CHAPTER 7

Simple Study Structures

7.1 INTRODUCTION

This chapter discusses the structure of some of the simplest observational-study plans that have been used in the medical and social sciences, when the objective is to measure the effect of a given treatment. A more-extensive catalog of such plans appears in the monograph by Campbell and Stanley (1963/1966), whose primary purpose was to appraise the strengths and weaknesses of such plans in educational research.

7.2 THE SINGLE GROUP: MEASURED AFTER TREATMENT ONLY

The situation in which this is the only kind of data can arise if all persons available for study were exposed to the treatment, if no unexposed group at all comparable can be found, and if no response measurements previous to the application of treatment exist. With this plan there is no basis for a judgment about the effect of treatment, unless we can guess accurately enough what would have happened in the absence of treatment. This could be so, for instance, with an unusual natural calamity. An estimate of the number of deaths caused by an earthquake, based on a count of dead bodies, might be highly accurate, the principal error being the status of missing persons not accounted for, where the number expected to die during the time in question in the absence of an earthquake may be safely regarded as minor.

In the absence of any external comparison group, it may still be possible to learn something if (1) different persons were exposed to the treatment in different degrees, and if (2) it is possible to assign some score or measure to each person representing degree of exposure. This was roughly the situation

in the important studies of the aftereffects of the atomic bomb on survivors in Hiroshima. By combining a person's memory of location (distance from the epicenter) and of local shielding by buildings, the survivors could be divided into four exposure groups. Subsequent four-group morbidity and mortality is, of course, no longer a "single-group" study. The group furthest from the epicenter could, in fact, be regarded as scarcely exposed at all to unusual radiation. The main difficulty with this group was that ethnically and socially they appeared somewhat different from the other groups.

In fact, any method that occurs to me for attempting an inference about the effect of the treatment amounts to changing this plan into a different one, either by obtaining comparable measurements before the treatment was applied or by guessing or finding comparison groups exposed to different levels of treatment.

7.3 THE SINGLE GROUP: MEASURED BEFORE AND AFTER TREATMENT

This situation is probably more common than the preceding one as a device from which conclusions about the causal effects of a treatment are attempted. Some new program or law intended to be beneficial in certain respects, for example, fluoridation of water or a change in working conditions, is such that all people in a given community or agency are exposed. In an attempt to evaluate the effects of the program, relevant response variables are recorded either for the community or agency as a whole or for a sample of people, both before the introduction of the treatment and at a time afterward when the treatment is expected to have produced its major effects. It is important that the sample be a random sample and that strong efforts be made to keep nonresponse to a minimum. If the treatment is quick-acting we may plan to use the same sample of people before and after; if not, the samples may be drawn independently or with perhaps some matching, as described in Chapter 5, to secure greater comparability.

With this plan we at least obtain an estimate $\bar{y}_a - \bar{y}_b$ of the time change that took place in a response variable and can attach a standard error to $\bar{y}_a - \bar{y}_b$. There remains the problem of judging how much this change was due to the treatment or to other contributing causes.

With only two sets of observations—before and after—a basic difficulty is that even if $\bar{y}_a - \bar{y}_b$ is clearly statistically significant by the appropriate t test, this difference may be merely the size that commonly occurs due to the multiplicity of other causes that create time changes in the time interval τ involved. Thus even if the investigator can think of no specific alternative source that might provide a rival hypothesis as to the reason for the $\bar{y}_a - \bar{y}_b$

difference, this provides only a subjective basis for a claim that the difference was caused by the treatment. There must be some objective support of the judgment that $\bar{y}_a - \bar{y}_b$ is larger than commonly occurs in a time interval of this size. Seeking evidence for this judgment usually involves getting data from populations not subject to the treatment during this time interval; in other words, changing from a single-group to at least a two-group comparison. Data from such other populations is often helpful in judgment even if they cannot be regarded as strictly comparable. With a single group or population it helps also if a series of observations at intervals τ have been made both before and after treatment, since we get some information on the sizes of changes in $\bar{y}$ that occur in this time interval in the absence of treatment. This plan is discussed in the following section.

7.4 THE SINGLE GROUP: SERIES OF MEASUREMENTS BEFORE AND AFTER

We therefore need a presumption that $\bar{y}_a - \bar{y}_b$ is of a size or direction that calls for explanation by major causes peculiar to the interval in question, of which the treatment may be one. The nature of alternative contributing causes will, of course, depend greatly on the type of study and the time interval elapsed between the pretreatment and the posttreatment measurements. The following are some examples.

Campbell and Stanley report Collier's (1944) study (actually a two-group study) of the effect on students of reading Nazi propaganda in 1940. The fall of France, which occurred during the time interval, may have had a major effect on attitude changes.

With an economic program intended to improve people's well-being, their final reports on the program may differ in a period of generally rising prosperity from those in a period of declining prosperity.

In evaluation of a program intended to increase worker satisfaction with working conditions in an agency, the questions asked may suggest to the worker the type of response acceptable to the management, which may, in turn, affect some reported changes in attitudes. Ingenuity and effort to ensure that reports are both anonymous and believed to be anonymous and skill in selection and wording of questions may help.

One precaution, especially when considerable time elapses between the pretreatment and the posttreatment measurements, is to keep the definition of measurement used for the standard and that for the level of measurement of response the same, particularly where human judgment or relations keep us involved in the measuring process.

Possible effects of the pretreatment measurements or of the knowledge that such measurements are being made, must also be considered. This issue is familiar in studies of teaching methods or of new devices in learning, where the response measurements are some kind of examination and the same persons are measured before and after. Quite apart from any treatment effect, students may perform better in a second posttreatment test because they are more mature or more familiar with the type of test; they could perform somewhat worse in the test if they have become bored or antagonistic. In instructional material to farmers on farm-management practices where the measurements are made by a skilled economist, a comprehensive initial questionnaire may suggest ideas to some farmers that induce changes quite apart from those at which the instructional material is directed.

With regard to the knowledge that pretreatment measurements would be made, Emmett (1966) reports an effect on the measures rather than on the subjects. The study involved the effect of radio programs in London which encouraged parents to bring their children to doctors or clinics for the standard children's inoculations. The numbers of inoculations per week during the two weeks before the radio programs and the three weeks after the programs were aired were as follows:

	Inoculations Per Week	
	Two Weeks Before	Three Weeks After
118 general practitioners	209	130
5 clinics	85	79

A factor here was apparently that the doctors and clinics, alerted that the radio programs would take place, made special efforts in the two weeks before the programs to get children who were due for inoculations to come in so that the doctors or clinics would be free to handle the anticipated rush of patients into the office after the radio program.

If nonoverlapping samples are drawn for the pretreatment and posttreatment measurements, then distortion from the pretreatment measurement is unlikely to be large. The elapsed time between pre- and post-measurements or other considerations may dictate that independent measurements be used. For example, in a fluoridation study, we would probably want the data to refer to children of specific ages, who would be different individuals pretreatment and posttreatment. If there is worry about the effect of pretreatment measurement in the case of a treatment of short duration to

which all are exposed, it might sometimes be feasible to draw initially a sample of size $2n$—twice the planned size. This sample is then divided into random halves, one to be pretreatment measured and one posttreatment measured, or even into n matched pairs, of which one member at random is pretreatment measured and one posttreatment measured. The primary penalty from using independent samples is a decreased precision in $\bar{y}_a - \bar{y}_b$. Its variance under the simplest model is $2\sigma^2(1 - \rho)/n$ with identical subjects, where ρ is the intrasubject correlation, versus $2\sigma^2/n$ or $2\sigma^2(1 - \rho_m)/n$, where ρ_m is the correlation between members of a matched pair.

If the cost of measurement is not excessive, another possibility with a treatment to which all are exposed is as follows. Having created the two groups of n random or paired samples, measure the first n samples before treatment and all $2n$ samples after treatment. Let the means of the three groups of n measurements be $\bar{y}_{1b}$, $\bar{y}_{1a}$ and $\bar{y}_{2a}$. The comparison $\bar{y}_{1a} - \bar{y}_{2a}$ provides a test of the effect of pretreatment measurement. If an effect is detected, the effect of the treatment is estimated by the comparison $\bar{y}_{2a} - \bar{y}_{1b}$. If no pretreatment measurement effect is found, the mean $(\bar{y}_{2a} + \bar{y}_{1a})/2$ may be compared with $\bar{y}_{1b}$ to estimate the treatment effect. Apart from other possible sources of bias mentioned in this section, these estimates are subject to some bias because the type of estimate used is determined by the result of a preliminary test of significance, but this bias should be small in large samples. When no effect of pretreatment measurement exists, it is theoretically possible to obtain a more-precise estimate of the treatment effect by using a weighted, instead of an unweighted, mean of $\bar{y}_{1a}$ and $\bar{y}_{2a}$ to take fullest advantage of the correlation between $\bar{y}_{1b}$ and $\bar{y}_{1a}$ which refer to the same subjects. If ρ were known, under a constant variance σ^2, the best weighting would be $[(1 + \rho)\bar{y}_{1a} + (1 - \rho)\bar{y}_{2a}]/2$. But the potential gain over equal weighting does not become worthwhile until ρ reaches 0.6, the variance of the estimated treatment effect being $(3 + \rho)(1 - \rho)\sigma^2/2n$ with optimum weighting and $(3 - 2\rho)\sigma^2/2n$ with equal weighting.

To summarize, avoid the single-group before–after study unless nothing better is practicable. By itself, the difference $\bar{y}_a - \bar{y}_b$ provides no basis for speculation about a treatment effect, unless there are strong grounds for concluding that this difference is of a size that would not have occurred in the time interval involved in the absence of a treatment effect or other major cause. In attempting conclusions, the investigator has the responsibility of using imagination and help from colleagues in listing alternative major contributing causes. For each such cause the investigator should consider any evidence that assists a judgment about the size and direction of the resultant effect. Discussion of these alternative causes and the investigator's

judgment about the biases that they create should be included in the report of study results.

Given a time series of measurements at intervals τ both before and after the introduction of the treatment, we are in a better position to judge whether an unusual change in the time series coincided with the introduction of the treatment. Campbell and Stanley (1963/1966) illustrate a variety of situations that may occur with four pretreatment and four posttreatment measurements. Some situations strongly suggest an unusual change in level or direction; in some the verdict on this point is dubious, while in others a glance shows no sign of anything unusual.

Thus in Figure 7.4.1, situation A indicates an unusual rise in level that persists after introduction of the treatment, while situation B not only

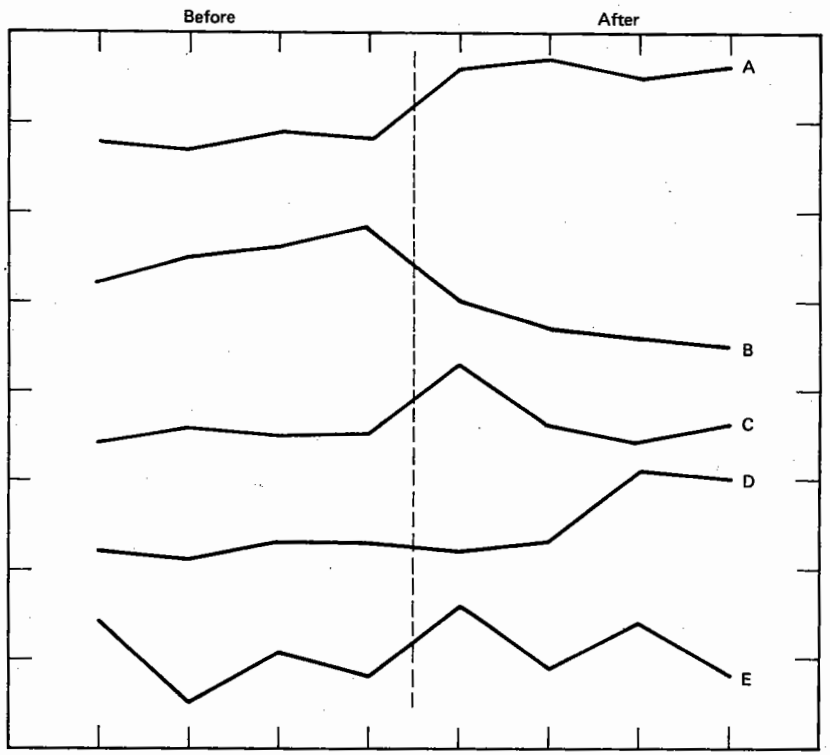

Figure 7.4.1

indicates a shift in level, but hints at a reversal of the time trend. In situation C, the shift in level lasts for only one time interval, which would be consistent with a treatment effect that was predicted beforehand to be of short duration. Situation D is more ambiguous, a shift occurring two intervals after introduction, which might either represent a delayed effect or some later causal force having nothing to do with the treatment. Situation E is one of many indicating nothing unusual occurring in the key interval.

It would help if visual impression of the before and after measurements could be made more objective by a valid test of significance of the postulated effect of the treatment. In short series the possibilities are limited to the simplest situations. In situation A, if successive observations in the time series could be regarded as independent, a standard t test of $\bar{y}_a - \bar{y}_b$ with 6 degrees of freedom could be used. However, in most time series, successive observations are correlated, the correlations depending on the time interval between the observations. Box and Tiao (1965) have given two tests for a persistent shift in level, based on two different models about the nature of the time correlations, but these tests require knowledge of the sizes of the relevant correlations. With a time series that clearly has rising or falling trends, tests of significance based on linear-regression models (e.g., changes in levels, slopes, or both) are easily constructed if residuals can be assumed independent, the chief trouble being the paucity of degrees of freedom likely to be available for estimating the residual variance. I do not know any likely tests related to linear trends when residuals are correlated. One temptation which must be avoided is to make the postulated treatment effect that is tested depend on the appearance of the time series (e.g., assuming a persistent shift δ in level in A in Figure 7.4.1, a shift δ only in the first posttreatment measurement in C, and delayed-treatment effect in D.) This tactic destroys the objectivity of the test.

As an illustration of the methodological approach in handling "interrupted" time-series data, Campbell and Ross (1968) consider an evaluation of the effects, on fatalities in motor accidents, of a crackdown on speeding by suspension of licenses, introduced by the state of Connecticut in December 1955, using five "before" measurements for 1951–1955 and four "after" measurements for 1956–1959.

From 1951–1955 the number of fatalities indicated an upward trend and from 1956–1959 a downward trend, the graph suggesting a definite reversal in the fatality-rate trend following the crackdown. Year-to-year variations in fatalities during 1951–1959 were substantial, however, and an application by G. V. Glass (1968) of the Box–Tiao test for a persistent shift in level to *monthly* data provided more observations and gave significance just short of the 10% level.

If we accept this as partial evidence for an unusual force at work beginning in 1956, Campbell and Ross proceed to consider the crackdown and possible alternative contributors as causal factors. With regard to the crackdown treatment, they plot, from 1951 to 1959, the suspensions of licenses for speeding as a percentage of all license suspensions. This graph shows a marked rise in license suspensions to a new level in 1956, as evidence that the crackdown was actually applied. Similarly, a graph of the percentage of speeding violations to all traffic violations showed a marked drop from 1956 onward.

Among *possible* alternative contributors as causal factors, Campbell and Ross consider a dramatic improvement in the safety features of cars and a decrease in the amount of hazardous driving conditions, concluding that neither of these constitutes a plausible rival hypothesis in this case. Another source to be considered is any possible consequence of the 1955 pretreatment measurement. As it happened there was about a 35% increase in deaths from 1954 to 1955, the latter figure being the highest ever in Connecticut. There may have been two effects of the initiation of the crackdown following this maximum. First, publicity given to the 1955 increase in deaths may have induced a period of more careful, defensive driving. Second, in a time series in which there are annual ups and downs due to a multiplicity of influences, an exceptionally high value, as was 1955, is likely to be due in part to an unusual combination of upward influences in 1955 and therefore to be followed by several lower values. As often happens in this type of study, it is difficult to estimate whether either of these effects was substantial. Another source to be investigated is any change in the standard of measurement or record keeping. This does not seem to be a contributor in this study, but can be major in some evaluations of new administrative practices, where a marked change in record-keeping habits, accuracy, and detail may be introduced as a part of the new program.

Even if there is no *planned* comparison group from which the treatment in question was absent, it is worthwhile to consider whether something can be learned from data available for other populations not strictly comparable. Campbell and Ross use this idea to compare traffic deaths per 100,000 persons during 1951–1959 for Connecticut and the nearby states of New York, New Jersey, Rhode Island, and Massachusetts. These states also show substantial year-to-year variations, but overall comparison of the post-1956 with a pre-1956 record of fatalities is more favorable in Connecticut than in the other states.

The concluding judgment of Campbell and Ross is "As to fatalities, we find a sustained trend toward reduction, but no unequivocal proof that they

were due to the crackdown. The likelihood that the very high prior rate instigated the crackdown seriously complicates the inference." These statements illustrate the type of conclusion that an observational study permits —one that nearly always involves an element of the investigator's judgment.

The objective of this study was to measure the effect of the crackdown on a single response variable—number of traffic deaths. In the broader aspects of program evaluation the investigator must choose response variables that reflect not only the intended effects of the program, but any other effects that he judges important. In this connection Campbell and Ross note, incidentally, from an examination of the 1951–1959 figures, two other indicated changes in 1956 that *may* represent less desirable effects of the crackdown. Relative to suspensions there was a marked increase in the proportion of arrests while driving with a suspended license, and also in the percent of speeding violation charges judged not guilty in the courts.

In a situation somewhat analogous to the time-series comparison, it may occasionally be possible to obtain some independent comparisons with and without the treatment in question. An example is the attempt to estimate how much the advertised appearance of a famous pitcher A adds to the attendance at a baseball game. In the professional leagues, two teams normally play on three or four successive days in the same park. The approach used is to compare the attendance on the days when A is pitching with the attendance on neighboring days that are similar as to weather, time of day or night, day of the week, and so forth. The power of the pitcher to draw a crowd for the other team is a confounding factor, but repetitions are obtained because pitcher A appears in numerous parks in different cities against different teams. The type of approach needed is similar to that in the Connecticut crackdown.

Although Cochran planned several more sections for this chapter, his writing stopped at this point. [The editors]

REFERENCES

Box, G. E. P. and G. C. Tiao (1965). A change in level of a non-stationary time series. *Biometrika*, **52**, 181–192.

Campbell, D. T. and H. L. Ross (1968). The Connecticut crackdown on speeding: Time series data in quasi-experimental analysis. *Law and Society Review*, **3**(1), 33–53.

Campbell, D. T. and J. C. Stanley (1963). Experimental and quasi-experimental designs for research on teaching, in N. L. Gage, Ed. *Handbook of Research on Teaching*. Rand McNally, Chicago. (Also published as *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, 1966.)

Collier, R. M. (1944). The effect of propaganda upon attitude following a critical examination of the propaganda itself. *J. Soc. Psychol.*, **20**, 3–17.

Emmett, B. P. (1966). The design of investigations into the effects of radio and television programmes and other mass communications (with Discussion). *J. Roy. Statist. Soc. Ser. A*, **129**, 26–59.

Glass, G. V. (1968). Analysis of data on the Connecticut speeding crackdown as a time-series quasi-experiment. *Law and Society Review*, **3**(1), 55–76.

Additional Reference

Had he been aware of this reference we believe that Cochran would have added it to the list for this chapter. [The editors]

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Rand McNally, Chicago.