

# PART II

## INTRODUCTION TO SCIENCE AND THE SCIENTIFIC METHOD

---

### CHAPTER 2

#### SCIENTIFIC METHOD: THE LOGIC OF DRAWING INFERENCES FROM EMPIRICAL EVIDENCE

*Table of Sections*

- Sec.**
- 2-1.0 Introduction.
    - 2-1.1 How Science Finds Answers.
    - 2-1.2 The Nature of Empirical Questions.
    - 2-1.3 Empirical Questions and Normative Questions.
    - 2-1.4 Purposes of Research.
    - 2-1.5 Settings for Research.
  - 2-2.0 Defining and Measuring Things.
    - 2-2.1 Conceptual and Operational Definitions.
    - 2-2.2 Scales of Measurement.
    - 2-2.3 Reliability and Validity.
    - 2-2.4 Roles Variables Play.
    - 2-2.5 Statistical Aspects.
  - 2-3.0 Sampling.
    - 2-3.1 Units of Analysis.
    - 2-3.2 Types of Sampling.
    - 2-3.3 Selection Bias.
    - 2-3.4 Statistical Aspects.
      - 2-3.4.1 Sample Size.
      - 2-3.4.2 Confidence Limits.
  - 2-4.0 Relationships Among Variables.
    - 2-4.1 Minimum Conditions for Inferring a Relationship.
      - 2-4.1.1 Single Cell Missing Data Pattern.
      - 2-4.1.2 Single Row (or Column) Missing Data Pattern.
      - 2-4.1.3 Main Diagonal Missing Data Pattern.
    - 2-4.2 Predictive Relationships.
      - 2-4.2.1 Statistical Aspects.
    - 2-4.3 Causal Relationships.
      - 2-4.3.1 Squeezing Causation From Correlational Data.
      - 2-4.3.2 Threats to Internal Validity.
        - [1] History.
        - [2] Maturation.
        - [3] Testing.
        - [4] Instrumentation.
        - [5] Statistical Regression.
        - [6] Selection.
        - [7] Experimental Mortality.
        - [8] Selection-Maturation Interaction, Etc.
        - [9] Chance.

## Sec.

[10] Distortions in Independent and Dependent Variables.

2-5.0 Threats to External Validity.

2-6.0 Research Designs.

2-6.1 Pre-experimental Designs.

2-6.2 True Experiments.

2-6.3 Quasi-experiments.

2-6.4 Statistical Aspects.

2-7.0 Conclusions.

Appendix 1. References on Research Methods and Philosophy of Science.

**WESTLAW Electronic Research**

See WESTLAW Electronic Research Guide preceding the Summary of Contents.

**§ 2-1.0 INTRODUCTION**

This chapter is a primer on scientific method. Its aim is to provide the non-scientist judge or lawyer an easily accessible understanding of how scientists learn about how the world works. At the root of scientific method is the application of logic to the problem of how to observe an empirical phenomenon in a way that will allow one to draw valid inferences about that phenomenon. In that sense, this chapter is really about a particular branch or application of logic.

The subject of scientific method is a necessary foundation of *every* discipline that seriously attempts to gain knowledge of the world through systematic empirical inquiry—the physical, biological, behavioral and social sciences alike—and as such it has been described innumerable times, in a multitude of works on manifold subjects, from elementary school textbooks to post-graduate treatises. And yet it remains a subject that is foreign to most lawyers and judges. This is a subject about which most lawyers and judges are, to put it bluntly, quite illiterate. For these reasons, this chapter will be presented in a simple and direct form, without footnotes, though it will conclude with a general bibliography of sources to which the reader can turn for more detailed learning about these concepts.

**§ 2-1.1 How Science Finds Answers**

How does a scientist learn the answer to a question? For examples: Which of several proposed treatments for cancer, inflation, or pilot error work best? Does Vitamin C prevent colds, does one surgical technique work better than another, which methods of teaching, giving judicial instructions, or farming work best? Or, to ask less “applied” and more “basic” questions: What is the nature of motion? What accounts for the inheritance of traits? How does memory work?

God does not whisper the answers into the ears of scientists, as though they were members of a modern priesthood. The only way a scientist can reach an answer to an empirical question is to conduct an empirical inquiry. This means observing the phenomenon of interest, though usually in an especially disciplined way. Mere observation is a part of the scientific method, but it usually is not nearly enough.

For example, how many teeth are in a horse's mouth? While Platonic philosophers believed that the answer was to be found by reasoning or debating what the proper number should be, a modern scientist's immediate instinct is that the answer is to be found in a horse's mouth. But one look into the mouth of one horse may be misleading. Any one horse might have a defective set of teeth, due perhaps to injury; or the number may change with age; or different breeds or sexes may have different numbers. So the modern scientist would systematically sample horses, and would most likely report the average and the range of variation for major subgroupings of horses. This illustrates how even the simplest of questions needs to be approached in a systematic and thoughtful way to avoid arriving at incorrect answers.

Many of the questions posed in the three paragraphs above can be answered only with a more specialized and more disciplined way of observing, which involves comparing things under different conditions. For example, for a century many surgeons believed that the best treatment for breast cancer was to remove the entire breast and considerable additional tissue along with it, so as to get downstream of any cancer cells and head them off (so the theory went). Suggestions that less destructive surgery might do as much good were met with defenses based on faith, rather than evidence, that it is only logical that radical mastectomies provided the greatest protection against spread of the cancer. In the 1970's experiments were performed which finally compared the efficacy of radical mastectomies to the far more conservative procedure of lumpectomy plus radiation. A sample of breast cancer patients was randomly divided so that some were given the traditional mastectomy while others were assigned to receive the more breast-conserving treatment. The latter proved to be at least as good as the former at stopping the cancer. This illustrates an experimental comparison, where the value of a treatment is discerned by comparing it to another treatment.

To real scientists a finding of fact is only as good as the methods used to find it. Scientific method is the logic by which the observations are made. Well designed methods permit observations that lead to valid, useful, informative answers to the questions that had been framed by the researcher. For scientists, the key word in the phrase "scientific method" is *method*. Methodology—the logic of research design, measures, and procedures—is the engine that generates knowledge that is scientific. While for lawyers and judges credibility is the key to figuring out which witnesses are speaking truth and which are not, for scientists the way to figure out which one of several contradictory studies is most likely correct is to scrutinize the methodology.

We conclude this introduction on a note of caution. Although this chapter is focused on the challenges of doing science well, the reader should be alert to the problem of assertions made on the basis of no science at all. Not all knowledge asserted by people who are commonly thought of as scientists is the product of the scientific method. It will help to think of science as a verb, not a noun. Science is what one does to build knowledge, not what someone is.

Some people or groups who call themselves scientists do not use the scientific method. That is, their beliefs have not been subjected to systematic empirical testing. Their own and their field's beliefs are based on casual

observation, or intuition, or faith, or the authority of past generations of members of their field exercising *their* intuition. Masquerading as science, such claims are likely to be defended by statements that the truth of the assertion rests on “my many years of experience,” or “generations of study by my field.” Were the findings based on evidence produced by the scientific method, the expert should be able to present those studies to any audience, including a court, along with the methodology and the results of the studies.

By shifting the attention of federal judges from the consensus of opinion among the field’s members to the underlying science of a field, *Daubert v. Merrell Dow Pharmaceuticals* has caused a number of fields to begin to re-examine themselves. At a panel discussion at a conference, defenders of one field responded to critics with such answers as these:

We who practice in this field know that what we are doing is correct, though we have no way of demonstrating that truth to outsiders.

You critics have focused your attack on our one weak spot, the lack of data about our claims.

To anyone with even a modest understanding of how scientific knowledge is generated and tested, these “defenses” will be recognized as admissions that science is absent from this field.

Sometimes there is a zone of genuine scientific knowledge possessed by a field, but some or many of its members step outside of that zone and make assertions that exceed their field’s empirically tested knowledge. Or they are answering questions that are based in part on well tested knowledge and in part on speculation. Before research had been conducted on the relative efficacy of breast cancer surgical strategies, the cancer surgeons performing radical mastectomies were not applying scientific knowledge about which treatment was most effective, because such knowledge did not exist. With respect to other surgical techniques, they may have been applying knowledge tested by the scientific method. A review of experiments testing a number of the most popular surgical techniques found that about a third of them were indeed effective, about a third were worthless, and about a third were doing more harm than alternative techniques that were available. Thus, it is less helpful to ask whether a person making an assertion is “a scientist” (such as a research chemist or research psychologist) or is “applying science” (such as a physician or engineer or clinical psychologist), and more helpful to ask how well grounded the assertion is on well-designed and well-conducted empirical studies testing the propositions contained in the assertion.

### § 2-1.2 The Nature of Empirical Questions

Not all questions are “empirical questions.” Empirical questions and the search for their answers are the special, and limited, contribution of scientific inquiry. Empirical questions can and should be tested empirically. Other kinds of questions do not call for empirical answers and are not subject to empirical testing.

Consider the following examples:

“Is the moon made of green cheese?” This is an empirical question that at one time had no answer and today the answer is known.

“Is there life on other planets?” This question remains unanswered. Note that we will never be able to say, “no, there is no life other than on Earth,” unless and until we can examine every other planet in the universe. Some people may be content to examine several thousand planets without life and make the inferential leap that there won’t be life on any of the other billions either. But the moment life is found elsewhere, if it ever is, we can say, “yes.”

“Is astrology valid?” There is nothing unscientific about the question. Numerous empirical studies have been conducted to test the claims of astrology, and virtually all of them have found astrological predictions not to be correct and astrology therefore not to be valid. The results might have come out the other way. That is how the scientific method sorts out the valid from the invalid.

“Does God exist?” “Is the death penalty moral?” “Does the word ‘vehicle’ in the statute include tricycles?” These are three questions that are not empirical questions and can have no empirical answers. Science cannot help. The first question asks about the supernatural, while empirical questions are confined to the natural world. The second question is a normative judgment, a value judgement. Again, that is not something scientists can study directly. A social scientist could, however, conduct a survey to see how many people believe the death penalty is a proper punishment, or a study of whether it in fact deters, or whether it is applied disproportionately more to some groups of people convicted of capital crimes and less to others. Those are empirical questions. But they are very different from the question of whether the death penalty is or is not moral. The third question is one of interpretation, of the lawmaker’s intentions, no doubt in the context of a court’s application to a particular situation. Interpretation of such meanings is not the realm of empirical question asking and answering.

### § 2-1.3 Empirical Questions and Normative Questions

Empirical questions are questions of “is.” Normative questions address issues of “ought.” In one sense they exist in two different intellectual worlds and the responsibility for dealing with them is divided into different professions. Scientists, engineers, and therapists handle the former; philosophers, theologians, and literary scholars deal with the latter. But sometimes there are important connections among them.

Sometimes normative decision-makers make decisions based on empirical questions to which answers are assumed. A lawmaker may have a goal in mind, such as, “we want to save lives by closing the gap between the supply of and demand for transplantable organs.” How to accomplish that is an empirical question. Methods may be chosen that cause the opposite to happen. In such a situation, an organ procurement organization might challenge the statute in court as one that fails the rational relationship test. Lawmakers may want to limit the availability of pornography because they believe it causes some harm. Whether it causes harm, has no effect on anything, or does some good is an empirical question. Thus, normative decision-making usually cannot be made entirely free of knowledge of the factual world.

The other side of the coin is that normative assumptions may guide, or mislead, empirical research. For example, in testing which of two medical treatments works "better," the researchers are likely to measure how long those receiving the different treatments live. The one that produces longer life is inferred to be the better treatment. But implicit in that empirical study is the value judgment that longer life, and not something else (such as freedom from pain, ability to function normally) is what matters. Patients might have made a different value choice and concluded that by their criteria the other treatment was "better."

### § 2-1.4 Purposes of Research

Research may be conducted for a variety of different purposes, among them curiosity and exploration, applied testing, description, formal theory development (basic research), and, occasionally, generating evidence relevant to resolving disputed factual questions in litigation.

Research may be conducted to answer a question that has piqued a researcher's curiosity. For examples: Benjamin Franklin wondered about the nature of those flashes in the sky during storms. Someone else wondered what might happen if one tried to combine one of the "noble" elements with some other element. An early astronomer may have wondered what would be seen if a telescope were aimed in a certain direction in the sky. To be human is to have questions about the world, and research can be conducted to answer many of those questions.

Closely related would be more systematic exploration in the pursuit of some phenomenon. For example, the early dentists who discovered pain killing drugs did so by trying one compound after another and observing their effects.

"Applied research" is aimed at answering immediate, practical questions. In developing a new jacket for sailors so as to increase their chances of survival if they fell into the sea, researchers wanted to know, first, from what parts of the human body was most heat lost (so extra insulation could be placed in those areas) and, second, what color is most easily perceived by people looking out into water from the angles at which rescuers tend to be looking (so the jackets would be made in that color). The search for anesthesia, mentioned above, fits this category as well. Or, does a certain drug cross the placental barrier? Which of several available paving compounds is stronger and lasts longer? Do analog or digital displays convey information more quickly and accurately to pilots and nuclear plant operators?

Some research is conducted to provide a thorough description of something. Thus, 19th century naturalists tried to collect every kind of specimen of plant or insect or bird that they could, in order to describe what they found. Later scientists would try to explain the patterns of similarity and variation that had been found.

"Basic research" is performed in order to develop theoretical understanding of a phenomenon of interest. In everyday parlance, theory often is taken to mean the opposite of fact. But in science it means an explanation for a set of observed facts. Theory is not contrary to fact, it is the abstract or

conceptual account for why the observed facts exist as they do. These may or may not lead to practical applications. The goal of basic research is knowledge and understanding for its own sake. The steps in the process of theory development and testing—hypothetico-deductive research—is oft-repeated in textbooks, and runs along these lines:

1. Observations of some phenomenon are made. For example, the movements of planets (which move in more complex orbits than the stars).

2. Possible explanations (theories) are proposed for what is observed. (For the movement of planets, one such theory, radical at the time of its first suggestion, was that the movements of planets could be explained by a theory that placed the Sun and not the Earth at the center of our solar system.)

3. Hypotheses are logically derived from the theories. (If the Sun is the center of the solar system, then certain other observations should be true. If the Earth is the center of the solar system, that would lead to different predictions.)

4. Studies are designed to test the hypotheses. In essence, the study makes new observations that might disconfirm the hypothesis and thereby falsify the theory. Different theories have different implications and lead to different hypotheses. (Ideally, a study can be devised whose outcome will disconfirm one theory's hypotheses and not the other's. This is called a "critical experiment" because it permits a head-to-head test of two or more theories, and helps to determine which has done the best job of accounting for the relevant phenomena. Sometimes scientific controversies persist for a very long time because no commonly agreed upon critical experiment can be conducted.)

5. The results of such empirical tests lead to the revision or abandonment of older theories or the creation of still newer and hopefully better theories.

6. The process repeats itself as more empirical tests are conducted and theories undergo continued re-evaluation.

Note that a hypothesis or a theory is never proven or confirmed to be true. Testing is capable only of disconfirming. But theories that withstand such attempts at falsification better and longer become accepted, at least until something better comes along. The opposite approach can readily be seen in non-scientific activities of numerous kinds, where investigators engage in a search for evidence that confirms their suspicions. This "confirmatory bias" is based on the erroneous assumption that a theory is confirmed by the accumulation of facts consistent with the theory. The logic of science is that innumerable consistencies can be found, all of which tell us little. It is the diligent search for inconsistencies, for falsification, that really puts a theory to the test. A theory that can withstand such scrutiny is one that deserves credence.

Finally, research sometimes is conducted to address disputed questions of fact "in anticipation of litigation." Such "tailor made" research affords the

benefit of being focused more directly on the issues before the court than “off the shelf” research, which had been conducted for one of the other purposes discussed above. On the other hand, research conducted specifically to address litigated factual questions is more suspect for having been prompted by the litigation and designed and carried out in consultation with counsel for one party.

### § 2-1.5 Settings for Research

Research can take place in a number of different settings, for example, in the laboratory, in the field, or through a simulation or models of various kinds. There is no one best place to study the phenomena of interest. Each choice involves trade-offs. The question, as always, is whether the circumstances of the research are appropriate for the focus of the study, and whether the conclusions drawn are sensible in light of the data collected and everything about the manner in which the data were collected.

Research conducted in a *laboratory* has the advantage of allowing the researcher to exercise the utmost control over extraneous influences on the phenomenon being studied. It also is more convenient and less expensive. Laboratory research sometimes involves a simulation of the phenomenon of interest. The disadvantage is that studies conducted in laboratories may involve more artificial instances of the phenomenon of interest, and may therefore be less generalizable to the more natural situation in which the researcher (and the law) may be interested. Further, laboratory studies of biological phenomena can be conducted *in vitro* (that is, in glass, in a test tube) or *in vivo* (in a living organism, such as a laboratory animal). Though both occur in laboratories, the difference between *in vitro* and *in vivo* studies is obviously a considerable one.

*Field research* is one solution to the problem of wanting to make sure that the phenomenon studied behaves as much like the natural versions to which the researcher would like to generalize any findings. The problem with field research is that it is harder to keep the variation from extraneous influences from intruding and masking the effects of interest. A review of research comparing studies of attitude change in laboratory settings versus field settings found that in answer to the same research questions, the two settings did not so much produce contradictory results as that often a phenomenon that was detected in the laboratory setting (where control was high) was not able to be detected in the field setting (where control was low).

*Simulations* of the real thing can make research more efficient, but perhaps also more dubious, depending upon how well the simulation captures the essential features of the phenomenon of interest. Simulations can consist of computer models (mathematical representations of the phenomenon of interest), animal models (study something in animals before doing so in humans), building a physical model of a process or object to be studied (e.g., testing the characteristics of a new aircraft design), simulations of social situations (e.g., how people react to emergencies), games (as in political science, business, or economics studies of how people interact), and so on.

Finally, there is *survey research* conducted by live interviewers or written questionnaires asking people to answer questions about their own past behavior (e.g., how much they purchase of a particular product), or how much they will do something in the future (e.g., if we build it, will you come?), or their attitudes and beliefs (e.g., whether they think a defendant who is to be tried in a court in their community is guilty or not), or about their reactions to something presented to them at the time of the research (e.g., who they think manufactured a certain product shown to them).

Such studies give the researcher a great gain in efficiency. It is easier to ask someone how often they drive while intoxicated than it is to try to follow them around and directly observe the behavior. But, as the example makes apparent, the price of ease of inquiry may be decreased accuracy. People may give untrue answers in order to try to appear more socially acceptable. Or their sincere memories may be imperfect. Moving through the list in the preceding paragraph from the top to the bottom brings us from the sorts of questions that may produce the least accurate information (recall of past behavior and prediction of future behavior) to what probably are some of the most trustworthy responses (being shown something and asked relatively innocuous-sounding questions about it). Thus, the content of the question needs to be scrutinized.

Of course, as every lawyer and judge knows, the way a question is framed, the choice of words, and other features of it can skew the answers obtained. For example, people give different answers when asked about, say, rights of “the pre-born” versus of “the products of conception.” For more discussion of the topic of self report survey research, see Chapter 5—Legal Applications of Survey Research.

## § 2-2.0 DEFINING AND MEASURING THINGS

The first step in the scientific method is to define what is to be observed. Though this may seem terribly basic, it is the first chance for the research to seriously misfire.

### § 2-2.1 Conceptual and Operational Definitions

*Conceptual definitions* are abstract statements of the phenomena of interest. *Operational definitions* are the concrete procedures one must undertake to observe the things being discussed. Talking about concepts in the abstract is one thing (e.g., aggressiveness, intelligence, reasonable decisions, disability). Defining precisely what observations are to count as an instance of the concept and what is not—that is, an operational definition—at least for the purposes of the research, is far more difficult. Moreover, one can make the world appear to be quite a different place merely by using different operational definitions of something. Consider several examples:

One study set out to find out how much intra-family violence occurred in a given locale and found it to be epidemic. A second study concluded, by contrast, that family violence was quite a rare occurrence. By their conceptual definitions, the studies were interested in the same thing. The discrepancy between their findings can be understood by examining their respective

operational definitions. The first study defined family violence to include anything from a raised voice onward. The second study did not count as violence anything that did not require hospitalization.

The manufacturer of a headache remedy once tried to convince consumers that “all aspirin is not alike” by insisting that theirs was “better” than the competition and flashing an address on the screen where viewers could write for a report on the supporting details of that claim. That report revealed that the product was superior along such dimensions as: clarity of labeling, accuracy of the count of pills in a 100-count bottle, stick-to-it-ness of the glue on the label, number of broken tablets in the average bottle, and so on. While the advertiser’s operational definition of “better aspirin” included those sorts of attributes, potential consumers no doubt assumed the manufacturer was informing them about the pharmacologic efficacy of the aspirin.

Years ago, a presidential administration was able to improve its civil rights record overnight by redefining “a desegregated school district” from being more than half of the schools in a district desegregated to being at least one school in a district. Similarly, indices of crime, cost of living, economic growth, educational achievement, and just about anything else can appear to change merely by changing the operational definition used, even though the world did not change a bit.

Changes in operational definitions sometimes can be quite subtle, as in the following example. Some researchers have concluded that apparent progress in cancer treatment is illusory, because increases in standard 5-year survival rates reflect nothing more than improvements in early detection. That is, by starting the clock sooner, more people reach the five year point, even though the natural history of their disease was unchanged and the same number of people died at the same time. This example involves an almost unnoticeable change in the operational definition of when the cancer “begins,” which changes the definition of whether one “survived” it or not.

A consumer of research knowledge has to satisfy himself that the researcher’s operational definitions adequately capture the concept that the research consumer is interested in knowing.

### § 2-2.2 Scales of Measurement

Once the operational definition is judged to be acceptable, the thing to be observed has to be measured. There are four basic “scales of measurement” that can be used. The importance of these scales is that the statistical procedures to be used to analyze the data depend upon what type of scale was used. A researcher who uses a statistical procedure that is not suited to the scale of measurement used can obtain results that are not valid.

*Nominal* scale measurement involves nothing more than “naming,” often placing the thing to be measured into one category or another. Hair color, gender, disease diagnosis, numbers on athletes, or phone numbers all are examples of nominal scale data. With this kind of measurement, one can do no more than to count the number of objects in a category, use as a measure of central tendency the mode (that is, the most frequently occurring category),

and employ statistical tools that work with categorical data, such as chi square tests.

*Ordinal* scale measurement adds order to naming, that is, notions of greater or lesser on some dimension. Ordinal scaling involves ranking. Runners finish first, second, third, and so on. We know that the one who finished first was faster than the one who finished second, but we cannot tell from the ranking how much faster the first place finisher was than the others. The gap between first and second may have been an hour, and between second and third only a minute. With ordinal measurement, one can calculate medians and use statistical tools such as rank correlations, in addition to those suited to nominal measurement.

*Interval* scale measurement adds quantity to order. Interval scaling involves rating rather than ranking. It tells us that the distances between scores consist of equal intervals. A person with an IQ score of 130 is as much smarter than a person with a 120 as the person with a 60 is compared to a person with a 50. Attitude scales, ratings of disease seriousness, and ratings of employee performance are other examples of interval scale measurement. But interval scales do not possess true zero points, so one cannot say that a person whose job rating is 100 performs twice as well as someone with a 50. Such a statement distorts what is possible with interval scale measurement. With interval measurement one can calculate means, standard deviations, Pearson correlations, and use statistical significance tests such as t-tests and F-tests. One can also use any of the statistical techniques permitted for nominal or ordinal data.

*Ratio* scale measurement adds a true zero origin to the above and permits ratio statements to be made. With ratio scales it becomes possible to say, for example, that a person who runs a mile in five minutes is twice as fast as someone who runs the mile in ten minutes. Quantities such as weight, cost, distance, and time are examples of ratio scale measurement.

The governor of a particular state once argued that it had improved its expenditures on education, and offered as evidence the fact that compared to the previous year, it had increased its expenditures by some considerable amount. An opponent said the state had, by contrast, slipped in its treatment of education, and offered as evidence the fact that its rank among all the states had not increased but fallen. Very likely both had their basic facts correct, but were using different scales of measurement. Accordingly, they can arrive at different conclusions. The ratio measure of absolute expenditures can go up at the same time that relative rank can go down.

At a committee meeting, the chair asked members to rank order the preferences for certain policy options before them. The chair then computed the mean rank of each option. But ranked data cannot validly be used to calculate means, because the mathematical process employed requires data that are at least interval. When properly analyzed, the median (the middle score) of the ranked data, revealed a different set of preferences among the committee members than the other, incorrect, technique for averaging the data had.

### § 2-2.3 Reliability and Validity

Scientists draw a sharp distinction between reliability and validity. In *Daubert v. Merrell Dow*, Justice Blackmun took pains to reject that distinction for the law of evidence, and to combine both reliability and validity into what he and many lawyers and judges before him referred to as the reliability of evidence. This is more than a semantic distinction, and perhaps it could be made more apparent by using three different words to refer to the concepts. For example, we might say that measures need to be “trustworthy” before we can put much confidence in them, and the main components of such trustworthiness are repeatability (reliability) and accuracy (validity). Both have to do with how good a measure is, and they tell us different things about the measure.

To a scientist or statistician, *reliability* refers to the ability of a measure to produce the same result each time it is applied to the same thing. Reliability refers to consistency, or reproducibility. If each time a person steps on to a bathroom scale it gives a different reading (while the person’s weight has not changed) then the scale is said to lack reliability. The reliability of a bathroom scale may be tested, for example, by having 50 people who weigh different amounts step on to it twice each, and then comparing the 50 pairs of readings. If the two sets of scores are highly correlated (that is, a person’s first reading is highly predictive of the second reading), then the bathroom scale can be said to be a reliable measure. Reliability is a necessary but not sufficient condition for a good measure. A measure can be reliable without being valid. Suppose someone decided to use the bathroom scale to measure intelligence. Even a bathroom scale with perfect reliability would have no correlation with whatever we mean by intelligence. So as a measure of intelligence the bathroom scale would have perfect reliability but no validity. *Validity*, then, is the extent to which something measures what it purports to measure. A measure can have no more validity than it has reliability. Think about the bathroom scale: if the readings bounce around (low reliability), we cannot know which reading is correct (low validity).

The reliability and validity of measures generated by human subjective judgment, as well as by mechanical instruments, laboratory tests, and paper-and-pencil (these days, computer-assisted) tests can be assessed by appropriate studies of reliability and validity. For example, there have been studies of the reliability (and sometimes the validity) of the clinical judgment of psychologists, teachers, trial judges, juries, radiologists, sonar operators, and others. For an interesting example of the complete divergence of reliability and validity, see Chapter 22—Handwriting Identification (reporting a study in which all document examiners reached the same answer to a problem (perfect reliability) but they were all wrong (zero validity)).

### § 2-2.4 Roles Variables Play

Variables may play a number of different roles in a study, depending on the nature of the study and the questions posed by the research. Some studies are designed merely to measure and describe something, but not to explain or predict that something. Where the question posed by the research is one of

cause and effect, the variable that is the cause will be termed the *independent variable* and the one that is the effect that responds to the cause is termed the *dependent variable*. Sometimes this relationship is a complex one, with other variables mediating between the cause and the effect, and these are called *intervening variables*. Extraneous variables which themselves systematically influence the dependent variable, and create the illusion of a cause-effect relationship between an independent and dependent variable, are called *confounding variables*, or sometimes merely *confounds* or *confounders*. In a frankly predictive study, which has no aspiration to explaining cause and effect, the variables doing the predicting may be called *predictor variables* and the variables being predicted will be called either *criterion variables* or, again, *dependent variables*.

Nothing inherent in the variables leads to these designations. Rather, the roles they play in different kinds of studies lead to their being referred to by different designations that reflect those roles.

### § 2-2.5 Statistical Aspects

Once variables are operationally defined, reliable and valid measures of that variable have been chosen, and the objects of study have been measured, those data can be analyzed using statistical tools. Rather than a study resulting in a long list of numbers, descriptive statistics are computed to provide summaries of the distribution of those data. The central tendency (also known as the "average") of the distribution may be given by the mean, the median, or the mode, depending on the scale of measurement used and the shape of the distribution. The variability (the spread-out-ness) of the distribution can be expressed by the variance, the standard deviation, or the range, among other ways. (For further details about these descriptive statistics, see Chapter 3—Statistical Proof.)

## § 2-3.0 SAMPLING

Researchers rarely collect data on every single instance of the objects of study. That is called a *census*. They usually *sample* those objects. Agricultural researchers sample the corn in a field; they do not measure each and every ear. The same goes for every other kind of empirical researcher. Sampling not only is less expensive and less time-consuming, under most circumstances it is more accurate than a census. With a proper sampling design, resources can be directed at collecting the most accurate data on a smaller number of people, things, or events. Indeed, demographers evaluate how well the United States census is conducted by comparing the results of the census to samples.

### § 2-3.1 Units of Analysis

The first step in sampling is to decide what is to be sampled, that is, what is the unit of analysis, what is the level of aggregation. For example, does one collect data about individual people or about aggregations such as cities or nations? About workers, organizations, or industries? About rocks, planets or solar systems? Some things exist only at higher levels of aggregation. For example, the way a corporation is organized cannot be discerned by examining

individuals, but only the structural relationships of groups of individuals to other groups.

These choices may have consequences for the statistical analyses that can be conducted and the conclusions that are drawn. Sometimes, when the phenomenon of interest can be studied by observing things at different levels of aggregation, different conclusions result from one using one unit of analysis rather than another.

### § 2-3.2 Types of Sampling

Typically, the goal of sampling is to learn about an entire population of things by looking at a subset of them. The key to accomplishing this is to select the sample in such a way that it is representative of the population. Then, what one learns about the sample is likely to be true also for the population.

The methods used to do this are known collectively as probability sampling. *Probability sampling* involves selecting cases from the population in such a way that there is a known probability of any case appearing in the sample. This permits the use of probability theory to draw inferences about the nature of the population. Following are some common kinds of probability sampling.

*Simple random sampling* involves drawing a sample from the relevant population so that every member of the population has an equal chance of being selected into the sample. For example, if one wanted to measure the incidence of a certain disease among students at a school, a sample of students could be chosen at random from a hat or a computer could generate a random subset of students. Then the sample could be contacted for whatever testing was needed. *Systematic sampling* is similar, but instead of choosing at random, a random starting point would be selected and then every, say, 10th student in the student directory would be chosen (providing a 10% sample of the student body). These methods work fine if every member of the population of interest is known and the population is homogeneous.

A *stratified sample* is one in which subgroups of the population have been specified in advance, and then sampling takes place from within each stratum. This is helpful when some groups within the population occur in small numbers, and a researcher wants to be sure that enough of them are drawn so that a large enough sample of them is obtained to draw trustworthy inferences about the subgroups in the population. One could draw samples proportionate to the size of each stratum. For example, 40% of the sample could be drawn from the stratum that contains 40% of the population and 60% of the sample from the stratum that contains 60% of the population. Alternatively, one might wish to draw a "disproportionate stratified sample." For example, a sample of 250 from the 10% minority and 250 from the 90% majority. Simple random sampling would have yielded only about 50 for the minority, and that might have been too few. The two subsamples from this stratified sampling would then be weighted (one receiving nine times the weight of the other) so that in the end the statements made about the population would be accurate.

The two preceding methods can be combined in various ways to deal with more complicated circumstances, such as where the population is far flung, heterogeneous, and the identities of elements (usually, names of people) are not known. An example of this is *multi-stage cluster sampling*. Imagine that we wanted to study the health of people within the U.S. We do not have a list of every individual to choose them at random, and even if we did it would be inefficient to try to visit those selected, dispersed all over the country. But we could obtain a list (a "sampling frame") of all counties in all states. We could draw a representative sample of counties from that list. The list might even be stratified in some way, such as by region of the country. (That provides the first stage of the sampling design.) From within the selected counties, we could draw a random sample of people to be tested, perhaps through random digit dialing of telephone numbers and inviting them to be tested. (Selecting those individuals is the second stage.) Thus, the people who become the sample are "clustered" in selected counties around the country.

Two lessons should be drawn from the sampling designs described. A particular sampling design can be devised to suit the nature of what is to be sampled by combining different sampling methods (random and stratified, in different stages). And each such design is, nevertheless, a probability sample because one would know the likelihood of selecting any element at any stage.

Various methods of *non-probability* sampling have been developed. One of these is *purposive sampling*, reflecting that the researcher has a particular purpose in mind for the way the sample is chosen. For example, in a study to discover how physicians learn about new drugs, researchers began with pharmacy records showing which physicians prescribed the drug in a given community. Then they interviewed the physicians, asking, among other things, which of their professional friends they also socialize with, and asked those friends who their professional friends were. The particular method described here is called "snowball sampling." As the sample grew, the researchers could track the diffusion of awareness of the new drug through the emerging friendship network. (It was found that doctors were more likely to learn about new medications from their friends than from medical journals.) It should be apparent that one of the more traditional probability sampling methods would not have been as useful to answering the research question.

More detailed discussion of several aspects of sampling can be found in Chapter 5—Legal Applications of Survey Research.

### § 2-3.3 Selection Bias

The major defect of any sampling project would be that it fails to select representative elements from the population of interest. Most commonly this results from selection bias.

*Selection bias* refers to a sample being drawn in a way that makes it unrepresentative of the population to which inferences are to be made. The problem of selection bias is most easily understood by considering several examples. The reader might ponder why conclusions from each of the follow-

ing samples are likely to be misleading. (Following in parentheses are likely answers.)

A criminologist set out to study criminals by conducting extensive interviews with a random sample of inmates in a state's prisons. (Whatever the study finds can tell us only about those criminals who were caught and incarcerated.)

Doctors learned about the nature of the disease histoplasmosis by studying patients who came to their hospitals with the disease, and concluded that it was a rare disease which was almost always fatal. (By contrast, public health researchers sampled the public at large and found the disease to be far more common and to lead only rarely to serious harm. Doctors in hospitals saw only the few patients who suffered from the disease seriously enough to require medical attention.)

During World War II a study was done to determine where additional armor might be placed on planes to protect them from anti-aircraft fire. To do the study, planes returning from missions were examined to see where they had taken the most hits from enemy fire. (The planes that were shot down but could have been saved with additional armor were the ones that would have been the most informative, but, of course, those never returned.)

A graduating class organizing committee sent an anonymous questionnaire to their members prior to their 25th reunion. It included the question of the classmate's income. When the responses that came back were averaged, the committee was surprised to learn how successful their graduating class members had been in life. (A "response rate" of less than 100%, even a low one, is not in itself a problem. The issue, as usual, is whether those who responded are representative of the population. In the present example, if the more successful alumni are more likely to send back their questionnaires, the average income for the class will appear higher than it actually is.)

## § 2-3.4 Statistical Aspects

### § 2-3.4.1 Sample Size

How large should a sample be to yield valid results? The answer is somewhat counter-intuitive. First of all, the absolute size of the sample is what is important, not the size of the sample relative to the size of the population. (The mathematics of this fact are presented in most textbooks on the statistics of sampling.) Second, the researcher usually has a better idea of what the needed sample size was after the data are collected than before. Third, the larger the sample size the better. The problem researchers face, however, is that while a real gain in accuracy can be achieved when going from  $n=10$  to  $n=25$  or to  $n=100$ , as larger numbers are added to the sample, the marginal gain in accuracy shrinks quite considerably. So a researcher has to ask whether the extra trouble and expense involved in adding another 100 or 500 is worth the gain in accuracy that will be achieved.

How large a sample needs to be depends upon three things:

1. The homogeneity of the variable to be measured in the population. For example, suppose a warehouse holding cans of soup had been flooded and all the labels washed off. But it was known that only a single flavor (that is the variable of interest) of soup was stored per warehouse. How many cans have to be opened to answer the question of what flavor soup was in all of the cans in the warehouse? The obvious answer is: one. (But if the warehouse held numerous different kinds of soup, a larger sample would have to be gathered.) The homogeneity of the variable often cannot be known until after the data are collected.

2. How narrowly the researcher needs to zero in on the answer. That is, how narrow the researcher wants the "confidence limits" to be around the statistical statements that can be made about the population. (Depending on the research question, sometimes plus or minus 10 or 20 percentage points will do well enough; other times it will not do at all.)

3. How confident the researcher needs to be that the obtained range around the population parameters is correct—logically, this depends upon the nature of the research questions, although typically researchers follow the convention of aiming for 99% or 95% confidence.

### § 2-3.4.2 Confidence Limits

Since samples are collected not for their own sake, but for the purpose of inferring back to the population from which the samples came, methods have been developed to allow researchers to perform that task of statistical inference. Suppose a regulator needed to determine the average amount of carbon monoxide being discharged per day per automobile in a given city. The researchers might sample cars and days. Suppose they found the mean amount in their sample to be "50 units." By knowing (a) the variation (heterogeneity) in their sample, (b) the sample size, and (c) the level of confidence the researcher wishes to have, the researchers can also calculate, for example, that the true population parameter falls within the range of, say, 50 units plus or minus 8 units, with 95% confidence. That is, there is a 95% probability that the true population mean falls somewhere between 42 and 58.

Further discussion about sample size and confidence limits can be found in Chapter 3—Statistical Analysis and Inference and Chapter 5—Legal Applications of Survey Research.

## § 2-4.0 RELATIONSHIPS AMONG VARIABLES

To this point we have discussed methodological issues related only to the question of measuring one variable at a time—from how many teeth are in a horse's mouth to incidence rates of a disease to consumer perceptions relevant to a possible trademark infringement to the amount of auto exhaust—without trying to relate one variable to another.

Quite often, however, people are interested in the relationship of variables to other variables: Which of several treatments is most likely to cure a disease? What variables predict who will do well in law school, astronaut training, etc.? What management techniques make workers more productive? What policies can cause the economy to grow? What programs are most likely

to reduce crime? Which educational methods are most effective? What methods of communicating ideas makes them most persuasive? What causes tornadoes? Life is filled with such questions, and researchers of all kinds are at work searching for the answers.

### § 2-4.1 Minimum Conditions for Inferring a Relationship

In order to draw any inference about a relationship among variables, one must have data from at least two levels of at least two variables. Anything less makes it impossible to say anything about a relationship. Figure 1a illustrates these minimum conditions. Suppose the question is whether scores on an intelligence test are related to performance in a particular job. Suppose a study were done on 1000 workers doing that job. Suppose that by some valid and reliable measures their job performance was evaluated and their intelligence was measured. The data in Figure 1a show that people tended to cluster in the two cells where (a) intelligence was high and performance was good and (b) intelligence was low and performance was poor. So we would infer from this pattern of evidence that higher intelligence was associated with better job performance. But the differences between these two cells and the remaining two are small, so the relationship is a small one: intelligence does not have a great deal to do with performance on this job.

Consider Figure 1b. Here the relationship is the opposite: high intelligence is associated with poor job performance. Moreover, the relationship is stronger than the previous one: on this job, high intelligence is an impediment to good performance.

Examine the data in Figure 1c. These reveal no relationship whatsoever between intelligence and job performance. Among workers of high intelligence, the ratio of those who perform well to those who perform poorly is 4:1. Among workers of low intelligence the ratio is exactly the same.

Note that any less data than these two measures on two variables would not permit us to draw any inferences at all about the relationship between the two variables of intelligence and work performance. Nevertheless, people often can be led to believe an assertion that lacks data to support it. This point can be understood by considering missing data patterns.

**Figure 1**  
**Minimum conditions for finding a relationship**

A

		Job Performance	
		Good	Poor
Intelligence	High	260	240
	Low	225	275

1000

B

		Job Performance	
		Good	Poor
Intelligence	High	225	275
	Low	300	200

1000

C

		Job Performance	
		Good	Poor
Intelligence	High	400	100
	Low	400	100

1000

### § 2-4.1.1 Single Cell Missing Data Pattern

Consider a situation where only one cell out of the minimum four cells is provided, and an inference is sought to be drawn from that evidence. Here is an example.

In a debate over the value of trying to form a strong trade relationship with one's enemies of long-standing, as a way of trying to reduce the chances that war would break out between them, one of the debaters offered as evidence against the proposition a list of eight pairs of nation-states that had been primary trade partners just prior to going to war against each other.

**Figure 2**  
**Single cell missing data pattern**

		Primary Trade Partners?	
		Yes	No
War	Yes	8	
	No		

During the debate, this had a strong impact against the proposition that trade relationships reduce the chances of war. But an examination of Figure 2 shows us that, because we do not know what data exist in the remaining three cells, we do not actually know anything yet about the relationship between trade and armed conflict. The relationship could be discovered only by collecting representative data, filling those remaining cells, and examining the pattern revealed.

In short, the debater offered a single-cell missing data pattern, and managed to convince most of the audience that it told them something important about the relationship at issue, when it did not tell them anything about the relationship.

### § 2-4.1.2 Single Row (or Column) Missing Data Pattern

Sometimes the data that are offered in support of some proposition consist of only one row or column from the basic 2x2 matrix that is the minimum necessary for drawing inferences of a relationship.

For example, occasionally people have suggested that marijuana, even if not harmful in itself, is dangerous because it leads (by some pharmacological or psychological or sociological route) to the use of harder drugs. They have offered the suggestion that if a substantial number of heroin addicts were found to have used marijuana when they were younger, that would confirm the hypothesis. Figure 3 depicts the pattern of data such commentators have in mind. The hypothetical data in the figure show 60% of a sample of 500 heroin addicts to have used marijuana at an earlier time.

Figure 3

Single row (or column) missing data pattern

Heroin Addict at Time-2	Smoke Marijuana at Time-1	
	Yes	No
Yes	300	200
No		

Because these data constitute a single-row missing data pattern, they cannot reveal whether or not a relationship exists. More specifically, without comparison data we cannot know whether fewer than 60%, about the same 60%, or more than 60% of people who are not heroin addicts earlier smoked marijuana. And it is on that comparison that the existence or non-existence of a relationship depends.

Suppose, for example, that someone had proposed that drinking milk as a child led to heroin addiction as an adult. The same table with marijuana replaced by milk would reveal that more than 99% of heroin addicts drank milk as children. Would that reveal that milk was to blame? Filling in the data for the rest of the table would make clear that 99% of non-heroin addicts drank milk as children, and therefore no relationship between milk drinking and heroin addiction existed. Until the rest of the data were supplied for Figure 3, one could not tell whether a relationship existed between marijuana smoking and heroin addiction, other than by speculating on what the missing cells contained.

§ 2-4.1.3 Main Diagonal Missing Data Pattern

The final pattern of missing data is where the data supplied fill only the cells along the diagonal, usually the main diagonal. As an illustration, suppose an asserted expert in the prediction of dangerousness is asked to report his track record, and his reply is that on 50 occasions he predicted that individuals would be dangerous and that they did do something harmful, while on 200 occasions he predicted that individuals would not be dangerous and that they did not do anything harmful. Figure 4 displays these data. By now the reader should readily see that until we know what occurred in the remaining cells, it is impossible to know from the limited data given whether the expert is highly accurate or highly inaccurate. If the expert made 2000 other predictions, which wound up in the two remaining cells, then we would know that these data portray an expert who was wrong on nearly 90% of his predictions.

**Figure 4**  
**Main Diagonal Missing Data Pattern**

Prediction of Dangerousness	Harmful Conduct	
	Yes	No
Yes	50	
No		200

### § 2-4.2 Predictive Relationships

When the minimum necessary data are available, as explained above, a predictive relationship, if one exists between the variables, can be discerned. Mere predictive relationships—also known as correlational relationships or findings that are the product of observational studies—tell us only whether one variable is associated with another, and how strong that association is. They do not tell us that changes in one of those variables *causes* changes in the other variable. Establishing causal relationships, a much more difficult task, is discussed in the sub-section following this one.

Here are some illustrations of predictive relationships. They are offered to illuminate the point that correlation does not prove causation, although often mere correlations are spoken of as if they established a causal relationship.

**Bugs, beards, and spurious correlations:** Researchers once found that the beards of some indigenous men of a third world venue tended to be inhabited by lice but not the beards of other men, and that the men with the lice tended to be healthier than those without the lice. Thus, a correlation was found such that lice and health went together. On these data alone, numerous explanations are possible: lice promote health, whatever these men are doing to remain healthy also promotes lice, whatever sickness has overcome the ill men also overcomes the lice, healthy men attract lice, and so on. Correlational relationships do not enable one to distinguish among a variety of different possibilities.

**Marriage and crime, and the direction of the causal arrow:** Observing that single men were more likely than married men to commit crimes, one commentator recommended that society try to marry off as many single men as possible so as to reduce crime and a whole array of other problems that these men presented to society. Single men also were more likely to have more illnesses, lower incomes, less education, and various other problems. But it is at least as likely that the causal arrow runs in the opposite direction. That is, men who are actively engaged in crime are more likely to suffer from health problems, poverty, and so on, and therefore are less likely to be found to be acceptable marriage partners and so have lower rates of marriage.

**Smoking during pregnancy and children's intelligence, and the "third variable" problem:** A study found that women who smoked during

pregnancy bore children with lower IQ's, on average, than women who did not smoke. This finding allows the *prediction* that the offspring of women who smoked during pregnancy would have lower intelligence. But it does not by itself permit a conclusion that smoking *caused* the lowered intelligence. A plausible alternative explanation is that mothers with lower IQs are more likely to smoke (which is true) and also are more likely to produce children with lower IQs (also true). Thus, it is a third variable (mother's IQ) which causes both of the other variables (smoking and children's IQ).

Here are two additional, and more obvious examples of the third variable problem:

A study finds that states with higher levels of pickle sales also have students who attain higher scores on tests of educational achievement. The researchers conclude that pickles improve school performance, and recommend that school cafeterias serve more pickles to students. A more likely explanation is that the economic situation of the state is responsible for both the level of pickle sales (in good economies more of everything gets sold, including pickles) and the performance of students (good economic conditions allow stronger tax bases, more expenditures on education, and better schools).

An observer from Mars who is studying the earth notes that cars tend to turn left after a light on the car's left side begins to blink, and they tend to turn right when a light on the right side begins to blink. Using the same faulty logic that earthlings often use, the martian concludes that the light *causes* the cars to turn (rather than the "third variable" of the driver, who causes both of those other events). A good predictive relationship has been found, but the observer has the causation wrong.

In none of these illustrations can causation be ruled out (until information is collected on the third variable), but mere correlation does not establish the causation.

### § 2-4.2.1 Statistical Aspects

Predictive relationships can be examined graphically with the help of "scatterplots" and "crosstabulation" tables (Figure 1 provides examples of crosstabulations). They can be measured using a variety of correlational statistics (the Pearson Product Moment Correlation and Spearman's Rank Correlation being two of the more common). Correlation coefficients vary between 0 and 1. The greater the coefficient the stronger the association between the two variables. Correlations also have signs accompanying them: Positive correlations mean that as one variable increases in magnitude, so does the other. Negative correlations mean that as one variable increases in magnitude, the other decreases. Slightly different ways of calculating correlations are used depending upon the nature of the data: whether the data in one or both variables are measured using nominal, ordinal, or interval scales. (See discussion of these, *supra*.)

Following are several illustrative correlations from various areas of research, which cover a wide span of the range of strengths of relationships:

Aspirin and heart attacks	.033
Psychotherapy effectiveness	.320
First year law school grades predicted from the LSAT	.410
Polygraph accuracy	.670
Civil jury awards predicted from medical specials	.714
Distance from hole and putting success for pro golfers	-.940

The nature of a predictive relationship between two variables can be described more fully by the use of Regression Analysis. This kind of analysis enables one to describe the relationship between the two variables with the formula for a straight line. Of particular interest is the slope of that line: the steeper the slope, the more change is “produced” in one variable by a change in the other variable. Having such a formula allows a direct and literal prediction of a score on one variable by knowing the score on the other variable. For example, how well is a student likely to do in the first year of law school predicted from the student’s LSAT score? Correlation and regression permit that prediction to be made with some accuracy.

Often, several predictor variables can be brought to bear on predicting some outcome variable. For example, how well can a law student’s first year grades be predicted by knowing the student’s undergraduate grade point average and age as well as LSAT score? The technique of multiple regression analysis permits several predictor variables to be combined in order to improve the accuracy of the prediction. One of the important additional statistics that accompanies a multiple regression analysis is the “multiple R squared,” symbolized as  $R^2$ . This is the square of the correlation between the scores *predicted* using the several predictor variables and the actual *observed* scores. It tells us the proportion of variance in the criterion (or predicted) variable that is predicted (or accounted for) by the predictor variables.

Some researchers have come to believe that “proportion of variance accounted for” gives an impression of a relationship that understates the magnitude of relationships. They have suggested converting correlations and multiple correlations into a “binomial effect-size display,” or “BESD,” which is more intuitively meaningful. Figure 5 illustrates such a display. Take a correlation of  $r = .32$  (which is the correlation between receiving or not receiving psychotherapy and showing at least substantial improvement in the patient’s condition based on a large number of studies). From the correlation alone, researchers would tend to call this a modest or moderate relationship. Squaring the  $r$ , to find the proportion of the variation in symptoms ameliorated by the psychotherapy, the relationship appears even more modest: 10% of the variance in symptoms is associated with the treatment. Figure 5 shows, however, that this correlation (.32) is equivalent to a change in cure rates from 34% to 66%.

**Figure 5**  
**Illustration of Binomial Effect Size Display**

$r = 0.320$

	Substantial Improvement	No Substantial Improvement	
Psychotherapy	66	34	100
No Psychotherapy	34	66	100
	100	100	

Some research, such as epidemiology, deals with relationships of more subtle magnitude. For example, look at the correlation between taking or not taking aspirin and having or not having a heart attack. Figure 6 shows the data on which this correlation is based. The vast majority of people in the study had no heart attacks whether they took aspirin or not. The correlation is .033 and the proportion of variance accounted for is .001 (or one-tenth of one percent). Even the BESD shows only a small change in survival rate (a change in heart attack occurrence from 52.6% to 47.4%). But in this area of research, one will find the relationship expressed neither in terms of correlations, variance accounted for, or BESD, but in terms of “relative risk,” or RR, that is, the likelihood that a person in the exposed group will suffer from the condition compared to a member of the unexposed group. In this example, the  $RR = .55$ , which is a protective relationship (exposure reduces the likelihood of the condition). Using this way of describing the data, one can say that aspirin cuts the risk of having a heart attack in half. That sounds quite different from saying that one tenth of 1% of the variance in heart attacks is associated with aspirin taking. Yet both are accurate statements of the relationship.

**Figure 6**  
**The Effect of Aspirin on Heart Attacks**

$RR = 0.5499$

	Heart Attack	No Heart Attack	Attacks per 1000
Aspirin	104	10,993	9.42
Placebo	189	10,845	17.13

More detailed discussion of the statistical aspects of predictive relationships can be found in Chapter 3—Statistical Proof. More detailed discussion of

the statistical aspects of epidemiological research can be found in Chapter 28—General Concepts of Epidemiology.

### § 2-4.3 Causal Relationships

Most of the time researchers are more interested in understanding causation than they are in discovering merely predictive relationships. (Though prediction is itself quite important, as testing for admissions in higher education, economic forecasting, and meteorology demonstrate.)

#### § 2-4.3.1 Squeezing Causation from Correlational Data

If the reader takes nothing else away from the preceding section, it should be that “correlation does not prove causation.” Nevertheless, the first step in trying to establish that a relationship is causal rather than merely correlational often is to try to extract causal inferences from correlational data.

The first temptation is to examine the correlation, and argue that particularly large correlations reflect underlying causation. This argument cannot succeed on its own. Many strong correlations have been found to be spurious, often due to “third variables.” (See the discussion of spurious correlations, *supra*.) Conversely, truly causal relationships come in various strengths and therefore can have correlations of various sizes. The logic of establishing causation requires more.

Numerous statistical techniques have been developed to try to help researchers draw better causal inferences from correlational data. These involve “controlling for” potentially confounding variables. These techniques include partial correlations, certain applications of multiple regression analysis, path analysis, cross-lagged panel correlation, and others. Discussing these is beyond the scope of this chapter. Suffice it to say that each of these techniques requires relatively sophisticated quantitative analysis aimed at statistically removing the effects of possible confounding variables. They also entail the hope that all of the important extraneous variables have been measured so that their effects can be statistically controlled (removed from having any influence) so that the real effects, if any, of the independent variables of interest can be seen. At the end of the day, causal inferences from correlational data must always be received with a healthy respect for the possibility that the statistical adjustments have been incorrect or the true causal variables have been omitted from the model and therefore could not be adjusted for at all. (For more discussion of these issues, see Chapter 3—Statistical Proof.)

#### § 2-4.3.2 Threats to Internal Validity

By far the simpler and more powerful solution to the problem of drawing causal inferences is to design a study in such a way that cause can be inferred more directly, without having to resort to a panoply of statistical fixes. In fact, research designs vary in the extent to which they allow unconfounded inferences to be drawn about what is causing the observed changes in the dependent variable. The logical structure of a research design is known as its

“internal validity.” Designs which minimize threats to internal validity allow the clearest inferences of causation. Designs with poor internal validity do not permit sound causal inferences to be drawn.

If designs high in internal validity exist, why don't researchers use them all of the time, and avoid the need for mathematical manipulations to try to clear up the confusion in their data? When circumstances permit, competent researchers are eager to use such designs. But for reasons of practicality or ethics, the best research designs may not be possible. The research designs that allow the strongest inferences are true experiments, which require that people or things be assigned at random to differing treatment conditions. (This and other designs will be discussed in more detail, in § 2-6.0, *infra*.) Astronomers, for example, simply have no ability to randomly assign different celestial bodies to experimental and control groups. They must take them as they find them. Researchers of the biological effects of toxic substances cannot instruct people to spend their lives exposed to certain substances and other people to remain unexposed. On the other hand, researchers in fields such as medicine, physics (though not astrophysics or geophysics), psychology, and agriculture, among others often can avail themselves of the most powerful research designs.

Where there is a threat to internal validity, plausible rival hypotheses exist which the research design is unable to rule out. The existence of plausible rival hypotheses that cannot be discounted means that inferences about causation are ambiguous. In order to understand the strengths and weaknesses of different research designs, in terms of their relative ability to permit unambiguous causal inferences, we must first acquire some appreciation for the various threats to internal validity, or confounds. To understand the following threats to internal validity, think of a simple study in which a group of patients is given a new treatment.

### **[1] History**

History is the threat to internal validity that refers to events to which the people or things are differentially subjected in addition to the independent variable. Thus, if patients who receive the treatment also are placed on a special diet and are given physical therapy, one cannot unambiguously infer what is affecting the outcomes for the patients, the treatment or the confounds (in this example, diet and physical therapy).

### **[2] Maturation**

Maturation refers to processes going on within the respondents, such as growing older, growing tired, growing hungry. Thus, for example, the body's natural healing processes are confounded with any treatments that are given (from which comes the old saying that a physician should make haste to treat a patient before the illness goes away on its own).

### **[3] Testing**

The second time a person takes a test the scores will be different from what they would otherwise have been, merely because of changes in the

person's experience of the test. Changes due to these effects should not be mistaken for changes from the first to the second testing due to the independent variable.

#### [4] Instrumentation

Instrumentation refers to changes that take place in the measuring instruments, including human observers, which can be mistaken for changes in the people or things being observed. For example, patients attended by nurses on one shift may be evaluated differently than by nurses on a different shift; the differences may be due not to changes in the patients but due to differences in the nurses. A whole area of social psychological research has developed that has illuminated these "subject-experimenter artifacts." For example, there are expectancy biases, in which observers, tend to see what they are given an expectation to see. In addition to affecting the perceptions of the observers, expectations change their behavior toward the persons or animals being studied and the person under study may himself change due to expectations about the effects of the treatment. A classic example of this problem (known as the "Hawthorne Effect") occurred in an industrial plant, where any and every change introduced by organizational effectiveness researchers raised the productivity and morale of the employees. What was happening was partly that the employees thrived on the attention and partly that they shared the expectations of the researchers that the changes being introduced would improve the circumstances and the work of the employees. Because of such expectancy effects many studies are conducted in a "blind" fashion, that is, where a person receiving an experimental treatment (for example, medications) is kept uninformed of whether what is being administered is the active substance or a placebo. Where circumstances permit, the studies are conducted in "double-blind" fashion: not only are the people being studied kept blind, but the researchers administering the treatment also are kept blind to which treatment they are giving and, when possible, blind as well to the hypotheses being tested.

#### [5] Statistical Regression

Statistical regression occurs when objects of study, including people, are selected because of their extreme scores on some measure. With no treatment at all, they tend to "regress" toward the mean of the distribution from which they came. Therefore, observed changes due to that regression effect can be mistaken for changes due to the treatment.

#### [6] Selection

Selection confounds occur where some of the people or objects placed into experimental and control conditions differ initially in some way. For example, suppose relatively healthier people are chosen for an experimental medical treatment (because it is thought that they can better withstand the rigors of the treatment) and less healthy people with the same illness are used as a comparison group. Distinguishing the effects of the treatment from the initial differences in health will be difficult or impossible.

**[7] Experimental Mortality**

Experimental mortality refers to the differential loss of participants from different experimental conditions. It introduces the same problem later in a study that selection artifacts introduce at the outset of a study.

**[8] Selection-maturation Interaction, Etc.**

Some of the confounds defined above can work together and thereby further obscure the results. A likely one is the interaction of selection and maturation: the people selected to be in different conditions of a study differ between experimental and control conditions in that those in one group are “maturing” at a different rate or to a different degree than those in the other group. For examples: patients of different age or condition are likely to heal at different rates, children of different ages will learn at different rates, nations at different stages of economic development can react to new challenges with different degrees of success.

**[9] Chance**

Chance is an artifact that results from random fluctuations in sampling and measurement. Two groups can be identical in every respect, and yet at the end of the study their measured scores still will not be identical. A difference due to chance might be mistaken for a difference due to the independent variable. While all of the artifacts defined above can be protected against by proper research design, chance is the one that cannot be. It is to deal with the artifact of chance that significance testing was invented. For a very brief summary of statistical significance testing, see § 2-6.4 on statistical aspects of research designs, *infra*. For more detailed discussion, see Chapter 3—Statistical Proof.

**[10] Distortions in Independent and Dependent Variables**

In thinking about things that interfere with the capacity of research to permit valid inferences to be drawn about the possible effects of an independent variable on a dependent variable, problems with the independent and dependent variables themselves should not be overlooked. (The nine subsections immediately above define types of confounds, which are factors *other than an independent* variable that may be causing apparent changes in a dependent variable.)

Faulty operational definitions: Recall from our earlier discussion of operational definitions, in § 2-2.1, *supra*, that if a variable is not meaningfully operationalized, then the findings it produces would not answer the question a consumer of that research may have thought was being asked.

Failure to induce the experimental manipulation: Even a well operationalized variable may not reach the respondents or may not be presented to them or done to them or, having reached them, may not be perceived in the intended way. For examples: In a study of the effects of toxic substances, a group presumed to have been exposed may not in fact have been. A prisoner rehabilitation program may not actually have been carried out. A study might seek to examine how people react to “rich” corporate defendants, while the

people in the study do not perceive the defendants as being any wealthier than average. In each of these instances, it would not be accurate to conclude that the independent variable had no effect if that independent variable was, in reality, not tested. Researchers concerned about these matters often perform "manipulation checks," that is, they test whether the independent variable was induced as intended.

**Floor and ceiling effects:** When a manipulation is so extreme in its strength or weakness that there is no variation possible between groups—that is, all people in all groups respond to it the same way—then it is impossible to test the effects of some other variable on it. For example, any study of the subtle effects on verdicts of innumerable variables that affect judge or jury decision-making will be incapable of detecting differences if the basic case facts are so extreme that everyone viewing the case reaches the same verdict. What good researchers try to do is to pre-test their procedures to make sure that floor or ceiling effects do not occur and that adequate variability exists in the dependent variable.

### § 2-5.0 THREATS TO EXTERNAL VALIDITY

In the preceding section we discussed internal validity: the logic of the structure of a study that permits or impedes drawing causal inferences about the effect of independent variables on dependent variables. Once we are satisfied that a study is internally valid, the next issue is whether it is externally valid, that is, whether it can be generalized beyond itself. This order of appraising a study is not arbitrary. Without internal validity there is nothing to generalize.

External validity refers to the representativeness of a study. If a study is externally valid, its findings can be generalized to other populations (of people, objects, organizations, times, places, etc.). Usually one does research at a specific time and place on a particular population, but hopes to be able to generalize the findings beyond the immediate people and circumstances of the study.

For examples: Can the findings of a study done in Minnesota be generalized to people in Arizona? Does a study of the efficacy of information presented in written form generalize to the same information presented in verbal, video, or computerized form? Does a study of manufacturing organizations have application to financial services organizations? If people are exposed to a battery of treatments, will any one of them work on people who have not been exposed to the others? Can we generalize to humans the findings of a study of the effects of a drug tested on laboratory rats? Or tested on monkeys? Or from adults to children? Or males to females?

Particular kinds of threats to external validity have been defined more precisely. By way of illustration, we will examine more closely only one of them: Reactive effects of testing are the change-inducing effects of taking a test on the person being tested. For example, suppose a firm wants to test the effects of an advertising campaign on people's attitudes. If it begins by testing public attitudes and plans to call the respondents back after the advertisements have been airing for a period of time, the pre-testing itself

may cause respondents to pay more attention to the ads than other members of the public, who had not been pre-tested. Thus, the findings can be generalized only to people who had been interviewed about the topic before the ads were aired.

While the logic of research design is a great help in evaluating the internal validity of research, concerns about external validity cannot be dealt with so easily. With respect to any of the questions posed two paragraphs above, readers could exercise their intuition: how similar do we feel that different settings or types or organisms are (with respect to the independent and dependent variables of interest)? But these are merely guesses. At the end of the day the only way to know with any rigor whether an effect observed in Minnesota will hold in Arizona, or whether an effect observed in rats will hold for humans, is replication. This is one reason that researchers rarely place much faith in any single study, or even any single type of study. For reasons of generality, among other reasons, they prefer to see findings replicated in other places, using other participants, under various other conditions. The more different circumstances a phenomenon can be replicated in, the greater its generality, and the more confidence researchers as well as consumers of research should have in the phenomenon.

## § 2-6.0 RESEARCH DESIGNS

Various research designs are more or less vulnerable to threats to internal and external validity. The following discussion is not an exhaustive review of research designs, but provides the reader with the ability to appreciate the power of some research designs to provide answers and the weakness of others.

Our discussion will proceed by way of an extended example. Let us take as our research project a study of the question whether Vitamin C cures colds or not. Following are a variety of approaches a researcher could take in attacking this empirical question. Through these we will be able to see the strengths and weaknesses of different research designs. The emphasis of our discussion will be on threats to internal validity.

### § 2-6.1 Pre-experimental Designs

First, suppose the researcher employs a *case study*. A person who is observed to have a cold is given Vitamin C, and soon thereafter the cold goes away. Is this convincing evidence of the curative powers of Vitamin C? Let's think about the confounding variables that may be operating in this study (see § 2-4.3.2). Some other factor may have cured the cold, such as the chicken soup or bed rest the cold sufferer also was taking (history). The patient's immune system may have cured the cold rather than the Vitamin C (maturation). To these let us add the lack of generality produced by a study with a sample size of  $n = 1$ .

Suppose, then, that the researcher rounds up 100 people to conduct essentially the same study. This is a *one-group pre-test post-test design*. At the outset they all have colds, they are given vitamin C, and a week later, of those who could be examined, only 40% still have their colds. Is the decline from

100% ill to 40% attributable to the vitamin C? The confounds mentioned above have not gone away. In addition: Perhaps those rating the cold symptoms had become so insensitive to them from time<sub>1</sub> to time<sub>2</sub> that it took worse symptoms to register as still having a cold (instrumentation). Perhaps those whose colds turned to pneumonia and were now in the hospital or quit the study in frustration were the ones who could not be found, thus exaggerating the observed cure rate (mortality). As a result of the confounds in this design, this study also leaves us less than persuaded.

Our persistent researcher next tries a *static group comparison*. In this study people are found who either already take vitamin C regularly or who do not, and their colds are monitored over time. Suppose a relationship is found such that those who take vitamin C get fewer colds and the colds they get are milder and shorter. Can we now conclude that vitamin C cures (and prevents) colds? The reader may recognize this as a fairly primitive correlational (or observational) study (discussed earlier, in § 2-4.2). It still is vulnerable to a variety of threats to its internal validity. Those who take vitamin C regularly may systematically differ from those who do not in other things they do: nutrition, exercise, rest (history confounds). Those who regularly take vitamin C may, either constitutionally or because of their other health habits, be basically healthier to begin with than their counterparts (selection), or have more active immune systems that kill cold viruses more effectively (interaction of selection and maturation). Note that instrumentation is, in this design, not likely to be a problem, because both groups are being examined at approximately the same times, so any such problems would be equal for both groups.

### § 2-6.2 True Experiments

Let us now contrast the above designs to a true experiment. It is worth taking a moment to define the term "experiment." In common parlance an experiment simply means testing something out, trying something out. To scientists it refers to a specific type of research design through which such testing is done. It means a testing situation in which a basis for comparison is provided such that conditions are identical between two or more groups (such as an experimental group and a control group) except for the feature that is the focus of the study.

For social, behavioral, and biological scientists, accomplishing this requires *randomly assigning* the participants of the study to experimental and control conditions so as to maximize the probability that the two groups do not differ. True experiments are known by several synonyms. They may be called, simply, experiments. In medical research they often are termed "randomized controlled trials," or "controlled trials," or "clinical trials." Whatever the name, the meaning is that people are randomly assigned to be exposed to different conditions or treatments, so that confounds are eliminated and the groups do not differ except with respect to the independent variable. Cause and effect inferences will be unambiguous.

Let us see how this works with the vitamin C inquiry. Imagine that 200 cold sufferers are randomly assigned to two groups. One group receives

vitamin C, the other receives nothing, or a placebo. (Ideally, the vitamin C and placebo are given in double blind fashion, so that neither the person administering nor the person receiving the drug can know which is the active drug and which the placebo. Identifying numbers will be decoded later by the researchers.) A week later the people are examined and it is found, let us say, that 50% of the control group still have their colds, but only 40% of the vitamin C group does. On the face of things, that difference of 10 percentage points suggests that the vitamin C did more good than the placebo.

Moreover, because of the use of a true experimental design, we will see that confounds have been eliminated. History: Because of random creation of the two groups, the proportion of those eating chicken soup or staying home in bed is highly likely to be equal between the two groups. Maturation: The immune system response will be the same, on average, in the two groups. We can see that although even the control group lost half of its colds due to spontaneous remission (something that, in the one-group pre-test post-test design could not be distinguished from an alleged vitamin C effect) the difference of 10 percentage points represents the effect of vitamin C over and above maturation. Instrumentation: If patients were sent for examinations at random before and after administration of the independent variable, any changes in the perceptions of the examiners will be evenly distributed between the experimental and control groups and could not have caused an artifactual difference. Selection: Had young adults been placed in the vitamin C group and older people in the control group, or if people had been allowed to select themselves into the vitamin C or control groups, then any observed difference might have been due to those differences between the people in the two groups. But random assignment of people to groups maximizes the likelihood that, on any dimension we could name, we would find the same proportion of such people in one group as the other, and those differences would not have been able to exert an effect on one group and not the other. Mortality: Similarly, people dropping out of the two groups would be expected to occur at the same rates and not differentially. This would be true with respect to any characteristic we could name.

If we had doubts about any of these things, we could check them and see if they were present differentially in one group compared to the other.

### § 2-6.3 Quasi-experiments

Quasi-experiments are research designs that are not as clean and straightforward as true experiments, but which provide enough control, or some randomization, or can be supplemented or corrected so as to eliminate many, and occasionally all, threats to validity. In short, partly by virtue of the design, partly by collecting additional data to address shortcomings, a quasi-experimental design can be made to approximate an experiment, and allow cause-effect inferences that may approach the clarity of experiments.

For example, in an *interrupted time series* design, data are collected on the dependent variable (e.g., highway death rates) over a period of time (e.g., five years), then the independent variable is induced (e.g., the introduction of seat belts) and then data are collected for a period of time following the

introduction of the innovation (e.g., another five years). Suppose it is found that after the introduction of seat belts the accident death rate fell. The only confound that is not protected against in this design is history. Perhaps other changes occurred at the same time as the treatment. Suppose it is suspected that at the same time as seat belts were introduced: the price of gasoline rose so people drove less, speed limits were lowered and enforced better, the weather was less icy, and so on. What the researcher would do would be to collect data on those things, to see if any of the suspected history confounds really occurred. If they did not, the inference that it was the seat belts that lowered the death rate is strengthened. If they did, statistical adjustments might be possible to see if, after the effects of the confounds are statistically removed, a seat belt effect can still be discerned.

Sophisticated correlational (observational) studies, which include enough data to test or control statistically for possible confounds, can be thought of as members of the quasi-experimental class of research designs. Crude correlational studies are better regarded as members of the pre-experimental class.

Case-control studies, found in epidemiological research, may be thought of as quasi-experimental designs. For each case in the group exposed to the substance of interest, a comparison case is chosen which is similar on many characteristics, except that the control case has not been exposed to the substance. It should be obvious to the reader that this is an attempt to reduce the effects of possible confounding variables, and to try to approximate the power of a true experiment.

Research designs come in great variety. Both the researcher and the consumer of research knowledge have to think carefully and clearly about the strengths and weaknesses of the research design, and whether the inferences sought to be drawn from the study are possible in light of the design. When the findings of a number of studies are combined, either by old fashioned reading of the studies or by a formal, quantitative, meta-analysis, the various findings can be given more weight or less weight, depending upon the quality of the research design through which the findings were obtained.

### § 2-6.4 Statistical Aspects

As noted earlier, the confound of *chance* cannot be ruled out, even by the most ideal experimental design. In the example above, is the difference between 40% and 50% in the number of patients in the experimental and control groups who still have colds a real difference or is it the product of random fluctuation? This question is answered by statistical hypothesis testing. The data permit the researcher to calculate the probability that, were the starting assumption of no difference between the experimental and control groups (known as the null hypothesis) to be rejected, what is the likelihood that the researcher would be making a Type I error (erroneously rejecting a true null hypothesis). Conventionally, unless this probability falls below .05 (fewer than five chances in 100 of committing a Type I error), the researcher refrains from rejecting the null hypothesis. For more detailed discussion of inferential statistics and hypothesis testing, see Chapter 3—Statistical Proof.

**§ 2-7.0 CONCLUSION**

Science is neither mechanical nor magical. It is a process of drawing inferences from evidence. The evidence for those inferences is generated by research which necessarily employs a selection of research methods. A finding is only as good as the methods used to find it. There is no one best way to study a phenomenon of interest. Each methodological choice involves trade-offs. The issue, always, is whether the methodology of the research is appropriate for the questions posed by the study, and whether the conclusions drawn are justifiable in light of the data collected and everything about the methods by which those data were generated. The choices of methods require careful thought, both by researchers and consumers of the research. The purpose of this chapter has been to arm legal consumers of scientific research with concepts that will facilitate that critical and thoughtful appraisal.