

## CHAPTER 2

### SETTING AND SIGNIFICANCE OF THE CAUSAL STATISTICS PROJECT

#### Sections

- 2.1 Introduction
- 2.2 The Research Process
- 2.3 The Setting of Causal Statistics
- 2.4 The Need for Causal Statistics
- 2.5 Extant Causal Inquiring Systems
  - 2.5.1 Summary
  - 2.5.2 Path Analysis
  - 2.5.3 Econometrics
  - 2.5.4 The Simon-Blalock Approach
  - 2.5.5 Comparison of Extant Causal Inquiring Systems
- 2.6 Comparison of the Causal Statistics Project with Extant Causal Inquiring Systems
- 2.7 The Need for Further Research

#### 2.1 Introduction

The social and medical sciences are in need of analytical techniques which will enable them to draw causal inferences from non-experimental data. The same is also true for quasi-experimental and imperfectly experimental data. But non-experimental data represents the extreme case and, therefore, will be the focus of our attention.

It is fairly easy to build causal theories, if one is allowed to experiment; but very difficult for non-experimental studies. To the chagrin of most social scientists their fields are, to a great extent, non-experimental sciences.

It is desirable to know the causal relationships between variables because, in making decisions for action, we generally need to know the various effects of alternative actions--i.e., the various effects of alternative manipulations of one or more variables. In other words we need the ability to make causal inferences from non-experimental data in order to make causal predictions.

Associative statistics (e.g., correlation analysis, regression analysis, analysis of variance, etc.) will give us this ability only if we impose severe and unnecessarily restrictive assumptions. In practice the social scientist seldom realizes the number and/or restrictiveness of the assumptions he is implicitly making when drawing causal inferences.

## 2.2 The Research Process

For purposes of this discussion, the research process will be visualized as diagrammed in Figure 2-1. To some extent Figure 2-1 is an over simplification in that feedbacks could exist at almost every stage. But this added complexity is omitted since Figure 2-1 is sufficient to show the places of causal statistics, causal theories, and causal predictions in the research process.

Box 1 represents the real world of actual objects, phenomena, and "laws" about which we wish to learn.

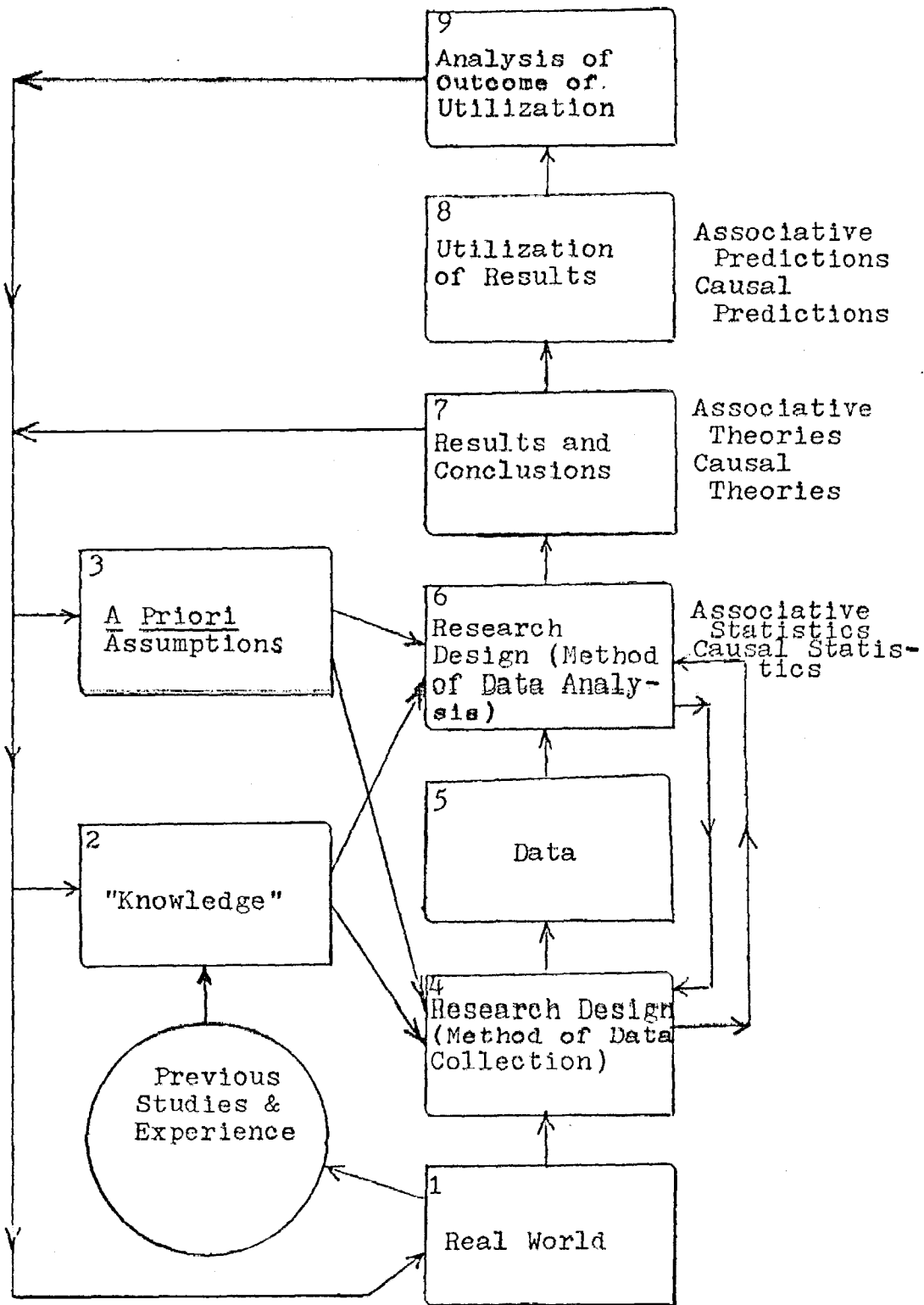


Figure 2-1

The purpose of the research process is to test hypotheses and construct theories about the real world.

The "knowledge" which is employed in the research process is contained in box 2. The word knowledge is placed in quotation marks to indicate that the "knowledge" referred to in box 2 may, in fact, be judged as incorrect or inaccurate at a later time. The "knowledge" in box 2 has been obtained from previous studies of and experience with the real world. These are denoted by the circle in Figure 2-1.

Box 3 denotes the a priori assumptions which are inserted into the research process. A priori assumptions are employed when some part of the research process requires a piece of information, but that information is not a part of our body of "knowledge." For example, when we correlate  $X_1$  and  $X_2$ , we assume that whatever association exists between them is linear.

The method of data collection is represented by box 4. This along with the method of data analysis (box 6) compose the research design. Both previous "knowledge" and assumptions are employed in designing the method of data collection. The method of data collection can be viewed as a filter through which sensible intuitions (Kant's term), emanating from the real world, are processed (i.e., categorized, quantified, and positioned in time) to yield data (box 5). It should be noted that data collection is strongly prone

to error. The sensible intuitions will almost always be distorted (by categorization, lack of precision, covering up, etc.) in one way or another before they become (cause) data.

The method of data analysis (box 6) is designed on the basis of previous "knowledge" and a priori assumptions. There is also a feedback relationship between the methods of data collection and data analysis. A change in one may force a change in the other. The method of analysis can be viewed as a filter through which the raw data is processed and shaped to make it more understandable. Its output is results and conclusions (box 7).

The results and conclusions feedback to increase our "knowledge" of the real world (box 2) and change (maybe) the a priori assumptions (box 3) which would be employed in a later, related study. Also, the structure of the real world (box 1) is, to some extent, changed by our additional "knowledge." A self-fulfilling prophecy is an example in which added "knowledge" changes the structure of the real world.

There is a philosophical point here. In a precise (philosophical) sense no change occurs. This is because the microcausal laws of the universe remains unchanged; only the macrocausal "laws," inferred by man, change.

After the results are obtained and the conclusions made the decision maker, planner, or practitioner uses

them to project various alternatives into the future and chooses the most desirable. This utilization of results (box 8) results in decisions, actions, and outcomes.

Box 9 denotes the analysis of these outcomes. How do the outcomes deviate from the predictions and why? The results of this analysis feeds back into "knowledge," assumptions, and the real world as did the results and conclusion obtained in box 7.

### 2.3 The Setting of Causal Statistics

The common method of data analysis (box 6) is associative statistics. For non-experimental research studies, the application of associative statistics to data analysis yields results (box 7) which are associative theories. For example closeness of supervision ( $Z_1$ ) is negatively correlated with productivity ( $Z_2$ ). But this does not tell us if  $Z_1$  is the negative cause of  $Z_2$ ,  $Z_2$  is the negative cause of  $Z_1$ , or  $Z_3$  (say, company policy) is the cause of both.

When it comes to utilizing the results (box 8), associative theories are sufficient for making associative predictions but not causal predictions. For example, if  $Z_1$  is high, we can predict that  $Z_2$  will be low. But we cannot necessarily change  $Z_1$  from a high to a low value and then causally predict that  $Z_2$  will change from low to high. More abstractly, associative predictions are made when the user--based upon the

initial values of variables in the predictive system and the results obtained in a previous study (or any empirical "knowledge," for that matter)--forecasts the unknown value of a certain variable.

If causal statistics were used as the method of data analysis, the results would be causal theories. Causal theories enable us to make causal predictions. Causal predictions are made when the user--based upon initial values of variables in the predictive system and the results obtained in a previous study (or any empirical "knowledge")--forecasts the effects of a manipulation of one of the variables on another variable. Causal theories enable us to control the future rather than just forecast it.

#### 2.4 The Need for Causal Statistics

Figure 2-2 shows the relationship between the type of study (i.e., experimental or non-experimental), the method of data analysis, and the type of theory obtained.

	Experimental	Non-experimental
Associative Statistics	Associative Theories Causal Theories	Associative Theories
Causal Statistics	Associative Theories Causal Theories	Associative Theories Causal Theories

Figure 2-2

Experimental data is obtained when the researcher manipulates one or more of the variables and observes

the effects on other variables in the system. Non-experimental data is obtained by observation alone, with no manipulation or control by the researcher. Most physical sciences data are of the experimental type, while most social sciences data are of the non-experimental type.

Although here the top margin of Figure 2-2 is viewed as a dichotomous axis, it is, in fact, a continuum. At one extreme on the continuum is the perfect experiment and at the other extreme is the perfect non-experiment. In the perfect experiment the researcher is able to control all relevant variables, whereas in the perfect non-experiment the researcher does not affect or control any of the relevant variables. Any other case falls somewhere on the continuum between.

Causal theories are easy to obtain from experimental research studies. For example, consider the following experiment, which is over simplified for brevity and clarity. Take 50 rats which have a certain, normally fatal disease. Inject chemical, X, into the blood stream of all the rats. If all of the X injected rats recover, we can say, using associative statistics or causal statistics and based upon comparatively mild assumptions, that X caused them to recover.

Causal theories are not so easily obtained from non-experimental studies. For example, consider the same 50 rats, but do not inject them with any chemical.



Just observe them. Say that they all recover and in testing them we find that each contained a chemical, Y, in its blood, i.e., we find a correlation between recovery and Y. Can we reasonably conclude that Y caused their recovery? No, because the disease could have caused Y or another chemical, Z, could have caused both the recovery and Y or Y could be present in the blood of all rats. Analysis of non-experimental data by associative statistics yields associative theories. Although, in some special cases and/or under very stringent assumptions, associative statistics can yield valid causal theories; generally associative statistics is quite inefficient in obtaining them.

If we used causal statistics for analysis in this non-experimental study, it would demand more information. Once the additional information was obtained, causal theories could be inferred, based on the least restrictive assumptions possible.

Causal theories are much more desirable than associative theories because we typically wish to control (i.e., affect, change, manipulate) the future rather than just forecast it. Research in the social sciences is typically of the non-experimental type. Based on Figure 2-2 and the two preceding statements, we can conclude that causal statistics is an extremely important and valuable tool for the social sciences.

## 2.5 Extant Causal Inquiring Systems

### 2.5.1 Summary

At present, there are three, more-or-less distinct, causal inquiring systems. They are path analysis, econometrics, and the Simon-Blalock approach. In actual fact they are virtually identical to one another.

### 2.5.2 Path Analysis

Path analysis was introduced in a phenomenally innovative paper by Sewall Wright\* in 1921. Since that

---

\*Wright, Sewall: "Correlation and Causation," J. of Agricultural Research, Vol. 20, 1921, pp. 557-85.

time additional innovations and, also, acceptance have been amazingly slow. Path analysis considered only one-way causation until 1954 when John Tukey\*\* introduced

---

\*\*Tukey, John Wilder: "Causation, Regression, and Path Analysis," in Oscar Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush, eds., Statistics and Mathematics in Biology, Ames: Iowa State College Press, 1954, pp. 35-66.

two-way path analysis. This is the only innovation in path analysis of major importance since 1921.

Basically, path analysis is a linear regression or simultaneous linear regression technique in which the coefficients are causal, assuming that the basic

assumptions of the model are valid. These coefficients ( $c_{0i}$ 's) are called path regression coefficients. See Wright\* (1960) for a summary of path analysis.

---

\*Wright, Sewall: "Path Coefficients and Path Regressions: Alternative or Complementary Concepts?" Biometrics, Vol. 16, 1960, pp. 189-202.

---

### 2.5.3 Econometrics

Econometrics employs regression and simultaneous equation models. It is far more advanced mathematically than path analysis, but there are comparatively few papers in the field which consider the causal implications of these mathematical techniques.

Econometricians try to avoid the word "cause" because of their misinterpretation of Humian philosophy on the subject. Due to their avoidance of this word, econometricians have failed to consider sufficiently (a) many of the causal implications and properties of econometrics and (b) many of the problems and benefits connected with causal prediction.

Two good econometric references are Johnston\*\* and Goldberger\*\*\*.

---

\*\*Johnston, J.: Econometric Methods. New York, McGraw-Hill, 1963.

\*\*\*Goldberger, Arthur S.: Econometric Theory. New York, John Wiley & Sons, 1964.

---

#### 2.5.4 The Simon-Blalock Approach

The Simon-Blalock approach began with a 1954 paper by Herbert Simon\*. This paper served as the foundation

---

\*Simon, Herbert A.: "Spurious Correlation: A Causal Interpretation," J. of the American Statistical Association, Vol. 49, 1954, pp. 467-479.

---

for a great deal of later work by Hubert Blalock.

Basically, this approach gives causal interpretation to some of the more elementary formalizations of econometrics. An exception is Blalock\*\* (1969) in which

---

\*\*Blalock, Hubert M., Jr.: Theory Construction: From Verbal to Mathematical Formulations, Englewood Cliffs, Prentice-Hall, 1969.

---

he gives preliminary consideration to some simple differential equation models.

#### 2.5.5 Comparison of Extant Causal Inquiring Systems

Mathematically, these three causal inquiring systems are either identical in form or deducible, one from the other. There are sometimes slight differences in causal interpretation, but these are not important.

#### 2.6. Comparison of the Causal Statistics Project with Extant Causal Inquiring Systems

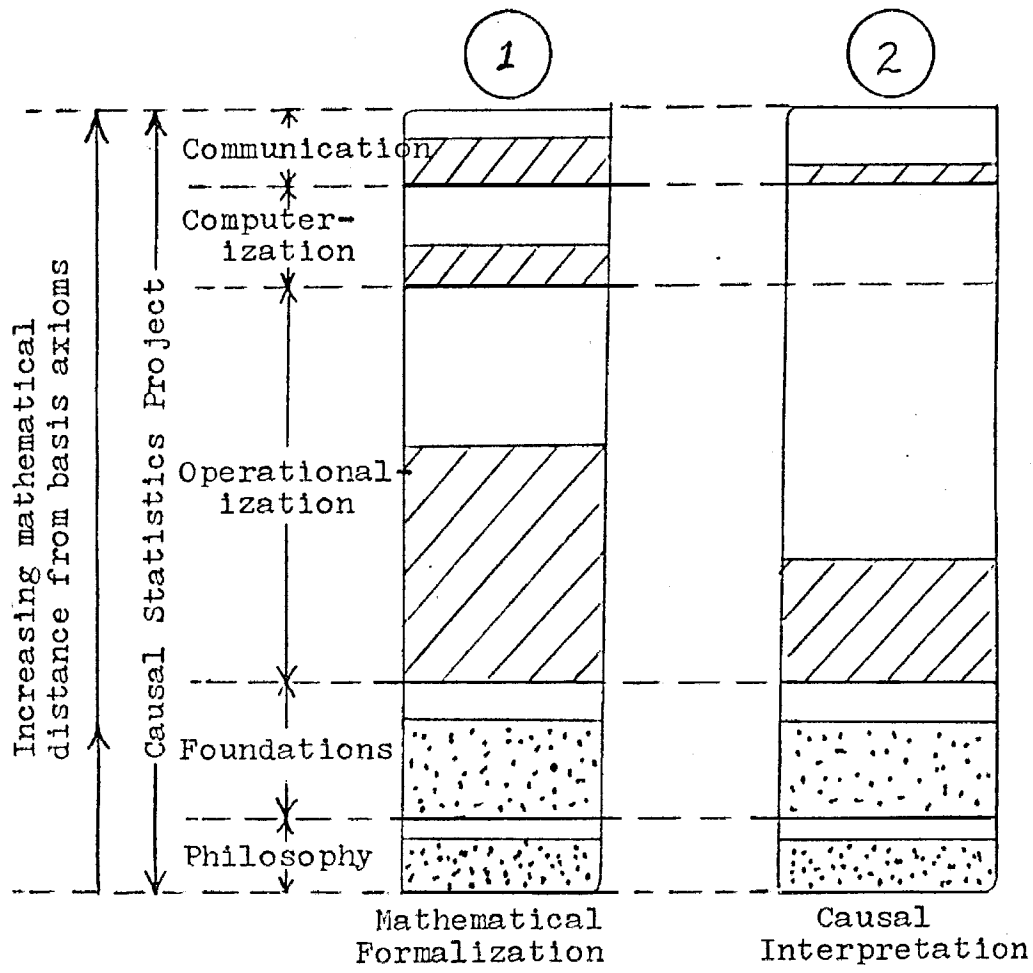
To see the relationship between the causal statistics project and existing causal inquiring systems,

consider Figure 2-3. The figure is, to some extent, over simplified. Its form is a compromise between complete accuracy and simplicity of explanation. But its complexity is certainly sufficient to show the comparison between the causal statistics project and present causal inquiring systems.

The figure is composed of two bar graphs, each representing the entire causal statistics project. The lower parts of the bar graphs represent the analysis of the philosophical underpinning of the concept of causality and foundations research in causal statistics. The middle portion represents the operationalization research. The computerization of causal statistics and the communication of the techniques of causal statistics to others (e.g., books and articles) are represented by the top parts of the bar graphs.

As one moves up the bar graphs, he gets further--in the sense of distance along a mathematical derivation--from the basic axioms upon which causal statistics is based. The left (first) bar graph represents the mathematical formalization in causal statistics and the second, the causal interpretation of these mathematical forms.

The ledger shows that the cross hatched areas represent subject matter which has already been researched. In the first bar graph these areas represent the mathematical formalisms of path analysis,



Ledger:




-  subject previously researched
-  subject to be researched in this dissertation
-  subject remaining to be researched

Figure 2-3

econometrics, and the Simon-Blalock approach and their computerization and communication. In the second they represent the causal interpretations which have been made of these mathematical formalizations. It is clear from the graphs that only a portion of the present formalization has been causally interpreted.

The areas with dots in them denote subject areas to be researched in this dissertation. As can be seen, causal philosophy will be analyzed. Also, a large part of both the mathematical formalization and causal interpretation of the foundations will be researched.

The white areas indicate research which remains to be done. The portion of the foundations which has not been researched is the logical derivation (i.e., the deduction via symbolic logic) of causal statistics. The derivation in the dissertation is a less rigorous, mathematical derivation. The unresearched parts of the operationalization and computerization are listed in Parts V and VI of Table 1-1, respectively.

The causal statistics project, as proposed, will derive, operationalize, computerize, and communicate a generalized causal inquiring system. As one can see from Figure 2-3, this generalized system encompasses the present causal inquiring systems as subsets.

## 2.7 The Need for Further Research

From Figure 2-3 it is obvious that research beyond the present state of knowledge (the contents of this

dissertation not yet being considered as part of the present state of knowledge) of causal inquiring systems is needed. The further research which is needed is listed in Table 1-1, under Parts II, III, V, and VI.

The further research needed under Parts II and III is in the areas of philosophical and logical foundations of causal statistics. The often misinterpreted philosophy of causality needs to be analyzed and the philosophical basis for the concept of causality set forth. Also, the general formulation of causal statistics needs to be derived and the fundamental axioms and assumptions sets noted. These research areas are the task of this dissertation. Their importance will be discussed in detail in the next chapter.

There are several areas of needed research listed under Part V. The mathematical techniques of econometrics need to be generalized to handle the generalized formulation of causal statistics. There is a need for more causal inference oriented mathematical formalization. It is desirable for causal statistics to be integrated with other statistical techniques, like factor analysis. As generalizations, mathematical formalization, and integration proceed; causal interpretation will have to be extended to them. Specific application problems should be investigated and alleviated. A sensitivity analysis should be performed to determine the effects of various degrees of



dissatisfaction of assumptions. Also a simple algorithm for the application of causal statistics needs to be formulated.

After causal statistics has been developed, it should be computerized (Part VI) for easy access by empirical researchers. Finally, an organized, coherent presentation of causal statistics should be written.