POPULAR STATISTICS

a series edited by

D. B. Owen

Department of Statistics
Southern Methodist University
Dallas, Texas

Nancy R. Mann

Biomathematics Department
University of California at Los Angeles
Los Angeles, California

1. How to Tell the Liars from the Statisticians, Robert Hooke
2. Educated Guessing: How to Cope in an Uncertain World, Samuel Kotz and Donna F. Stroup
3. The Statistical Exorcist: Dispelling Statistics Anxiety, Myles Hollander and Frank Proschan
5. Misused Statistics: Straight Talk for Twisted Numbers, A. J. Jaffe and Herbert F. Spirer
6. Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety, Chamont Wang

Sense and Nonsense of Statistical Inference

Controversy, Misuse, and Subtlety

Chamont Wang

Department of Mathematics and Statistics
Trenton State College
Trenton, New Jersey

1993

Marcel Dekker, Inc.  New York • Basel • Hong Kong
Chapter 3

Statistical Causality and Law-Like Relationships

I. INTRODUCTION

In empirical science and decision-making, inferences regarding cause-and-effect occur all the time. Statistical tools, as one can expect, provide a big boost to a causal argument. As a scientific tool, however, statistical reasoning offers a mixed blessing to inferences about causes and effects. One reason is that statistical causality toes a thin line between deduction and induction, and often fuels confusion on the part of researchers and decision-makers.

This chapter is thus devoted to a discussion of the use and misuse of statistics in causal inference. It begins, in Section II, with several examples of abuses and compares orthodox statisticians’ view with causal inferences drawn by lay practitioners.

In Sections III–IV, the chapter looks at some scholarly exchanges on the subject of causality. Strictly speaking, the term “probabilistic causality” is self-contradictory. As a consequence, debates on the foundation of probabilistic causality are sporadic in the literature. Recently, in his paper entitled “Statistics and Causal Inference,” Holland (1987, JASA) formulated the “Fundamental Problem of Causal Inference” and stated the motto: “No Causation without Manipulation.” Holland’s article has since drawn attention (and criticism) from other respected authorities.

With Holland’s motto in mind, the discussions in Sections III and IV delineate the boundary between “controlled experiments” and “observational studies,” a distinction which seems obvious but is often hard to maintain. Section V also discusses the problems of drawing causal inference in sample surveys. (Certain portions of Sections III and IV are rather technical. Some readers may prefer to skip these parts in their first reading.)

In sharp contrast to Holland’s motto is the position taken in the book, Discovering Causal Structures: Artificial Intelligence, Philosophy of Science, and Statistical Modeling (Glymour et al., 1987). Glymour advocates the use of statistical models for nonexperimental science; the project was funded by three National Science Foundation grants and will be briefly reviewed in Section VI. A case study is presented in Chapter 4 to demonstrate the potential defects in Glymour’s ambitious program.

II. SENSE AND NONSENSE IN CAUSAL INference:
EXAMPLES

EXAMPLE 1 A Harvard study (The New York Times and the Trenton Times, February 10, 1988), conducted by Dean K. Whita, director of instructional research and evaluation, concluded that Harvard students who took courses to prepare themselves for the SAT (Scholastic Achievement Tests) scored lower than students who did not take the classes. The study concluded that SAT coaching provides no help, and that “the coaching industry is playing on parental anxiety” (Harvard University Admissions Director, W. Fitzsimmons).

The study was based on questionnaires distributed to Harvard freshmen in Fall 1987. Of the 1409 participants, the 69% who had taken no coaching classes scored an average of 649 on the verbal portion of the SAT and 685 on the math. Students who had been coached (14% of the 1409 participants) scored an average of 611 on the verbal and 660 on the math:

<table>
<thead>
<tr>
<th></th>
<th>No coaching</th>
<th>Coached</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSAT</td>
<td>649</td>
<td>611</td>
<td>+38</td>
</tr>
<tr>
<td>MSAT</td>
<td>685</td>
<td>660</td>
<td>+25</td>
</tr>
</tbody>
</table>

To the uninitiated, the data definitely support the claims of the Harvard officials. But on the second look, the study was obviously flawed; you don’t need a statistician to figure out that good students do not need coaching and will still score higher than poorer students who were coached. In a similar situation, people who go to hospitals are in general less healthy than those who
don’t go to hospitals; but this fact does not imply that going to hospitals tends to weaken a person’s health condition.

It appears that the Harvard officials’ original intention to discourage students from going to coaching school is admirable. But the statistics they used to promote their argument simply do not bear scrutiny.

EXAMPLE 2 In the fall of 1987, a study was conducted by the American Council on Education and the Higher Education Research Institute at the University of California, Los Angeles. The study involved 290,000 college freshmen and found that a record number (75.6%) identified the goal of “being very well off financially” over the goal of “developing a meaningful philosophy of life” as “essential” to them, or “very important.” This is up from 73.2% one year earlier, 70.9% in 1985, and nearly double the level in 1970, 39.1%. Fig. 1 shows the graphic representation of these numbers.

The survey was done scientifically. But how to interpret the findings? In a news story, The New York Times (January 14, 1988) concluded that there is a rising “trend of materialism.” This conclusion seems to be valid, but Professor Merrill Young of Lawrence University argued otherwise (The New York Times, February 19, 1988).

According to Professor Young, money was of little concern to the students before 1977, because they considered that if they graduated, they would find it “practically lying in the street.” In contrast, since 1977 Professor Young has witnessed increasing doubt on the part of students about their chances of getting jobs after graduation. He concluded that “the rising trend has not been in materialism, but in pessimism.”

Professor Young also attributes most comments about the greed of today’s college students to the now middle-aged members of the censorious college generation of the late 1960s and early 1970s. “They spent their 20s scolding their elders,” the Professor complained. “Now they are spending their 40s lecturing the young.”

Reading this penetrating analysis, one may speculate still further. First, why does it have to be either materialism or pessimism? Isn’t it possible that both factors are responsible for the new trend, in a dynamic fashion? Second, what happened in 1977? Can we learn something about this specific year and use the information in a constructive way? So far no definitive answers to these questions appear to be available.

EXAMPLE 3 In an attempt to help people assess the quality of health care, the U.S. Government issued data showing the mortality rate at nearly 6,000 hospitals (The New York Times, December 18, 1987). The new statistics, contained in seven large books, were derived from data on 10 million hospital admissions and 735,000 deaths among Medicare beneficiaries in 1986. The lesson of the study was this: Watch out if the mortality rate of a hospital is higher than expected.

But is mortality rate a relevant indicator? If one wants to know the quality of health care, there are many obvious factors to be considered: the equipment in a hospital, the staff, the environment, the infectious control program, and the quality control procedures (nursing, pharmacy, standardization, etc.).

In addition mortality rates are deeply confounded by the fact that good hospitals tend to attract the severely ill, who in turn inflate a hospital’s mortality rate. Also, hospitals are known to send people home to die to keep them off the mortality list. Thus, an alternative conclusion might be: Watch out if the mortality rate of a hospital is lower than expected.

EXAMPLE 4 In this example we discuss the effect of statistical training on reasoning. It is believed by certain behavioral scientists that the teaching of statistical rules in an abstract way, as it is normally done in statistics courses, has an effect on the way people reason about problems in everyday life (Holland, Holyoke, Nisbett, and Thagard, 1987, p. 256). The scientists cited a study by Fong et al. (1986; funded by Grants from NSF, SES, ONR, etc.). A similar study by Fong et al. (1987) was published in Science and was cited in Statistical Science as good evidence that “the study of statistics, as taught by
psychologists, improves students’ ability to deal with uncertainty and variability in quite general settings” (Moore, 1988).

The Fong (1986) study examined four groups of subjects differing widely in statistical training: (1) college students who had no statistical training at all, (2) college students with one or more statistics courses, (3) graduate students in psychology, most of whom had had two or more semesters of statistical training, and (4) Ph.D.-level scientists who had had several years of training.

Subjects were presented with a problem about restaurant quality. There were two versions: randomness and no randomness. Within each group tested, half of the subjects received the cue and half did not.

In the no-randomness cue version, a traveling businesswoman often returns to restaurants where she had an excellent meal on her first visit. However, she is usually disappointed because subsequent meals are rarely as good as the first. Subjects were asked to explain, in writing, why this happened.

In the randomness cue version, an American businessman traveling in Japan did not know how to read the menu; so he selected a meal by blindly dropping a pencil on the menu and ordering the dish closest to it. As in the other version, he is usually disappointed with his subsequent meals at restaurants he originally thought were superb. Why is this?

Answers were coded into two categories: (1) nonstatistical, if the subject assumed that the initial good experience was a reliable indicator that the restaurant was truly outstanding, and attributed the later disappointment to a definite cause (e.g., “Maybe the chef quit”); and (2) statistical, if the subject suggested that meal quality on any single visit might not be a reliable indicator of the restaurant’s overall quality (e.g., “odds are that she was just lucky the first time”). The results were summarized in Fig. 2 (where p = % of statistical responses). Fong (p. 277) concluded that

the chart demonstrates clearly that the frequency of statistical answers increased dramatically with the level of statistical training, chi-square(6) = 35.5, p < .001. [emphasis supplied].

As a statistics educator, I am unable to be so optimistic in the interpretations of the chart: (1) If group 1 “NO STATISTICS” is relabeled as “NO COMPUTER SCIENCE,” and groups 2–4 are reclassified accordingly, then we get Fig. 3. A statement now can be drawn in parallel to Fong’s conclusion: the above chart demonstrates clearly that the frequency of statistical answers increased dramatically with level of computer science training, chi-square(6) = 35.5, p < .001. The same trick might work if we replace “computer science training” with “art and humanity,” or “English and history,” etc. In other words, the chart proves nothing except that when one gets older, he or she tends to be more aware of the variability in real life. (2) It is a shame that 80% of group 2 (who had had one or more statistics courses) and 60% of group 3 (graduate students) failed to give statistical responses if they didn’t
receive the cue. (3) If the problem contained the randomness cue, there are only tiny differences among groups 1, 2, and 3. Also, members of group 1 (who received the probabilistic cue) fare better than group 3 (without the cue): 50% vs. 40%. Evidently years of training on statistical formulas were wiped out by a single cue. (4) Twenty percent of Ph.D. scientists failed to think statistically in this study, and the percentage may be much higher in everyday life.

Reasoning in everyday life is, in my observation, almost independent of (or negatively correlated with) a person's academic achievement. On numerous occasions I have witnessed academicians (myself included) drawing ludicrous conclusions if the subject-matter is outside their technical focus. The situation can easily turn ugly if ego, emotion, or self-interest set in. For example, when students fail an exam, they tend to blame the instructor, and the instructor (often a Ph.D.) tends to blame the students for their laziness, poor backgrounds, etc. Causal inference of this sort has repeated itself since the Garden of Eden.

In a more believable study, behavioral researchers at the University of Pennsylvania found that people tend to judge decision-makers on the basis of factors that have more to do with luck than ability (Baron and Hershey, 1988). The researchers interviewed a group of students and asked them to rate hypothetical decision-makers according to variable outcomes. For instance, case 1 read as follows:

A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pain also interfered with other things, such as travel and recreation.

A type of bypass operation would relieve his pain and increase his life expectancy from age 65 to age 70. However, 8% of the people who have this operation die from the operation itself. His physician decided to go ahead with the operation. The operation succeeded.

Evaluate the physician's decision to go ahead with the operation.

Case 2 was the same except that the operation failed and the man died.

In the face of uncertainty, it is obvious that good decisions can lead to bad outcomes, and vice versa. Decision makers thus cannot infallibly be graded by their results. However, in cases 1 and 2, the student ratings reflect the opposite of this reasoning (n = 20 undergraduates at the University of Pennsylvania):

<table>
<thead>
<tr>
<th>Case</th>
<th>Outcome</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Success</td>
<td>0.85</td>
<td>1.62</td>
</tr>
<tr>
<td>2</td>
<td>Failure</td>
<td>-0.05</td>
<td>1.77</td>
</tr>
</tbody>
</table>

Statistical Causality and Law-Like Relationships

Out of 140 pairs of similar cases that differed only in success or failure, higher ratings were given by students to the cases with success in 62 pairs and with failure in 13 pairs. In sum, the study found that the students show an outcome bias, even though they think they should not (or even though they think they do not). The study also found that this tendency colors decisions in law, politics, regulation, institutions, and nearly every other facet of human life.

In certain branches of science, percentages (and statistical jargon) alone appear to be enough to establish the credibility of a causal link. In such disciplines, nothing can be perfect (as argued by some), therefore everything goes. To clean up certain confusions, the next two examples are designed to illustrate the origin and content of hard-line statisticians' views on statistical causality. (This kind of statistician is becoming, in my judgment, an endangered species.)

EXAMPLE 5 In the mid-1950s, the standard method for the treatment of stomach ulcers involved surgery. In 1958, a researcher developed a new method to treat stomach ulcers, not involving surgical intervention. He applied the method to 24 patients, and all were reported cured. The success rate was 100%.

If your relatives or close friends had stomach ulcers, would you recommend the new treatment? (In a recent class, all of my students answered yes.)

In order to draw a causal inference, an experienced statistician first considers how the data were generated. In this case, the technique used for treating ulcers is called gastric freezing (FPP, 1978, p. 7). The procedure goes like this: A balloon is placed in a patient's stomach, and coolant is pumped through the balloon; this treatment freezes the stomach, stops the digestive process, and the ulcer begins to heal.

After you know how the data were generated, do you still recommend the new treatment? (At this point all of my students say no.)

In 1963, a double-blind randomized controlled experiment was conducted to evaluate gastric freezing. There were 82 patients in the treatment group and 78 in the control group, assigned at random. The conclusion based on this controlled experiment was that gastric freezing was worthless (see details in FPP, p. 8). The balloon method eventually slipped into the black hole of history.

In scientific endeavors, the task of drawing conclusions about cause-and-effect is often confounded by unexpected lurking variables. Powerful techniques for ironing out the collective effect of confounding factors were developed by Sir Ronald A. Fisher. The techniques were used by scientists in the evaluation of gastric freezing and now are widely used in agriculture, biomedical research, and engineering applications.
The contribution of R. A. Fisher was the origin of a new branch of science, now called statistics. The foundations of the Fisherian tradition are not only mathematical formulas but also the physical act of randomization. This tradition thus advanced from medieval statistics (which was solely based on enumeration) to a branch of modern science.

A big problem in statistical causal-inference is that randomization may not be possible for physical or ethical reasons. A case study follows.

**Example 6** Since about 1920 there has been a dramatic increase in the rate of lung cancer in men. This rise also corresponds to a great increase in cigarette smoking in men. But can we conclude that cigarette smoking is a cause of the increase in lung cancer?

To hard-pressed statisticians, randomization is not feasible for this instance, and confounding factors (such as genetic defects) are always possible. Here is an example (FPP, 1978, p. 10–11): "Does smoking cause cancer? Is smoking the guilty factor? Smoking probably isn't good for you, but the verdict must be: not proved." A very fine mathematician was anguished at this verdict. "Given all the evidence how can FPP say this?" The mathematician worried that FPP's verdict might spoil the young people and concluded: "One of these three authors must be a smoker!"

Among certain statisticians, the issue of smoking and lung cancer was bitter and unsettling. Its history can be traced back to R. A. Fisher's (1959) book, *Smoking, The Cancer Controversy*. Tons of "evidence" were presented but dismissed by hard-line statisticians (see, e.g., Eysenck, 1980) on the ground that those studies were based on statistical correlation, not controlled experiments.

For instance, some researchers pointed out that the mortality ratio of lung cancer for smokers vs. nonsmokers is 10.8, seemingly strong evidence that smoking does cause lung cancer. But other researchers quickly found a mortality ratio of 0.66 for colorectal cancer and a ratio of 0.26 for Parkinson's disease. Another scientist also found that among 138 patients with Parkinson's disease, only 70% had ever smoked, compared with 84 of 166 nonsmokers who do not have Parkinson's disease (Eysenck, 1980, p. 17–18), a finding which seems to suggest that smoking will reduce the rate of Parkinson's disease.

For statisticians, it is established that many diseases are linked with smoking. But it is troublesome if we take it for granted that lung cancer, heart disease, and many other lethal illnesses are the direct consequence of smoking cigarettes.

In a private conversation in 1985, Freedman said that he had finally accepted that smoking causes lung cancer, because tobacco smoke condensate painted on the skin of mice produces skin cancers. (See also Freedman, 1987). This led a friend of mine to comment that "after all, the conclusions in Fisher (1959) and FPP (1978) were wrong."

**III. Rubin's Model and Controlled Experiments**

Causal statements often take, explicitly or implicitly, the form "if-then-else," a mode of reasoning similar to the deductive logic formulated by Aristotle in about 400 B.C. However, it is easy to confuse "one thing causes another" with "one thing follows another."

As an example, Aristotle asserted that cabbages produce caterpillars daily—a seemingly "logical" conclusion only to be refuted by controlled experiments carried out by Francesco Redi in 1668.

Aristotle, often ridiculed for the causal analyses in his book *Physics*, might have been saved from drawing erroneous conclusions if he had learned the following motto formulated by Rubin and Holland (see Holland, 1986):

*NO CAUSATION WITHOUT MANIPULATION.*

This motto echoes Fisher's stand on the issue of smoking and lung cancer.

In addition, Holland (1986) went further to conclude that "examples of the confusion between attributes and causes fill the social science literature" (p. 955), and that "the causal model literature [in social science] has not been careful in separating meaningful and meaningless causal statements and path diagrams" (p. 958).

Glymour (1986) responded to Holland's assertion by saying that "there is no need for this sort of care." Glymour's major argument is that probabilistic causality (as advocated by Holland) is "counterfactual." The argument is quite cute and has gained popularity among certain statisticians (e.g., Freedman, 1987; Hope, 1987; Holland, 1987, 1988). On the other hand, some statisticians have found it too abstract to penetrate. According to Glymour, statistical causalities are **counterfactual**, in the sense that

1. **Counterfactuals can be logically false:**
   - If X were the case then X and not X would be the case.
   [For example, with 95% probability, event A will happen.]

2. **Counterfactuals can logically entail one another:**
   - "If X were the case then Y would be the case" entails "if X were the case then Y or Z would be the case."
   [For example, with 95% confidence, A causes B.]

Statistical Causality and Law-Like Relationships
3. Counterfactuals have different entailment relations than do ordinary material conditionals: [emphasis supplied]

[In probabilistic causality,] "If X then Y" entails "If X and Z then Y."

But in material conditionals, "If X were the case then Y would be the case" does not entail "If X were the case and Z were the case then Y would be the case."

For example, "If I had struck the match just now it would have lighted" is true, but "If I had struck the match just now and there had been no oxygen in the room, it would have lighted" is false. (Glymour, 1986, p.964).

Simply put, counterfactuals are not logical; neither are probabilistic causalities. An implication of Glymour’s analysis is that Holland’s accusation that statistical causality in social science often is meaningless) is itself meaningless. Glymour (1987, p. 17) further asserted that

we do not assume that causal relations can be reduced to probabilistic relations of any kind, but neither do we contradict such assumptions.

Counterfactual accounts of causality, according to Glymour (1986, p. 965), “have the advantage that they seem to make it easy to understand how we can have knowledge of causal relations, and equally, to ease our understanding of the bearing of statistics on causal inference.” The disadvantage is that

they appeal to unobservables—to what would be true if . . . , and to what goes on in possible worlds we will never see. They, therefore, present us with a mystery as to how we can know anything about causal relations.

Glymour’s analysis of counterfactuals is not intended to discredit probabilistic accounts of causality; instead, he attempts to dismiss Holland’s motto and the Fisharian tradition that correlations cannot all be taken as causalities.

However, Glymour’s arguments for such statistical anarchy are not acceptable for the following reasons. First, his “material conditionals” belong to the category of “inductive logic,” which simply is not logic (but indeed counterfactual). When it comes to physics or chemistry (e.g., lighting a match), things can be so stable that they mimic deductive logic. But no matter how many times you have failed to light a match in a room without oxygen, no logic will guarantee that you will fail again the next time.

In our physical world, stability exists, but there is no certainty. This assertion does not mean that science is impossible or that our world is in chaos. Rather, it means that a logical justification of scientific laws is impossible (see Chapter 2, Section II, “Hume’s problem”). During the course of human history, physical laws that describe nature have proven stable and have thus provided a sound foundation for all sciences. Both “material conditionals” and “probabilistic counterfactuals” rest on this same foundation, although they may have different strengths.

Statistical Causality and Law-Like Relationships

My second objection to Glymour’s anarchism is that his account of probability is inductive, not deductive. Here “inductive probability” means the fiducial probability calculated from a batch of encountered data, and “deductive probability” means that we can deduce the logical consequences of a statistical procedure. For example, if a box contains a red marble and a blue marble, then the probability of getting a red marble is 50%. This is deductive, not inductive.

In real-world applications, of course, things are not so clear-cut. For instance, a nationwide study shows that taking an aspirin every other day can reduce the risk of a first heart attack by 47 percent (the New England Journal of Medicine, January 28, 1988).

The study has been widely publicized. In hindsight, many people claim that they knew the beneficial effect a long time ago. For example, Kevin Begos (The New York Times, March 18, 1988) complained that in a 1966 Louisiana medical journal a study had demonstrated the effect but it “was objected that the study was not definitive.”

Undoubtedly the results in both studies are all counterfactuals. But why does the scientific community favor one and discriminate against another? The reason is that the new study utilized randomization and involved 22,071 physicians. A box model containing red and blue marbles can thus be used for this study, and the justification of the model is the physical act of randomization. The prediction accuracy from the model now can be controlled, and the statistical procedure can be evaluated by precise mathematics.3

In short, there are two kinds of statistical causalities: physical causality based on statistical manipulation (i.e., randomization), and psychological causality based on enumeration from non-experimental data. The algebra for both causalities is the same, but one follows the methodology of Francesco Redi, the other of Aristotle.

In logic, a syllogism is called an “abduction” if the minor premise has no proof. In the case of Glymour’s anarchism, the masquerading of statistical correlation as causation is neither deduction nor induction, but abduction (i.e., plain kidnapping).

Holland (1986) proposed Rubin’s model as a way of dealing with the problem of causation. In literature, the complexity of Rubin’s model keeps growing. However, one could argue that the model, if stripped down, is essentially Fisher’s randomized comparison model.

Holland was aware that the topic of causal inference is messy and mischievous (p. 950–958). In addition, he formulates explicitly “the Fundamental Problem of Causal Inference,” which means that (strictly speaking) we cannot observe the effect of a treatment on a population unit. This inability is, at bottom, Hume’s problem of induction.

Nevertheless, our knowledge of nature advances generation after generation. According to Holland there are two general solutions by which we can
get around "the Fundamental Problem of Causal Inference": the scientific solution and the statistical solution.

The scientific solution is to exploit various homogeneity or invariance assumptions.

The statistical solution is different and makes use of the population U [and the average causal effect over U, the universe] in a typically statistical way. (p. 947) [emphasis supplied]

According to Holland (pp. 948–949), the statistical solution (or Rubin's model) includes the following special cases: (1) unit homogeneity, (2) constant effect, (3) temporal stability, and (4) causal transience. Apparently this framework includes "the scientific solution" as a special case.

But this conclusion is misleading. In practice, scientists explore causal patterns in nature. They also use established theories and mathematical tools (e.g., differential equations) to expand the territory of human knowledge. It is beyond me how Rubin's model can accommodate the rich techniques used by general scientists. In particular, I don't know how Rubin's model can help scientists to decipher a genetic code, to develop the mathematical equation for a chemical reaction, or to pin down the cause of sunspot activities.

In statistics, the purpose of randomization is to achieve homogeneity in the sample units. Further, randomization itself is independent of any statistical formula. For these reasons, the physical act of randomization, in our opinion, should be classified as a "scientific solution," rather than a "statistical solution."

More importantly, it should be spelled out that stability and homogeneity are the foundations of the statistical solution, not the other way around.6,7

Here the stability assumption means that the contents in the statistical box-model are as stable as marbles, an assumption stated in elementary statistics books but seldom taken seriously by statistics users. The box models in many "statistical analyses," as we shall see in the subsequent sections, contain not marbles, but only cans of worms.

IV. Rubin's Model and Observational Studies

In order to survive, we all make causal inferences. Even animals make causal inferences; just imagine how a mouse will run when he sees a cat.

When drawing a causal inference, most of the time neither we nor the mouse have a controlled experiment to rely on. Rubin and Holland's motto "no causation without manipulation" then sounds impractical and in fact untrue in numerous situations.

Here are some examples: (1) A few years ago, my life turned upside-down because of the death of a close friend—a cause I couldn't and wouldn't manipulate. (2) A smart move by a competitor causes a businessman to panic. (3) The meltdown of the stock market causes severe financial losses to some of my acquaintances.

Examples like these are plentiful, and the causal links are undeniable. But how far we can generalize and apply these facts to other parts of life is a totally different story. The generalization issues can easily take us into Hume's problem of induction.

Rubin and Holland are aware of the limitation of their motto. In an active attempt to cover both experiments and observational studies under the single umbrella of Rubin's model, they stress that the key element of their motto is "the potential (regardless of whether it can be achieved in practice or not) for exposing or not exposing each [population] unit to the action of a cause" (Holland, p. 936). It appears that this interpretation might open the door to numerous as-if experiments in social science studies. If so, then anyone could draw illegitimate conclusions without guilt, and the motto might become much ado about nothing.

Fortunately, Rubin and Holland are statisticians who wouldn't give up the Fisherian tradition easily. In their attempt to enlarge Fisher's method to cover observational studies, Rubin and Holland (and Rosenbaum, 1984, 1968; Rosenbaum and Rubin, 1984a, 1984b) have tried very hard to maintain the rigor of their model.

For instance, Rubin and Holland insist that a characteristic that cannot be potentially manipulated is not a cause. Under this restriction, race and gender, for example, cannot be considered as causes (Holland, p. 946). Tapping the same theme, Rosenbaum and Rubin (1984a, p. 26) insisted that age and sex do not have causal effects because it is not meaningful to discuss the consequences of altering these unalterable characteristics. [emphasis supplied]

It is interesting to see the reaction if you ask someone whether race, sex, and age are causes of certain events. Prepare yourself. The answers may not be trivial at all.

Among academicians too, opinions on race, sex, and age as causes are unsettling. Here are some examples. Pratt and Schlaifer (1984) maintained that sex and age can have a causal effect and that the effect of sex (in some cases) can be consistently estimated because "sex is randomized" (p. 31).

Freedman, Pisani, and Purves (1978) accepted the notions of "the effect of race" (p. 16), "the effect of age" (p. 39), and "the psychological effect" (p. 8). In a later work, Freedman (1988) again uses language like "aging affects metabolic processes."

Clark Glymour (1986) expressed that he was not convinced that race and gender should be excluded from the category of causes. He cited an example: "My parents told me that if I had been a girl, I would have been named Olga, instead of Clark."
Chapter 3

Statistical Causality and Law-Like Relationships

According to Rosenbaum (1984, p. 42), a treatment is strongly ignorable if (1) the responses, r(0), r(1), ..., r(T), are conditionally independent of the treatment assignment z given the observed covariates X, and (2) at each value of X, there is a positive probability of receiving each treatment.

In mathematical notations, conditions (1) and (2) can be formulated as follows:

\[ \Pr[z | X, r(0), r(1), \ldots, r(t)] = \Pr[z | X], \]

\[ 0 < \Pr[z = t | X] < 1, \text{ for } t = 0, \ldots, T, \text{ for all } X. \]

In plain English, "strong ignorability" means that the way treatments were assigned to units can be ignored in statistical computation (if X is given). For example, in a randomized comparison, X denotes the codes of different populations, and causal inference can be drawn straightforwardly by the use of statistical formulas.

In observational studies, the process by which treatments were assigned to units often is poorly understood, and thus "strong ignorability" may not apply. It would be wonderful if "strong ignorability" could be tested by certain statistical formulas. Regarding this important task, Rosenbaum (1984, pp. 45) wrote:

If the causal mechanism is correct and, in particular, if

\[ f(j) = f(k) \text{ for some } j, k \]

where \( f(j) \) is a function of \((r(j), X)\), then rejection of the hypothesis of strong ignorability given \( X \) is also rejection of the hypothesis that the unobserved covariates \( U \) have the same distribution in the various treatment groups.

After all these intellectual investments in Rubin's model (or the Rubin-Holland-Rosenbaum model), we deserve some reward.

EXAMPLE (Rosenbaum, 1984, p. 45–46). This example concerns the effect of fallout from nuclear testing on childhood leukemia. It is a follow-up of a study by Lyon et al. (1979).

From 1951 to 1958, at least 97 above-ground atomic devices were detonated in the Nevada desert. Data on mortality rates in high-fallout counties are shown as follows (n = number of incidents, r = mortality rate per 100,000 person-year):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 7</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>r = 2.1</td>
<td>4.4</td>
<td>2.2</td>
</tr>
</tbody>
</table>
The data show that the mortality rate doubled from 2.1 to 4.4 and later set back to 2.2. These results seem to indicate that the nuclear fallout caused about 16 children to die of leukemia (other cancers aside). This conclusion is compatible with laboratory experiments, in which the carcinogenic effects of radiation are well established.

However, the issue becomes puzzling if we also look at the mortality rates of low-fallout counties.

<table>
<thead>
<tr>
<th>Low-exposure counties</th>
<th>High-exposure counties</th>
<th>Low-exposure counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-fallout</td>
<td>r = 4.0</td>
<td>3.9</td>
</tr>
<tr>
<td>High-fallout</td>
<td>r = 2.1</td>
<td>4.4</td>
</tr>
</tbody>
</table>

From this table, the previous conclusion that nuclear fallout has an effect on leukemia in high-fallout areas now appears illusory. And it becomes even more puzzling if we compare the mortality rates of other cancers:

<table>
<thead>
<tr>
<th>Low-exposure counties</th>
<th>High-exposure counties</th>
<th>Low-exposure counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-fallout</td>
<td>r = 4.6</td>
<td>4.2</td>
</tr>
<tr>
<td>High-fallout</td>
<td>r = 6.4</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Taken at face value, the data in the above table suggest that the mortality rate was lowest (at 2.9) in high-fallout counties during the high-exposure period!

Lyon et al. (1979) offered several possible explanations for these numbers:

1. Migration: The residents of the high-fallout area might have migrated to the low-fallout area.
2. Wind blow: Various weather conditions might have altered the fallout to unmonitored areas.
3. Bad luck: A chance clustering of leukemia (or cancer) deaths might have occurred in a short time.

The investigators (Lyon et al.) could not control these factors, nor did they have data related to these issues.

Without collecting more data, can Rubin’s model shed new light on the problem? Rosenbaum (1984, p. 41) stated:

Statistical Causality and Law-Like Relationships

If treatment assignment is strongly ignorable, then adjustment for observed covariates is sufficient to produce consistent estimates of treatment effects in observational studies. A general approach to testing this critical assumption is developed and applied to a study of the effects of nuclear fallout on the risk of childhood leukemia. (emphasis supplied)

This statement is quite a promise. And statistical methods may score another triumph in a very difficult real-life application.

In order to understand thoroughly Rosenbaum’s new analysis of a batch of old data, one has to be familiar with the notation of strong ignorability. In addition, one has to understand the Poisson kernel of the log likelihood of the raw counts. Also, the reader has to follow through a “log linear model on a flat” (p. 46) for certain counts:

\[
\log (m_{xy}) = \log (N_{xy}) + u + u_{C(c)}
\]

The likelihood ratio chi-square statistic for testing the fit of the model is 13.6 for 6 degrees of freedom, with a significance level of .03. To obtain a better fit, Rosenbaum considered another log linear model:

\[
\log (m_{xy}) = \log (N_{xy}) + u + u_{Y(y)} + u_{E(e)} + u_{C(c)} + u_{EC(ec)}
\]

This model provided a satisfactory fit, with a significance level of .50. Finally Rosenbaum reached the following conclusions (p. 46):

1. there have been temporal changes in reported cancer mortality aside from any effects of fallout, and
2. the high and low exposure counties had different mortality rates both before and after the period of above-ground testing and, moreover, these differences followed different patterns for leukemia and other cancers.

Well, after all the hard work, the results are not much different from commonsense conclusions one might draw by simply looking at the tables. In other words, it is hard to see the logical necessity of using Rubin’s model to arrive at these results (and, in my judgment, there is none).

In summary, in the attempt to derive causation from observational studies, the so-called Rubin’s model (a new rising star in the statistical literature) is still in its infant stage, and it is highly probable that the model will remain so forever.

V. CAUSAL INFERENCE IN SAMPLE SURVEY AND OTHER OBSERVATIONAL STUDIES

In this section we will take up the issue of causal inference in sample surveys. Specifically, we will discuss the problem of using sampling techniques to jus-
tify a causal link in observational data. To begin with, let’s examine a case history from Freedman, Pisani, and Purves (1978, pp. 38-41).

A drug study was funded by NIH (the National Institute of Health) to assess the side effects of oral contraceptives, “the pill.” One issue considered by the study was the effect of the pill on blood pressure. About 17,500 women aged 17 to 58 in the study were classified as “users” and “nonusers.”

Note that randomization was impossible in this case, and a direct comparison of the blood pressures of users and nonusers can be tricky. For instance, about 3,500 women had to be excluded from the comparison, because they were pregnant, post-partum, or taking hormonal medication other than the pill. This action of throwing away data appears to go against the standard teaching of “the Law of Large Numbers,” but it brings us closer to the focus of the problem.

Other controls for confounding factors are needed. For instance, about 70% of the nonusers were older than thirty while only 48% of the users were over thirty. “The effect of age is confounded with the effect of the pill,” FPP wrote. “And the two work in opposite directions.”

As a result, it is necessary to make a separate comparison for each age group. The following table exhibits the percentage of women with systolic blood pressure higher than 140 millimeters.

<table>
<thead>
<tr>
<th>Age</th>
<th>17-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonusers</td>
<td>5%</td>
<td>5%</td>
<td>9%</td>
<td>19%</td>
</tr>
<tr>
<td>Users</td>
<td>7%</td>
<td>8%</td>
<td>17%</td>
<td>30%</td>
</tr>
</tbody>
</table>

The interpretations offered for the above percentages is as follows (FPP, p. 40):

To see the effect of the pill on the blood pressures of women aged 17 to 24, it is a matter of looking at the percents in the columns for users and nonusers in the age group 17 to 24. To see the effect of age, look first at the nonusers column in each age group and see how the percents shift toward the high blood pressures as age goes up. Then do the same thing for the users. [emphasis supplied]

The “effects” can also be summarized in Figure 4. The chart indicates that age and blood pressure are strongly correlated. But before we wrap up the whole story, note that the chart is similar to Fig. 5 which was used (see Section II, Example 4) by Fong and other scientists to establish “the effect of statistical training” in everyday reasoning. Note that in Section II we have refuted...
Chapter 3

Fong’s study. On what ground can we now accept Freedman’s conclusion? Specifically, at least two questions should be asked:

1. What does one mean by “the effect of age” on blood pressure?
2. Does the study under discussion establish “the effect of the pill” on blood pressure?

For instance, in a laboratory we can manipulate the amount of oral contraceptives, but how are we going to manipulate different levels of age? Under the Rubin-Holland-Rosenbaum doctrine (see Section IV), expressions like “the effect of age” are smoke in logical eyes.

The second question appears easier to answer. “The physiological mechanism by which the pill affects blood pressure is well established,” FPP wrote. “The Drug Study data discussed here document the size of the effect.” In other words, this study does not establish the effect of the pill; rather, it gives rough estimate of the size of a known effect.

But the logical foundation of this conclusion is still very shaky. Recall the study about the effect of nuclear fallout on childhood leukemia (Section IV). In one case, scientists obtained the following result ($r = \text{mortality rate per 100,000 person-year}$):

<table>
<thead>
<tr>
<th></th>
<th>Low-exposure</th>
<th>High-exposure</th>
<th>Low-exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r = 2.1$</td>
<td>4.4</td>
<td>2.2</td>
<td></td>
</tr>
</tbody>
</table>

This table is compatible with laboratory experiments, in which the carcinogenic effects of radiation are well established. However, in another case, the scientists obtained a conflicting result:

<table>
<thead>
<tr>
<th></th>
<th>Low-exposure</th>
<th>High-exposure</th>
<th>Low-exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r = 4.6$</td>
<td>4.2</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>High-fallout</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>counties</td>
<td>$r = 6.4$</td>
<td>2.9</td>
<td>3.3</td>
</tr>
</tbody>
</table>

which suggests that the mortality rate was lowest (at 2.9) in high-fallout counties during the high-exposure period.

Statistical Causality and Law-Like Relationships

As we can see now, Freedman’s conclusion is not based on logic. Otherwise, the same logic would apply to the effect of nuclear fallout.

Nevertheless, we must concede that Freedman’s conclusion is valid. The reasons are that (1) it is based on a large study that involved 14,000 women, and (2) the conclusion has not been challenged by other research findings.

The large sample in the drug study makes the result more convincing. But more importantly, the result is consistent with existing knowledge. By contrast, in the nuclear-fallout example, earlier results conflict with later results. Earth, as one may claim, is a very dirty laboratory, at least dirtier than human bodies.

In brief, this is how knowledge in observational studies progresses: a causal statement (such as the effect of age) is formulated to describe a general characteristic that is consistent with existing knowledge and is deemed useful for other applications. Still, the causal statement is provisional and is subject to further refutation. If other research findings conflict with the statement, then the statement has to be reexamined or discarded. Otherwise, the statement will be regarded as valid and useful.

This process of refutation is, in our opinion, the only way to justify a causal statement in observational study. Rubin’s model or any other exotic statistical schemes provide, in most cases, not help but confusion. Cross-tabulation, regression models, and other statistical techniques are useful for the exploration of causal relationships. But they too do not provide justification of a causal statement.

Note that randomization is frequently used in sample surveys. But its function has nothing to do with causal inference. Rather, it is for the generalization from the sample to the target population. For this reason, causal statements in a randomized sample survey have to be classified as a subcategory of observational study.

In scientific inquiry, controlled experiments usually produce more reliable causal conclusions than observational studies. But this is not a rule. There are numerous exceptions.

For example, in astronomy we cannot manipulate any quantities whatever, yet predictions in astronomy are often more accurate than those produced by controlled experiments in clinical trials.

In a personal conversation at the 1987 ASA Annual Meeting, Holland stunned me by saying that modern astronomers usually carry out a lot of experiments before they draw conclusions. This keen observation unfortunately does not apply to geology. For example, drilling for petroleum is an enormous gamble. When a geological study indicates the place where petroleum might have accumulated, there is less than a 10% chance that oil is actually present; and there is only a 2% chance that it is present in commer-
cially useful amounts. (Professional drillers sometimes joke that “random drilling may be better.”)

The intrinsic unpredictability of many observational studies is beyond any logic. The gap between prospective and retrospective analyses is also unbridgeable. We therefore disagree with Holland’s contention that Rubin’s model (or anybody’s model) is a framework for both experiments and observational studies. Holland has also asserted that “the idea that Rubin’s model is somewhat incapable of accommodating time-series data is misleading.” In my opinion, it is Holland’s statement that is misleading. For example, if Holland hasn’t forgotten his own motto, I hope he can tell us how he can literally manipulate time-series data or any encountered data in geoscience.

For randomized experiments, causal inferences can be drawn straightforwardly by statistical bulldozer. For observational studies, causal inferences have to be assessed on a case-by-case basis; a generic approach like Rubin’s model may create more confusion than the clarity the model originally intended.

For those who are serious about identifying the cause from encountered data, a piece of advice from R. A. Fisher (quoted from Rosenbaum, 1984, p. 43) usually proves helpful: Make your theories elaborate.

This advice may be foreign to some statistics users. However, the practice of the maxim is an integral part of process quality control, an orphan in traditional statistical training.

For instance, when problems emerge in a manufacturing process, how do engineers set forth to find the cause and solve the problem? They explore related scientific knowledge. More specifically, the techniques include brainstorming, Ishikawa charts, Pareto diagrams, and statistical control charts.

Here is a beautiful example from Grant and Leavenworth (1980, p. 95). In this case study, statistical control charts were maintained to monitor the quality of specially treated steel castings. For a number of months, the final products of castings did meet rigid requirements. But all of a sudden points on X-charts went out of the control limits.

The supervisors in the production department were sure no changes had been made in production methods and at one point attempted to blame on the testing laboratory. Unfortunately, independent tests showed the points continuing to fall out of the control limits.

Pressure mounted on production personnel, and they were forced to explore potential causes further. This activity of finding the cause or causes usually is very frustrating; but for competent scientists, it can also be as challenging (and rewarding) as Sherlock Holmes’ adventures.

To conclude this case study, finally somebody noted that a change in the source of cooling water for heat treatment had been made just before the time the points fell out of control charts. Nobody believed that a change of cooling water could have created so much trouble. However, as a last resort, the origi-

Statistical Causality and Law-Like Relationships

nal source of quench water was restored. Immediately the points fell within the control limits, and the process remained under control at the original level. To sum up the whole story, the great detective had it all (Sherlock Holmes, The Sign of the Four):

How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth.

VI. CAUSES, INDICATORS, AND LATENT VARIABLES

In social or behavioral sciences, unmeasured latent variables often are created to account for a large portion of variability in the data. The statistical techniques used to construct these latent variables are primarily principal components and factor analysis (see, e.g., Morrison, 1976). In the case of factor analysis, the trick goes like this. Let \( X_1, X_2, \ldots, X_p \) be observable random variables; next let \( Y_1, Y_2, \ldots, Y_m \) \( (1 \leq m < p) \) be unmeasured latent variables that will be used to construct a simpler structure to account for the correlations among the Xs:

\[
X_1 = a_{11}Y_1 + \cdots + a_{1m}Y_m + c_1
\]

\[
\ldots
\]

\[
X_p = a_{p1}Y_1 + \cdots + a_{pm}Y_m + c_p
\]

Occasionally one is lucky: \( m = 1 \). Some examples of latent variables are the so-called “cognitive ability” or “socioeconomic status”; the latter is routinely used in sociological studies as a latent variable that affects education, income, occupation, and other observable variables (Glymour, 1987).

Latent variables of this sort appear to be outside the sphere of Rubin’s model. In numerous cases, these variables are useful constructs for policy- or decision-making. But a nasty question has been haunting the research workers in the field: Can the latent variables be considered “causes?”

To answer this question, some examples from natural science are worth retelling. To begin with, let’s consider the story of John Dalton, an Englishman who has been widely recognized as the father of modern chemistry (for his contribution to the theory of the atom).

Greek philosophers (such as Democritus) are sometimes credited for their invention of the notion of “atom.” But Dalton’s contribution (1803) was different, because it was based on the laws of conservation of mass and definite proportions—laws that had been derived from direct observation and had been tested rigorously. The theory can be expressed by the following postulates (Brady and Humiston, 1982):

- All elements are made up of atoms which cannot be destroyed, split, or created.
• All atoms of the same element are alike and differ from atoms of any other element.
• Different atoms combine in definite ratios to form molecules.
• A chemical reaction merely consists of a reshuffling of atoms from one set of combinations to another. The individual atoms remain, however, intact.

Dalton’s theory was wrong on numerous counts; but the theory proved successful in explaining the laws of chemistry and was embraced by the scientific community almost immediately. We now know that in one drop of water there are more than 100 billion atoms. With a quantity this small, how did Dalton manipulate an atom? Specifically, how did he gauge the weight of an atom, or measure the distance between atoms in a crystal of salt? The answer is that none of these tasks was accomplished by Dalton or his contemporaries.

As a matter of fact, the doubt of the very existence of atoms and molecules did not rest completely until the systematic measurements of the sizes of atoms were carried out in 1908 by J. B. Perrin (the winner of the 1926 Nobel Prize for Physics). This confirmation occurred 105 years after Dalton’s original proposal.

Dalton’s concept of atom was, in his time, an unmeasured latent factor; but it was accepted as a fruitful working hypothesis by “reasonable” scientists. Without Dalton’s theory, modern chemistry and physics might have had to wait until the next century.

The second example of a latent factor is related to the discovery of an important elementary particle. In the subatomic world, protons (p) may transmute into neutrons (n), and vice versa. The reactions are called beta-decay, which involves the release of an electron (e) or antielectron (\(\bar{e}\)):

\[
\begin{align*}
n &\rightarrow p + e \\
p &\rightarrow n + \bar{e}
\end{align*}
\]

A curious feature of beta-decay is that the emitted electrons or antielectrons sometimes emerge with one energy and sometimes with another. This phenomenon prompted some theorists to speculate on the failure of the conservation law of energy. In 1931 Wolfgang Pauli proposed that an elementary particle (now called “neutrino”) is the cause of the variability of the energy of beta-rays. At the time there was no experimental evidence for the existence of any extra particle in beta-decay, and Pauli’s proposal was only an act of faith, based on nothing but a firm belief in the conservation laws (Ohanian, 1987, p. 407). Experimental evidence for the existence of neutrinos was not available until 1953, 22 years after Pauli’s original proposal.

A more spectacular example than Pauli’s neutrino is Maxwell’s electromagnetic waves. Prior to Maxwell’s work, Faraday found that moving a magnet through a coil of copper wire caused an electric current to flow in the wire. The discovery led to the development of electric generator and motor. In a striking fashion, Maxwell used the experimental discoveries of Faraday to arrive at a set of equations depicting the interaction of electric and magnetic fields. On the basis of these equations, he predicted (in 1884) the existence of electromagnetic waves that move through space with the speed of light. This prediction was confirmed 23 years later by H. Hertz (World Book, 1982). In his experiment, Hertz used a rapidly oscillating electric spark to produce waves with ultra-high frequencies; he also showed that these waves cause similar electrical oscillations in a distant wire loop. The manipulation of these waves later led to the invention of radio, television, and radar.

In the above examples, the latent factors could not be directly measured or manipulated when they were initially introduced to modern science. For this reason, Rubin and Holland’s motto, “no causation without manipulation” may need a more flexible interpretation. In short, if a theory is sound and indirect confirmations do exist, a man-made causal schema like Dalton’s atomism deserves the respect and attention of working scientists.

In social and behavioral sciences, the statistical techniques for extracting latent variables have been a constant target of criticism. One of the most famous rejections of latent variables came from B. F. Skinner (1976), the godfather of behaviorism. Simply put, Skinner maintained that anything that can be explained with latent variables can be explained without them (a la Glymour, 1987).

In opposition to Skinner’s and Holland’s rejections of latent variables as legitimate causes, Glymour (1987) responded that “the natural sciences are successful exactly because of their search for latent factors affecting the phenomena to be explained or predicted.” To substantiate this claim, Glymour argued:

Newtonian dynamics and celestial mechanics, the theory of electricity and magnetism, optics, chemistry, genetics, and the whole of modern physics would not have come to pass if natural scientists [had] behaved as the critics of latent variables prefer.

Gravitational force, electrical fluids and particles, electromagnetic fields, atoms and molecules, genes and gravitational fields, none of them could be directly measured or manipulated when they initially became part of modern science.

Glymour is quite right on this mark. But one might feel uncomfortable when he extrapolates the success of latent factors in natural science to other branches of science:

Critics may grant that the introduction of latent variables has been essential in the natural sciences, but maintain that they are inappropriate in the social and behavioral sciences. It is hard to think of any convincing reason for this view.
Glymour is convinced that heuristic search, utilizing artificial intelligence and statistical correlation matrices, could produce latent variables for important scientific discoveries (Glymour, 1987, p. 23):

In Part II of this book we will show that linear causal models can explain in the same fashion that, say, Daltonian atomism explained the law of definite proportions, that Maxwell explained electromagnetic phenomena, or that Copernican theory explained regularities of planetary motion. [emphasis supplied]

Scientific laws, as it appears, can now be easily cloned by statistical correlation matrices. Figure 6 is a typical example of Glymour's search for a causal structure about industrial and political development (p. 148). GNP is gross national product, Energy is the logarithm of total energy consumption in megawatt hours per capita, Labor is a labor force diversification index, Exec is an index of executive functioning, Party is an index of political party organization, Power is an index of power diversification, CI is the Cutright Index of political representation, and the $e$'s are the added error terms.

To shorten the notations, let $ID = \text{industrial development}$ and $PD = \text{political development}$. Then the path from $ID$ to $GNP$ and the path from $ID$ to $Energy$, for example, are equivalent to the linear statistical equations:

$$GNP = a_1 * ID + e_1$$
$$Energy = a_2 * ID + e_2$$

A revised model (Fig. 7) can be achieved by adding more paths to the original skeleton. The path regression for $ID > GNP$ and $ID > Exec$ read as follows:

$$GNP = a_4 * ID + e_1$$
$$Exec = a_4 * ID + b_4 * PD + e_4$$

Note that in this model, ID and PD are latent variables, while the rest are measured quantities. The last equation is thus a linear statistical law relating a measured variable with two unmeasured variables.

The fit of the revised model is marginal, with a P-value = .053 for a chi-square test. Naturally one would look for alternatives or to improve the model by adding further connections. With the aid of a computer program called TETRAD (Glymour, p. 162), 38 models were located and ordered by the probability of their chi-square statistic see Fig. 8). In this figure, TTR is a TETRAD measure of the goodness-of-fit, which, according to Glymour, is negatively correlated with the chi-square test. But one can easily detect the inconsistency of these measures in the pairs of models such as $(M2,M3), (M4,M5), (M3,M37)$, etc.; where $M2$ stands for model 2.

In one of his concluding remarks, Glymour wrote,

To justify the assumptions of the [best] model, we need either substantive knowledge or grounds for thinking that no other plausible model explains the data as well.

This conclusion certainly should be underlined many times by all statistics users. But the bottom line is: "What is the causal structure of all these variables?" In Glymour's discussions of the computer outputs, we find nothing clear except the following: (1) the substantive knowledge was weak, (2) the measurements of the variables (such as Exec, an index of executive functioning) and the measures of goodness-of-fit are rough, (3) one may find infinitely many other plausible models by adding more variables, more connections, time-lag effect, and more sophisticated transformations.

In sum, one can hardly see any causal structure from these models. And one will be better off if one does not take these models seriously as scientific laws. To support this position, in Chapter 4, under the title of "Amoeba Regression," we will present further examples to demonstrate the structureless nature of such statistical models.

Similar observations about path analysis were made in Freedman (1985, 1987). After careful and interesting elaborations, Freedman rejected flatly the idea that linear statistical laws in social science are at all comparable to
### All Models Located with TETRAD

<table>
<thead>
<tr>
<th>Model</th>
<th>TTR</th>
<th>Equations Explained</th>
<th>(p(X^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) en-&gt;po, en-&gt;ex, ID-&gt;po</td>
<td>0.666</td>
<td>10</td>
<td>0.6082</td>
</tr>
<tr>
<td>2) en-&gt;po, en-&gt;ex, ID-&gt;ex</td>
<td>0.666</td>
<td>10</td>
<td>0.6004</td>
</tr>
<tr>
<td>3) en-&gt;po, en-&gt;ex, gp-&gt;pa</td>
<td>0.270</td>
<td>5</td>
<td>0.5752</td>
</tr>
<tr>
<td>4) en-&gt;po, en-&gt;ex, pa-&gt;en</td>
<td>0.246</td>
<td>4</td>
<td>0.5680</td>
</tr>
<tr>
<td>5) en-&gt;po, en-&gt;ex</td>
<td>0.666</td>
<td>10</td>
<td>0.5504</td>
</tr>
<tr>
<td>6) en-&gt;po, en-&gt;ex, ci-&gt;la</td>
<td>0.251</td>
<td>4</td>
<td>0.5198</td>
</tr>
<tr>
<td>7) en-&gt;po, en-&gt;ex, la-&gt;pa</td>
<td>0.202</td>
<td>6</td>
<td>0.4792</td>
</tr>
<tr>
<td>8) en-&gt;po, la-&gt;pa, po-&gt;ex</td>
<td>0.202</td>
<td>6</td>
<td>0.4085</td>
</tr>
<tr>
<td>9) ID-&gt;po, ID-&gt;ex, PD-&gt;en</td>
<td>0.560</td>
<td>5</td>
<td>0.2940</td>
</tr>
<tr>
<td>10) ID-&gt;po, en-&gt;ex, pa-&gt;en</td>
<td>0.328</td>
<td>5</td>
<td>0.1526</td>
</tr>
<tr>
<td>11) ex-&gt;en, en-&gt;po</td>
<td></td>
<td></td>
<td>0.1434</td>
</tr>
<tr>
<td>12) en-&gt;po, la-&gt;ex, pa-&gt;en</td>
<td>0.213</td>
<td>3</td>
<td>0.1234</td>
</tr>
<tr>
<td>13) ID-&gt;po, en-&gt;ex, ex-&gt;gp</td>
<td>0.358</td>
<td>5</td>
<td>0.1222</td>
</tr>
<tr>
<td>14) en-&gt;po, gp-&gt;ex, en-&gt;pa</td>
<td>0.220</td>
<td>6</td>
<td>0.1189</td>
</tr>
<tr>
<td>15) en-&gt;po, la-&gt;ex, ex-&gt;pa</td>
<td>0.223</td>
<td>5</td>
<td>0.1108</td>
</tr>
<tr>
<td>16) en-&gt;po, gp-&gt;ex, po-&gt;pa</td>
<td>0.220</td>
<td>6</td>
<td>0.1002</td>
</tr>
<tr>
<td>17) ID-&gt;po, po-&gt;en, ex-&gt;en</td>
<td>1.670</td>
<td>8</td>
<td>0.0931</td>
</tr>
<tr>
<td>18) en-&gt;po, la-&gt;ex, la-&gt;pa</td>
<td>0.222</td>
<td>5</td>
<td>0.0819</td>
</tr>
<tr>
<td>19) en-&gt;po, gp-&gt;ex, la-&gt;pa</td>
<td>0.146</td>
<td>7</td>
<td>0.0551</td>
</tr>
<tr>
<td>20) ID-&gt;ex, ID-&gt;po</td>
<td>1.149</td>
<td>11</td>
<td>0.0530</td>
</tr>
<tr>
<td>21) ID-&gt;ex, po-&gt;en, ex-&gt;en</td>
<td>1.364</td>
<td>8</td>
<td>0.0477</td>
</tr>
<tr>
<td>22) en-&gt;po, pa-&gt;po, la-&gt;pa</td>
<td>0.272</td>
<td>5</td>
<td>0.0438</td>
</tr>
<tr>
<td>23) en-&gt;po, po-&gt;en</td>
<td></td>
<td></td>
<td>0.0336</td>
</tr>
<tr>
<td>24) en-&gt;po, la-&gt;po, la-&gt;pa</td>
<td>0.198</td>
<td>4</td>
<td>0.0246</td>
</tr>
<tr>
<td>25) en-&gt;po, pa-&gt;po, en-&gt;pa</td>
<td>0.286</td>
<td>6</td>
<td>0.0211</td>
</tr>
<tr>
<td>26) en-&gt;po, la-&gt;po, ci-&gt;la</td>
<td>0.198</td>
<td>4</td>
<td>0.0161</td>
</tr>
<tr>
<td>27) T-&gt;ex, T-&gt;en, T-&gt;po, T-&gt;ID, T-&gt;PD</td>
<td>0.565</td>
<td>8</td>
<td>0.0145</td>
</tr>
<tr>
<td>28) en-&gt;po, gp-&gt;po, pa-&gt;en</td>
<td>0.237</td>
<td>4</td>
<td>0.0135</td>
</tr>
<tr>
<td>29) en-&gt;po, gp-&gt;po, la-&gt;pa</td>
<td>0.206</td>
<td>5</td>
<td>0.0133</td>
</tr>
<tr>
<td>30) en-&gt;po, gp-&gt;po, ci-&gt;en</td>
<td>0.237</td>
<td>4</td>
<td>0.0108</td>
</tr>
<tr>
<td>31) en-&gt;po, pa-&gt;po, ex-&gt;pa</td>
<td>0.286</td>
<td>6</td>
<td>0.0096</td>
</tr>
<tr>
<td>32) T-&gt;ex, T-&gt;en, T-&gt;po</td>
<td>2.345</td>
<td>7</td>
<td>0.0015</td>
</tr>
<tr>
<td>33) ex Cen, en Cpo, ex Cpo</td>
<td>2.345</td>
<td>7</td>
<td>0.0015</td>
</tr>
<tr>
<td>34) gp-&gt;po, pa-&gt;po, en-&gt;pa</td>
<td>0.220</td>
<td>6</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td>35) ex-&gt;en, po-&gt;en</td>
<td></td>
<td></td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td>36) gp-&gt;po, gp-&gt;po, la-&gt;pa</td>
<td>0.220</td>
<td>6</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td>37) gp-&gt;po, gp-&gt;po, la-&gt;pa</td>
<td>0.209</td>
<td>6</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td>38) Skeleton</td>
<td>5.300</td>
<td>13</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Figure 8** All models located with TETRAD. Copyright © 1987 by Academic Press, Inc. Reprinted by permission.

### Statistical Causality and Law-Like Relationships

Maxwell's equations or Copernicus-Kepler astronomy. One reason is that the stochastic assumptions embodied in path regression do not stand on water. Another reason is that few investigators in social science do careful measurement work. "Instead, they factor-analyze questionnaires. If the scale is not reliable enough, they just add a few more items" (Freedman, 1985).

\[ \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \]

In many observational studies, causal inferences are plain speculations. In the history of science, such speculations have often turned out to be the seeds of important discoveries. For this reason, we believe it is worthwhile to provide another motto:

**NO CAUSATION WITHOUT SPECULATION**

This motto is a complement, not a rival, to Holland's motto. In short, speculation breeds scientific theory, but the theory has to be tested rigorously against reality. As any insider can testify, in the history of science, many beautiful theories had been put together, only to be shot to pieces by a conflict with experiment (see Will, 1986). Experimentation is often called "the Queen of Science." Certainly no exaggeration.

But a word of caution is that statistics users often equate the word "test" with statistical testing. As an example, Granger (a towering figure in econometric research) wrote (1986): "If causality can be equated with some measurable quantity, then statisticians should be able to devise tests for causation" [emphasis supplied].

As a statistician, I feel honored that statistics users have this kind of confidence in my profession. On the other hand, I am also deeply troubled, because in most cases a scientific theory of cause-and-effect has to be tested by the scientists themselves, not by statisticians. Statisticians have indeed made important contributions to numerous instances of scientific testing; but their role is often blown out of proportion.

Consider this example about the effect of gravitational force on a falling object (Scientific American, March 1988). Newton and Einstein both maintained that an object's gravitational acceleration is independent of its mass and substance. New theories challenge this fundamental notion. But how can we put the issue to a test? Can we do so by calling in a statistician to perform a hypothesis testing? Not a chance.

Not only in physics and chemistry, but also in social science studies, statistical methods alone yield no golden promise as Granger hopes. As an example, sending people to school is often believed to increase their income. But how can we measure the effect and "devise tests for causation" (a la Granger)? The solution is not straightforward, if it exists at all.\(^{10}\)
Chapter 3

NOTES


2. In fact, a recent study (Science News, June 4, 1988) provides evidence that an absence of tumor-suppressor genes may lead to lung cancer.

3. Another study involving 5,000 British physicians failed to find any similar benefits from daily aspirin intake (Science News, February 6, 1988). In comparison, the U.S. study involved more than 22,000 test subjects and three times as many heart attacks. The law of large numbers gives the U.S. study a winning edge.

4. The study does not provide information on the side effects of aspirin. For this purpose, one needs other studies.

5. This definition of "the statistical solution" to the problem of causal inference is correct. But one has to be aware that "the universe" of a typically randomized study consists only of a handful of subjects.

6. For instance, in a clinical trial, applications of a randomized study to new patients rely on both the stability and homogeneity assumptions of our biological systems.

7. One may argue that the stability assumption can be (and should be) tested by statistical methods. This is true in cases that involve reliability theory and life testing. In most applications, stability assumption is established by enumeration, and does not involve any formal statistical inference.

8. The issues involving race and sex are more controversial, because in some cases the notions that race and sex are causes do have a certain genetic basis.

9. Ingenious applications of this kind are plentiful. See, for example, Ott (1978). Process Quality Control, Trouble-Shooting and Interpretation of Data, pp. 2-6, 34-37, 57-58, 80-82, 110, 112-115, 116-118, 130-132, 146, and 171-173.


REFERENCES


I. DISCOVERING CAUSAL STRUCTURE:
SCIENCE NOW CAN BE EASILY CLONED

In 1987, as we have mentioned in Chapter 3, Glymour and coworkers (Scheines, Spirtes, and Kelly) published a book entitled Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. This ambitious work was funded by three NSF grants. A major goal of the book is to uncover causal structure. For this task, the group headed by Glymour developed a computer program called TETRAD, which is purported to generate scientific theory (p. 58) by peeking into a specific set of data where substantive knowledge of the domain is whistling in the dark (Glymour, p. 7). This attempt is admirable, but it is certainly in sharp contrast to a popular belief (Albert Einstein, 1936) that "theory cannot be fabricated out of the results of observation, but that it can only be invented."

It is well known that experiments are not always practicable (or ethical) and that even when they are, they may not answer the questions of interest. "Faced both with urgent needs for all sorts of knowledge and with stringent limitations on the scope of experimentation," Glymour (p. 3) wrote, "we resort to statistics."

With great confidence in statistical modeling and TETRAD, Glymour (p. 13) asserted: "A TETRAD user can learn more in an afternoon about the existence and properties of alternative linear models for covariance data than could an unaided researcher in months or even years of reflection and calculation."

Some scientists and statisticians oppose TETRAD or the like. Their arguments and Glymour's forceful defense (see Glymour, 1987, pp. 9, 15-59, 234-246, etc.) constitute a lively debate, and the battles of different schools of thoughts are likely to continue for years to come.

With the aid of Glymour's powerful computer package and his dazzling arguments of epistemology (the theory of knowledge), such statistical models may become a new fashion among certain researchers. But as a statistician, I have mixed feelings about Glymour's ambitious program. On the one hand, I am willing to accept the TETRAD activities under the new motto, "No causation without speculation." On the other hand, if TETRAD delivers only an empty promise, the reputation of my discipline will suffer. For this reason, I conducted an experiment to test Glymour's promise.

A set of triplets was created and given to the students in my regression class:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.75</td>
<td>3.42</td>
</tr>
<tr>
<td>2</td>
<td>2.50</td>
<td>3.21</td>
</tr>
<tr>
<td>2</td>
<td>2.25</td>
<td>3.01</td>
</tr>
<tr>
<td>2</td>
<td>2.00</td>
<td>2.89</td>
</tr>
<tr>
<td>2</td>
<td>1.80</td>
<td>2.75</td>
</tr>
<tr>
<td>2</td>
<td>1.50</td>
<td>2.55</td>
</tr>
<tr>
<td>2</td>
<td>1.25</td>
<td>2.38</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>2.30</td>
</tr>
<tr>
<td>2</td>
<td>0.70</td>
<td>2.15</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>2.10</td>
</tr>
<tr>
<td>3</td>
<td>2.50</td>
<td>3.95</td>
</tr>
<tr>
<td>3</td>
<td>2.00</td>
<td>3.63</td>
</tr>
<tr>
<td>3</td>
<td>1.50</td>
<td>3.40</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>3.18</td>
</tr>
<tr>
<td>4</td>
<td>1.75</td>
<td>4.38</td>
</tr>
<tr>
<td>4</td>
<td>1.50</td>
<td>4.29</td>
</tr>
<tr>
<td>4</td>
<td>1.20</td>
<td>4.18</td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
<td>4.10</td>
</tr>
<tr>
<td>5</td>
<td>1.75</td>
<td>5.30</td>
</tr>
<tr>
<td>5</td>
<td>1.50</td>
<td>5.22</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>5.15</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>5.08</td>
</tr>
</tbody>
</table>
I told the students that the background of the data would not be revealed and that they were allowed to use any method to find an equation to fit the data. A few hints were available: (1) \( Y \) is a function of \( X_1 \) and \( X_2 \). (2) There is no serial correlation among the data; thus the Durbin-Watson test and time-series modeling of residuals were not necessary.

The students were excited about the project; it was like a treasure hunt. To begin with, they all plotted the data in scatter diagrams (see Fig. 1). The Pearson correlation coefficients are .9476 for \((X_1, Y)\) and 0.054 for \((X_2, Y)\), with corresponding P-values of .0001 and .810, respectively. Yet a possible relationship between \( X_2 \) and \( Y \) could not be rejected out of hand, because an enhanced diagram reveals a certain pattern (see Fig. 2). The scatter diagrams indicate that a simple algebraic function might suffice. Nevertheless, my students tried different transformations of the variables. After elaborate screening procedures, each student presented his/her best models. Here are some of them (the null hypothesis for each P-value being that the coefficient is zero):

![Figure 1 Scatter diagram for \( Y \) vs. \( X_1 \) and \( Y \) vs. \( X_2 \).]

![Figure 2 Enhanced scatter diagram for \( Y \) vs. \( X_2 \).]

<table>
<thead>
<tr>
<th>Model</th>
<th>( Y = (1.99)X_1^2 + (0.420)X_2^2 )</th>
<th>R-square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P: .0001</td>
<td>96.36%</td>
</tr>
<tr>
<td>2.</td>
<td>( Y = (0.897)X_1 + (0.525)X_2 )</td>
<td>P: .0001</td>
</tr>
<tr>
<td>log(( Y )) = (-.60) + (.9)v( \sqrt{X_1} ) + (.18)X_2</td>
<td>P: .0001</td>
<td>98.29%</td>
</tr>
<tr>
<td></td>
<td>log(( Y )) = (.174) + (.25)X_1 + (.18)X_2</td>
<td>P: .0001</td>
</tr>
<tr>
<td></td>
<td>log(( Y )) = (.282) + (.257)X_1 + (.057)X_2</td>
<td>P: .0001</td>
</tr>
<tr>
<td>6.</td>
<td>( Y = (.247) + (.914)X_1 + (.0897) \cdot \exp(3X_2) )</td>
<td>P: .0001</td>
</tr>
<tr>
<td>7.</td>
<td>( Y = (.678) + (.867)X_1 + (.659) \cdot \log(3X_2) )</td>
<td>P: .0001</td>
</tr>
<tr>
<td>log(( 1/(Y - 1) )) = (.70) + (-.36)X_1 + (-.29)X_2</td>
<td>P: .0001</td>
<td>97.08%</td>
</tr>
<tr>
<td>9.</td>
<td>( Y = (1.03) + (.82)X_1 )</td>
<td>P: .0001</td>
</tr>
</tbody>
</table>

The students were required to perform a comprehensive diagnostic checking of their models, which includes residual plots, stem-leaf and box-whisker plots, normal probability plot, Shapiro-Wilk normality test, F-test, Cook's influence
statistics, correlation matrix of the estimated parameters, variance inflation factors, cross-validation (D. M. Allen's PRESS procedure), etc.

These diagnostic statistics indicate that models 1-8 are all good models. Further, the R-squares of these models are higher than 96%—a lot better explanatory power than most models in social science studies. It appears that any of these models should perform well, at least for the purpose of prediction. "But what is the underlying mechanism that generated the data?" my students asked.

Well, a few days before the assignment, I sat at home and drew 22 right triangles like those in Fig. 3. The assignment the students received was to rediscover the Pythagorean theorem, and they all failed.¹

This experiment led me to coin a new term "amoeba regression," meaning that the causal structures "discovered" by such regression techniques may be as structureless as an amoeba. The experiment also forced me to ponder the role of statistical modeling in social science research.

It was not an overstatement when someone said that "mathematics is the mother of science." Without Greek mathematics (especially Euclidean geometry), I cannot imagine what Newtonian mechanics would be. Without Newtonian mechanics, I cannot imagine what modern physics and technology would be. Moving back to the fundamentals, without the Pythagorean Theorem, I cannot imagine what Euclidean geometry would be.² In short, if a scientific discipline indulges in linear regression but lacks anything fundamental, the discipline may have serious problems.

The above experiment did not accommodate sophisticated transformations (for example, the celebrated Box-Cox transformation). Out of curiosity, I conducted another regression analysis, assuming that a fictitious investigator squares the Y-variable and then fits the second order Taylor expansion:

\[
Y^2 = -0.125 + (0.12)X_1 + (0.28)X_2 + (0.99)X_1^2 + (0.96)X_2^2 + (-0.07)X_1 \cdot X_2
\]

\[
P: \quad 0.80 \quad .67 \quad .46 \quad .0001 \quad .0001 \quad .28
\]

Note that only the quadratic terms are statistically significant. The investigator then refits the data as follows:

\[
Y^2 = (1.007)X_1^2 + (1.02488)X_2^2
\]

\[
P < \quad .0001 \quad .0001
\]

\[
SE: \quad .00308 \quad .01229
\]

This model passes the scrutiny of diagnostic checking with an R-square of 99.99%. The model almost recovers the Pythagorean formula.³

But there are other profound troubles in using Taylor expansion and regression analysis to "discover" fundamental laws. First, by what criterion can we be sure that the last model is better than the others? To my knowledge, current technology provides no clear-cut solution. Second, if the legs of the triangles are not orthogonal, then

\[
Y^2 = X_1^2 + X_2^2 - 2 \cdot X_1 \cdot X_2 \cdot \cos \theta
\]

where \( \theta \) is the angle between the vectors \( \mathbf{X}_1 \) and \( \mathbf{X}_2 \). I think the chance that a shotgun regression would rediscover this formula is minuscule.

Like many statistics users, I used to believe that if a model passed a comprehensive diagnostic checking, the model should be useful, at least for the purposes of prediction. The above amoeba regressions shake this belief. The experiment indicates that regression models cannot all be taken as seriously as scientific laws. Instead, they share similarities with the "common sense" in our daily life: informative but error-prone, depending on the skills of the users. Since statistics are often called "educated guesses," I propose to call these regression models "educated common sense."

Glymour's TETRAD produces, at best, "descriptive models," which may or may not be useful, depending on the model, the data, and what happens next (Freedman, 1987, p. 220).

In brief, there are two kinds of regression models: (1) structural models that represent a deep commitment to a theory about how data were generated (Freedman, 1987); and (2) descriptive models, which are the mere refinements of common sense. Although, as a statistician, I have a high respect for educated (and uneducated) common sense, I believe that it is profoundly mistaken to put the two types of models in the same bag.

Glymour's effort to build TETRAD reminds me of a fabled Chinese, Sen Nong (fl. between 3000 and 4000 B.C.), a physician who was reported to have tried several hundred species of herbs on himself in order to find useful medicine for the Chinese people.

Before modern medication was introduced to Asia, generations of Chinese were saved (and perhaps in many cases poisoned) by Sen Nong's herbal medi-
cies. Today some oriental medications remain valuable, and their functions are still mysterious to western scientists. However, the real revolution (and the real lifesaver) in oriental history is the adoption of western medication.

The main driving force behind modern science is the quest to understand the underlying mechanisms. In medication, this means solid knowledge of physiology, pathology, biochemistry, anatomy, etc., which were developed mainly by substantive knowledge. Regression analysis certainly helps, but it is used in highly focused subject matters, under tightly controlled experiments.

By contrast, in many branches of social, economic, and behavioral sciences, solid theory and good measurement are both lacking. Research workers then take statistics (and computers) as their last and best hope. Their desperation and methodology were shared by ancient Chinese physicians. Such methods as TETRAD are, in my opinion, legitimate and useful research tools; but their roles have been greatly exaggerated. Indeed, I would be horrified if medical researchers proposed linear statistical laws, instead of physiology and pathology, as the foundation of medical practices.

As pointed out in Glynour et al. (p. 16), some social scientists prefer to avoid causal language. This attitude is highly recommended but was unfortunately dismissed by Glynour.

II. REGRESSION AND TIME-SERIES ANALYSIS: SCIENCE OR LUNACY? (Part I)

A major benefit of scientific knowledge is prediction. Since the 1950s, regression and time-series analysis have been used for economic forecasting, which is reportedly a 100–200 million dollar business. This new industry has an impressive history, which even includes the prestige of Nobel Prizes.

In 1969 Norwegian economist Ragnar Frisch (a professor in social economy and statistics) won the first Nobel Prize in Economic Science, partly for his contribution of introducing random disturbances to deterministic dynamic systems.

In 1980 Lawrence Klein won the Prize partly for pioneering the development and use of econometric models to forecast economic trends in the United States. The Klein-Goldberger (1955) model further gave rise to numerous econometric research efforts that rely heavily on two-stage least squares, three-stage least squares, and time-series analysis.

In 1989 Trygve Haavelmo also won the Nobel Prize “for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structure” (the Royal Swedish Academy of Sciences; Trenton Times, October 12, 1989). According to Paul Samuelson, another Nobel laureate at the Massachusetts Institute of Technology.

The whole modern age in economics has tried to use statistics to test out different theories. It was Haavelmo who made the great breakthrough, so we can know we’re testing our particular theory and are not getting a spurious relationship. [emphasis supplied]

Today the models derived from Haavelmo’s work “are widely used in predicting the course of national economies and formulating government policies” (The New York Times, October 12, 1989). But how good are these models? And what are other experts’ opinions of these models?

In a meta-analysis, Armstrong (1984/1978) provided interesting findings related to the above questions. The study involved a survey of 21 experts from some of the leading schools in econometrics (MIT, Harvard, Wharton, Michigan State, etc.) and from well-known organizations that sell econometric forecasts.

The survey was designed to examine two popular beliefs:

1. Econometric models provide more accurate short-term forecasts than do other methods; and
2. More complex econometric methods yield more accurate forecasts than do simpler methods.

Of the 21 experts surveyed, 95% agreed with the first statement, and 72% agreed with the second.

Respondents were asked how much confidence they had in their opinion on accuracy. Confidence was rated on a scale from 1 (“no confidence”) to 5 (“extremely confident”). The average response was about 4. No one rated confidence lower than 3.0 (p. 23).

The experts were also asked to rate their own professional competence. Those who rated themselves as more expert felt that econometric methods were more accurate.

To assess the reliability of the experts’ opinions, Armstrong went through a search of 30 studies from research journals, and the verdict on the first belief is disheartening: “Econometric forecasts were not found to be more accurate” (1984, p. 25).

Moving on to the second belief that more complex econometric methods yield more accurate forecasts, empirical evidence gathered by Armstrong from 11 direct and 5 indirect studies contradicted the folklore. Armstrong’s conclusion also strengthens an argument by Milton Friedman, another Nobel laureate (1951), that “the construction of additional models along the general lines [as Klein’s model] will, in due time, be judged a failure.”

Many researchers have devoted their whole lives to complex econometric models. But so far the forecast accuracy of these models is disappointing, unless the projections are adjusted subjectively by the modelers. In many cases, forecasts by simultaneous-equation (econometric) models “were beaten by those of a simple AR(4) model” (Granger, 1986). Moreover, such AR models themselves do not outforecast naive extrapolation models (Makridakis and Hibon, 1984). Makridakis’s conclusion was drawn from an exhaustive
comparison of the models on 111 time-series data; the study provides strong evidence that naive extrapolation is all we need in economic forecasting. But then it is natural to ask the next question: What is the difference between a naive extrapolation model and no model at all?

In one study (Dreman, 1984), professors at Harvard and MIT kept track of stock forecasts made by at least five or six analysts. The numbers of companies monitored were 769 in 1977 and 1,246 in 1981. The conclusion was that what forecasters had been touting simply did not pay off. The dismal performance of forecasting models was well summarized by Dreman: "Astrology might be better!"

This situation resembles the often aberrant predictions in geology. People who are desperate for forecasting usually put blind faith into anything associated with sophisticated mathematical formulas. However, a lesson I have learned during years of time-series modeling is this: Forecasting is too important to leave to time-series analysts.

In short, current statistical forecasting simply does not live up to its catchy title. The wholesaling of such techniques to general scientists or to ordinary users will eventually result in mistrust of our profession. For example, I would recommend for the reader's attention an article entitled "Statistical Forecasting Made Simple, Almost" in The New York Times, C4, June 24, 1986. One of my colleagues (an influential administrator) was excited about it and was planning to purchase a site license so that we could predict student enrollments and plan better for course offerings. The administrator called me in and asked how many copies of the software we should buy. "Zero; the software is a dud," I replied, showing him a demo-diskette that I had bought one year earlier. The Times article was, in my judgment, an outrageous promotion of state-space modeling. With great dismay, I crossed out the title of that article and replaced it with the following: "Statistical Forecasting Made Simple? Never!"

III. REGRESSION AND TIME-SERIES ANALYSIS: SCIENCE OR LUNACY? (Part II)

An important distinction was proposed by Dempster (1983) to contrast technical statistics and functional statistics. Dempster made it clear that technical statistics deals exclusively with tools and concepts that have been applied widely across many professions and sciences, while functional statistics requires blending adequate skill and depth in both technical statistics and substantive issues. The two sometimes overlap, but often technical statistics simply do not function.

Dempster, an eminent Harvard statistician, observed that formal analysis of the connection between technical statistics and the real world is "almost nonexistent" and that "very little of this is taught to statisticians." But when scientists need advice on data analysis, whom do they turn to? Further, when research journals are littered with misleading statistics, is it the scientific community or the statistical community that should be held responsible?

EXAMPLE 1 In the spotlight here is a user's guide called SAS For Linear Models (Freund and Littel, 1985/1981, pp. 92-100), a very popular "guide" among industrial and academic statistics users. In this case, the authors used the General Linear Model to analyze the test scores made by students in three classes taught by three different teachers. The sample sizes of the three classes were 6, 14, and 11. Randomness of the data was, as usual, not indicated. With this kind of data, stem-and-leaf plots are about the only techniques that are appropriate. Yet the authors spent 8 and 1/2 pages on lengthy discussions of the technical statistics, which include the general linear model, the matrix operations, the estimation of coefficients, the one-way design and balanced factorial structure, the adjusted and unadjusted means, the uniqueness of the estimates, and so on and so forth.

Nowadays statistics users are preoccupied with grinding instead of intellectual reasoning. This kind of practice gives my profession a bad name, and famed statistics educators like Freund and Littel owe us much explanation for the misleading example they have presented to novice statistics users.

In a private conversation, a SAS sales representative defended the example by saying that the authors were trying to show statistics users a valuable algorithm, not a causal link. I responded: "If Freund and Littel are short of any meaningful example, they should simply label the data sets as A, B, C, treating them like pure digits." Statistical comparison is a powerful tool for data analysis, but spending 8 and 1/2 pages on the printout of a General Linear Model for students' test scores is both an abuse of statistics and an abuse of the computer as well.

EXAMPLE 2 Ever since its early development, statistics has frequently been called upon to answer causal questions in nonexperimental science. As pointed out by Glymour (1987, p. 3):

Today, statistical methods are applied to every public issue of any factual kind. The meetings of the American Statistical Association address almost every conceivable matter of public policy, from the safety of nuclear reactors to the reliability of the census. The attempt to extract causal information from statistical data with only indirect help from experimentation goes on in nearly every academic subject and in many nonacademic endeavors as well. Sociologists, political scientists, economists, demographers, educators, psychologists, biologists, market researchers, policy makers, and occasionally even chemists and physicists use such methods.
Glymour did the statistical community a good service in giving us a clear picture of how popular statistical causal inferences are in nonexperimental sciences. The popularity, however, does not by itself establish the soundness of the applications of these methods. In the case of the safety of nuclear reactors, the reader is urged to consult Breiman (1985, pp. 203–206) for an insider’s account of statisticians’ failure on this issue. For the heated debate on the reliability of the census, an illuminating source is Freedman and Navidi (1986).

It is interesting that Glymour cited the meetings of ASA (American Statistical Association) to support his advocacy of drawing causal inferences from correlation data. Anyone who frequents professional gatherings would testify that the presentations in these meetings cannot all be taken as serious scientific investigations. This phenomenon occurs not only in ASA meetings but also, I suspect, across all academic disciplines.

In my observation, numerous applications presented in the ASA meetings are like that of Freund and Littell in Example 1: the logical link between the real-life data and the technical statistics presented is simply not there. A notable example is the crude application of statistical models in legal action involving issues of gender and racial discrimination in salaries. In the case of sex discrimination, the following variables are often used:

\[ Y = \text{salaries of the employees in a certain institution} \]
\[ G = 1, \text{ if the employee is male} \]
\[ = 0, \text{ otherwise} \]
\[ X = \text{a vector of covariates such as educational background, seniority, etc.} \]
\[ e = \text{error term}. \]

A “direct” regression model of the discrimination is:

\[ Y = a + b \cdot G + c \cdot X + e. \]  \hspace{1cm} (1)

Some experts in the field also use the following “reverse” regression (see, e.g., Conway and Roberts, 1986):

\[ X = a + b \cdot G + c \cdot Y + e. \]  \hspace{1cm} (2)

Sex discrimination is then alleged if the coefficient \( b \) in expression (1) or (2) is significantly different from zero.

These practices are now common and recently drew criticism from hardline statisticians. In a special session of the ASA Annual Meeting, Holland (1988) discussed the weakness of these models in the context of Rubin’s model and concluded that both direct regression and reverse regression provide only sloppy analysis in the case of sex discrimination.

In a similar attack on the use of regression models for employment discrimination, Fienberg (1988, *Statistical Science*) mentioned a case history involving hiring, compensation, and promotion:

When the case went to trial, the centerpiece of the plaintiff’s argument was the testimony of a statistical expert who carried out multiple regressions galore. . . . She concluded that the coefficient for gender was significantly different from zero and proceeded to estimate the damages. . . .

To rebut this evidence the defendant put on the stand a statistical witness who . . . added a term . . . to the models of the plaintiffs’ expert and noted the extent to which the estimated gender coefficient changed.

The judge, in his written opinion, stated that neither of the experts’ regressions had anything to do with the realities of the case, but ruled in favor of the plaintiffs nonetheless.

Appalled by the statistical installations in this case study, Fienberg asked the defense attorney:

Why did the company’s lawyers allow their expert to present such mindless regression analyses in response to the equally mindless ones of the plaintiff’s expert?

The defense attorney responded:

You don’t understand. If the plaintiffs’ expert hadn’t been busy running multiple regressions she might have taken a closer look at the employee manual which describes what in essence is a two-tiered job system. . . .

When our expert responded by running his own regressions, the lawyers were quite pleased. They believed that the outcome would have been far worse if he had explained to the court what we really do because then the judge could easily have concluded that our system was discriminatory on its face.

As an expert in the field, Fienberg wrote,

Those of us with interests in the legal arena continue to look with horror on the ways in which statistical ideas and methods are misused over and over again by expert witnesses (often not statisticians or even those trained in statistics) or misinterpreted by judges and juries. [emphasis supplied]

In his resentful comment on the misuse of econometric models for employment discrimination, Fienberg concluded that a statistician will be better off “to focus on employment process decisions and to learn about the input to those decisions.”

In the 1988 ASA Annual Meeting, Holland challenged the audience with the following question: “Is sex a cause in salary discrimination?” Before he gave his answer, Holland quipped: “Sex probably is the greatest of all causes.” But in the case of salary equity, Holland maintained that being a man or a woman does not cause the salary inequalities, but that “it is discrimination that is the cause.” Bravo, Holland.

But how are we going to assess the magnitude of the discrimination? Holland’s answer is downright honest: “I don’t know.”
EXAMPLE 3 The third example in this section is taken from a book entitled *Applied Time Series Analysis for the Social Sciences* (McCleary and Hay, 1980, pp. 244-270). This example reflects very much the way statistical analyses were performed by the expert witnesses in Fienberg’s case study.

In an application of time series to population growth, McCleary and Hay bent on a lavish investment of energy in technical statistics, which include prewhitening, ACF (autocorrelation function), CCF (cross-correlation function), residual analysis, diagnostic statistics, and so on and so forth. Finally the authors derived the following multivariate time-series model:

\[
\rho_t = \frac{.23(1 - .50B) - .952}{1 - .62B} h_{t-1} + \frac{a_t}{1 - .26B}
\]

where \( p_t \) is the annual Swedish population; \( h_t \) is the harvest index series; \( f_t \) is the fertility rate; \( a_t \) is the white noise; and B is the backward operator.

After a 26-page Herculean effort to build the above “formidable model” (quoted from McCleary and Hay, 1980), the authors devoted only two lines to the model interpretation: “The model shows clearly that, during the 1750-1849 century, the harvest had a profound influence on population growth.”

Naive statistics users often perceive number-crunching as the ultimate goal of scientific inquiry. The above examples are only the tip of an iceberg. (The reader will find more examples in Freedman, 1987.) As a rule of thumb, often the more computer printouts empirical studies produce, the less intellectual content they have.

EXAMPLE 4 In the 1960s, London suffered a smoke episode that was estimated to have claimed several thousand lives. Data from such episodes and from persistently high smoke areas have been compiled over many years. These data were available to U.S. health officials, who follow a routine procedure to calibrate the current standard of clean air. The procedure contains a document that summarizes all available relevant information. The document was submitted for review and written public comments, before it became final.

In this U.S. study, the British data were useless for an obvious reason: The health effects observed took place at levels that were much higher than the U.S. standard. Also, there is no reliable way to convert the British data into United States measurement, because at that time scientists on the two sides of the Atlantic used totally different methods to measure the amount of smoke in the air.

Given this kind of data, no statistical magic can produce useful information. But according to Breiman (1985, p. 208), the written public comments consisted of over 200 pages. Well over half of this text was consumed by discussions of statistical techniques, such as the correct way to model with multiple time series, transformation of variables, lags, standard errors, confidence intervals, etc.

The technical statistics in these reviews were impressive but totally irrelevant. Breiman refers to these statistics as an “edifice complex,” denoting the building of a large, elaborate, and many-layered statistical analysis that covers up the simple and obvious. In this specific example, the edifice building was, according to Breiman (1985), a less admirable product of “a number of very eminent statisticians.”

IV. REGRESSION AND TIME-SERIES ANALYSIS: SCIENCE OR LUNACY? (Part III)

In 1912, an American statesman, Elihu Root, won the year’s Nobel Peace Prize. The Nobel laureate was also widely considered as one of the ablest lawyers the United States has ever produced (the New York Times). He once wrote:

> About half the practice of a decent lawyer consists in telling would-be clients that they are damned fools and should stop.

In my observation, many statisticians follow a completely different tactic: They are enthusiastic about promoting the use of statistics and are often reluctant to tell scientists (or would-be scientists) about the severe limitation of statistical inference. Many veteran statisticians are aware of this problem, but they either are busy in producing esoteric articles or simply do not want to rock the boat.

A prominent exception is Professor David Freedman of the University of California at Berkeley. In a series of articles, Freedman (see the references at the end of this chapter) mounts a full-fledged attack on law-like statistical models that rely on regression and/or time-series techniques. The models are among the dominant research methodologies in econometrics and social science research. One of the models under fire was the celebrated study by Blau and Duncan (1967) on the American occupational structure (Freedman, 1983). Blau and Duncan’s classic work was cited by the National Academy of Sciences (McCAdams et al., 1982) as an exemplary study in social science.

> “Despite their popularity,” Freedman (1987, p. 101) declared, “I do not believe that they [the statistical models in question] have in fact created much new understanding of the phenomena they are intended to illuminate. On the whole, they may divert attention from the real issues, by purporting to do what cannot be done—given the limits of our knowledge of the underlying processes.” Freedman (p. 102) also asserted: “If I am right, it is better to abandon a faulty research paradigm and go back to the drawing boards.”
Freedman (1987, p. 217) insists that multiple regression, as commonly employed in social science, "does not license counterfactual claims." His objections to these models can be summarized as follows. First, the models are devoid of intellectual content. The investigators do not derive the models from substantive knowledge; instead, the models are either data-driven or simply assumed (1983, p. 345).

Second, nobody pays much attention to the stochastic assumptions of the models. In most social-science applications, these assumptions do not hold water. Neither do the resulting models (1983). Some investigators are aware of the problem of the stochastic assumptions in their models and therefore label the computer outputs as merely descriptive statistics. "This is a swindle," Freedman declared. "If the assumptions of the regression model do not hold, the computer outputs do not describe anything" (1985, p. 353).

Third, statistics as a science must deal explicitly with uncertainty. But in practice, complicated statistical models often are the dominant source of uncertainty in a serious investigation (Freedman and Zeisel, 1988). Instead of solving problems in real life, such models sweep the problems under the carpet.

Fourth, these models do not render meaningful predictions; they only invite misinterpretations (1987). In comparison to natural science models (e.g., Newtonian mechanics and Mendelian genetics), social science models do not capture the causal relationships being studied. In sharp contrast, the natural science models work "not only by log likelihood criteria, but for real" (1987, p. 220).

Freedman's unflattering attack (1987) on social-science statistical models attracted the attention of 11 scholars. Some lent their support, while others expressed their dismay. In response, Freedman wrote:

We have been modeling social science data since the 1930s, taking advantage of modern data collection and computing resources. The bet may have been a reasonable one, but it has not paid off. Indeed, I have never seen a structural regression equation in economics or sociology or education or the psychology of drug addiction, to name a few fields (not entirely at random). If the discussants have examples, they have been too shy to disclose them.9

Freedman's (1987) criticism of the poor performance of statistical models in social (and economic) science is similar to that of Richard Feynman, a famous physicist who placed much of experimental science in the category of "cargo cult science." After World War II was over, according to Feynman, certain Pacific islanders wanted the cargo planes to keep returning. So they maintained runways, stationed a man with wooden headphones and bamboo for antennas, lit some fires, and waited for the planes to land (The New York Times, February 17, 1988).

Amoeba Regression and Time-Series Models

It is the same, Feynman said, with cargo cult scientists. "They follow all the apparent precepts and forms of scientific investigation, but they are missing something essential because the planes don't land."

Social science researchers may feel that physicists and hard-line statisticians are too critical. But as a friend of mine (a psychologist) once told me, "In natural science, people step on each other's shoulders. But in my field, we step on each other's faces."

Despite all these negative comments, social science remains indispensable. In such fields, there are plenty of informative studies (see examples in Freedman, 1987, p. 206; Freedman et al., 1978, pp. 41-43). But these useful studies are based on good measurement, insightful analysis, and simple statistical methods that emphasize the raw data. Sophisticated statistical models, by contrast, tend to overwhelm common sense (Freedman, 1987) and misdirect the course of the research. Good design and measurement are needed in order to develop good theory. "Pretending that data were collected as if by experiment does not make it so." In conclusion, Freedman (1987) wrote:

Investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models.

Amen.

V. STATISTICAL CORRELATION VERSUS PHYSICAL CAUSATION

Dempster (1983) suggested that although statistical model-building makes use of formal probability calculations, the probabilities usually have no sharply defined interpretation, so the whole model-building process is really a form of exploratory analysis.

Dempster is quite right, and I think it will be healthier if we take away the aura of these models and use them as just additional tools of Tukey's EDA. (The reader will find a similar conclusion in Freedman, 1987, p. 220.) Some researchers may feel insulted and fear a loss of prestige if their models are labeled as EDA; but this is a hard reality they will have to swallow.

A popular conception is that EDA does not involve probability models or complicated formulas. This is entirely mistaken. Consider Tukey's Fast Fourier Transform. The transform is highly sophisticated and widely used in scientific applications. But its main purpose is simply to hunt for hidden patterns, a typical activity of EDA.

As is well known in the history of mathematics, algebra can often yield more than one expects. Therefore, not all statistical modeling that involves intricate algebraic manipulations should be considered useless.
Chapter 6

On Objectivity, Subjectivity, and Probability

In this chapter we will devote ourselves to certain aspects of objective science and subjective knowledge. For this purpose, we begin with an appreciation of the advantages and limitations of logic and mathematics—the unchallengeable "objective science."

The first advantage of logic is that reasoning alone is capable of leading to new discoveries—despite a common belief that deductive reasoning produces no new knowledge. Here is an example of how scientists are able to reach for new discoveries by manipulating abstract symbols.

According to Einstein's special theory of relativity, the energy of a particle that has mass m and momentum p is given by

$$E^2 = m^2c^4 + p^2c^2.$$  

(1)

If the momentum p is zero, the equation reduces to the well-known

$$E = mc^2.$$

By taking the square root of the equation (1), mathematically we should have

$$E = mc^2 \quad \text{and} \quad E = -mc^2.$$

Unfortunately, negative energy was considered as not supported by physical reality and was dismissed by most physicists, Einstein included (Gribbin, 1984).
In the late 1920s, Paul Dirac formulated a new wave equation for electrons, which incorporated relativity into quantum mechanics. Dirac's equation is very complicated, but it explains the spin of the electron as an angular momentum arising from a circulating flow of energy and momentum within the electron wave. His equation also predicts a relationship between the magnetic moment and the spin of the electron (Ohanian, 1987). However, Dirac's monumental work was considered by most physicists as marred by one "defect": he attempted to explain the negative solution in his equation by the existence of particles in a sea of negative-energy state.

Dirac's theory of negative-energy particles was born out of a conviction that his complicated calculations were perfect in almost every aspect and thus should not result in producing a nuisance negative number. Therefore he invented negative-energy particles to account for an "imperfect feature" in his equation. But his theory was completely anti-intuitive and no physicist could take the idea seriously. In 1932, as a surprise to everybody, Carl Anderson discovered the tracks of electronlike particles of positive charge in a cloud chamber. Subsequent measurements concluded that the new particles have the same mass, same spin, and same magnetic moment as electrons—they are identical with electrons in every respect, except for their electric charge. The new particle was later called the anti-electron. In the mid-1950s, the discovery of anti-proton and anti-neutron further confirmed Dirac's original ideas that were mainly based on the faith of mathematical perfection. (Anti-matters are now routinely used in particle accelerators that gobble up billions of dollars of federal money and involve hundreds of physicists.)

Dirac was known as a man whose only scholarly activity was playing with mathematical equations. In 1933 he was honored with a Nobel Prize for his equation and a bold idea that literally opened up a new universe for humanity. His story also indicates that fooling around with mathematical equations may pay off in totally unexpected ways.

I. STATISTICAL "JUSTIFICATION" OF SCIENTIFIC KNOWLEDGE AND SCIENTIFIC PHILOSOPHY

The second advantage of deductive reasoning is that if a theorem is proved true, then the conclusion of the theorem will never change—it is an ultimate truth. If one desires pure objective knowledge, then one has to re-examine mathematics for its ability to produce unbounding knowledge. In the early part of the 20th century, logical positivism did just that. The movement also strove to develop the logic of science (Carnap, 1934) and to "propagate a scientific world view" (Schlick, 1928). A major goal of the movement was to establish a "scientific philosophy," in contrast to the traditional "speculative philosophy" of Plato, Socrates, Descartes, Hume, etc. (Reichenbach, 1951).

On Objectivity, Subjectivity, and Probability

The movement soon dominated the field of the philosophy of science. One of their tasks was to reduce all objective knowledge to formal logic. A monumental achievement in this direction was the publication of *Principia Mathematica*, Volumes I–III (Russell and Whitehead, 1910–1913). Russell's goal was to develop a symbolic logic and to show that all pure mathematics can be reduced to formal logic. The attempt was driven by a desire to eliminate contradictions that threaten the foundations of set theory and other branches of mathematics. It was believed that mathematics and logic are inseparable, and that if we want to obtain purely objective knowledge, then first we have to reconstruct mathematics as a body of knowledge that is contradiction-free (Davis and Hersh, 1981).

Fig. 1 is an example of Russell's legendary work on how the arithmetic proposition 1 + 1 = 2 is established (after 362 pages of elaboration). This mammoth exercise is likely to be viewed by non-mathematicians as plain stupidity. But it is an attempt (a praiseworthy attempt) to establish mathematics as a
model of absolute truth. The attempt was indeed a grand achievement of axiomatic reasoning. It also inspired the great mathematician David Hilbert to put forth a program that he claimed would banish all contradictory results in mathematics. In retrospect, Hilbert's faith in this program was admirable: "What we have experienced twice, first with the paradoxes of the infinitesimal calculus and then with the paradoxes of set theory, cannot happen a third time and will never happen again" (Regis, 1987).

In 1931, six years after Hilbert's manifesto, Kurt Godel proved, by the very method of logic, that all consistent axiomatic formulations of number theory include undecidable propositions. In plain English, the theorem showed that no logical system, no matter how complicated, could account for the complexity of the natural numbers: 0, 1, 2, 3, 4, ..., (Hofstadter, 1979). The theorem thus puts an end to the search for logical foundations of mathematics initiated by Russell and Hilbert.

After decades of elaboration on the logic of science and on "the ultimate meaning of scientific theories" (Schlick, 1931, p. 116), Russell (1948) finally conceded that "all human knowledge is uncertain, inexact and partial." This conclusion may be true, but the tone is too pessimistic. It may also lead to the anarchism that nothing is knowable and therefore anything goes. For instance, when a space shuttle blows up in the sky and kills all of the astronauts in it, the scientists in charge cannot simply shrug off the event by saying that "all human knowledge is uncertain, inexact and partial."

Moving back to mathematics and logic, although Godel had proved that logical systems are intrinsically incomplete, he never proved that mathematics as a branch of human knowledge is inconsistent or uncertain.

It should be spelled out that, outside the domain of pure mathematics, there are two types of knowledge: local and global knowledge. While local knowledge can be quite certain, global knowledge is in general partial and hard to justify. For example, when Newton released an apple from his hand, the apple dropped to the floor. That is certain. But what is the reason free objects all drop to the ground? According to Aristotle, that is the fate of everything. According to Newton, that is the result of gravity. According to Einstein, that is the result of relativity.

Given the competing theories, a big question is, how do we know which one is the best? Furthermore, a theory is a generalization of finitely many observations. What is the scientific justification of this generalization? According to logical positivists, the answer is the assignment of a degree of confirmation to each scientific theory. This so-called "degree of confirmation," according to Reichenbach (and his Berlin school of philosophy), rests upon the calculus of probability, which is the "very nerve of scientific method." Also, the notorious Hume's problem of induction "can thus be solved" (Reichenbach, 1951).

In addition, when conflicting theories are presented, probability and "inductive logic" are used to settle the issue. More precisely, "the inductive inference is used to confer upon each of those theories a degree of probability, and the most probable theory is then accepted" (Reichenbach, 1951). This constitutes the core of the "scientific philosophy" advocated by Reichenbach and his Berlin school of philosophy.

However, the justification of scientific knowledge is often not an easy matter, let alone the justification of "scientific philosophy." For example, what is the probability that a specific nuclear-power plant will fail? In such a tightly controlled situation, if scientists cannot come up with a reliable probability, then there is little hope of assigning a meaningful probability to each scientific theory. As a matter of fact, many famous statisticians have plunged into the problem of the nuclear-power plant only to retreat with disgrace. (See Breiman, 1985; Speed, 1985).

It is now clear that the "speculative philosophy" of Plato, Descartes, and Hume contributed more to human knowledge than Reichenbach's voluminous writings on the so-called "scientific philosophy."

Nevertheless, the influence of logical positivism has been enormous in many branches of science in such distinctive manners as: (1) using mathematical probability to justify certain unjustifiable "scientific" results; (2) masquerading "subjective knowledge" under the cover of "objective estimation."

This kind of unfortunate influence poses less problem in the natural sciences than in social-behavioral studies. One reason for this phenomenon is that the natural sciences are usually built upon a solid foundation and investigators are too busy to worry about philosophical debate on the nature of their research conducts.

Social and behavioral scientists, on the other hand, are insecure about their research findings. Therefore, they need "justification" like the one advocated by Reichenbach. This justification, as we have seen in the previous chapters, is misleading and is detrimental to their research efforts.

In general, it is easier to develop the scientific theories of nature than to grasp the nature of scientific theories. But fortunately, Thomas Kuhn (1970), a physicist, provided us with a better picture in this regard. According to Kuhn, scientists work under different paradigms that are "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners." Under a paradigm, a scientific community devotes itself to solving problems that are narrowly focused or fragmented, or even problems that apparently need not be solved.
On Objectivity, Subjectivity, and Probability

A distinctive feature of the Einstein paradigm is that many of the new theories based upon general relativity are just “theories”: they are not empirically testable, and the way theorists are pumping up suppositions is very much in the manner of Aristotle framed up his views of nature.

For instance, in order to test the super-string theory or certain Grand Unified Theories, the particle accelerators would have to be light years across in size. This is hardly in the spirit of Newtonian physics, whose trademark was firmly established when Galileo started doing experiments to test Aristotle’s claims on falling objects. But now with these new theories about Grand Unification, physicists are just sitting there thinking about exotic theories and not doing experiments.

Another prominent feature of general relativity is that it is not a well-tested theory itself, at least not until recent years. As a matter of fact, throughout his life, Einstein saw only two confirmations of the general theory of relativity. Both confirmations rely on observational data, not controlled experiments. In addition, the confirmations are distinctively statistical. Together, the confirmations constitute two case studies about the troublesome nature of statistical analysis that are common in non-experimental sciences.

One confirmation of general relativity is about the peculiar movement of Mercury, a strange phenomenon that cannot be explained by Newton’s law of gravity. Numerous proposals were put forward to account for the discrepancy. But it was Einstein’s framework that appeared to explain the movements of Mercury and other planets in the most elegant fashion. Table 1 shows some statistics in this respect (Marion, 1976, p. 294):

<table>
<thead>
<tr>
<th>Planet</th>
<th>Theoretical calculation</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>43.03 ± 0.03</td>
<td>43.11 ± 0.45</td>
</tr>
<tr>
<td>Venus</td>
<td>8.63</td>
<td>8.4 ± 4.8</td>
</tr>
<tr>
<td>Earth</td>
<td>3.84</td>
<td>5.0 ± 1.2</td>
</tr>
</tbody>
</table>

But just like any statistical artifact in observational studies, the confirmation was challenged by other observations. In 1966 Dicke and Goldenberg found that the Sun is not a sphere and that the flatness of the Sun contributes about 3 arcseconds to Mercury’s strange movement. This observation cast doubt on general relativity and seemed to favor other competing theories such as the Brans-Dicke theory. (For more details, see, e.g., Will, 1986.) But then others soon pointed out that the Dicke-Goldenberg results might be due to the intrinsic solar brightness between poles and equator, instead of a true flatness of the Sun. Debates surrounding these matters abounded in the scientific literature.

Today, scientists appear confident that the issue will eventually be resolved in...
favor of general relativity. But so far no decisive piece of evidence has been produced.

Another confirmation of general relativity was the well-touted Eddington expedition of 1919. The event has been argued by some (e.g., the British historian Paul Johnson) as the beginning of the "modern era," in that the old concepts of absolute space and absolute time were lost forever.

But the fact is that Eddington's confirmation is somewhat scandalous. Here we will relate the story as appeared in Earman and Glymour (1980) and in Fienberg (1985). We will also give some of our calculations and comments.

According to the general theory of relativity, starlight is bent by the gravitational field of the Sun. Einstein's theory predicted a deflection of 1.74", as opposed to the value of 0.87" predicted by Newtonian theory. To settle the issue, Eddington organized a pair of expeditions to make appropriate measurements during the eclipse of the sun. Three sets of data were collected (all measurements are in arcseconds):

<table>
<thead>
<tr>
<th>Data set</th>
<th>Mean deflection</th>
<th>Probable error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.61&quot;</td>
<td>0.444&quot;</td>
</tr>
<tr>
<td>2</td>
<td>0.86&quot;</td>
<td>0.48&quot;</td>
</tr>
<tr>
<td>3</td>
<td>1.98&quot;</td>
<td>0.178&quot;</td>
</tr>
</tbody>
</table>

As anyone can tell, the second data set yielded a mean deflection that is almost identical to the Newtonian value of 0.87". But this set of data was not reported when Eddington returned to an extraordinary joint meeting of the Astronomical and Royal Societies. Einstein's prediction was thus "confirmed" and a modern era of curved space and time was born.

In a critique of the Eddington expedition, Earman and Glymour (1980) pointed out many troublesome features in the whole affair. To begin with, it is very tricky (or error-prone) to measure what ought to be measured during an eclipse. For example, comparison photographs were taken at different times, one set during the day and the other at night. Temperature changes, the relocation of the telescope, and slightest mechanical deviations in the machine could have, in theory and in practice, significant impact on the measurements.

For another example, the first data set in the table were calculated from the images of only two plates. Further, the images were so blurred that Eddington disregarded the standard least-squares technique and expended extra effort to derive a scale factor indirectly from other check plates. In addition, the rotational displacement of the images was obtained by assuming the validity of Einstein's predicted deflection.

Behind the stage, the way Eddington deleted or adjusted his data does not appear to be a model of scientific conduct. Nevertheless, the Eddington "confirmation" of Einstein's theory is, by any standard, a great success in media and in the scientific community. The "confirmation" indeed made the name Einstein a household word.

After the 1919 confirmation, there were other expeditions to test Einstein's theory. The results varied, but most of them were a bit higher than the deflection predicted by the general relativity. In their comments about these confirmations, Earman and Glymour (1980) wrote, "It mattered very little. The reputation of the general theory of relativity, established by the British eclipse expeditions, was not to be undone." To support this position, Earman and Glymour related the affair to the prediction of gravitational red shifts.

The red shift was the first of famous predictions based on the notion of curved space-time. The prediction was made in 1907 by Einstein himself. The confirmations of this prediction, however, had always been elusive. Before 1919, no one claimed to have obtained red shifts of the size predicted by the theory. But within a year of the announcement of the Eddington results several researchers reported finding the Einstein effect. According to Earman and Glymour (1980), there had always been a few spectral lines that could be regarded as shifted as much as Einstein required; all that was necessary to establish the red shift prediction was a willingness to throw out most of the evidence and the ingenuity to contrive arguments that would justify doing so.

In addition, Earman and Glymour maintained:

The red shift was confirmed because reputable people agreed to throw out a good part of the observations. They did so in part because they believed the theory; and they believed the theory, again at least in part, because they believed that the British eclipse expeditions had confirmed it.

Such criticisms, amazingly, have not eclipsed the fame or the "validity" of general relativity. Modern observations, according to certain experts in the field (see, e.g., Hawking, 1988, p. 32), have accurately confirmed certain predictions; and the theory so far still holds the truth about space, time, and gravity. But other physicists, such as Kuhn, may not be as convinced as these experts.

We now turn to a comparison between general relativity and special relativity. First, special relativity deals specifically with physics in the situations where gravitational force is negligible. In addition, unlike general relativity, special relativity has been checked and rechecked and confirmed time and time again (by controlled experiments). Indeed the theory has been accepted by physicists as a reality beyond any shadow of doubt. For instance, just by the sheer power of the atomic bomb, you don't want to argue with the equation E = mc² (Will, 1986).
Second, special relativity gave rise to Dirac’s theories about the spins of elementary particles and about the existence of anti-matters, all of which have been confirmed in tightly controlled experiments over and over again.

Third, special relativity also paved the way for the theory of quantum electrodynamics (QED), the most accurate theory in all of physics (Ohanian 1987, p. 444; Will, 1986).

Kuhn himself is an expert in physical science. It is thus a big surprise that he did not credit the coherent progress in physical science exemplified in the ontological development from Newtonian physics and special relativity to Dirac’s theory and the theory of quantum electrodynamics.

Nevertheless, his brilliant observation about general relativity reinforced a suspicion that in certain branches of non-experimental (and semi-experimental) science, “progress,” despite mountainous publications, might be less than it seems.

II. CLASSICAL PROBABILITY, COMMON SENSE, AND A STRANGE VIEW OF NATURE

In the previous section we discussed some advantages of deductive reasoning. In this section we will further discuss another advantage, and later in this section the limitations of logic and mathematics. The advantage is that mathematical reasoning often produces results that are contradictory to our intuition and common sense. The clash of mathematics and intuition, to my knowledge, always results in the surrender of intuition to mathematical reasoning—because experiments always turn out on the side of logical reasoning. To illustrate this point, consider the following example:

There are three fair dice whose $3 \times 6$ sides will be labelled from 1 to 18 by Mr. A (a mathematician). After the dice are labelled, Mr. B (another mathematician) chooses one of the dice. Finally Mr. A chooses another die from the remaining ones. The two players roll their dice and the person with the larger number wins $200.

In this game, our intuition says that in the long run Mr. B will have advantage over Mr. A, because Mr. B has the first choice of die. Careful calculation says that our intuition is wrong. For instance, if Mr. A arranges the 18 numbers as follows: $I = \{18,10,9,8,7,5\}$, $II = \{7,16,15,4,3,2\}$, and $III = \{14,13,12,11,6,1\}$. Then

Pr[$I > II$] = $(6 + 3 + 3 + 3 + 3 + 3)/36 = 21/36$,
Pr[$II > III$] = $(6 + 6 + 6 + 1 + 1 + 1)/36 = 21/36$,
Pr[$III > I$] = $(5 + 5 + 5 + 5 + 1)/36 = 21/36 > 50%$.

In other words, die I is better than die II; die II is better than die III; and die III is better than die I. Hence Mr. A will always beat Mr. B, no matter how smart Mr. B is!
This may sound impossible. If you don’t believe this conclusion, we can go somewhere and bet.

The previous dice are capable of producing more surprises. For example,

$E(I) = $ the mathematical expectation of rolling die I $= (18 + 10 + 9 + 8 + 7 + 5)/6 = 9.5 = E(II) = E(III)$.

That is, the expected values of the three dice are the same, even if we have just concluded that the probability die I will beat II is larger than 50%.

If you have enjoyed the above example, the next will be equally fascinating:

Game 1: Flip a coin; the gambler will bet $1 on either head or tail (the probability of winning is .5).

Game 2: the gambler will bet $1 either red or black on a roulette (the probability of winning is 18/38, or .474).

Assume that in game 1 the gambler has $900 and his goal is $1,000,000 (one million), and that in game 2 the gambler also has $900 but his goal is only $1,000 (one thousand). In which game is the gambler more likely to reach his goal? Game 1 or 2?

In graphic form, the above information can be summarized as follows:

In game 1: $900 \rightarrow $1,000,000 ($p = .50$)

In game 2: $900 \rightarrow $1,000 ($p = .474 \approx .50$)

Intuitively, it is a long way to go from $900 to one million dollars. Therefore it is more likely to reach the goal in Game 2 than Game 1. However, if we let

$h(K) = Pr[reaching the goal of $N$, giving that the initial capital = $K]$,

then standard calculations of Markov absorption probabilities yield the following formulas:

Game 1: $h(K) = K/N, N =$1,000,000

(1)

Game 2: $h(K) = (q/p)^K - 1

(2)

(l/q^p)^N - 1$,

$q = 1 - p, N =$1000

If $K =$900, we obtain the following probabilities:

Game 1: $h(900) = .09%$

Game 2: $h(900) = .003%$

Note that .09/.003 = 30. This means that the gambler is 30 times more likely to reach the goal in Game 1 than in Game 2, a conclusion quite contrary to intuition for most people.

In their book, How to Gamble If You Must?, Dubins and Savage (1965) mentioned a strategy that can be applied to the situation of Game 2: Bet $100
the first time; if you lose, then bet $200 the second time, etc. Using this strategy,

\[
h(900) = p + \frac{pq + p(q^2)}{1 - (pq)^2} = 88\%
\]

In other words, if one uses the Dubins-Savage strategy in Game 2, then the probability he will reach the goal is about 88%. Without any strategy, the probability is 0.0003.

It now appears that the Dubins-Savage strategy is a sure way to win in the long run. But this is not true. For instance, if we calculate the expected values, the answer is $-20 with the strategy and $-899.97 without a strategy. These negative values lead to a question: In the long run, which is better to play with or without using the Dubins-Savage strategy? To answer this question, we let

\[
m(K) = \text{the expected time the gambler either reaches his goal or gets ruined.}
\]

Then for Game 2 (without strategy),

\[
m(K) = 1 + p \cdot m(K + 1) + q \cdot m(K - 1) ; \quad K = 1, 2, \ldots, N - 1;
m(0) = 0, m(N) = 0.
\]

The solution to these equations yields

\[
m(K) = 22492 = 22,500.
\]

This means that it will take about 22,500 spins of the roulette wheel for the gambler (without strategy) to reach his goal or get ruined. On the other hand, it takes, on the average, only 2 spins to end a game if one uses the Dubins-Savage strategy. Recall that the expected gain of the Dubins-Savage strategy is $-20. Therefore,

\[
(-20) \cdot \frac{22500 \text{ spins}}{2 \text{ spins}} = -$225,000
\]

In other words, the gambler who uses the Dubins-Savage strategy repeatedly will lose about $225,000 by the time the other gambler (without strategy) loses $900.7

Similar puzzles are plentiful in probability and many branches of mathematics (see, e.g., Coyle and Wang, 1993). These puzzles may appear “contradictory” to novice students or amateur scientists. If well explained, these seemingly contradictory results are consistent; more importantly, they are supported by experiment. In other words, the “paradox” is only a conflict between reality and your feeling of what reality ought to be. This is one of the most valuable assets of mathematics over common-sense reasoning.

On Objectivity, Subjectivity, and Probability

We now turn to the limitation of logic and mathematics. In human reasoning, the major problem with logic is that logic itself cannot check the assumptions in a chain of arguments. To illustrate, consider the following example:

A pencil has five sides, with only one side carved with a logo. If you toss the pencil, what is the chance that the logo will come out on top?

So far almost all answers I have heard are the same: 1/5. The correct answer is 0 or 2/5, depending on the definition of “top.” In this example, no amount of logic or mathematical formulas can help to reach the correct answer unless the investigator looks at the pencil in his hand and thinks about the pencil in the stated problem.

The moral of this example is that a total reliance on logic is not enough. This may sound like an old cliché, but isn’t much of the activity of logical positivism merely using logic to justify all human knowledge? The following example is personal, but it is quite typical. A group of Ph.D. candidates from Asia were involved in a debate on the political future of their country. After heated exchange, a mathematics candidate who was trained in doing abstract algebra lectured another anthropology candidate: “I am a mathematician. I am trained in logic. So my conclusion is more scientific than your conclusion.”

In my observation, many quantitative researchers are like that naive mathematics candidate—they believe that pure logic (e.g., statistical formulas and electronic computers) will produce more scientific results than anecdotal evidence. Worse, they are very rigid about the “rigor” and the “objectivity” of their statistical results. Since they have learned the wrong things too well, we will put forth additional criticism of such mentality in Sections III, IV, and V of this chapter.

In the previous example of tossing a pencil, the question can be more complicated:

Toss the pencil 100 times. What is the probability that the logo will show up at least 30 times?

In this case, one can first examine the shape of the pencil and then apply standard mathematics to solve the problem. However, in many scientific applications of the probability calculus, the phenomenon under study is not as easy as the shape of a pencil. For example, the assumptions in a social-behavioral regression often do not hold water, but researchers seldom hesitate to calculate the P-values of the coefficients and declare that the equation is “scientific” and “objective.”

Well, if you blindfold yourself, the statistics generated by a fixed procedure are certainly “objective.” But “objectivity” does not by itself guarantee “reliability,” as the five-sided pencil has taught us.
Einstein's article was intended to point out an impossibility, but it turned out to be prophetic. Experiments (see, e.g., Shimony, 1988, Scientific American) appear to have strong evidence that two entities separated by many meters can exhibit striking correlations in their behavior, so that a measurement done on one of the entities seems instantaneously to affect the result of a measurement on the other.

On another front, Schrödinger challenged the theory of "collapsing waves" by a thought experiment now widely known as "Schrödinger's cat." The example neatly penetrates the weak spot of the orthodox view. Imagine that a cat is sealed in a box, along with a Geiger counter, a small amount of radioactive material, a hammer, and a bottle of poison (such as cyanide). The device is arranged so that when an atom of the radioactive substance decays, the Geiger counter discharges, causing the hammer to crush the bottle and release the poison (Gribbin, 1984).

According to the orthodox interpretation, before we look inside the box, nothing can be said about the radioactive sample—it is not true that the sample either decayed or did not decay, because the wave function collapses only when we look at it. But if the state of the radioactive sample is indeterminate (before we look at it), so are the bottle of poison and the cat. In other words, the radioactive sample has both decayed and not decayed, the bottle is both broken and not broken, and the cat is both dead and alive.

This thought experiment (as well as other paradoxes in quantum mechanics) has generated numerous scholarly articles and Ph.D. dissertations. But the whole issue is far from settled. Among the different proposals, the most interesting ones are put forth by a group of Italian theorists who challenge directly the validity of using standard probability theory in the micro-universe. Specifically, they challenge the validity of the conditional probability in formula (3) and thus the Bayes formula in (5). In doing so, they propose non-Kolmogorovian probability model for micro-phenomenon. The ground of their proposal is that the very act of "looking at it" means an interference of light (photons) with electrons. Therefore the Kolmogorovian model of probability may not hold in the micro-universe. Here is one of the theorems in this new direction.

**THEOREM** (L. Accardi, 1982)

Let \( p, q, r \) be real numbers in the open interval \((0,1)\).

Let

\[
    P = \begin{bmatrix} p & 1 - p \\ 1 - p & q \end{bmatrix}, \quad Q = \begin{bmatrix} q & 1 - q \\ 1 - q & q \end{bmatrix}, \quad R = \begin{bmatrix} r & 1 - r \\ 1 - r & r \end{bmatrix}.
\]

Then

1. A Kolmogorovian model for \( P, Q, R \) exists if and only if

\[
    |p + q - 1| < r < 1 - |p - q|.
\]

On Objectivity, Subjectivity, and Probability

2. A complex Hilbert space model for \( P, Q, R \) exists if and only if

\[
    1 < \frac{p + q + r - 1}{2(pqr)^{1/2}} < 1.
\]

3. A real Hilbert space model for \( P, Q, R \) exists if and only if

\[
    \frac{p + q + r - 1}{2(pqr)^{1/2}} = \pm 1.
\]

4. A quaternion Hilbert space model for \( P, Q, R \) exists if and only if a complex Hilbert space model exists.

Accardi and his co-workers believe that Kolmogorovian model is not universally applicable, and that paradoxes created by orthodox views of quantum mechanics may eventually be eliminated by Hilbert space models. Their works are very interesting, but their conviction so far has not been widely accepted.

Historically, physics was part of philosophy. The two separated when Galileo used experimental methods to test doubtful claims initiated by Aristotle. By this standard, the rising non-Kolmogorovian models of quantum mechanics have to be deemed philosophically. If the models are further developed and tested by a rich array of experimental data, then the models eventually may become part of physics.

To the insiders of academic endeavors, science is not always logical. It is interesting to see how physicists live with all the uncertainties and paradoxes surrounding the foundation of their discipline. Feynman, for instance, feels perfectly comfortable with the uncertainty. He deserves it. His discipline is so far the most solid and accurate among all natural sciences. Further, his QED (quantum electrodynamics) stands as the most accurate theory in all physics. In short, if your discipline has good theory, good measurement, and accurate prediction, then you are entitled to be a little arrogant, like Richard Feynman.

III. INTUITION AND SUBJECTIVE KNOWLEDGE IN ACTION: THE BAYES THEOREM (AND ITS MISUSE)

One of the tasks of logical positivism is to outlaw all speculative statements as meaningless. A consequence of this "scientific philosophy" is a disrespect of intuitive and subjective knowledge. In statistics, this influence is reflected by the strict teaching of the Neyman-Pearson school of statistical reasoning. In social-behavioral sciences, this influence also "led to the declaration that concepts like mind, intuition, instinct, and meaning were 'metaphysical' survivals unworthy of study" (Harris, 1980, p. 15).

Under the shadow of both logical positivism and Neyman-Pearson's rigid teaching, certain researchers are afraid of talking about subjective knowledge (and anecdotal evidence). To them, next to the word "science," the most sacred term may very well be "statistical test." (Beardsley, 1980, p. 75).

In this section we will try to point out that there are numerous examples which show that human knowledge is not acquired by objective methods, let
alone so-called statistical estimation. For example, how does a journal editor decide to accept or reject a research paper in mathematics? Can he do so by conducting a hypothesis testing on a batch of “objective” data? For another example, how can a teacher be objective when he or she goes into a classroom? Copy everything from the book to the blackboard? For yet another example, how does a scientist select a research topic? Take a simple random sample from all papers that have ever been published? Examples like these are everywhere if one is willing to open up his mind.

It is interesting to note that although “objective” researchers do not believe subjective knowledge, they do believe in mathematical formulas and statistical procedures. Hence our next task is to use mathematical formulas to prove that often intuition and subjective knowledge override a 100% objective estimation. Readers who are application-oriented may skip the mathematical details in Example 1.

EXAMPLE 1  Flip a coin 100 times. Assume that 99 heads are obtained. If you ask a statistician, the response is likely to be: “It is a biased coin.” But if you ask a probabilist, he may say: “Wooow, what a rare event!”

My students usually push me to take a stand on this issue. They are accustomed to cut-and-dried solutions, just like certain simple-minded quantitative researchers. Here is my answer: If the coin came from K-Mart, I would tend to stick with the belief that the coin is unbiased. But if the coin came from Las Vegas, then it is a different story.

As a matter of fact, a Bayesian assessment indicates that the coin (from K-Mart) is only moderately biased. A formal analysis proceeds as follows. Let \( p \) be the probability that the coin lands heads, and \( y \) be the number of heads in \( n \) tosses. It is reasonable to assume that the prior of \( p \) is a symmetric Beta distribution with parameter \( a \). The Beta densities with \( a = 20 \) and \( 85 \) are shown in Fig. 3. Note that if \( a \) is very large, then the prior will concentrate near \(.5\). The mathematical form of the Beta density is

\[
f(p) \propto p^{a-1}(1-p)^{n-a-1}.
\]

Hence,

\[
E(p) = .5, \\
Var(p) = 1/\{4(2a + 1)\}.
\]

We further assume that most coins from the U. S. Mint will land heads with probability \( p \) ranging from .4 to .6. This assumption implies that the standard deviation of \( p \) is about .2/6. Therefore

\[
a = 112.
\]

Since the likelihood function of a coin tossing is binomial, standard calculation of the posterior of \( p \) yields another Beta density with parameters \((y+a)\) and \((n-y+a)\). Hence

\[
E(p|y) = (y+a)/(n+2a),
\]

\[
Var(p|y) = E(p|y)(n-y+a)/(n+2a)(n+2a+1).
\]

Some applications of the formulas (6) and (7) are given in the following table.

| \( n \) | \( y \) | \( y/n \) | \( E(p|y) \) | \( SD(p|y) \) |
|---|---|---|---|---|
| 30 | 29 | 96.7 | 55.5 | 3.1 |
| 40 | 39 | 97.5 | 57.2 | 3 |
| 50 | 49 | 98 | 59 | 3 |
| 100 | 99 | 99 | 65 | 2.6 |
For the case n = 30 in this table, the Bayesian estimation of the true p is only 55.5%, which is not significantly different from 50%; on the contrary, the naive 95% confidence interval for p is about [90%, 100%]—a quite misleading conclusion if the coin indeed came from K-Mart. Even for the case where n = 100 and y = 99 (or the extreme case where y = n = 100), the Bayesian assessment of the coin does not support the conclusion that the coin is severely biased.

Remark 1 In this example, the coin-tossing is objective, but the Beta prior density is not—it was chosen mainly for the sake of mathematical convenience. Other densities may do as well.

Remark 2 The calculation of the parameter a is also based on another piece of knowledge that is not objective at all: we assume that most coins will land heads with probability ranging from .4 to .6. These numbers were pulled right off my head; I don't have any hard data to justify these "objective" numbers.

Remark 3 If big capital is involved in the betting, then the possibility that the coin is indeed severely biased may become a daunting nightmare. A frequentist solution is available for this special instance: Flip the coin another 10,000 times; the data will eventually swamp the prior. In technical terms, this means that E(p|y), which equals (y + a)/(n + 2a), will approach the true p if n goes to infinity.

EXAMPLE 2 Some empirical scientists may consider that the above example of coin-tossing is a contrived example remote from real-world applications. Therefore, the next example is chosen for its broad social implications in the testing for AIDS, drugs, legal affairs (Scientific American, March, 1990), and polygraph (lie detector) tests.

Assume that in an AIDS screening program, the reliability of a medical test is 95%. That is, if a person has AIDS, the test will show positive results 95% of the time; if a person does not have AIDS, then the test will be negative 95% of the time. Now suppose a test on Johnny is positive, what is the chance that Johnny has AIDS?

By common sense the probability that Johnny has AIDS is 95%. But people with good intuition and high respect for humanity may argue against the total reliance on the objective medical test: Assume that there are 20,000 students in a big university, and that there are 40 students who have AIDS. The error rate of medical testing is 5%, so (20,000 - 40) x 5% = 998. This means that 998 students could be wrongly identified as having AIDS, which is equivalent to a death sentence. How can "objective science" be so crude?

Such arguments have been used to combat the promotion of mass testings for drug use and AIDS (e.g., the editorial board of The New York Times, November 30, 1987). In most social issues, the disputes can only be settled by a vote. But fortunately, in this case a mathematical analysis can be enlisted to resolve the controversy.

Specifically, let's consider the probability tree in Fig. 4, where Pr[AIDS] = a, Pr[No AIDS] = 1 - a, Pr[Positive|AIDS] = .95, and Pr[Negative|No AIDS] = .95. By using formulas (1) and (2) in Section II, we can derive the following Bayes formula:

\[
\text{Pr[AIDS|Positive]} = \frac{.95a}{.95a + .05(1 - a)},
\]

which is the probability that Johnny has AIDS, given the information that his test result is positive. Note that if a = 1/2, then Pr[AIDS|Positive] = .95. In other words, if a = 50%, then the common-sense conclusion that P(Johnny has AIDS) = 95% is correct. However, if a is not .5, then the answer can be quite different. For example, if a = 1/10,000 (i.e., in the whole population, out of 10,000 people, there is only one AIDS patient), then by formula (8), Pr[AIDS|Positive] = 0.2%. That is, the chance Johnny has AIDS is only 0.2%, not 95%. Even if Johnny is from a risk group where a = 2/1,000, the probability calculated from formula (8) is only 3.7%, not 95%.

At this moment mathematics appears to support the argument against mass testing. However, this appearance will evaporate upon close examination. Note that formula (8) can be applied recursively. For instance, if a = 1/10,000, and the medical test is positive, then by applying formula (8) once, Pr[Johnny has AIDS] is about 0.2%. Now if an independent test on Johnny is still positive, then we can use .2% as the new prior probability (i.e., a) and formula (8) will produce a probability of 3.7%. Repeating the same calculation 3 more times, we can conclude that P(Johnny has AIDS, given 5 positive tests) = 99.6%.

This is another example showing that repeated experiments overwhelm the prior. But repeated experiments are, in most cases, not available in soft sciences; therefore one should always double-check calculations against one's instincts. And if the two collide, then one should shelve the calculation. That is, go with your instinct, or wait for more evidence.

![Figure 4 Tree diagram for the AIDS problem.](image-url)
On Objectivity, Subjectivity, and Probability

and in *JASA* (1987b, with eight discussants). The prominent status of the discussants (and the journals involved) make the discussion an unforgettable event in statistical community.

Berger's main thesis is that P-values overstate the evidence against the null hypothesis. In a zealous attempt to discredit the current practice, his quest is thus to "pound nails into the coffin of P-values." (Berger, 1987b, p. 155, *JASA*). This kind of statement is about the strongest language one can expect in scholarly publications.

Berger's argument is that Bayesian evidence against a null hypothesis can differ by an order of magnitude from P-values (1987b, p. 112):

For instance, data that yield a P-value of .05, when testing a normal mean, result in a posterior probability of the null at at least .30 for any objective prior distribution. ("Objective" here means that equal prior weight is given the two hypotheses and that the prior is symmetric and nonincreasing away from the null.) [emphasis original]

As a good mathematician, Berger gave a solid justification to his claim (using his "objective" priors). But upon further reflection, one may be curious why equal probabilities are assigned to both hypotheses. Note that the null hypothesis in Berger's derivation is

\[ H_0 : \theta = \theta_0, \]

which represents a single point; while the alternative hypothesis is

\[ H_1 : \theta \neq \theta_0, \]

which represents two half lines containing infinitely many points. This assignment of 50% probability to a point null reminds us of a mistake committed frequently by freshman students in introductory statistics courses. For example, you ask the students: "What is the probability a fair coin will land heads?" They will answer: "50%." So far this answer is correct. But why? Many students will say that it is because there are two possible outcomes, so the probability for each to happen ought to be 50%. But this argument is simply not true. Further, if you ask them:

Flip a fair coin 100 times. What is the probability of getting 50 heads?

Most students would say 50%. But binomial formula (or a normal approximation) gives a probability of only 8%. If the number of tosses approaches infinity, then the probability it will land *exactly* half heads and half tails will approach zero.

If we apply these elementary arguments to the testing of parameters in normal or binomial distributions, we cannot accept the probability of a point null to be 50%. But this is precisely what Berger has used as one of the premises in his mathematical derivations.
Berger's conclusion reveals a typical confusion of "statistical hypothesis" and "scientific hypothesis." A statistical hypothesis describes a distinct feature of a population that is narrowly defined—it is either defined for a population at a specific time (e.g., sample survey), or for a highly focused situation (e.g., repeated measurements of microwave in a very small piece of the sky). This is its limitation, but also its strength.

A scientific hypothesis, on the other hand, is something like Darwin's theory of evolution or the Big Bang theory of our universe. This kind of hypothesis does carry weight (or probability) for our belief. But a scientific hypothesis is in most cases not tested by formal statistical inference. For example, why does the scientific community accept evolution theory but reject the religious creation theory? By setting up null and alternate hypotheses, and then conducting a Neyman-Pearson (or Bayesian) hypothesis testing?

IV. BAYESIAN TIME-SERIES ANALYSIS AND E.T. (EXTRA TIME-SERIES) JUDGMENT

There are two different kinds of Bayesian statisticians: subjective Bayesians and objective Bayesians. In this section we will first examine the content of an “objective” Bayesian analysis in business forecast. We will then compare the “scientific" philosophy of such “objective Bayesians” to other schools of statistical analysis.

In the late 1970s, a new type of Bayesian model for economic forecasting was developed in the Research Department at the Federal Reserve Bank of Minneapolis (Litterman, 1986a, 1986b, J. of Business and Economic Statistics). The performance of these new models was very impressive (see Figs. 5 and 6), when compared to three professional forecasts (Chase Econometric Associates, Wharton Econometric Forecasting Associates, and Data Resource, Inc.). Both figures showed that the traditional forecasts deviated from Litterman’s, while the realized values were remarkably close to the predictions made by the new method. The model that generated the previous forecast is based on a time-series technique called vector autoregressive modeling, which has been around for many years. The novelty of the new models is that the investigators (see, e.g., Litterman, 1986a, 1986b) assign a Gaussian prior distribution to the autoregressive coefficients.

The modeler (Litterman, 1986b) emphasizes that his prior distribution is not derived from any particular economic theory. The reason is that the modeler does not find “a good deal of consensus on the economic structures involved.” Instead, the modeler decides to assign uniform prior (i.e., same standard deviation) to the first lag of the dependent variables in the equations. He also assigns decreased standard deviations of further coefficients in the lag distributions in a harmonic manner (Litterman, 1986b, p. 30). Forecasts are generated mechanically from the resulting Bayesian models.
It is hard to describe the details of the heated debates between those for and against Popper's statement. But many who opposed Popper turned silent after they realized that theory and observation are like the question of which comes first, the chicken or the egg, and that observations are always selective. In other words, science does not begin with the gathering of data. Rather, it begins with the belief that certain data should be collected in the first place.

Furthermore, Freedman (1985) observed that good measurements are critical to good science, and to make good measurements one needs good theory.

Here is another example that would further support Popper's position. The example involves Dirac's theory and the discovery of the anti-electron as discussed in the introduction to Chapter 6. Before Dirac put forth his theory of anti-matters, physicists had studied the tracks of subatomic particles for more than 20 years. But the tracks of anti-matter had always been dismissed as statistical aberrations or things of that nature. It was Dirac's theory that guided physicists to look at our universe from a totally different perspective (Gribbin, 1984).

By comparison, soft sciences usually lack good theories; research workers therefore resort to data gathering and statistical manipulations as the goal of science. But a safe bet is that, without a good theory, even if the data are thrown in the face of a "data analyst," the chance that he will find anything significant is next to zero.

In certain disciplines the reality is that measurements are rough, theories are primitive, and randomization or controlled experiments are impossible. In such cases, what can one do (other than data gathering and statistical manipulation)? This is a question frequently asked by soft scientists.

A quick answer to this question is indeed another question: "Why don't you use your common sense?"

**EXAMPLE 1** In a University lecture, a speaker tried to use time-series analysis to model the signal shown in Fig. 7.

A univariate ARIMA model certainly won't do; so he introduced a distribution to an autoregressive parameter. A "Bayesian" model then smoothed the signal and the investigator was very happy with the result. The mathematical constructs enlisted to attack the problem (such as Kalman filter and state-space modeling) were truly impressive. But in this case, questions raised by common-sense knowledge may be more revealing than the razzle-dazzle of the mathematical manipulations. According to that speaker, the signal is an earthquake data. A good scientist would try to find out the cause of the extreme variability in the signal; but the above statistician is satisfied if the data look smoothed.12

**EXAMPLE 2** A couple who want a son have given birth to 4 girls in a row. To the couple involved, having an extra child can be a big burden physically and financially. However, if the chance of having a son is not low (e.g., not less than 20%), then they are willing to give it a try. The issue is of much gravity to this couple. Therefore, they asked the following questions: (1) What is the probability of having 4 girls in a row? (2) What is the probability that the couple's next child will be a boy?

Textbook answers to these questions are very straightforward: (1) \( .5^4 = 6.25\% \), which is not significant to reject the null hypothesis that \( p = .5 \); therefore the answer of (2) is 50\%. And this is indeed the answer I got from a biologist.

However, common sense indicates that sex-ratio may be associated with many other factors, such as race, birth order, parental ages, etc. And this is indeed the case. (See, e.g., Erickson, 1976.) Furthermore, a physical examination of the couple might reveal that the chance their next child is a boy is one in ten million, not 50 percent.

In other words, one has to recognize (1) that statistical inference homogenizes population units and hides their differences and (2) that statistical laws do not apply to individuals. Therefore, in the process of making an important decision, one has to search for other information that may be relevant to the issue at hand. A useful technique to facilitate this process is "brain-storming" (see, e.g., Evans and Lindsay, 1989, p. 441). This technique, however, is not in the domain of statistical inference.

It is important to note that no data analysis should be accepted without a healthy dose of skepticism about what the data might have concealed. In other
words, if the data are not consistent with your common sense, check it out: Your common sense may not be all wrong; the experiment might have been biased or rigged; and the data might have been contaminated or deliberately misinterpreted. After all, statistical truth is only 95% truth, or 99% truth. A measured skepticism thus may lead to useful information that was previously unnoticed.

EXAMPLE 3 A question hovers. What should you believe if your common sense conflicts with “official statistics” reported by experts in the field? Here is an example in this regard.

A federal air report ranked New Jersey 22nd nationally in the release of toxic chemicals. The report was based on a comprehensive study sponsored by the U.S. Environmental Protection Agency (the Trenton Times, March 24, 1989). A spokesman for the New Jersey Department of Environmental Protection was very happy about the ranking: “It speaks well of our enforcement, we feel.” But New Jersey residents who have driven through northern New Jersey and gotten choked may not feel so.

As a matter of fact, the statistics in this example didn’t really lie, they only got misinterpreted: New Jersey releases 38.6 million pounds of toxic chemicals annually, well behind many other states. But New Jersey is a small state, and its ranking jumped from 22nd to 4th when the tonnage release was divided by square mileage.

EXAMPLE 4 In theory, if one has a random sample from a population, then one can draw “scientific” inference based on standard formulas (assuming that no bias has crept into the sampling procedure). On the surface, the information in a random sample is “complete,” and the estimated values of population parameters are thus “objective,” “deductive,” and “scientific.” However, a pursuit of information beyond the randomized data may be very revealing.

For example, in a clinical trial, the investigator conducts a randomized experiment, but finds that there are 35 males (out of 100 male patients) in the treatment group, while the remaining 65 males are in the control group. Note that this is 3 SE (standard error) from what one would expect about the number of males in the experimental group. Also, with more males in the control, the study may give the investigator a substantial edge to “prove” the effect of the treatment. In such an awkward situation, should an honest scientist go ahead to do the experiment? Also, should a referee of a medical journal accept the randomized study as “scientific?”

Throughout this book, we have been advocating the deductive nature and the scientific advantages of randomized studies. The above discussion now appears to be pulling the rug from under this stance.

An “easy” solution to this problem is to separate the patients into groups of different gender, and do the randomization afterwards. But critics to randomization may argue that age is another confounding factor and that the investigator has to separate the groups of the patients further. In order to satisfy the critics, the investigator also has to take into account race, health condition, family background, and genetic differences, etc. A cruel reality is that there may exist an infinite number of confounding factors. And pretty soon one would run out of patients to perform a “valid” experiment.

But fortunately, Mother Nature has been kind to the investigators: there are only a few factors that are important, and the rest would not affect the outcome significantly. That is, only few factors are vital, and the rest are trivial. This phenomenon—vital few, trivial many—is indeed omnipresent and is one of the reasons why statistical method seems to work: If the experimental design captures the effects of major factors, then randomization will lead to the conclusion whether the treatment is beneficial.

However, findings from such experiments are still incomplete in the following sense. First, the legitimacy of the randomized experiment rests on the law of large numbers; i.e., the findings will have credibility only if they are consistent with independent experiments. Otherwise, the earlier results may have to be re-evaluated. Second, the experiment by itself usually cannot determine the side effects and optimum uses of the treatment.

For instance, a recent study concluded that new medicines approved by the Food and Drug Administration (FDA) caused serious reactions that required changes in warning labels or in a few cases withdrawal from the market (Star-Ledger, Newark, New Jersey, May 28, 1990).

The study was conducted by the General Accounting Office (GAO), a congressional investigating agency, which reported that 51% or 102 of 198 drugs studied had “serious post-approval risks” that often were not detected or disclosed by the FDA until several years after the drugs were on the market. Such unsuspected adverse reactions included heart failure, birth defects, life-threatening breathing difficulty, kidney and liver failure, blindness, and severe blood disorders, etc.

In response to the GAO report, the spokesman of the Pharmaceutical Manufacturers Association said that the report did not address the benefits of using the drugs (including those with potential risks), and that some of the drugs identified by the GAO are either the only available treatment or the best available treatment.

The U. S. Department of Health and Human Services also took issue with the GAO report. The department, in a formal reply to the GAO, objected to the methodology and the soundness of the study and expressed concern that it will “unnecessarily alarm consumers, causing some to reject the use of lifesaving drugs out of fear of adverse events that might occur only in extremely rare instances.”

The general public may feel uncomfortable with the disagreement among scientists. But the fact is that such episodes occur from time to time. Science,
after all, does not claim certainty. Rather, it is an intellectual pursuit for a better understanding and better description of nature. As history has witnessed, when scientists think they understand everything, nature might very well give us a few more surprises. Constant challenges to the scientific establishment, if well-grounded, are thus the true spirit of science and will benefit us all.

VI. WOMEN AND LOVE: A CASE STUDY IN QUALITATIVE/QUANTITATIVE ANALYSIS

In 1972 a behavioral researcher, Shere Hite, initiated a large-scale study of female sexuality. The result was a monumental work considered by many the third great landmark in sex research, after The Kinsey Report (1948, 1953) and Human Sexual Responses (1966) by Masters and Johnson. In this section, we will examine scholarly debates over Hite's work. In addition, we will discuss the weakness of a common practice (used in Hite's work) that compares statistical breakdowns between the sample and the target population.

A motivation behind Hite's project on female sexuality was that (Hite, 1976):

Women have never been asked how they felt about sex. Researchers, looking for statistical "norms," have asked all the wrong questions for all the wrong reasons—and all too often wound up telling women how they should feel rather than asking them how they do feel.

The purpose of this project is to let women define their own sexuality—instead of doctors or other (usually male) authorities. Women are the real experts on their sexuality; they know how they feel and what they experience, without needing anyone to tell them.

For this purpose, Hite believed that:

- a multiple-choice questionnaire was out of the question, because it would have implied preconceived categories of response, and thus, in a sense, would also have "told" the respondent what the "allowable" or "normal" answers would be. (Hite, 1987)

As a consequence, Hite decided to use essay questionnaires and claimed that they "are not less 'scientific' than multiple-choice." By 1976, a book based on 3,000 women's responses to her essay questionnaires was published and soon became a huge success in both sales volume and the candid revelations of how women really feel about sex.

The study was replicated (and confirmed) in at least two other countries. The book and her second volume on male sexuality have been translated into 13 languages, and are used in courses at universities in this country and

around the world. Because of her work, Hite was honored with a distinguished service award from the American Association of Sex Educators, Counselors and Therapists.

An important revelation in Hite's 1976 report is that women had been compelled to hide how they feel about lack of orgasm during intercourse. A natural conclusion (Hite, 1976) is that the traditional definition of sex is sexist and culturally linked:

Our whole society's definition of sex is sexist—sex for the overwhelming majority of people consists of foreplay, eventually followed by vaginal penetration and then by intercourse, ending eventually in male orgasm. This is a sexist definition of sex, oriented around male orgasm and the needs of reproduction. This definition is cultural, not biological.

In the early 1980s, Hite further sent out 100,000 questionnaires to explore how women were suffering in their love relationships with men. The response to this survey was a huge collection of anonymous letters from thousands of women disillusioned with love and marriage, and complaints about painful and infuriating attitudes on the part of men.

Hite's book, Women and Love (1987), thus provided a channel of release for women who experienced frequent degradation and ridicule by men. The book advocates another "cultural revolution" (instead of "sexual revolution") and was praised by some scholars:

The Hite report on Women and Love explodes the current myth of the "insecure women" who seek out destructive relationships...it documents a hidden, "socially acceptable" pattern of behavior in relationships which puts women's needs last. 

Naomi Weisstein, Ph.D.
Professor of Psychology
SUNY, Buffalo

A massive autobiography of women today.

Catherine R. Stimpson
Dean of Graduate School
Rutgers University

The Hite Report trilogy will long be regarded as a landmark in the literature of human behavior...respect will continue to grow for these works in years to come.

Richard P. Halgin, Ph.D.
Department of Psychology
University of Massachusetts

The most important work on men and women since Simone de Beauvoir's The Second Sex.

Albert Ellis, Ph.D.