were nonsmokers and it would not be surprising to find peaks around 10 and 20 cigarettes per day. Thus, if we formed two classes of smokers by amount smoked and a third class of nonsmokers, a rough calculation suggests that $V(\hat{\beta}_3)/V(\hat{\beta})$ is around 1.10–1.15 rather than 1.23 as shown in Table 3.4.1.

In practice the reasons for using a few classes at different levels in preference to a full regression analysis have seldom been explicitly stated by investigators. One may guess the reasons as being either (1) for simplicity in the measurement problem and in analysis and presentation, or (2) in some cases, for a judgment that if the amount a_i received by the ith person can be measured only roughly, full regression analysis would not gain much in precision over analysis by use of a few classes. This question has not been fully explored, but to a first approximation it looks as if the presence of errors of measurement in the a_i hurts both the regression and the classification methods to about the same extent, so that the ratios $V(\hat{\beta}_g)/V(\hat{\beta})$ do not change much.

3.5 OTHER POINTS RELATED TO TREATMENTS

This section considers a few miscellaneous points related to the definition of the treatments. Often the level of the treatment varies from person to person because use of the treatment is to some extent voluntary. A well-known example is the distribution of a contraceptive device to married women in a program in which it is planned to estimate the effectiveness of the program. In fact, although the women agree to cooperate, some women in the study sample may not use the device at all, some may use it minimally and inconsistently or use it for a time and then cease, and some may use it as recommended by the planners. From the viewpoint of public policy the main interest may lie in the subsequent birth rate for the sample as a whole, but it is obviously relevant to know to what extent the contraceptive device was actually used, to try to learn the reasons for limited use, and to obtain the birth rate for both the consistent and the inconsistent users. The plans for the study should include regular monitoring of the women in the study sample in order to obtain this information. The studies on smoking were careful when gathering data to distinguish between those who had never smoked and the ex-smokers (those who were not smoking at the time of receipt of the questionnaires, but who had smoked in the past). Comparison of death rates for ex-smokers with death rates for nonsmokers and current smokers, provided much useful information.

The description of the treatment may require setting up a special record-keeping system during the course of the study. To evaluate the worth of making available a limited amount of psychiatric guidance (psychiatric social workers in addition to some expert psychiatric counseling) to a number of families in a health plan, the investigator needs to know with which family members the guidance staff worked, what kinds of mental problems were revealed, the amount of guidance that was given, and so forth. Construction of a good record-keeping system for such purposes must be part of the initial plans.

Sometimes the treatment combines several agents. For example, educational campaign to encourage voter registration or inoculation of children might include announcements on radio and television, distribution of leaflets to houses, and some public lectures. The planning group will want to consider whether they can learn something about the part played by different aspects of the treatment, though this is usually difficult. At least, questions could be framed to ask whether people had been reached by different components, what they remember of the contents of these components, and what they think influenced them in one way or another.

The treatment may take a different form for some people in the study sample than for others. In a study of the effectiveness of wearing seat belts, some people involved in auto accidents may have been wearing both the lap belt and the over-the-shoulder–type belt. Head injuries to children at birth were found to vary in type. The type should obviously be recorded. Existence of more than one type raises the question: Shall we attempt to measure the effects of each variant? A relevant statistical aspect of this question is the following. If alternative forms of the treatment have effects in the same direction, a much larger sample size is needed to distinguish clearly between the sizes of the effects for different types than to measure the overall effect of the treatment, especially if some forms are used by only a small minority. The situation is discussed more specifically in Section 4.1 on sample size.

The best decision is a matter of judgment in a situation like this. If a treatment is used in two forms, some investigators prefer to include both forms, even if the sample size permits only a very imprecise comparison between the effects of the two forms. They argue that they will measure the overall effect of the two forms as they are used, and can at least check that the two forms do not have widely different effects. Others feel that if their sample size is such that the difference between the effects of the two forms is almost certain to be nonsignificant, a report of this finding may lead readers to assume that the two forms have been shown to have equivalent effects, and discourage further research on the minority form. They prefer a "clean" study restricted to the majority form only.

3.6 CONTROL TREATMENTS

Often the investigator first selects the treated group. In fact, an imaginative investigator may notice the existence of an interesting treated group, which gives rise to the study. The investigator then looks for a suitable group on whom this treatment is not acting. Such a group is quite commonly called a "control group," though some investigators prefer the more general term "comparison group," perhaps because the word "control" has some of the features of an advertising slogan, hinting at more power to remove biases than is usually possessed.

Ideally, the requirement for a control group is that it should differ from the treatment group only in that the treatment is absent. The choice of the control group should lead us to expect that if the treatment has no effect, the responses y should have the same mean and shape of distribution in the control group as in the treated group. In particular, the control group should be subject to any selective forces that are known to affect the treatment group and are not themselves possible consequences of the treatment. This last condition can be important. In a study of the effect of birth-related head injuries on children's performance in school at age 11, it would be unwise to have an uninjured control group deliberately chosen so that performance in class at age 6 was similar to that of the treatment group at age 6, unless it were known with certainty that birth-related head injuries had no effect on performance at age 6. Otherwise, a comparison between the treated and control groups at age 11 might have removed part or even all of the treatment effect.

The most common difficulty in the search for a control is that we cannot find a group known to be similar in all other respects to the treated group, particularly when use of the treatment is to some extent voluntary. When making a choice, it is advisable to list the apparent deficiencies of any proposed controls and to try to judge which control seems least likely to produce a major bias in the treated–control comparison. If no single control free from the danger of serious bias can be located, there is merit in having more than one control, particularly if the different controls are suspected of being vulnerable to different possible sources of bias. In this event, a finding that the treated group differs from all of the control groups in the same direction strengthens any claim that there is a real effect of the treatment.

Another requirement sometimes overlooked for a control is that the quality of measurement should be the same in treated and control groups. In a study of tuberculous and nontuberculous (control) families, the same caseworker was assigned to take the measurements in both families so as to guarantee similar quality of measurement. She warned, however, that her measurement would be both more complete and more accurate in the

tuberculous families, with whom she had worked for years, than with the control families who were specially recruited for this study and did not know her.

Some investigators like to use data taken from published statistics for the general public as a control, possibly because the general public represents the target population to which they would like the results to apply. Such controls are seldom satisfactory. Many treated groups that appear in studies have some special features apart from the treatment that make them not comparable to the general public. Further, with routinely gathered statistics the meaning of the measurements and their completeness and quality are often quite different from those that apply in a carefully planned study.

To carry this point further, an investigator may claim that a certain treatment is deleterious to health in males because the sickness rate in the treated group is say 20% higher than in the control group, a significant difference statistically. An opponent may state that this claim is unjustified because the sickness rates in the investigator's treated and control groups are *both* definitely lower than those for males with the same age distribution in the general public. The investigator would then retort that the treated group of men differed from men in the general public in certain known respects, that the control group was carefully chosen to differ in the same respects in order to be comparable with the treated group, and that comparison with data for the general public is logically irrelevant.

The investigator is correct, but the question arises: To what target population does the 20% increase in sickness rates apply? The investigator may reply that it applies to the kind of population of which the treated and control groups can be considered a random sample; but as mentioned in Chapter 2, this population may be of no particular interest from the viewpoint of public policy because of the selective forces that affect it, and may even be difficult to envisage. If the investigator claims that the 20% difference is applicable to a more general population of males, he is making a claim without producing supporting data, particularly if this is the first study of this treatment. This point affects a great many observational studies, because the groups that get studied are often specialized in several respects, for quite sound reasons. The same remarks apply to any statement of confidence limits, which account for only the amount of variation in the sampled population.

As studies in different locations by different workers accumulate, there is an opportunity to examine whether this 20% increase is also found in different populations. This is what happens in research on many problems, though often in a haphazard and informal way. The author of the study believes that although the levels of performance in other populations may differ, the size of treatment effect, difference or percent, will apply to other

populations. This hope has often, but by no means always, been realized in the past as we mention in the next section.

3.7 THE RESPONSES

The choice of response measurements (measurements to be used to compare the performances under different treatments) is obviously important. In selecting the responses, several points should be considered. Relevance to the stated objectives of the study is an obvious one. To cite an example, reported by Yates (1968), measures of the percentages of buildings destroyed, obtained at great labor from aerial photographs, were used to assess the effects of bombing raids on German cities on Germany's industrial production. For the stated objective this response variable was not particularly relevant, since early British bombing raids concentrated on the town centers, whereas the factories were mainly on the outskirts. A similar statistic constructed for factories gave quite different results and, in order to approach the objective more closely, could be combined with British data on the relationship between damage to factory buildings and actual decrease in factory production.

Different aspects of the responses may be relevant. In the above example, in considering measures of morale in studies of the effects of the bombing raids on morale, one proposed indicator was the proportion of daily absences from work without an obviously operative reason. Another approach involved using a battery of questions about the respondent's opinion on the state of the war, feelings of ability to cope with day-to-day problems, attitudes toward government, and so forth. Either or both types of approach might be advisable, depending on the study's range of objectives. In a study with an obvious primary response measurement, it is worth asking: Are there other aspects of the possible effect of the treatment that should be measured?

The same remark applies to subsidiary measurements that may provide insight on how a causal effect is produced or may strengthen or weaken the evidence that there is a causal effect. For example, an immunization campaign against diphtheria in young children in Britain in the 1940s was followed by scattered reports of paralytic poliomyelitis in some children. At that time there was no polio vaccine.

These reports might suggest that the usual inoculations in children increased their risk of subsequently contracting polio; yet, since many children had been inoculated, the appearance of *some* polio cases among the inoculated was not of itself surprising in the absence of any standard of comparison. Accordingly, in a study conducted in 1949, Hill and

Knowelden (1950) obtained two samples each of 164 children; one sample was of children with polio, and the control sample was of children without polio. The samples were matched for age, sex, and place of birth. The aim was to see whether a much higher proportion of the polio cases than the control sample had been inoculated. The investigators found that 96 of the polio cases and 83 of the control samples had received inoculations—a difference that was not striking and certainly not statistically significant.

As additional information, the investigators recorded (1) the dates of any inoculations, and (2) the sites (left arm, left leg, etc.) of the inoculations. This information enabled two further comparisons to be made. Of 17 inoculations made during the month immediately preceding the polio attack, 16 occurred in polio cases and 1 in the control sample. Of the inoculations made more than one month previous, 80 occurred in polio cases and 82 in the control sample. A second piece of information was provided by the polio cases alone. For inoculations made *more* than one month previous to the onset of paralysis, the site of the inoculation injection was also one of the sites of paralysis in 13 out of 65 cases, or 20%. For inoculations *less* than one month previous, the corresponding figures were 29 out of 36 cases, or 81%. These two results both pointed to an increased risk of paralytic polio from inoculations in the month preceding the time when polio becomes epidemic each year. Taken together with other data, these results led to a recommendation that doctors should avoid giving standard inoculations to children during this period.

The investigator should of course be aware of what is known about the quality of any proposed measuring process. Often, this aspect presents no problem. At the other extreme, there are instances in which a satisfactory method has not yet been developed for measuring a given type of response; therefore, sound research studies cannot proceed until some breakthrough in measurement has occurred. In such cases a common situation is that several different measuring techniques (e.g., by a battery of questions) have been developed, but it is not clear exactly what is being measured by each technique. If information on the agreement between different techniques is scanty, use of two of them on each person in a study is worth considering as a means of picking up useful comparative information for future studies.

Sometimes the process of measurement requires use of several similar instruments (e.g., interviewers, judges, raters, laboratories). A standard precaution is to have each judge measure the same proportion of subjects, preferably selected at random, from each treatment group. Whenever feasible, it is also worth the inconvenience to ensure that each judge is ignorant of the treatment group to which any subject belongs at the time when the judge is rating the subject. Otherwise, consistent differences between judges in levels of rating, or the judge's preconceived ideas of how the treatments

should rank in performance, may produce fallacious differences between treatments. These precautions are easily overlooked, as experience has shown, in a study that involves numerous detailed operating decisions.

The size and range of the study may determine the type and, therefore, the quality of the measuring instrument for responses. For instance, in a study of nonhospitalized mental ill-health, the choice might lie, depending on its size, between measurement by trained psychiatrists, by psychiatric nurses, or by a standard questionnaire administered by lay interviewers. In a large-sample study of school education, information about schoolwork conditions in the home and the parent's attitudes and aspirations with respect to their children's education, might have to be obtained from questions answered by the children in school because the available resources do not permit use of direct interviewing of parents in their homes.

It is difficult to advise to what extent the original aims and scope of a study should be deliberately restricted by sacrificing some of the original objectives in order to permit a higher quality of measurement. The weaknesses of the smaller-sample study are likely to be restricted range of questions, reduced precision, and limitation to a subpopulation much narrower than the population of interest. There is merit in a restricted study that can (1) indicate whether the case for an extensive study is strong or weak, (2) allow internal comparisons between the best-available measuring techniques and less-expensive ones that would have to be used in a broad study, and (3) provide valuable experience in the problems of conducting this type of study. The weaknesses of a large extensive study in which the investigators bit off more than they can chew are likely to be high nonresponse rates and measurements clearly vulnerable to bias, with the consequence that the main conclusions are subject to serious question. Its strength may be that it can attempt to sample the population of interest. Another weakness of the large study is that it may fall of its own weight and never be completed. Fortunately, these often die aborning.

In practice, the decision for or against a large study will be influenced also by the amount of public interest in the problem, the pressure for early results, and scale of operation that attracts the principal investigators.

3.8 TIMING OF MEASUREMENTS

Problems of timing the study measurements arise if the treatment is expected to have responses that last for some time into the future. Examples are the fluoridation of a town's water or a rise in the general tax rate.

Shortly before the treatment is applied, a relevant response variable is measured on a sample of children or families, for instance a survey of the numbers of decayed, missing, and filled teeth by age and sex, or of family spending and saving. The following question then arises: How long will these response variables be measured after the start of the treatment? If little is known in advance about the shape of the time–response curve, no definite answers can be given to this question, except perhaps to schedule at least two subsequent observations, to keep later study plans flexible, and to speculate about the likely nature of the time–response curve from a combination of theoretical ideas on the nature of the presumed causal process and of any related data from other studies.

Usually these first two measurements have to be scheduled, at least tentatively, at the time when the study is being planned and funds allotted or sought. When these measurements have been obtained, a mathematical model of the nature of the response curve may help in deciding on the desirability and timing of any subsequent measurements. Suppose that the response is expected to increase with time until it reaches some steady maximum value or asymptote. A response curve of this type is often well enough represented by the curve $Y = \mu_0 + \alpha(1 - e^{-\beta t})$, where μ_0 is the initial level at time $t = 0$. When t is large, $e^{-\beta t}$ becomes negligible, so that the parameter α represents the maximum increase. The parameter β indicates how quickly this maximum is reached. At time $t = 3/\beta$, for instance, the curve is within 5% of the maximum increase. If β is doubled, this 95% of the total increase is reached in half the time.

We assume a single treated group. From the initial sample mean and the means $\bar{y}_1$ and $\bar{y}_2$ of the two later responses, we can estimate μ, α, and β. From these estimates and the equation of the curve we can, in turn, estimate (1) how much benefit would be obtained from one or two additional future observations, and (2) the best later times at which to take these observations, in order to obtain as precise estimates as possible of α, β, or the course of the response curve. This use of theory might be particularly helpful if the calculations revealed that the first two posttreatment measurements had been taken much too soon, as might happen with a treatment whose effects are slow to appear.

If we had initially a rough idea (perhaps from other studies) of the time taken to reach 95% of the increase, from which the value of β could be estimated, this information would be useful in planning the times τ and 2τ of the first two posttreatment observations. For sketching the course of the curve, a good compromise is to take τ as the time when the increase has reached about 70% of its maximum. Taking τ at 50% of the maximum is a bit early for this curve.

3.9 SUMMARY

This chapter deals with preliminary aspects of studies intended to compare a limited number of groups of individuals. The groups are exposed to different agents or experience, called *treatments*, whose effects it is desired to study.

An initial written statement of the objectives of the study, with reasons for choosing these objectives, is essential. This may be in the form of questions to be answered, hypotheses to be tested, or effects to be estimated. This statement is helpful in selecting the locale of the study, the specific groups to be studied, the data needed to describe the treatments, the response variables, and the kind of statistical analysis required.

In exploratory studies there is a tendency to think mainly in terms of tests of significance, because the investigator feels that an objective of finding out whether there is *some* effect, and in which direction, is as much as can be accomplished. Thinking in terms of estimating the size of the effect and its practical importance may make the investigator more aware of potential sources of bias and may lead to a more informative statistical analysis.

When the level of treatment varies from group to group, a two-group study with levels widely apart is usually advisable in small initial studies. A third group at an intermediate level adds little or nothing to the estimation of a linear effect, but may either provide a test of linearity or serve as some reassurance about bias if the response is thought to be linear.

When the level of treatment varies from person to person and can be measured at least roughly for individual persons, a common practice is to divide the persons into two or more groups according to these individual levels. For the estimation of a linear effect of levels, the precision obtained with different numbers of groups, the best group sizes, and the relative precision given by groups of equal size are presented.

Special procedures or record keeping may be required when the treatment is complex or when exposure to it is to some extent voluntary. Issues relevant to the selection of an untreated comparison group, sometimes called a control group, are discussed.

In selecting measurements to be made of the responses in groups under different treatments, relevance to the objectives of the study is an obvious criterion. Another is the quality of the measurement. Different aspects of the responses, measurable by different variables, may deserve study. Subsidiary measurements may clarify the interpretation of the main results.

If the process of measurement requires several instruments of the same type (e.g., interviewers, judges, raters, laboratories), standard precautions are to have each judge measure the same proportion of subjects from each

treatment group and to conceal from any judge (whenever feasible) the particular group that he is measuring at any specific time.

Sometimes a treatment, such as fluoridation of a town's water or a rise in taxes, will be applied for a relatively long time. Decisions are required about the times at which the treated group is to be observed. If the general shape of the likely time–response curve can be envisaged, statistical theory may assist these decisions.

REFERENCES

Buck, A. A., et. al. (1968). Coca chewing and health. *Am. J. Epidemiol.*, **88**, 159–177.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295–313 [Collected Works #90].

Cox, D. R. (1957). Note on grouping. *J. Am. Stat. Assoc.*, **52**, 543–547.

Hill, A. B. and J. Knowelden (1950). Inoculation and poliomyelitis. *Br. Med. J.* **ii**, 1–16.

Seltser, R. and P. Sartwell (1965). The influence of occupational exposure to radiation on the mortality of American radiologists and other medical specialists. *Am. J. of Epidemiol.*, **81**, 2–22.

Yates, F. (1968). Theory and practice in statistics, *J. Roy. Statist. Soc.*, Ser. A, **131**, 463–477.