# USING QUANTITATIVE DATA TO STUDY SOCIAL PHENOMENA: SOME IGNORED ISSUES

**Stanley Lieberson**
**Department of Sociology, Harvard University, Cambridge, Massachusetts 02138**

**Keywords: True experiments, social research, observational studies, statistical estimates, design problems**

Social research is usually based on a statistical simulation of what would occur if a true experiment was possible. By "true experiment," I mean one in which subjects, whether they be individuals or social aggregates ranging from small groups to nation-states, are randomly assigned to the conditions of interest. In varying degrees, it is difficult in most social sciences to conduct such experiments (particularly in sociology, political science, anthropology, and history; to a lesser extent in economics and psychology). It is also a problem in medical studies on smoking, tension, food consumption, weight, and other issues of nutrition and lifestyle when subjects are not randomly assigned. When researchers are obliged to employ observational data to estimate the outcome that would have been obtained under true experimental conditions, a wide variety of statistical procedures are employed. It is appropriate that these assumptions be examined closely since they can easily lead to shaky and possibly misleading conclusions. By "misleading" I mean conclusions that could easily be different from those that would have occurred had true experiments been possible. Of course, the greater desirability of true experiments is implicitly recognized since experimentation is normally used whenever possible. However, some of the difficulties in using observational data as a substitute do not appear to be fully understood. Some of these are reviewed in this paper: the contamination problem, the assumption of causal symmetry, the misuse of variance, and the use of control variables.

## The Contamination Problem

Social research on humans or their groups cannot assume an isolation between the test and experimental conditions as readily as in the study of other subjects. A crucial feature of an experiment is confidence that the subjects receiving the control condition are not in any way influenced by those subjects receiving a test condition. If, for example, we heat up a metal (the test condition), the control would be placed far enough away so that it is not responding to the heat being administered. That is, of course, trivial and obvious. But consider the observational data for human behavior. After school desegregation was introduced by court order in many cities, a classic problem was whether its consequence was to accelerate white flight from these cities as a way of avoiding racial integration in the schools. The non-experimental solution was to compare white rates of exodus from central cities in systems that were desegregated with those where desegregation had not occurred. Differences in the rate of exodus were used to ascertain whether white flight had been increased by desegregation. Note, however, that there is a very demanding and shaky assumption here. Can we be confident that the "control" cities, i.e., those not experiencing court orders, tell us what would have happened in the desegregated systems had they not been desegregated? Two problems exist. First of all, residents in the segregated cities are aware of what is going on elsewhere, and we do not know if they are responding to what is obviously going to happen to them in the near future. Hence, do they move with this in back of their mind? Do white newcomers chose housing with possible desegregation as a consideration? To the degree that the answer is in the affirmative, to that degree the level of net exodus in segregated cities is greater than what would have otherwise occurred in the absence of the orders in other cities. And, to that degree, the control cities are misleading controls--and the gap between the two subsets of cities is underestimated. This is but one example of how a true control may not be as easily available when observational data are employed. This becomes clear when we consider experiments on medication even when they do involve random assignment. A placebo is given in order to rule out the possibility that the response of those receiving the test drug is due in part or whole to receiving any medication even if it is "ineffective." In this case, the comparison between the test condition and either the control condition of receiving no drug or the currently used drug would not necessarily measure the question at issue.

**The Assumption of Causal Symmetry**

To a striking degree, non-experimental social research operates under an assumption of what I have called "causal symmetry". By this I mean the assumption that the influence of a given value of $X$ on $Y$ is unaffected by the previous value of $X$. I am not referring to autocorrelation, but rather to the history of $X$. If $X_{t2}$ is $n$, then it is assumed that its influence on $Y$ is unaffected by whether $X_{t1}$ was higher or lower than $n$. In other words, it is as if we are dealing with something like Boyle's Law, where pressure can go up or down and the volume of gas will respond according to the specific pressure at the time, with the previous pressure having no influence. The pressure can go up and down ad nauseam and the volume will respond in each case regardless of direction. However in both the physical world and the social world, not all changes are reversible, as if there was something like Boyle's Law operating. In practice, causal situations are often asymmetrical; once a condition is established, the original causes can be removed but the condition will either remain unchanged or will change very differently from what would be predicted from observing the earlier causal pattern leading to its development. This has a bearing on many social processes that are studied with non-experimental data. For example, for many social policy issues it is not at all clear that learning the causes of a given problem is necessarily very helpful in understanding the utility of proposed solutions. Once a problem exists, removing its causes need not be an appropriate solution. Likewise, if the response of $Y$ to a shift back in $X$ is not harmonious to what was used to account for *Y's* earlier movement, the operation of an asymmetrical cause means this need not be taken as casting doubt about the initial explanation. Basically, the standard statistical procedures often fail to deal with the possibility of an irreversible process such that an outcome cannot be reversed by reversing the causes. An outcome may be reversed, but only through different causal factors. This is nicely illustrated when one considers, say, the causes of the feminist movement and whether later elimination of those causes would reverse the movement. Or what would happen if the forces leading English to become the great international language of communication were no longer operating. Very likely, once established, it would take a lot more for English to recede as the international language.

**The Misuse of Variance**

Non-experimental research often uses the standard of "explaining variance" as a way of determining the adequacy of a proposed explanation of some dependent variable. This standard, although widely used, is not always an obvious or appropriate one. In cases where there is not much variance, it is typically the view that there is not much to explain. But does this makes sense? Yes, if the intent is to use explained variance to answer questions about how well the theory or proposed explanation can account for the observed events. But this works only if the statistical tail wags the substantive dog. Suppose there are minimal differences between the observed units, say the educational level of various ethnic groups. Is there no question then? Hardly, since the key fact is that the descendants of various origin groups are virtually indistinguishable in terms of educational attainment. It is the uniformity of the distribution that is of striking interest, not the explanation of variance. And that is to be explained obviously through a very different procedure.

A useful example of the limits of variance as an explanation in the social sciences is provided by gravity. Assuming there is no vacuum (or there is an imperfect vacuum), if we simultaneously drop a variety of objects in different tubes, we will find that they travel to the bottom at different speeds. Why do the rock, the feather, a piece of paper, a book, and a slice of pizza all travel at different speeds? What explains the variance between them? Various factors such as density, surface, and design will play a role. If all the variance is explained, would this mean that we now know why they drop? No, not all. We know why they drop at different speeds. Obviously, two questions are being confused. And that is the case in many social science studies that use explained variance standards for causal analysis.

In social research often the investigator is obliged to work with only a small number of cases, say data for a limited number of nations. A "solution" sometimes used to increase variance is to examine subparts (states, provinces, counties, etc.) of each nation. This then boosts the number of cases and hence makes the data susceptible to analyses of the variance. But an explanation of variance between subareas of a nation is not the same as an explanation of variation between nations. Accordingly, tempting though it is, such an analysis does not serve as a suitable substitute. One question deals with differences between nations and the other deals with difference within a nation. What accounts for the murder rate being so much higher in the United States than in other countries? Is the answer necessarily to be found in the factors accounting for area differences within the nation? Possible, but hardly likely. The implications of such an assumption would require that, for the independent variable under consideration, one finds comparable values for it in subareas in different countries and--in turn--that the murder rate be comparable in those subareas. In other words, the nation effect must disappear. Again, then, accounting for variance per se is not necessarily sufficient for understanding social processes.

**The Use of Control Variables**

For the experimental model to work, there has to be reasonable confidence that there is no other explanation for the observed pattern. If either selectivity is an issue or the non-random assignment is of consequence, there must be some way of taking into account the condition. Control variables are probably the most common solution used in non-experimental social science. For each attribute in which the control and test populations differ (due to non-random assignment), a variable is introduced so that the differences between the two populations are taken into account--through one statistical procedure or another. If, on the aggregate, the parents of children attending Head Start programs are different from other parents, say in education or income or being marital status, then the results are adjusted accordingly. It is as if we were to match up each set of parents with a comparable set so that we are confident that characteristic is not responsible for the differences between the children in their educational outcome. The problem is that it is difficult to be confident that unmeasured selectivity is fully taken into account. For example, if we are in effect matching parents by education and income and marital status, is there some other difference between parents that leads some to send their children to Head Start and others to not do so? Remarkably, it is even possible for application of the controls to modify the outcome in a way that is further from the actual results that would be obtained in a true random sample. For this and other reasons, the usual assumption is questionable that the application of control variables is either beneficial (moving the outcome to results closer to what a true experiment would show) or at least benign (not improving the outcome, but causing no harm). In practice it is possible for the results, after controls are applied, to be further from what a true experiment would have shown. Moreover, the application of a large number of controls, as is often the case, often entails a multi variate array that strays from the usual statistical assumptions and, in turn, generates special problems in which the assumptions underlying the statistics are badly violated.

In summary, true experiments are an enormously valuable device for attaining knowledge. However, in many of the social sciences, true experiments (meaning random assignment of subjects) is virtually impossible or would otherwise be unethical. Accordingly, ersatz experiments are commonly used based on observational data. The results have far more difficulties than are commonly recognized among those obliged to work with these data.