THE LOGIC OF EVALUATION

W. EDWARDS DEMING

WHAT IS EVALUATION?

The point of view here will be that evaluation is a pronouncement concerning the effectiveness of some treatment or plan that has been tried or put into effect. The purpose of this chapter will be to explain some of the problems in the design and interpretation of a study whose aim is to evaluate the effectiveness of some treatment or plan; also to point out some of the difficulties of studying by retrospect the cause of success or failure, or the cause of a disease or of a specific alleged cure therefor. Emphasis will be placed on ways to improve the reliability of evaluation by understanding and avoiding possible misuses of statistical techniques in evaluation.

It is fascinating to look around us and to observe how often people apply some treatment in the hope of producing a desired effect, then claim success if events turn in their favor, but suppress the whole affair if they do not.

A governor put 200 additional policemen on the highways to decrease the rate of accidents (he hoped). Serious accidents dropped from 74 to 63 the month following his action. Was this decrease attributable to the policemen, as he claimed? The answer seems at first to be so obvious: yes, of course. But wait. If every accident be independent of every other accident, then the student of statistical theory would recognize the number of accidents in a given period of time as a Poisson variate. He would then accept the square root of the number of accidents as a random variable distributed normally with variance ¼. The difference between the square roots of the number of accidents in two months would be distributed normally about O with variance ¼ + ¼. On this basis, one would calculate

$$t = \frac{\sqrt{74} - \sqrt{63}}{\sqrt{\frac{1}{4} + \frac{1}{4}}} = .94$$

for the t-value of the observed difference.

Without any calculation at all, one could only say that (1) any two months will be different; and (2) the decrease in accidents was consistent with the hypothesis that the governor's efforts had some effect. What does the above calculation add to our knowledge? It tells us that it would be rash to conclude that the data establish the hypothesis, for the small value of t admits a competing hypothesis, namely, that the observed difference was simply a random fluctuation, the kind of difference that would turn up in scoop after scoop of black and white beans drawn from a bushel of black and white beans mixed and remixed between scoops. Lack of independence between accidents, such as icy roads that persist over several days, would only decrease t, and would weaken further any argument that the governor's efforts were successful. We therefore see no statistical evidence from the figures given that the governor's efforts had any effect. Maybe they did. We shall never know.

Examples that show results that went in the wrong direction are hard to find: they get buried, not published. No one is around to take the negative credit for a failure.

A mother tries to persuade a child, by precept, example, or punishment, to cease and desist from some practice or habit. How effective is she? A young man saves money and gives up his job for a year in order that he may go to school. He applies education to himself, in the hope of improving in the future his economic and social status in life. He may eventually evaluate his decision: he may be satisfied that he did the right thing, or he may decide otherwise. By what criteria should he evaluate his decision?

Do fluorides in the drinking water retard greatly the decay of teeth? Does smoking cause cancer? Is marijuana really harmful? How effective is Head Start? In what way? Do seat belts save lives? How effective are loss leaders in a grocery store? What can go wrong in a test market?

Did the Federal Reserve Board make some right moves in the depression of 1969-1972? Will Variety A of wheat, sown in some specified area next year, show a yield at least 5 more bushels per acre than Variety B? Is EXTHRX effective as an antidepressant? For what kind of patients? What are some of the side effects, and how long before they appear? Does a certain plan of parole and education achieve the goals claimed in advance?

How effective are incentives for reenlistment in the Navy? What is the loss to a grocer who runs out of stock Saturday noon of a popular item? What is the cost of a defective item that goes out from a manufacturer to a consumer?

A prototype of some assembly or machine (e.g., an airplane) is put together for test. Will tests of the prototype predict the performance of machines that will later come out of regular production? Why not?

May one estimate from the results of an accelerated test establish the length of life of a lamp, or of a vacuum tube, or of a vacuum cleaner, or the mean time to failure of a complex apparatus? Why not? Or if so, how?

A flash of lightning brightens the landscape. A clap of thunder hits our ears a few seconds later. We never raise a question about the cause of the thunder; we agree that the lightning caused it, and we do not try to convince anyone that the thunder caused the lightning. Innoculation for smallpox is effective. Cholera in London came from drinking water that came from wells. Certain treatments and drugs for tuberculosis are effective. Most of these statements, well accepted now, were learned without benefit of statistical design.

Social programs and wide-scale tests of treatments are unfortunately laid out almost always so that statistical evaluation of their effectiveness cannot be evaluated. Government regulations on safety of mechanical and electrical devices are meaningless. And what about the side effects from noxious by-products of catalytic converters?

No one can calculate by statistical theory in advance, or even afterward, the effect of changes in interest rates, the impact of a merger, or of a step taken by the Federal Reserve Board. A statistically designed test is impossible, though accidental comparisons may of course turn up.

A firm advertises in magazines and newspapers and other media, or by direct mail, to increase sales. Adequate design of the experiment is usually difficult, and not even attempted. As a result, the effectiveness of the campaign is still in doubt after the experiment, just as it was before. An increase in sales could be the result of the campaign, but there are usually half a dozen competing hypotheses such as the effects of nonresponse or of other failures in cooperation of respondents, errors in response, changes in economic conditions, impact of competition, new products, new models, any one of which could explain what was observed.

The advantages of evaluation with the help of a statistical designed experiment, when such a thing is possible, are better grounds for understanding the results, speed, and economy. But we have to learn to use statistical inferences that are conditional, relating only to special conditions.

When men arrive at a consensus on cause and effect, they have solved, temporarily at least, a problem in evaluation. Textbooks in statistics and in the social sciences are replete with methods and examples of evaluation (not necessarily called by this name), without warning of pitfalls. The most important lesson we can learn about statistical methods in evaluation is that circumstances where one may depend wholly on statistical inference are rare.

NEED FOR CARE IN DEFINITIONS OF TERMS

There has never in man's history been an era of greater effort toward safe drugs, safe automobiles, safe apparatus, safety on the job, decrease in pollution, war on poverty, aids to underprivileged children, and all sorts of well-meant social programs. The problems of evaluation of these efforts are compounded by failure to define terms operationally, as well as by failure to lay down criteria by which to weigh gains and advantages against losses and disadvantages. A drug that helps thousands may be harmful to a few people. Is it safe?

Any adjective that is to be used in evaluation requires an operational definition, which can be stated only in statistical terms. Unemployed, improved, good, acceptable, safe, round, reliable, accurate, dangerous, polluted, flammable, on-time performance (as of an airline or train) have no meaning except in terms of a stated statistical degree of uniformity and reproducibility of a test method or criterion.

There is no such thing as the true value of anything.

The label on a blanket reads "50% wool." What does this mean? Half wool, on the average, over a month's production? Or does it relate somehow to this blanket that we purchased? By weight? If so, at what humidity? The bottom half of the blanket is wool and the top half is something else? Is the blanket 50% wool? Does 50% wool mean that there must be some wool in any random cross-section the size of a half dollar? If so, how many cuts shall be tested? How must they be selected? What criterion must the average satisfy? And how much variation between cuts is permissible? Obviously, the meaning of 50% wool requires statistical criteria. Words will not suffice.

FOUR REQUIREMENTS FOR AN EFFECTIVE SYSTEM OF EVALUATION

The four requirements for an effective system of evaluation are:

1. A meaningful operational measure of success or of failure, satisfactory to experts in the subject matter, of some proposed treatment applied to specified material,[1] under specified conditions. (Examples: a medical criterion of recovery or improvement in some affliction: a criterion for recognition of a definite and notable increase in production of wheat or of rice; a criterion for recognition of a definite and notable improvement of quality of a textile or of a carburetor; a criterion for improvement in quality of transmission of signals; a criterion for recognition of a definite and notable increase in the speed of learning a language.)

2. Some satisfactory design of experiments, tests, surveys, or examination of data already recorded. The design of a new study will include selection of samples of the specified material; a record, for the duration of each phase of the study, of certain specified environmental conditions that appear to be important; procedures for carrying out the investigation; and statistical controls to aid supervision of the investigation.

3. Methods for presentation and interpretation of the results of the experiments, tests, survey, or other investigation, that will not lead to action different from the action that would be taken on the basis of the original data.[2] The data must include a record of the environmental conditions, including test method, questionnaire, perhaps the names of the observers. They must include a description of the frame.

4. Some official or some group of people authorized to take action (with or without evidence).

LIMITATIONS OF STATISTICAL INFERENCE

A statistical study, prospective or retrospective, proceeds by investigation of some or all of the material in a frame.[3] A complete investigation is called a census. The frame is an aggregate of tangible units of material of some kind, any or all of which may be selected and investigated. The frame may be lists of people, dwelling units, schoolchildren, areas, blocks and plots in agricultural trials, business establishments, materials, manufactured parts, or other units that would supposedly yield useful results if the whole frame were investigated.

A point often forgotten is that the results of statistical inference refer only to the material in the frame that was studied, the instrument of test, and the method of using it, and to the ranges of economic and physical conditions and stresses within which there was randomization. Statistical inference ends with the frame and the environmental conditions under which the frame was studied. The theory of probability cannot help us outside these limits.

All probabilities are conditional and all statistical inference likewise, being conditional on the frame and the environmental conditions of the experiment. Any probability calculated from an experiment, if it has any use at all, is a prediction that future experiments on samples of material drawn by random numbers from the same frame, tested in the same way, and under the same environmental conditions, would show about the same results within calculable limits. Unfortunately, in an analytic study (next section), where the aim is to provide a basis for action on a process (if we get any good at all out of the experiment), the environmental conditions will be different from those that governed the experiment. It follows that any estimate or other evaluation based on an experiment can be used in an analytic study only on the authority of an expert in the subject matter who is willing to offer a judgment on whether the results are applicable to other conditions.

A good question to ask in the early stages of preparation of a study is this: What will the results refer to? How do you propose to use them?

ENUMERATIVE STUDIES CONTRASTED WITH ANALYTIC STUDIES

Effective use of statistical methods requires careful distinction between enumerative studies and analytic studies, with continual recognition of the limitations of statistical inference. The aim of any statistical study is to provide a basis for action. There are two broad types of action:

Enumerative—Action on the frame.
Analytic—Action on the cause-system (process) that produced the frame and will produce more frames in the future.

The methods of statistical design and of statistical inference are different for the two types of action. Failure to make the distinction between them has led to uninspired teaching of statistical methods and to misguided inferences.[4]

In an enumerative study, action will be taken on the frame and will depend purely on the estimate of the number or proportion of the people or materials in the frame that have certain characteristics (sometimes on the maximum or minimum). The action does not depend on how or why man or nature produced the frame. Examples:

1. We may need to know how many children by age there are in a certain region whose diet is below a minimum tolerable level (perhaps in calories, perhaps in vitamin or protein content). The reason to make the count is to know how much food to supply and what kind.
2. A quick count of the number of people left without homes and without food by a flood or earthquake. A vital question is how many people, adults, infants, and infirm are in need of the necessities of life.

3. The census of the U.S. for congressional apportionment, district by district.
4. A census of a city taken as a basis for an increase in financial support from the state.
5. We may need to know the total debits and credits in dollars on the books of some railway for services that they performed jointly with other railways during the past year. The frame could be, for example, 3 million interline abstracts in the files of this railway.
6. An inventory of certain materials is to be taken to assess the total value of an inventory. This inventory may determine the selling price of the material, or it may find its way into the auditor's annual report, or it may be used for tax purposes.
7. Cores bored from bales of wool selected by random numbers from a shipload of wool as it is unloaded, and analyzed by a chemist for clean content, determine the price and the duty to be paid on the whole shipload.
8. A telephone company may make a field inspection of the equipment it owns to determine the present worth of this equipment as a basis for rates for service.

In an analytic study, the aim is to try to learn something about the cause-system (process) to be in a position to change it or to leave it alone, whichever appears to be better for the future benefit of man or of his pocketbook. The frame studied (material or people) in an analytic problem is not of interest in itself. A complete census or study of the entire frame (all the people in an area, or all of last week's product) is still only a sample of what the cause system can produce, and did.

There is no finite multiplier of the form 1/n - 1/N in an estimate of variance in an analytic study. This same multiplier is of course very important in an enumerative study, as it reduces the sampling variation to zero for a complete census, that is, when n = N.

Some studies serve both enumerative and analytic uses. The census of any country, aside from enumerative uses (number of representatives or number of councilmen for an area, allocation of water, electricity, teachers) furnishes information by which economists, sociologists, and agricultural experts construct and test theories of migration, fertility, growth of the population, aging of the population, consumption of food, the aim being to understand better the changes in fertility and longevity that take place in the distribution of the population by sex, age, education, income, employment, occupation, industry, and urbanization. One aim among other aims might be to alter the causes of poverty and malnutrition.

A study of accounts receivable, primarily for an enumerative purpose, namely, this year's financial statement, may also yield information that is helpful in reducing errors of certain types in the future.

TWO POSSIBLE MISTAKES IN AN ENUMERATIVE STUDY

One may make either one of two types of error in taking action on the basis of an enumerative study. To take a concrete example, we are about to purchase a load of ore. The price to pay will depend on the results of assay of samples of the ore. We may, as a result of the sampling and assay:

1. Pay more by an amount D than the ore is worth;

or

2. Sell it for less by an amount D′ than it is worth.

We must pause at these words. We talk as if it were possible to find out what the load of ore is worth. We can proceed only if we are willing to accept some method as a master standard. Thus, we might agree that the master standard shall be the result of assays that follow a specified procedure on a large number of samples of the ore, more than we think are necessary for our purchase to be made presently. In practice, we take enough samples to provide a useful estimate of the master standard.

Statistical theory enables us to minimize the net economic loss in such problems from too much testing and from not enough testing.[5]

Techniques that are useful in enumerative studies are theory of sampling, including, of course, theory for optimum allocation of effort, losses in precision in estimates for the whole of a frame when differential sampling fractions are specified in order to get separate estimates for a particular stratum. Confidence intervals and fiducial intervals are useful in inference. Controls by appropriate statistical techniques of the instruments and of the methods of using them, and control of field-work, are essential for reliability and economy, and to understand the results. Calculation of the risk of being wrong in an inference from a statistically designed study in an enumerative problem is in the nature of a mathematical consequence.[6]

Unfortunately, as we shall see, no such beauty of theory exists in an analytic study.

TWO POSSIBLE MISTAKES IN AN ANALYTIC STUDY

There are also two types of mistake in taking action in an analytic study. These mistakes are totally different in nature from the mistakes of using an enumerative study. In an analytic problem:

1. We may adopt Treatment B in preference to A based partly or wholly on a statistical study, only to regret later our action to adopt it;

or

2. We may fail to adopt B, retain A, only to regret later our failure to adopt B.

One may make either mistake, with or without the help of an experiment, and it requires no high degree of education to make them. It is easy to bet on the wrong horse, to use an ineffective method of advertising, to purchase and install a machine that turns out later on to be a mistake, to plant the variety of wheat with the lesser outturn, misjudge a drug, approve social legislation that turns out to backfire, and so on.

The aim in the use of statistical theory should be to develop rules that will minimize in the long run the net loss from both mistakes. How to use statistical inference in analytic problems has received, so far, scant treatment in the statistical literature.

We shall not pause here for an example, as one will appear later. Suffice it to say here that, in contrast with the possible errors of using an enumerative study, we cannot, in an analytic study, calculate or govern by statistical methods the risks of making either error. The reason is that our action will be tested on future material, not yet produced, and we know not in advance what these future conditions may be. Even if we knew, we do not know except by substantive knowledge how they would affect the cause-system (treatment) of the future.

The watchmaker works on your watch and claims after a few weeks that it keeps perfect time. You wear it under other conditions–other temperatures, movements, irregular winding–and it loses time or becomes erratic. The watchmaker evaluates himself on the performance of your watch on the job, not by the record in his shop. This is why he tells you to bring the watch back after a few weeks so that he may adjust it if necessary.

The season, date, climate, rainfall, levels, dosage, length of treatment, age, ranges of concentration, pressure, temperature, speed, or voltage, or other stresses that may affect the performance of the process will be different in the future. Two varieties of wheat tested at Rothamsted may show that Variety B delivers under certain conditions much greater yield than Variety A. But does this result tell you which variety would do better on your farm in Illinois? Can you evaluate from the experiment at Rothamsted the probability of going wrong in adopting Variety B in Illinois? No. Tests of varieties of wheat lead to valid statistical inference only for the climate, rainfall, and soil that the study was conducted on. We shall never meet these conditions again. Yet the results, carefully presented, may be useful in the hands of the expert in the subject matter.

We must face the fact that it is impossible to calculate from the data of an experiment the risk of making the wrong choice. The difficulty is that there is no statistical theory that will predict from data of the past what will happen under economic or physical conditions outside the range of the study. We can only be sure that conditions outside this range will be encountered. There is thus no such thing as the power of a statistical test. (These assertions conflict sharply with books and teaching on tests of hypotheses, to which I will return later with a comment.)

Generalization to people from results of medical tests on rats is a perennial problem. Statistical theory can only tell us about rats. Generalization to people is the responsibility of the expert in the subject matter (chemistry, or various specialisms in medical science).

The aim of evaluation is to provide a basis for action in the future, with the aim to improve the product, or to help people to live better, whatever be the definition of better. Evaluation is a study of causes. Evaluation is thus analytic, not enumerative.

USE OF JUDGMENT-SAMPLES

It is hazardous to try to estimate or generalize from a judgment-sample to a portion or all of the frame whence the sample was selected. Use of a judgment-sample instead of a percentage or total of the frame for this purpose is worth no more than the reputation of the man that signs it. The reason is that there is no way except by judgment to set limits on the margin of uncertainty of the estimate.

Nevertheless, judgment-samples serve at times a very useful purpose by throwing light on a comparison of treatments. In spite of the fact that we are permitted to carry out a comparison of treatments only on patients who are highly abnormal (usually patients who do not need either treatment, or which neither treatment can help), or at a selected location such as Rothamsted, it is comforting to note that if the two treatments appropriately randomized and tested under these special conditions turn out to show results different by as much as D, then we have learned something: we may assert that the two treatments are materially different in some way—chemically, socially, psychologically, genetically, or otherwise. This we may assert even though we may never again use the treatments with patients like the ones tested, nor raise wheat under the same environmental conditions. The establishment of a difference of economic or scientific importance under any conditions may constitute important new knowledge.

Such a result, however, does not permit generalization: we cannot assert by statistical inference that other patients, nor other pupils, nor two varieties of wheat raised in some other location would show similar differences. Further experimentation would be required.

Randomization within a judgment-sample of plots within blocks (for trials of wheat), or of patients (for comparison of treatments) removes an important area of doubt and justifies the use of probability for conditional inferences. To understand the power of randomization within a judgment-sample of plots, one need only reflect on the contributions to our knowledge and economy that have emanated from the Rothamsted Experimental Station.

One could even go so far as to say that all analytic studies are carried out on judgment-samples of materials and environmental conditions, because application of the results will be to conditions beyond the boundaries of the experiment. This is why substantive judgment is so important in an analytic study.

We may often minimize the doubts about a series of experiments by choosing conditions for the study that will approximate (in the judgment of substantive experts) the conditions to be met in the future. Or, there may be a chance to run tests over a wide range of conditions. Thus, for tests of a variety of wheat, we might be able to run comparative experiments under different conditions of rainfall, irrigation, soil, climate, and length of growing season. One might, by substantive judgment, not by statistical theory, feel safe in planting or in not planting one of the varieties under test. In other words, one might, by substantive judgment, in fortunate circumstances, claim that the risk of the error of type 1 in a given analytic study is very small.

Thus, the law in physics that F = ma requires no qualification. A student in physics learns it once for all time. Originated by Sir Isaac Newton in London, it appears to hold in Liverpool, Tokyo, Chicago.

The advantage brought into a state of statistical control—stable in the Shewhart sense[7]—is that we may use statistical theory to predict the characteristics of tomorrow's product.

EFFECTIVE STATISTICAL INFERENCE

The aim of statistical inference in an analytic problem should be to give the expert in the subject matter the best possible chance to take the right action, that is, to reduce to a minimum the losses from the two types of mistake. A careful description of the conditions of the experiment are, as Shewhart emphasized,[8] an important part of the data of the experiment: the expert in the subject matter requires this kind of information (unfortunately too often omitted by statisticians).

There is no knowledge without temporal spread, which implies prediction.[9] In most analytic problems, the substantive expert must contribute heavily to the conclusions, the knowledge, that can be drawn out of a study.

Statistical inference in an analytic problem is most effective when it is presented as conclusions valid for the frame studied and for the range of environmental conditions specified for the tests. It is important to make clear that conclusions drawn by statistical theory may not hold under other conditions, and that other conditions may well be encountered.

Tests of a medical treatment, to be useful to future patients, would specify ranges of dosage, length of treatment, severities and other characteristics of the illness treated, and observation of side effects; otherwise, there would be serious difficulties in evaluating of the test results. "The comparison was carried out over a period of three weeks. No side effects were observed." Consumer research on some products can be nigh meaningless without reference to the season, climate, and economic conditions, for example, studies on consumption of soft drinks, or of analgesics, or of intentions to travel.

The theory of sampling and design of experiment are important in analytic studies. Optimum allocation of effort in analytic studies often differs from optimum allocation of effort in enumerative studies, though there is no literature to cite. Analysis of variance is useful as a rough tool of inference, to be followed up with more careful analysis. The trouble with analysis of variance is that it obscures trends and differences between small segments. The same caution holds for factor analysis and for cluster analysis. Any technique can be useful if its limitations are understood and observed.

Techniques of analysis that are most efficient in analytic problems include run charts to detect trends and differences between small classes. A run chart is simply a plot of results in order of age, time, duration of test, stress, or geographic location. A scatter diagram is often helpful. A distribution, simple though it be, is a powerful tool. Extreme skewness and wiggles detect sources of variation and lead to improved understanding of the process. The Mosteller-Tukey double square-root paper is useful, even when results are moderately correlated and do not follow strictly the binomial distribution.[10]

STATISTICAL TESTS OF HYPOTHESES

Unfortunately, as already stated, no statistical technique will evaluate the risks in an analytic problem. A brief note in the negative about testing hypotheses belongs here. The sad truth is that so-called tests of hypotheses, tutored well but not wisely in books and in teaching, are not helpful in practical problems, and as a system of logic, are misleading.[11] Two different treatments or two different varieties are never equal under any set of conditions: this we know without spending a nickel on an experiment. A difference between two treatments, though far too small to be of any economic or scientific consequence, will show up as "significantly different" if the experiment be conducted through a sufficient number of trials. A difference may be highly significant, yet be of no economic nor scientific importance. Obviously, such a test conveys no knowledge.

Likewise, tests of whether the data of a survey or an experiment fit some particular curve is of no scientific or economic importance. $P(\chi^2)$ for any curve, for any system, approaches zero as the number of observations increases. With enough data, no curve will fit.

The question that one faces in using some curve or relationship is whether it leads to a useful conclusion for experience in the future, or whether some other curve would do better? How robust are the conclusions?

Examples in the books on tests of hypotheses and in teaching are usually analytic in nature, but are treated as if they were enumerative, with inferences applicable to neither type.

Likewise, the teaching of regression estimates usually makes no distinction between (a) estimates of a total count in a frame, or the average per unit (enumerative uses), and (b) estimates of parameters (analytic). The techniques are different, the theory of optimum allocation of effort is different, and the uses even more so.

To state usefully the analytic problem in symbols, we first require from the substantive expert the number D, the difference that he requires between the two treatments (processes) to warrant action, which might of course be to continue the experiment. He needs an answer to the question

$$\text{Is } B \geqq A + D\,?$$

What we really need to know is whether the difference D will persist under conditions other than those that govern the experiment. As the manager of a large firm put it to his statistician, in consideration of two possible sizes of product, how much would it cost to carry out experiments that would tell him with fair certainty whether size B of the product would bring in 15% more dollars in sales than size A would bring. Here, $D = .15$. If the difference is less than 15%, it would not be worthwhile (in the judgment of the manager) to change the size: above 15%, it would be.

The appropriate statistical design will depend on the value of D.

For an example, one need only open any book on mathematical statistics, or any journal in psychology or biometrics. To avoid innuendo, in Table 4.1 I give an example close to hand.[12] The characteristic is somnambulism in children.

There is no mention of what difference D might be important. Moreover, the results must surely be obscured by difficulties in observation: "The behavioral findings presented in this report were obtained from a parent or guardian, usually the child's mother" (p. 2). The questionnaire was left at the home, picked up later. There is no mention of any test on the reliability of such observations.

TABLE 4.1 UNPLEASANT DREAMS

Sex	Frequently	Not often	Never	Unknown
Both	1.8	41-.8	52.1	4.3
Boys	2.0	41.2	52.0	4.8
Girls	1.6	42.4	52.1	3.9

While 10% of children in the national study were reported to have done some sleepwalking, only about 1% did so frequently. The data in the table, however, are sufficient to clearly establish the statistical significance of the relationship (χ^2_4 = 35.3 for boys and 24.4 for girls, $P < .001$ for both boys and girls).

In my own experience, correlation between two informers or observers on such characteristics can only be described as disappointing, even at the extreme ends of the scale, where theory tells us that agreement should be good if both observers are independent and equal.[13] One could conclude that the differences between boys and girls in this study are measures of differences between observers, mostly mothers, instead of differences between boys and girls.

AN EXAMPLE OF AN ANALYTIC STUDY

Suppose that the problem is to decide whether the cause-system has the value p or p′, or how much it has changed over a period of time, and why. As an example, p might be the birthrate per schizophrenic female in the state of New York in one 3-year period (e.g., 1934-1936) and p′ the rate 20 years later (1954-1956), after drugs for schizophrenia had come into general use by most psychiatrists. The substantive problem is to find why the rate changed, if it changed. The plan is to study the records of patients that entered the hospitals in the state of New York over the two periods. The first step would be to screen the case notes of a sample of patients admitted in the specified periods, to decide which female patients within the prescribed range of age (i.e., 20 to 39) were schizophrenic. The results of the study are highly dependent on just who is classified in this screening as schizophrenic; hence the screening must be carried out by psychiatrists who are willing to abide by an accepted glossary. There must be controls in the form of independent judgments of a subsample of cases to measure the variance between psychiatrists and to develop an identifiable system of diagnosis. The statistical problem is more than to estimate $p - p'$.

The next step would be further examination of the case notes of the females classified as schizophrenic to discover whether they were on drugs in or out of the hospital, how many children had been born to them before admission, and to trace these females over a period of years to discover how many more children they had over a span of years, and how much time they spent in the hospital. It would be a simple matter, when the results are in, to calculate the overall change $p - p'$. But how would one use the standard error so calculated? Clearly, it would have little meaning and less use. The problem of interpreting the results would be difficult, even with the most skillful statistical design and interviewing of patients and informants. The problem is not one in statistical significance. It is for this reason that extreme accuracy in an analytic study is wasted effort.

A change in rate from p to p′ by an amount D would be established or refuted only by examination in detail by age, size of community, orientation of the hospital. Useful statistical tools would be scatter diagrams aided by the sign test, and comparison of cumulative distributions.

We could go wrong in our conclusion, but unfortunately there is no statistical test we can apply to the data of the study that can tell us the risk of ascribing the change in birthrate to the use of drugs which decrease the time spent in the hospital, increase the time spent at home, when the experts decide in later years that drugs were not the cause of the change in birthrate. Neither is there a statistical test to tell us the contrary risk of eliminating drugs as a cause, when the experts decide in later years that drugs were definitely a contributing factor.

ANOTHER EXAMPLE

There are two methods of packing coffee into tins. Method A is the machinery already on the floor and the customary way of using it. Method B is new machinery that its manufacturer claims will turn out the work more rapidly and hit closer to any prescribed weight, so that with his machinery it is not necessary to put as many additional grains of coffee into a can to meet requirements of minimum weight as it is with the machinery in use. One machine of the new type is to be set up along a production line next week, and it is proposed to test it against the standard method of the past. It is hoped to reach a decision within a few weeks on whether the new machinery would be sufficiently advantageous to warrant the cost of replacement.

Would it be good management to try to be guided entirely on the results? One could run the two methods side by side and get figures, but what could he infer from these figures? Would the figures predict unforeseeable events such as time out for repairs, ability of the manufacturer of B to supply parts and service? The new machinery may not require repairs for six months, at the end of which time it may start to deteriorate.

Another possible difficulty is that the test to be run during the next few weeks could be unfair to the new machinery because the men that will operate it will be either operators of the regular machinery; or if sent in from the outside, they will hardly have a chance to accustom themselves to the new environment before the test will be running. As a further point, in spite of the manufacturer's efforts, Machine B may not be installed properly: it may require adjustments over a period of weeks.

Certain decisive results are of course possible. The new machinery may break down continually, or it may distinctly outclass the standard machinery and methods, with little danger (on engineering judgment, not statistical) of running into heavy costs of maintenance.

If forced, anybody could, at the end of a test, or with no test, make one decision or the other: (1) adopt the new machinery; (2) stay with the old machinery. Management would perhaps decide later that they had made a wise decision, or an unwise one. More likely, they would never raise a question about their decision, nor be able to provide any information about it.

THE RETROSPECTIVE METHOD

This is a method of evaluation that is much used, even though the hazards of wrong conclusions are great unless the observations are interpreted with care. It may therefore be useful to explain the method in simple terms and to offer a few words of caution. In the retrospective method, one divides into groups (diseased, not diseased) a population as it exists today, and inquires into the past histories of the individuals in these groups. The aim is of course to discover whether the past histories are different in any meaningful way and thus to discover the causes of the differences observed today between the two groups. The method is tempting, by reasons of economy, speed, and simplicity: we do not need to follow over a long period of time the people or animals or plants that we wish to study, with all the problems of tracing people as they move about. Still more tempting, the retrospective method does not require us to try to divide a sample of people into two groups, A and B, and say to Group A, you people are not to smoke during the next 20 years, and to Group B, you people are to smoke 2 packs a day for the next 20 years (all of which is of course fantastic).

The following example explains the retrospective method (oversimplified, as every term requires a lengthy operational definition):

Cause 1 (C1): he was a smoker 20 years ago.
Cause 2 (C2): he was not a smoker 20 years ago.
Effect 1 (E1): alive now, diseased.
Effect 2 (E2): alive now, not diseased.

By examination of a proper sample of people living today, we may divide those that have attained a certain age, say 50, into four groups, shown in Table 4.2, into which we have entered the observed frequencies, x_{ij}. Now suppose that the frequencies off the diagonal were zero ($x_{12} = x_{21} = 0$). Every person diseased today was a smoker 20 years ago. Every person not diseased today was not a smoker 20 years ago. Could we conclude that smoking 20 years ago caused disease

TABLE 4.2

	Result (observed today)		
Cause (from history)	E 1 (diseased)	E 2 (not diseased)	Total
C 1 (smoked)	x_{11}	x_{12}	$x_{1.}$
C 2 (did not smoke)	x_{21}	x_{22}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{..}$

today? No, but the retrospective method can raise question marks for further study.

The trouble with the retrospective method is that it studies only the survivors. We can study today only the survivors of 20 years ago. One must admit the possibility, and investigate it, that all the deaths that occurred over the interval of 20 years were nonsmokers; that smoking toughens one's resistance to diseases other than the specified disease, and that a smoker thus has a better chance to live 20 years, even though, at the end of that period, he will already have contracted the specified disease.

It is easy to make a wrong inference by a computation of chi-square for the 2 x 2 table under discussion. The survivors we study today are not a sample of the people who were alive 20 years ago. The survivors alive today do not tell us all that we need to know about the effects of the suspected causes that operated 20 years ago. We need to know what happened to the nonsurvivors. Where are the rest of the people, not alive today, who were alive 20 years ago? What happened to them? We cannot calculate limits of uncertainty on conclusions concerning suspected causes of disease drawn purely from observations on today's survivors. This is the great failing of the retrospective method, and it is serious.

Must we throw away the information acquired in a retrospective study? No, do not throw it away; supplement it. The retrospective method raises questions, hypotheses to study. The next step is to fill in the gaps, perhaps by making use of small prospective studies, pointed directly at the target. Unfortunately, one must wait for results.

NOTES

1. Following Frank Yates, I use the word *material* to denote people, patients, business establishments, accounts, cases, animals, agricultural products, industrial products, or anything else.

2. Shewhart's Rule 2, from Walter A. Shewhart, *Statistical Method from the Viewpoint of Quality Control* (Graduate School, Department of Agriculture, Washington, 1938), p. 92.

3. The concept of the frame was first defined but without use of any specific term by F. F. Stephan, American Sociological Review 1 (1936): 569-580.

4. The contrast between enumerative and analytic studies is set forth in chapter 7 of Deming, *Some Theory of Sampling* (Wiley, 1950; Dover, 1966). See also chapter 31 in *New Developments in Survey Sampling* by Norman L. Johnson and Harry Smith (Wiley-Interscience, 1969).

5. Richard H. Blythe, "The Economics of Sample-Size Applied to the Scaling of Sawlogs," The Biometrics Bulletin [Washington] 1 (1945): 67-70. Leo Törnqvist, "An Attempt to Analyze the Problem of an Economical Production of Statistical Data." Nordisk Tidsskrift for Teknisk Økonomi 37 (1948): 263-274.

6. Some pitfalls in estimation are described in chapter 31 in *New Developments in Survey Sampling* by Norman L. Johnson and Harry Smith cited earlier.

7. Walter A. Shewhart, *Statistical Method from the Viewpoint of Quality Control* (Graduate School, Department of Agriculture, 1939), chap. 3.

8. Ibid.

9. C. I. Lewis, *Mind and the World-Order* (Scribners, 1929), chaps. 6 and 7.

[10] Frederick Mosteller and John W. Tukey, "The Uses and Usefulness of Probability Paper," Journal of the American Statistical Association 44 (1949): 174-212. The double square-root paper is manufactured by the Codex Book Company of Norwood, Mass.

[11] Joseph Berkson, "Tests of Significance Considered as Evidence," Journal of the American Statistical Association 37 (1942): 325-335; Carl Earhardt, "Statistics, a Trap for the Unwary," Obstetrics and Gynecology 14 (Oct. 1959): 549-554; J. Wolfowitz, "Remarks on the Theory of Testing Hypotheses," The New York Statistician 18 (March 1967); W. Edwards Deming, "Boundaries of Statistical Inference," being chapter 31 in *New Developments in Survey Sampling* by Norman L. Johnson and Harry Smith (Wiley, 1969); Denton E. Morrison and Ramon E. Henkel, *The Significance Test Controversy* (Aldine, 1970).

[12] Relationships among (sic) parent ratings of behavioral characteristics of children, National Center for Health Statistics, series 11, no. 121, Oct. 1972.

[13] John Mandel, "Flammability of Children's Sleep-Wear," Standardization News [Philadelphia], May 1973, p. 11.

Reprint from:

HANDBOOK OF EVALUATION RESEARCH, VOL. 1
(Elmer L. Struening and Marcia Guttentag, Editors)

SAGE PUBLICATIONS, INC.
275 South Beverly Drive
Beverly Hills, California 90212

SAGE PUBLICATIONS LTD
St George's House / 44 Hatton Garden
London EC1N 8ER