Statistical Science
1999, Vol. 14, No. 3, 243–258

From Association to Causation: Some Remarks on the History of Statistics

David Freedman

Abstract. The "numerical method" in medicine goes back to Pierre Louis' 1835 study of pneumonia and John Snow's 1855 book on the epidemiology of cholera. Snow took advantage of natural experiments and used convergent lines of evidence to demonstrate that cholera is a waterborne infectious disease. More recently, investigators in the social and life sciences have used statistical models and significance tests to deduce cause-and-effect relationships from patterns of association; an early example is Yule's 1899 study on the causes of poverty. In my view, this modeling enterprise has not been successful. Investigators tend to neglect the difficulties in establishing causal relations, and the mathematical complexities obscure rather than clarify the assumptions on which the analysis is based.

Formal statistical inference is, by its nature, conditional. If maintained hypotheses $A, B, C, \ldots$ hold, then H can be tested against the data. However, if $A, B, C, \ldots$ remain in doubt, so must inferences about H. Careful scrutiny of maintained hypotheses should therefore be a critical part of empirical work—a principle honored more often in the breach than the observance. Snow's work on cholera will be contrasted with modern studies that depend on statistical models and tests of significance. The examples may help to clarify the limits of current statistical techniques for making causal inferences from patterns of association.

Key words and phrases: Association, causation, regression, history of statistics, modeling significance, epidemiology.

1. INTRODUCTION

In this paper, I will look at some examples from the history of statistics—examples which help to define problems of causal inference from nonexperimental data. By comparing the successes with the failures, we may learn something about the causes of both; this is a primitive study design, but one that has provided useful clues to many investigators since Mill (1843). I will discuss the classical research of Pierre Louis (1835) on pneumonia and summarize the work of John Snow (1855) on cholera. Modern epidemiology has come to rely more heavily on statistical models, which seem to have spread from the physical to the social sciences and then to epidemiology (Sections 4 and 5). The modeling approach was quite successful in the physical sciences, but has been less so in the other domains, for reasons that will be suggested in Sections 4–6.

Regression models are now widely used to control for the effects of confounding variables, an early paper being Yule (1899); that is the topic of Section 4. Then some contemporary examples will be mentioned, including studies on asbestos in drinking water (Section 5), health effects of electromagnetic fields, air pollution, the leukemia cluster at Sellafield and cervical cancer (Section 7). Section 8 discusses one of the great triumphs of the epidemiologic method—identifying the health effects of smoking. Other points of view on modeling are briefly noted in Section 9. Finally, there is a summary with conclusions.

David Freedman is Professor, Department of Statistics, University of California, Berkeley, California 94720 (*e-mail: freedman@stat.berkeley.edu*).

2. LA MÉTHODE NUMÉRIQUE

In 1835, Pierre Louis published his classic study on the efficacy of the standard treatments for pneu-